# Optics

Yury Holubeu *

November 13, 2025

This note is not intended for distribution.

Optics and applications are discussed in detail. Links below show contents of solved problems, summary of special topics. Readers who have opened this note for the first time are highly recommended to read the preface.

## Contents

**Preface** **15**

**I —— Typical Optics in a Nutshell ——** **17**

**1 Overview of main laws and models** **17**
1.1 Typical optical laws and models in a nutshell . . . 17
1.1.1 Main laws and relations in optics . . . 17
1.1.2 Basics of geometrical optics and lenses . . . 18
1.1.3 Reflection and transmission of light . . . 30
1.1.4 Dispersion, Doppler effect . . . 32
1.1.5 Interference, coherence . . . 32
1.1.6 Diffraction . . . 46
1.1.7 Typical spectral devices and their properties . . . 56
1.1.8 Fourier optics: basics . . . 74
1.1.9 Radiation of light: basics . . . 90
1.1.10 Scattering of a light: basics . . . 91
1.2 Mathematics for optics . . . 91

**2 Special parts of optics in a nutshell** **91**
2.1 Quantum optics fundamentals in a nutshell (!??!) . . . 91
2.1.1 Laser optics: basics (?!) . . . 91
2.1.2 Quantum States of Light . . . 96
2.1.3 Radiative Transitions, Jaynes–Cummings model, cavities . . . 96
2.1.4 Open optical systems, decoherences . . . 96
2.1.5 Main formalism and equations of open systems . . . 96
2.1.6 Decoherence . . . 96
2.2 Some well-known optical setups . . . 97
2.2.1 Optical fibers in a nutshell . . . 97

**II Fundamentals of Optics** **98**

**3 Fundamentals of Optics** **98**
3.1 Philosophy of optics (?) . . . 98
3.1.1 The present moment . . . 98
3.1.2 Beauty of the world . . . 98
3.1.3 What is the nature of light? . . . 98
3.2 Electromagnetic waves for Optics . . . 98
3.2.1 Fundamentals of field theory for optics . . . 98
3.2.2 Polarization of waves . . . 100
3.2.3 Refractive index (???) . . . 102
3.2.4 Other about polarization . . . 103
3.3 Photometric concepts and units (by Landsberg) . . . 104

*https://yuriholubeu.github.io/, yuri.holubev@gmail.com

3.3.1 Basic concepts . . . 104
3.3.2 Transition from energy quantities to luminous quantities . . . 104
3.3.3 Units for luminous measurements . . . 104
3.3.4 Light measurements (photometry) . . . 104
3.4 Other Fundamentals of Optics . . . 105
3.4.1 Path Integral for Optics . . . 105
3.4.2 Limits of Applicability of Wave Optics . . . 105
3.4.3 Change of Phase While Propagation . . . 107

**4 Geometrical Optics 108**
4.1 Aberration of the Optical System by Landsberg . . . 108
4.2 Optical Instruments According to Landsberg . . . 108
4.3 General Laws of Geometric Optics . . . 108
4.3.1 Transition to Geometric Optics from Wave Optics . . . 108
4.3.2 The Eikonal Equation: Basics . . . 109
4.3.3 The Eikonal Equation: Special Features and Applications . . . 116
4.3.4 Limits of Geometric Optics . . . 116
4.4 Laws of Geometric Optics for Typical Setups . . . 116
4.4.1 Formula of a Thin Lens . . . 116

**5 Other Methods and Ideas of Geometrical Optics 116**
5.1 Geometric Theory and Modeling of Optical Systems . . . 116
5.2 1. General Consideration of Centered Optical Systems . . . 116
5.2.1 1.1. An Elementary Optical Cell . . . 117
5.2.2 1.2. Centered Optical System . . . 119
5.2.3 1.3. Building an Image in a Centered System . . . 122
5.3 2. The Addition of Centered Optical Systems . . . 124
5.4 3. Examples of Centered Optical Systems . . . 125
5.4.1 3.1. The Telescopic System . . . 125
5.4.2 3.2. Thick Lens . . . 125
5.4.3 3.3. Thin Lens . . . 127
5.4.4 3.4. A System of Two Thin Lenses . . . 127
5.5 4. Optical Devices . . . 128
5.5.1 4.1. Magnifying Glass . . . 128
5.5.2 4.2. The Microscope . . . 129
5.5.3 4.3. The Viewing Tubes . . . 130
5.6 5. Basic Aberrations of Optical Systems . . . 131
5.6.1 5.1. Spherical Aberration . . . 131
5.6.2 5.2. Chromatic Aberration . . . 133

**6 Interference. Coherence 134**
6.0.1 1.1. The Wave Equation. The Simplest Types of Waves . . . 134
6.0.2 1.2. Monochromatic Waves. The Complex Amplitude. The Helmholtz Equation 137
6.0.3 2.1. Two-Beam Interference. The Visibility of the Stripes . . . 138
6.0.4 2.2. Interference of Different Types of Waves. The Width of the Interference Fringes . . . 139
6.1 3. The Statistical Nature of Radiation. Coherence . . . 141
6.1.1 3.1. Statistical Model of Light Emission by a Source . . . 141
6.1.2 3.2. The Width of the Line. Mechanisms of Broadening . . . 143
6.1.3 3.3. The Time Coherence Function . . . 145
6.1.4 3.4. Interference and Coherence . . . 146
6.1.5 3.5. Spectral Analysis . . . 147
6.1.6 3.6. The Wiener-Khinchin Theorem . . . 150
6.2 4. Spatial Coherence . . . 150

6.2.1 4.1. Degree of Spatial Coherence . . . 150
6.2.2 4.2. The Van Zitter-Zernike Theorem . . . 151
6.2.3 4.3. Homogeneous Source . . . 153
6.2.4 4.4. Spatial Coherence and Interference Aperture . . . 154
6.2.5 4.5. Localization of the Interference Pattern . . . 155
6.3 5. Interferometers . . . 156
6.3.1 5.1. The Rayleigh Interferometer . . . 157
6.3.2 5.2. The Jamin Interferometer . . . 158
6.3.3 5.3. The Michelson Interferometer . . . 160
6.3.4 5.4. Newton's Rings . . . 161

**7 Diffraction** **162**
7.1 1. Diffraction . . . 162
7.1.1 1.1. Setting the Task. The Wave Parameter . . . 162
7.1.2 1.2. Boundary Conditions. The Transparency Function . . . 163
7.1.3 1.3. The Huygens-Fresnel Principle . . . 166
7.1.4 1.4. Fresnel Diffraction . . . 177
7.1.5 Fresnel Diffraction: Examples . . . 187
7.1.6 1.5. Fraunhofer Diffraction . . . 188
7.1.7 Fraunhofer Diffraction: Examples . . . 199
7.1.8 Numerical Methods for Diffraction . . . 206
7.2 2. Principles of Fourier Optics . . . 207
7.2.1 2.1. A Spectral Method for Solving the Diffraction Problem . . . 207
7.2.2 2.2. Examples of Calculations of Diffraction by Spectral Method . . . 210
7.2.3 2.3. Fresnel Diffraction on Periodic Structures. The Talbot Effect (Self-Production) 212
7.2.4 2.4. The Spectrum of Plane Waves in Slit Diffraction. The Uncertainty Relation 213
7.2.5 2.5. The Limit of Geometric Optics . . . 214
7.2.6 2.6. The Field in the Focal Plane of the Lens. Spatial Transformation. Fourier Analysis . . . 218
7.2.7 2.7. The Principle of Double Diffraction and the Formation of Optical Images (Abbe's Theory) . . . 220
7.2.8 2.8. Spatial Filtering . . . 221
7.2.9 2.9. The Resolution of Optical Systems . . . 225
7.2.10 2.10. Image Formation. Transfer Function. The Resolution of the Microscope . 228

**8 Typical Spectral Devices** **231**
8.0.1 1. The Principle of Operation of Spectral Devices . . . 231
8.0.2 2. Description of Spectral Devices . . . 232
8.0.3 3. Angular Dispersion of Spectral Devices . . . 240
8.0.4 4. Resolution Capacity . . . 241
8.0.5 5. The Dispersion Region . . . 245
8.1 Laser Radiation . . . 246
8.1.1 1.1. Elementary Processes in Laser . . . 247
8.1.2 1.2. Gain Factor . . . 248
8.1.3 1.3. The Generation Spectrum. Doppler Broadening . . . 249
8.1.4 1.4. The Condition for Reaching the Generation Threshold . . . 250
8.1.5 1.5. Reverse Occupancy. Pumping . . . 251
8.2 2. The Mode Composition of the Laser Radiation . . . 252
8.2.1 2.1. Longitudinal Modes . . . 252
8.2.2 2.2. The Spectrum of Laser Radiation . . . 253
8.2.3 2.3. Coherence of Laser Radiation . . . 254
8.2.4 2.4. Transverse Modes* . . . 256
8.2.5 2.5. Lamb's Failure and Mode Competition* . . . 258
8.2.6 3. Common Types of Lasers* . . . 259

**9 Optics by Born Wolf** **262**
9.1 Historical Introduction . . . 262
9.2 Basic Properties of the Electromagnetic Field . . . 270
9.2.1 The Electromagnetic Field . . . 270
9.2.2 1.2 The Wave Equation and the Velocity of Light . . . 279
9.2.3 1.3 Scalar Waves . . . 282
9.2.4 1.4 Vector Waves . . . 291
9.2.5 1.5 Reflection and Refraction of a Plane Wave . . . 304
9.2.6 1.6 Wave Propagation in a Stratified Medium. Theory of Dielectric Films . . . 320
9.3 Electromagnetic Potentials and Polarization . . . 339
9.3.1 2.1 The Electrodynamic Potentials in the Vacuum . . . 340
9.3.2 2.2 Polarization and Magnetization . . . 344
9.3.3 2.3 The Lorentz-Lorenz Formula and Elementary Dispersion Theory* . . . 352
9.3.4 2.4 Propagation of Electromagnetic Waves Treated by Integral Equations* . . . 366
9.4 Foundations of Geometrical Optics . . . 376
9.4.1 3.1 Approximation for Very Short Wavelengths . . . 376
9.4.2 3.2 General Properties of Rays . . . 388
9.4.3 3.3 Other Basic Theorems of Geometrical Optics . . . 395
9.5 Geometrical Theory of Optical Imaging . . . 401
9.5.1 4.1 The Characteristic Functions of Hamilton . . . 401
9.5.2 4.2 Perfect Imaging . . . 410
9.5.3 4.3 Projective Transformation (Collineation) with Axial Symmetry . . . 418
9.5.4 4.4 Gaussian Optics . . . 424
9.5.5 4.5 Stigmatic Imaging with Wide-Angle Pencils . . . 434
9.5.6 4.6 Astigmatic Pencils of Rays . . . 437
9.5.7 4.7 Chromatic Aberration. Dispersion by a Prism . . . 442
9.5.8 4.8 Photometry and Apertures . . . 449
9.5.9 4.9 Ray Tracing* (!!!????) . . . 459
9.5.10 4.10 Design of Aspheric Surfaces . . . 466
9.6 Geometrical Theory of Aberrations . . . 472
9.6.1 5.1 Wave and Ray Aberrations; The Aberration Function . . . 472
9.6.2 5.2 The Perturbation Eikonal of Schwarzschild . . . 476
9.6.3 5.3 The Primary (Seidel) Aberrations . . . 479
9.6.4 5.4 Addition Theorem for the Primary Aberrations . . . 487
9.6.5 5.5 The Primary Aberration Coefficients of a General Centred Lens System . . . 488
9.6.6 5.6 Example: The Primary Aberrations of a Thin Lens . . . 495
9.6.7 5.7 The Chromatic Aberration of a General Centred Lens System . . . 498
9.7 Image-Forming Instruments . . . 501
9.7.1 6.1 The Eye . . . 501
9.7.2 6.2 The Camera . . . 503
9.7.3 6.3 The Refracting Telescope* . . . 507
9.7.4 6.4 The Reflecting Telescope . . . 513
9.7.5 6.5 Instruments of Illumination . . . 519
9.7.6 6.6 The Microscope . . . 520
9.8 Elements of the Theory of Interference and Interferometers . . . 524
9.8.1 7.1 Introduction . . . 524
9.8.2 7.2 Interference of Two Monochromatic Waves . . . 525
9.8.3 7.3. Two-Beam Interference: Division of Wave-Front . . . 528
9.8.4 7.4. Standing Waves . . . 544
9.8.5 7.7 The Comparison of Wavelengths with the Standard Metre . . . 549
9.9 7.5. Two-Beam Interference: Division of Amplitude . . . 551
9.9.1 7.5.1 Fringes with a plane parallel plate . . . 551
9.9.2 7.5.2 Fringes with thin films; the Fizeau interferometer . . . 556

9.9.3 7.5.3 Localization of fringes † . . . 563
9.9.4 7.5.4 The Michelson interferometer . . . 572
9.9.5 7.5.5 The Twyman-Green and related interferometers . . . 574
9.9.6 7.5.6 Fringes with two identical plates: the Jamin interferometer and interference microscopes . . . 578
9.9.7 7.5.7 The Mach-Zehnder interferometer; the Bates wave-front shearing interferometer . . . 585
9.9.8 7.5.8 The coherence length; the application of two-beam interference to the study of the fine structure of spectral lines . . . 589
9.10 7.6 Multiple-Beam Interference . . . 596
9.10.1 7.6.1 Multiple-beam fringes with a plane parallel plate . . . 596
9.10.2 7.6.2 The Fabry-Perot interferometer . . . 602
9.10.3 7.6.3 The application of the Fabry-Perot interferometer to the study of the fine structure of spectral lines . . . 606
9.10.4 7.6.4 The application of the Fabry-Perot interferometer to the comparison of wavelengths . . . 611
9.10.5 7.6.5 The Lummer-Gehrcke interferometer . . . 616
9.10.6 7.6.6 Interference filters . . . 622
9.10.7 7.6.7 Multiple-beam fringes with thin films . . . 626
9.10.8 7.6.8 Multiple-beam fringes with two plane parallel plates . . . 636
9.11 Elements of the Theory of Diffraction . . . 643
9.11.1 8.1. Introduction . . . 643
9.11.2 8.2. The Huygens-Fresnel Principle . . . 644
9.11.3 8.3. Kirchhoff's Diffraction Theory . . . 648
9.11.4 8.4. Transition to a Scalar Theory* . . . 660
9.11.5 8.5. Fraunhofer Diffraction at Apertures of Various Forms . . . 665
9.11.6 8.6. Fraunhofer Diffraction in Optical Instruments . . . 675
9.11.7 8.7. Fresnel Diffraction at a Straight Edge . . . 702
9.11.8 8.8. The Three-Dimensional Light Distribution near Focus . . . 710
9.11.9 8.9. The Boundary Diffraction Wave . . . 724
9.11.10 8.10. Gabor's Method of Imaging by Reconstructed Wavefronts (Holography) . . . 729
9.12 The Diffraction Theory of Aberrations . . . 735
9.12.1 9.1. The Diffraction Integral in the Presence of Aberrations . . . 736
9.12.2 9.2. Expansion of the Aberration Function . . . 741
9.12.3 9.3. Tolerance Conditions for Primary Aberrations . . . 745
9.12.4 9.4. The Diffraction Pattern Associated with a Single Aberration . . . 749
9.12.5 9.5. Imaging of Extended Objects . . . 756
9.13 Interference and Diffraction with Partially Coherent Light . . . 763
9.13.1 10.1 Introduction . . . 763
9.13.2 10.2 A Complex Representation of Real Polychromatic Fields . . . 766
9.13.3 10.3 The Correlation Functions of Light Beams . . . 771
9.13.4 10.4 Interference and Diffraction with Quasi-monochromatic Light . . . 777
9.13.5 10.5 Some Applications . . . 788
9.13.6 10.6 Some Theorems Relating to Mutual Coherence . . . 799
9.13.7 10.7 Rigorous Theory of Partial Coherence* . . . 802
9.13.8 10.8 Polarization Properties of Quasi-monochromatic Light . . . 810
9.14 Rigorous Diffraction Theory . . . 822
9.14.1 11.1 Introduction . . . 822
9.14.2 11.2 Boundary Conditions and Surface Currents . . . 823
9.14.3 11.3 Diffraction by a Plane Screen: Electromagnetic Form of Babinet's Principle 825
9.14.4 11.4 Two-dimensional Diffraction by a Plane Screen . . . 826
9.14.5 11.5 Two-dimensional Diffraction of a Plane Wave by a Half-plane . . . 832
9.14.6 11.6 Three-dimensional Diffraction of a Plane Wave by a Half-plane . . . 845

9.14.7 11.7 Diffraction of a Localized Source by a Half-plane . . . 847
9.14.8 11.8 Other Problems . . . 855
9.14.9 11.9 Uniqueness of Solution . . . 859
9.15 Diffraction of Light by Ultrasonic Waves . . . 860
9.15.1 12.1 Qualitative Description of the Phenomenon and Summary of Theories Based on Maxwell's Differential Equations . . . 860
9.16 12.2 Diffraction of Light by Ultrasonic Waves as Treated by the Integral Equation Method 866
9.16.1 12.2.1 Integral equation for $\boldsymbol{E}$-polarization . . . 867
9.16.2 12.2.2 The trial solution of the integral equation . . . 868
9.16.3 12.2.3 Expressions for the amplitudes of the light waves in the diffracted and reflected spectra . . . 871
9.16.4 12.2.4 Solution of the method of successive approximations . . . 872
9.16.5 12.2.5 Expressions for the intensities of the first and second order lines for some special cases . . . 875
9.16.6 12.2.6 Some qualitative results . . . 876
9.16.7 12.2.7 The Raman-Nath approximation . . . 879
9.17 Optics of Metals . . . 881
9.17.1 13.1. Wave Propagation in a Conductor . . . 881
9.17.2 13.2. Refraction and Reflection at a Metal Surface . . . 885
9.17.3 13.3. Elementary Electron Theory of the Optical Constants of Metals . . . 894
9.17.4 13.4. Wave Propagation in a Stratified Conducting Medium. Theory of Metallic Films . . . 896
9.17.5 13.5. Diffraction by a Conducting Sphere; Theory of Mie . . . 903
9.18 Optics of Crystals . . . 933
9.18.1 14.1. The Dielectric Tensor of an Anisotropic Medium . . . 933
9.18.2 14.2. The Structure of a Monochromatic Plane Wave in an Anisotropic Medium 936
9.18.3 14.3. Optical Properties of Uniaxial and Biaxial Crystals . . . 948
9.18.4 14.4. Measurements in Crystal Optics . . . 962
9.18.5 14.5. Stress Birefringence and Form Birefringence . . . 979
9.18.6 14.6. Absorbing Crystals . . . 984

**III Fundamentals of Quantum Optics 1004**

**10 Basics of Laws of Quantum Optics 1004**
10.1 Photoelectric effect . . . 1004
10.2 Compton effect . . . 1004
10.3 Pressure of light . . . 1004

**11 Basics Models of Quantum Optics 1004**
11.1 Lasers: Basics . . . 1004
11.1.1 Basic Model of a Laser . . . 1004

**IV Problems 1012**

**12 Basic Parts of Optics 1012**
12.0.1 Basic problems of geometrical optics . . . 1012
12.0.2 Advanced problems on geometrical optics 1 . . . 1014
12.0.3 Advanced problems on geometrical optics 2 . . . 1025
12.0.4 Advanced problems on geometrical optics 3 . . . 1037
12.0.5 Problems about photometry . . . 1042
12.0.6 Problems about transmission and reflectance of light . . . 1046
12.0.7 Problems about interference of monochromatic light . . . 1050

12.0.8 Problems about quasi-monochromatic light, temporal coherence . . . 1055
12.0.9 Problems about spatially extended light sources . . . 1062
12.0.10 Problems about Fresnel diffraction . . . 1071
12.0.11 Problems about Fraunhofer diffraction . . . 1084
12.0.12 Problems about spectral devices 1 . . . 1092
12.0.13 Problems about spectral devices 2 . . . 1103
12.0.14 Problems about Fourier optics and holography . . . 1108
12.0.15 Problems about dispersion . . . 1112
12.0.16 Problems about polarization effects 1 . . . 1120
12.0.17 Problems about polarization effects 2 . . . 1124

**13 Questions about checking knowledge 1128**
13.0.1 Questions about understanding typical optics . . . 1128
13.0.2 Questions about understanding details in optics . . . 1129
13.0.3 Questions about understanding quantum optics . . . 1129

**14 Preparation to some parts of optics 1129**
14.1 Background for optics . . . 1129
14.1.1 Problems for mastering mathematics for optics . . . 1129
14.1.2 Problems for mastering field theory for optics . . . 1129
14.1.3 Problems about some generally obvious proofs . . . 1129
14.2 Easy optics for preparation . . . 1129
14.2.1 Problems about easy geometrical optics . . . 1129
14.2.2 Problems about proofs in geometrical optics . . . 1129

**15 Problems about specific setups, phenomena 1130**
15.1 Problems about specific well-known phenomena . . . 1130
15.1.1 Problems about Pressure of Light . . . 1130
15.1.2 Simple Problems about Nonlinear Optics . . . 1131
15.1.3 Problems about Radiation and Black Body . . . 1131
15.1.4 Problems about Specific Optical Devices . . . 1151
15.1.5 Problems of other parts of physics that uses concepts from optics . . . 1151

**V —— Special Optics in a Nutshell —— 1154**

**16 On Other Properties and Methods of Optics 1154**
16.1 Other properties and methods for typical phenomena . . . 1154
16.1.1 Diffraction: other properties and methods . . . 1154
16.2 Other well-known but not so useful phenomena . . . 1156
16.2.1 Radio waves from labnik . . . 1156
16.2.2 Polarization, Crystal optics from labnik . . . 1161
16.3 Special professional methods in optics . . . 1164
16.3.1 Geometrical Optics by Born and Wolf . . . 1164
16.3.2 Eikonal equation . . . 1165
16.3.3 Deeper Understanding of Corpuscular Optics . . . 1165
16.3.4 Specific interference . . . 1165
16.3.5 Specific dispersions . . . 1166
16.3.6 Specific reflectance . . . 1166
16.3.7 Green's function method for optics (?) . . . 1166
16.3.8 $S$-matrix for optics (?) . . . 1166
16.3.9 On PT-symmetry in optics (!!!?!?!??!?) . . . 1166
16.3.10 Fourier Optics and Holography (??) . . . 1167
16.3.11 Talbot Effect, Other Phenomena of Diffraction (???) . . . 1172

**17 Other Models** **1176**
17.1 Other Well-known Models of Optics . . . 1176
17.1.1 Optical Anisotropy (!??!) . . . 1176
17.1.2 Nonlinear optics: basics (????) . . . 1176
17.1.3 Photonic Crystals . . . 1177
17.1.4 Scattering on Crystals . . . 1177
17.1.5 X-ray Diffraction (?) . . . 1177
17.1.6 Enhancement of Reflection . . . 1178
17.1.7 Enhancement of Absorption . . . 1178
17.1.8 Amplification of Waves . . . 1178
17.2 Quantum Optics: Other Well-known Models . . . 1178
17.3 Special Theoretical Methods . . . 1178
17.3.1 Path integral for optics: formalism . . . 1178
17.3.2 Path integral for optics: diffraction . . . 1178
17.3.3 Pressure of Light, Optical Tweezers . . . 1178
17.3.4 Geometrical gravitational lensing . . . 1178
17.3.5 Attosecond pulses . . . 1178

**VI Optics in Continuous Media** **1179**

**18 Propagation of Light Across the Boundary of Two Media** **1179**
18.0.1 Fresnel Formulas . . . 1179
18.0.2 Total Internal Reflection . . . 1179

**19 Interaction of Waves with Matter and Molecular Optics** **1179**
19.1 Fundamentals of Optical Wave Scattering . . . 1179
19.1.1 Rayleigh Scattering . . . 1179
19.1.2 Scattering Cross Section and Angular Distribution . . . 1179
19.1.3 Absorption . . . 1180
19.1.4 Polarization of Scattered Waves . . . 1180
19.1.5 Other Effects in Scattering . . . 1180
19.2 Propagation of Waves Through Matter . . . 1180
19.2.1 Refractive Index . . . 1180
19.2.2 Reflection and Refraction at a Plane Boundary . . . 1183
19.2.3 Fresnel Formulas for Reflected and Transmitted Waves . . . 1185
19.2.4 Polarization Effects . . . 1187
19.2.5 Energy Flux Through the Boundary . . . 1189
19.3 Dispersion and Absorption of Light (According to Landsberg) . . . 1189
19.4 Dispersion . . . 1189
19.4.1 Phase and Group Velocity . . . 1189
19.4.2 Classical Theory of Dispersion . . . 1192
19.4.3 Plasma and Ionosphere . . . 1194
19.4.4 Advanced Theories of Dispersion (????) . . . 1194
19.5 Other About Scattering . . . 1195

**20 Effects and Models** **1195**
20.1 Optics of Crystals . . . 1195
20.1.1 Crystals . . . 1195
20.1.2 Uniaxial Crystals and Double Refraction . . . 1196
20.1.3 Crystal Plates . . . 1198
20.2 Artificial Anisotropy (According to Landsberg) . . . 1199
20.3 Nonlinear Optics and Electro-Optical Effects (??) . . . 1199
20.3.1 Frequency Conversion in Nonlinear Media . . . 1199

20.3.2 Nonlinear Permittivity . . . 1201
20.3.3 Self-Focusing . . . 1206
20.3.4 Strong Electromagnetic Waves . . . 1212
20.3.5 Stimulated Raman Scattering . . . 1216
20.3.6 Second Harmonic Generation . . . 1218
20.3.7 Phase-Matching Conditions (?) . . . 1223
20.3.8 Optical Rectification . . . 1223
20.3.9 Interference of Polarized Waves . . . 1223
20.3.10 Pockels Effect . . . 1224
20.3.11 Kerr Effect . . . 1224
20.3.12 Faraday Effect . . . 1224

**VII Optical Phenomena in Nature 1225**

**21 Some Phenomena 1225**
21.1 Optics in Everyday Life . . . 1225
21.2 Effects with Photography . . . 1225
21.3 Light in Atmosphere . . . 1225
21.4 Other Phenomena . . . 1225
21.4.1 Color of the sky . . . 1225
21.4.2 Rainbow . . . 1225
21.4.3 Double Rainbow . . . 1225
21.4.4 Optical Halo . . . 1225

**VIII Optical Equipment and Experiments 1226**

**22 Basic Laboratory Experiments 1226**
22.1 Experiments about Geometrical Theory and Modeling of Optical Systems . . . 1226
22.1.1 The Study of Centered Optical Systems . . . 1226
22.1.2 Modeling of Optical Devices and Determining Their Magnification . . . 1235
22.1.3 Abbe Refractometer . . . 1241
22.2 Experiments about Interference and Coherence . . . 1248
22.2.1 Newton's Rings . . . 1248
22.2.2 Jamin Interferometer . . . 1252
22.2.3 Rayleigh Interferometer . . . 1256
22.2.4 Michelson Interferometer . . . 1260
22.2.5 Coherence of Light . . . 1266
22.2.6 Diffraction of Coherent Radiation on a Rough Surface . . . 1274
22.3 Experiments about Diffraction. Fourier Optics . . . 1288
22.3.1 Study of Light Diffraction . . . 1288
22.3.2 Diffraction of light by an ultrasonic wave in a liquid . . . 1298
22.3.3 Investigation of Microscope Resolving Power by the Abbe Method . . . 1307
22.3.4 Fourier Transform Method in Optics . . . 1313
22.3.5 Study of a Hologram . . . 1317
22.3.6 Diffraction of Light on Periodic Structures (Self-Reproduction) . . . 1327
22.4 Experiments about Spectral Instruments . . . 1331
22.4.1 Amplitude Diffraction Grating . . . 1331
22.4.2 Phase Diffraction Grating . . . 1333
22.4.3 Study of a Prism Using a Goniometer . . . 1335
22.4.4 Fabry–Perot Interferometer . . . 1338
22.4.5 Appendix about Spectral Instruments . . . 1341
22.5 Experiments about Laser Radiation . . . 1346

22.5.1 Helium–Neon Laser . . . 1346
22.5.2 Interference of Laser Radiation . . . 1355
22.5.3 Scanning Interferometer . . . 1365
22.6 Experiments about Interference and Tunneling of Radio Waves . . . 1372
22.6.1 Interference of electromagnetic waves in the millimeter range . . . 1372
22.6.2 Tunneling of millimeter radio waves . . . 1377
22.7 Experiments about Polarization, Crystal Optics . . . 1381
22.7.1 Double Refraction . . . 1381
22.7.2 Pockels Effect . . . 1386
22.7.3 Polarization . . . 1392
22.7.4 Rotation of the Plane of Polarization . . . 1400

**23 Fundamentals of Experimental Optics 1404**
23.1 Optical Fibers (???) . . . 1404
23.1.1 Overview of Applications of Optical Fibers . . . 1404
23.1.2 Elements of Optical Fibers . . . 1404
23.1.3 Typical Phenomena in Optical Fibers (??) . . . 1404
23.2 Waveguides . . . 1405
23.2.1 Propagation of Light in a Waveguide . . . 1405
23.2.2 Estimations of Typical Effects in Waveguides . . . 1405

**24 Improvement and Creation of Optical Equipment 1405**
24.1 Improvement of Optical Equipment . . . 1405
24.1.1 How to make a better cameras? . . . 1405
24.1.2 How to make a better glasses? . . . 1405
24.2 Creation of Optical Equipment . . . 1405
24.2.1 How to Create a Lamp (???) . . . 1405

**25 Vision and Eyes 1406**
25.1 How Eyes Work . . . 1406
25.1.1 Basic Models . . . 1406
25.1.2 Proof that Light is Transferred to Electric Signal . . . 1408

**26 Other Devices 1408**
26.1 Screens and Displays (??) . . . 1408
26.2 Optoelectronics . . . 1408
26.3 Very Exotic Devices . . . 1408
26.3.1 Single photon counters . . . 1409
26.3.2 single-pixel cameras . . . 1409
26.4 Optical coherence tomography today: speed, contrast, and multimodality . . . 1409
26.4.1 Swept Source Lasers for Faster OCT Imaging . . . 1412
26.4.2 Parallel OCT: Pushing Speed Limits . . . 1417
26.4.3 4 Label-Free Optical Angiography . . . 1426
26.4.4 5 Multimodal Optical Coherence Tomography . . . 1431
26.4.5 Basic Terminologies and Key Performance Parameters in OCT . . . 1447

**27 Hypothetical Devices 1450**
27.1 Invisibility . . . 1450

**IX Other Topics and Applications 1451**

**28 Other Effects and Methods** **1451**
28.1 Interference and Tunneling of Radio Waves . . . 1451
28.1.1 1. Introduction . . . 1451
28.1.2 2. Generation of Microwave Radio Waves . . . 1453
28.1.3 3. Propagation of Em Waves in Radio Waveguides . . . 1454
28.1.4 4. Tunneling of Microwave Radio Waves . . . 1457
28.2 Polarization. Crystal Optics . . . 1461
28.2.1 1. Types of Polarization, Basic Concepts . . . 1461
28.2.2 2. Obtaining Polarized Light . . . 1463
28.2.3 3. Observation and Analysis of Polarized Light. Malus' Law . . . 1464
28.2.4 4. Double Refraction . . . 1465
28.2.5 5. Crystal Plates . . . 1468
28.2.6 6. Electro-Optical Effects . . . 1469
28.2.7 7. Optical Activity. The Faraday Effect . . . 1469
28.3 Radiation in Space . . . 1470
28.3.1 Black Body Radiation: Basic Laws . . . 1470
28.3.2 Black Body Radiation: Consequences . . . 1473
28.4 Radiation in Media . . . 1473
28.4.1 Vavilov-Cherenkov radiation . . . 1473
28.5 Chemical Action of Light (????) . . . 1474
28.6 Holography . . . 1474
28.6.1 3.1. Point Source Hologram (Gabor Hologram) . . . 1476
28.6.2 3.2. The Resolution of the Hologram . . . 1479
28.6.3 3.3. A Diagram with an Inclined Support Beam . . . 1480
28.6.4 Connection of Holography in Optics and Holography in High Energy Physic . . . 1481
28.7 Other Effects . . . 1481
28.7.1 Photonic Crystals . . . 1481
28.7.2 Luminescence by Landsberg (???) . . . 1482
28.7.3 Optical Tweezers (?) . . . 1482

**29 Special optical devices and applications** **1482**
29.1 Modeling of Optical Phenomena . . . 1482
29.2 Analysis of Images . . . 1482
29.3 3D computer graphics physical fundamentals . . . 1482
29.3.1 Creation a 3D Graphics Program . . . 1482
29.3.2 Ray tracing . . . 1483
29.4 Photonic Crystals . . . 1483
29.5 Attosecond pulses . . . 1483

**30 Exotic Theoretical Models and Methods** **1483**
30.1 Exotic Theoretical Methods . . . 1483
30.1.1 *S*-matrix Method for Optics (!?!??) . . . 1483
30.2 PT-symmetry in optics (!!!?!?!??!?) . . . 1497
30.3 Hamiltonian Approach to Optics (???????!!!) . . . 1497
30.4 Ray transfer matrix analysis (?) . . . 1501
30.5 Integrable Models in Optics . . . 1501
30.6 Optics in Curved Spacetime and Gravitational Lensing: Basics . . . 1501
30.6.1 Basic Setup of Gravitational Lensing . . . 1502
30.6.2 Observations of Gravitational Lensing and Phenomena . . . 1509
30.6.3 Wave Optics of Gravitational Lensing . . . 1509
30.6.4 Some Typical Models of GL . . . 1509
30.6.5 Other about GL . . . 1511

**31 Catastrophe Optics and Caustics by Berry, Upstill** **1511**
31.1 § 8. Concluding Remarks . . . 1511
31.2 § 1. Introduction . . . 1512
31.3 § 2. Catastrophe Theory and Stable Caustics . . . 1514
31.3.1 2.1. Caustics Are Singularities of Gradient Maps . . . 1514
31.3.2 2.2. The Classification of Catastrophes . . . 1518
31.3.3 2.3. Examples of the Elementary Catastrophes in Optics . . . 1520
31.4 § 3. Diffraction Catastrophes . . . 1523
31.4.1 3.1. Asymptotic Solution of the Wave Equation . . . 1523
31.4.2 3.2. Integral Representations for Short Waves . . . 1526
31.4.3 3.3. The Hierarchy of Catastrophe Diffraction Integrals . . . 1528
31.4.4 3.4. Diffraction Scaling Laws . . . 1534
31.5 § 4. The Geometry of Wavefronts and their Normals . . . 1538
31.5.1 4.1. Paraxial Optics of a Uniform Medium . . . 1538
31.5.2 4.2. The Curvature Landscape . . . 1539
31.5.3 4.3. Umbilics . . . 1541
31.5.4 4.4. Cuspoids . . . 1543
31.6 § 5. Liquid Droplet Lenses . . . 1544
31.6.1 5.1. The Irregular Droplet as a Lens . . . 1544
31.6.2 5.2. Droplets with Uniform Internal Pressure . . . 1545
31.6.3 5.3. Droplets with Non-Uniform Internal Pressure . . . 1548
31.7 § 6. Caustic Networks . . . 1551
31.7.1 6.1. Networks and Junctions . . . 1551
31.7.2 6.2. Sinuous Networks . . . 1551
31.7.3 6.3. Triple Junctions . . . 1553
31.7.4 6.4. Junctions Are Organized By $^\circ X_9$ . . . 1555
31.8 § 7. Statistical Catastrophe Optics . . . 1556
31.8.1 7.1. Gaussian Random Wavefronts . . . 1556
31.8.2 7.2. Statistical Geometry in Two Dimensions . . . 1558
31.8.3 7.3. Statistical Geometry in Three Dimensions . . . 1560
31.8.4 7.4. Random Diffraction Catastrophes in Twinkling Light . . . 1562
31.9 Appendix 1: Wavefronts . . . 1564
31.10 Appendix 2: Primary Aberrations . . . 1566

**X Dead, Useless, Pointless Theories in Optics** **1569**

**32 Different Ideas** **1569**
32.1 Some Hypothetical Ideas . . . 1569
32.1.1 Different Ideas . . . 1569
32.2 Ideas from History . . . 1569
32.3 Philosophical Ideas about Light . . . 1569
32.4 Special Approach to Basic Propagation of Waves . . . 1569
32.4.1 The corner eikonal by Landau . . . 1569
32.4.2 Thin beams of rays by Landau . . . 1572
32.4.3 Wide beam mapping by Landau . . . 1577

**XI Appendix** **1580**

**A Introduction to Optics** **1580**
A.1 Extra Motivation . . . 1580
A.1.1 Amazing Facts and Phenomena in Optics . . . 1580
A.1.2 Gaining Skills in General Physics . . . 1580
A.1.3 Variety of Applications in Optics . . . 1581
A.1.4 Motivation for Branches of Optics . . . 1581
A.2 Mindset of a Professional in Optics . . . 1582
A.2.1 Ways to Make Money and Overview of Applications . . . 1582
A.2.2 Which Methods Work Good and Which Work Bad? . . . 1582
A.2.3 Unresolved Issues in Optics . . . 1582
A.2.4 Техническая сторона подхода к оптике . . . 1582
A.2.5 Should One Study Exotic Topics in Optics? . . . 1582
A.2.6 Ways to Guess Some Main Ideas . . . 1582
A.2.7 Effective Way of Studying Optics . . . 1582
A.3 Acknowledgements . . . 1583
A.4 Literature on Optics . . . 1583
A.4.1 Basic Literature . . . 1583
A.4.2 Extra Literature . . . 1584
A.4.3 Literature about Special Topics . . . 1585
A.5 Overview of Optics . . . 1586
A.5.1 Popular Introduction to Optics . . . 1586
A.5.2 Optics in Everyday Life . . . 1586
A.5.3 Relations of Optics with Other Sciences . . . 1587
A.6 History of Optics . . . 1588
A.7 Description of the Note . . . 1588
A.7.1 The Goal and Features of the Note . . . 1588
A.7.2 Designations and Constants . . . 1588
A.8 Puzzles in Optics for Different Situations . . . 1588
A.8.1 Fundamental Puzzles for Fun Discussion . . . 1588
A.8.2 Puzzles for Interesting Discussion . . . 1590
A.8.3 Puzzles About Everyday Life . . . 1590
A.8.4 Selected Fun Typical Problems . . . 1590
A.8.5 Principal Puzzles . . . 1590
A.8.6 Puzzles about Details . . . 1590
A.8.7 Selected Puzzles for Learning Subjects . . . 1590
A.8.8 Puzzles to Become Demotivated . . . 1590
A.8.9 Puzzles to Prove that a Person Doesn't Know the Subject . . . 1590

**B Some Physics for Optics** **1590**
B.0.1 Quantum Mechanics . . . 1590
B.0.2 Quantum Field Theory . . . 1590
B.0.3 Mechanics . . . 1591

**C Some Mathematics for Optics** **1591**
C.1 Harmonic Analysis . . . 1591
C.1.1 Fourier Transformation and Decomposition into Modes . . . 1591
C.1.2 Kotelnikov's Theorem . . . 1591
C.2 Integrals in Optics . . . 1591
C.2.1 Integrals in Radiation Theory . . . 1591
C.2.2 Fresnel Integrals . . . 1591
C.3 The Calculus of Variations . . . 1591
C.3.1 1. Euler's equations as necessary conditions for an extremum . . . 1592
C.3.2 2. Hilbert's independence integral and the Hamilton-Jacobi equation . . . 1593
C.3.3 3. The field of extremals . . . 1595

C.3.4 Extremals via Hamilton–Jacobi Equation . . . 1597
C.3.5 5. Hamilton's canonical equations . . . 1598
C.3.6 Integrand Without Explicit Variable Dependence . . . 1599
C.3.7 7. Discontinuities . . . 1600
C.3.8 Sufficiency conditions for an extremum . . . 1602
C.3.9 9. Minimum of the variational integral if an end point is on a surface . . . 1604
C.3.10 10. Jacobi's criterion for a minimum . . . 1605
C.3.11 11. Example I: Optics . . . 1605
C.3.12 12. Example II: Mechanics of material points . . . 1607
C.4 Light Optics, Electron Optics and Wave Mechanics . . . 1610
C.4.1 1. The Hamiltonian analogy in elementary form . . . 1610
C.4.2 2. The Hamiltonian analogy in variational form . . . 1612
C.4.3 3. Wave mechanics of free electrons . . . 1615
C.4.4 4. The application of optical principles to electron optics . . . 1617
C.5 Asymptotic Approximations to Integrals . . . 1618
C.5.1 1. The method of steepest descent . . . 1618
C.5.2 2. The method of stationary phase . . . 1623
C.5.3 3. Double integrals . . . 1624
C.6 The Dirac Delta Function . . . 1626
C.7 A Mathematical Lemma used in the Rigorous Derivation of the Lorentz-Lorenz Law (§ 2.4.2) . . . 1630
C.8 Propagation of Discontinuities in an Electromagnetic Field (§3.1.1) . . . 1632
C.8.1 1. Relations connecting discontinuous changes in field vectors . . . 1632
C.8.2 2. The field on a moving discontinuity surface . . . 1634
C.9 The Circle Polynomials of Zernike (§ 9.2.1) . . . 1635
C.9.1 1. Some general considerations . . . 1636
C.9.2 2. Explicit expressions for the radial polynomials $R_n^{\pm m}(\rho)$ . . . 1637
C.10 Proof of an Inequality (§ 10.7.3) . . . 1641
C.11 Evaluation of T'wo Integrals (§ 12.2.2)* . . . 1642
C.12 Other, Special Mathematical Methods . . . 1644
C.12.1 Compressed Measurements by Tao (?) . . . 1644

**D Programming and Modeling** **1644**

**E References** **1646**
E.0.1 References from Theoretical Articles . . . 1646
E.0.2 References from Articles about Applications . . . 1649

# Preface



Fundamentals of Optics

Fundamentals of Quantum Optics

Optics in Continuous Media

Optical Phenomena in Nature

Optical Equipment and Experiments

Other Topics and Applications

Dead, Useless, Pointless Theories in Optics

Appendix

I used the following literature.

Let's discuss some minimum knowledge and motivation that would be good to understand for studying optics.

### Результаты оптики (???)

(потом напишу)

### Best theoretical puzzles

#### m-th.? But what is a light?

(later I'll add statement)

**Instructions**

**Best simple technical puzzles**

**m-te.? Finally, half of wavelength in reflection**

(later I'll add statement)

**Instructions**

**Best technical puzzles**

**m-te.? Light near star**

(later I'll add statement)

**Instructions**

# Part I
# —— Typical Optics in a Nutshell ——

## 1 Overview of main laws and models

### 1.1 Typical optical laws and models in a nutshell

#### 1.1.1 Main laws and relations in optics

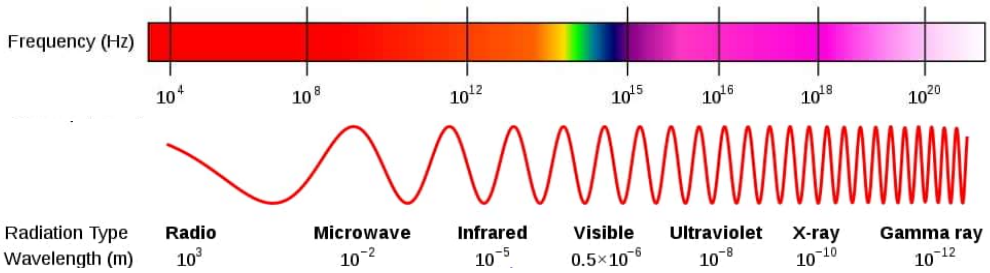
**Main parameters of light**

$$\lambda = \frac{v}{\nu} = vT \qquad v = \frac{\lambda}{T} = \lambda\nu \qquad \omega = 2\pi\nu = \frac{2\pi}{T}.$$

$$k = \frac{2\pi}{\lambda}, \qquad \nu = \frac{c}{2\pi}k,$$

$$p = \pm\hbar k = \pm\frac{h}{2\pi}k$$

$$E = |p|c = h\nu$$




Амплитуда
Интенсивность (!?!)
Фазовая и групповые скорости
Поляризация
(абзац про теорпольные свойства и появление поляризации)

**Погрешности оптических систем (??)**

(потом разберусь)

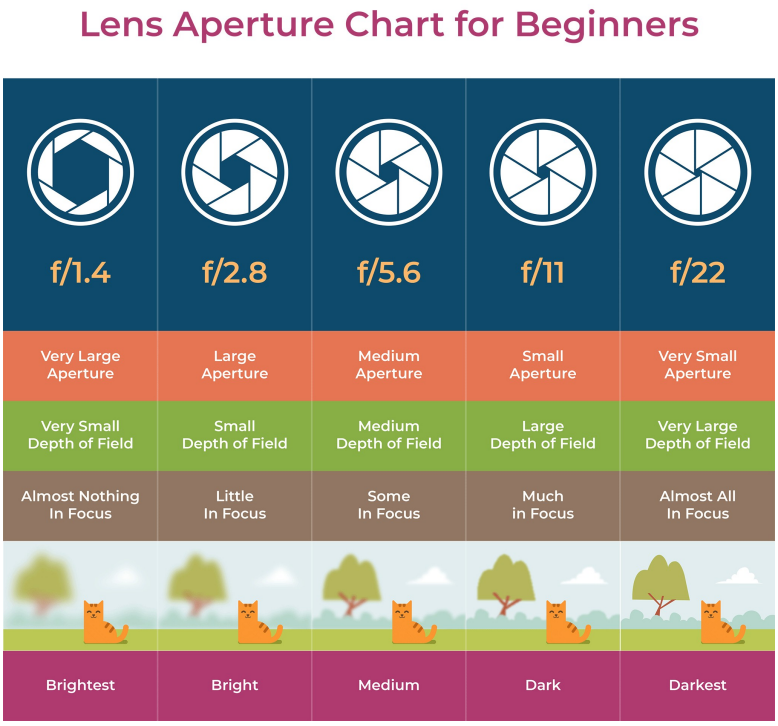
**Сферическая аберрация**

**Кома**

**Дисторсия**

**Хроматическая аберрация**

**Астигматизм**




## 1.1.2 Basics of geometrical optics and lenses

**Уравнение тонкой линзы (Формула тонкой линзы)**

$$D = \pm\frac{1}{F} = \pm\frac{1}{d} \pm \frac{1}{f},$$

***d*** - расстояние от предмета (П) до линзы, $f$ - расстояние от изображения (И) до линзы, $F - \phi$ кусное расстояние линзы, $D$ - оптическая сила линзы.

- Перед $\frac{1}{F}$ ставится знак $+$, если линза является собирающей, знак $-$, если линза является рассеивающей.

- Перед $\frac{1}{f}$ ставится знак $+$, если изображение является действительным, знак $-$, если изображение является мнимым.

- Перед $\frac{1}{d}$ ставится знак $+$, если предмет является действительным, знак $-$, если предмет является мнимым.

Для практического использования формулы линзы следует твердо запомнить правило знаков:

Для собирающей линзы, действительного источника и действительного изображения величины F, d, f считают положительными. Для рассеивающей линзы, мнимого источника и мнимого изображения — F, d, f считают отрицательными.

Заметим, что предмет или источник является мнимым только в том случае, если на линзу падает пучок сходящихся лучей.

Таким образом, линза с $F > 0$ является собирающей (положительной), а с $F < 0$ — рассеивающей (отрицательной).

Линейным (поперечным) увеличением Г называется отношение линейного размера изображения h' к линейному размеру предмета h. Из соотношения (3) находим линейное увеличение тонкой линзы

#### Особенности изображений в линзах

(чего особенного от 2го фокуса?)

#### Определения параметров линз

ГЛ опт ось -
полюс - точка на линзе
оптический центр - центр геометрический, где радиус

#### Определения свойств изображений

Мнимое / действительное -
Увеличенное / уменьшенное -
Прямое / перевернутое

#### Случай зеркала

(тут просто помним, что проходит луч дважды, так что умножаем на 2 силу линзы)

#### Способы решения задач геометрической оптики

(отработаю, если будет нужно.)
Дальше подставляем в формулу с правильными знаками расстояния (??)
получаются какие-то уравнения, которые должны давать нужный нам параметр.

#### Методы построения изображений

Проводим 2 луча.
Один параллельно оси линзы, другой к центру.
Дальше каждая линза по-своему направляет тот, который параллельно послан.
дальше не пересечении изображение и получается.

### толстой линзе (??!!!)

(??? потом напишу, пока вопросы про это, не научился еще с таким работать.)

### Фотометрия (!!!!!?)

#### Основные параметры фотометрии

(потом чуть что разберусь.)

Поток излуче

$$\Phi_{\text{e}} = \frac{W}{t}$$

Энергетическая сила света

$$I = \frac{d\Phi_{\text{e}}}{d\Omega}$$

Если источник изотропный (не зависит от направления): $I = \frac{\Phi_{\text{e}}}{4\pi}$ Энергетическая светимость

$$R_e = \frac{\Phi_{\text{e}}}{S}$$

Энергетическая яркость

$$B_e = \frac{I_e}{S}$$

Сила света - поток излучения источника, приходящийся на единицу телесного угла: $I = \frac{d\Phi}{d\Omega}, [I] = 1$ кд (кандела)

Световой поток - это поток лучистой энергии, оцениваемый по зрительному ощущению, $[\Phi] = 1$ лм (люмен)

Светимость - световой поток, испускаемый единицей поверхности наружу по всем направлениям в пределах угла $\left(0, \frac{\pi}{2}\right) : R = \frac{d\Phi_R}{dS}, [R] = \frac{\text{лм}}{\text{м}^2}$

Освещенность - величина, равная отношению светового потока, падающего на поверхность, к площади этой поверхности

$$E = \frac{d\Phi_f}{dS}, [E] = 1 \text{ лк (люкс)}$$

**Способы решения задач на фотометрию**

(прорешаю еще раз - напишу указания)

**Геометрическая оптика для олимпиад (??????)**

(прорешаю еще раз - напишу указания)

**Типичные характеристики оптических приборов**

(примеры приведу, чему они равны для той или другой установки.)

**угловая дисперсия** $D(\lambda) = \frac{d\varphi}{d\lambda}$

$$D = \frac{d\varphi}{d\lambda}$$

показывает, насколько с увеличением длины волны лучи будут шире отклоняться. (??)

Угловая дисперсия позволяет определить минимальное расстояние между ячейками приёмного устройства: если требуется разрешить две спектральные линии с разностью длин волн $\delta\lambda$, то расстояние между элементами приемного устройства должно быть заметно меньше $D\delta\lambda f$, $f$ - фокусное расстояние объектива зрительной трубы.

Сравним угловую дисперсию для амплитудной, фазовой дифракционной решёток, интерферометра Фабри-Перо и призмы. Возьмем $\lambda = 578$ нм (средняя длина волны для двух близких жёлтых линий ртути). Амплитудная дифракционная решётка с

плотностью штрихов $n = 5001$ мм для порядков спектра $m = 1, 2, 3$ имеет в соответствии с (4.11) угловую дисперсию $5,2 \cdot 10^3; 1,2 \cdot 10^4; 3 \cdot 10^4$ см$^{-1}$. Эшелетт для угла падения $\psi = 30^\circ$ имеет только один порядок спектра $m = 1$.

(??)

**разрешающая способность**

$$\frac{\lambda}{\delta\lambda}$$

характеризует возможность прибора различать две близкие спектральные линии с длинами волн $\lambda$ и $\lambda + \delta\lambda$.

**дисперсионная область**

$$\Delta\lambda$$

- самая малая разность между ближайшими длинами волн, при которой дифракционные максимумы от каждой не перекрываются.

**Построение изображений в линзах**

(тут про это все указания)

**Examples of lens systems**

**Тонкая линза**

$$\frac{1}{x} - \frac{1}{x'} = \frac{1}{f}$$

$$\frac{1}{f'} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)$$

(add picture, words)

*Proof.* $n = n', f = -f'$,

$$\frac{f - a'}{f} = \frac{f}{a + f} = -\frac{X'}{f} = \frac{f}{X} = \frac{y'}{y} = \frac{\alpha}{\alpha'}.$$

$$-\frac{1}{a} + \frac{1}{a'} = \frac{1}{f}, \quad XX' = -f^2$$

□

**Телескопическая система**

**Толстая линза**

(картинку добавлю)

$$f_1 = -F_1 = \frac{R_1}{n-1}, \qquad f_1' = -F_1' = -\frac{nR_1}{n-1}$$
$$f_2 = -F_2 = -\frac{nR_2}{n-1}, \qquad f_2' = -F_2' = \frac{R_2}{n-1}$$

Пусть $d$ - толщина линзы, при этом оптический интервал

$$\Delta = d + f_1' - f_2 = d - \frac{n}{n-1}\left(R_1 - R_2\right).$$

С учётом полученных выражений по формулам (1.28), (1.29) находим фокусные расстояния и оптическую силу толстой линзы как сложной системы, а также координаты её главных и фокальных точек:

$$f_{1\Sigma} = -f_{2\Sigma} = \frac{1}{\Phi} = \frac{nR_1R_2}{(n-1)^2\Delta}$$
$$H_{i\Sigma} = -\frac{R_i d}{(n-1)\Delta}, \quad F_{i\Sigma} = H_{i\Sigma} + f_{i\Sigma}$$

Если радиусы $R_i$ различны и выполняется неравенство $d \ll \frac{n}{n-1}\left|R_1 - R_2\right|, qrq$ в этом приближении «не очень толстой» линзы

$$\Phi_\Sigma = \frac{1}{f_{1\Sigma}} = -\frac{1}{f_{2\Sigma}} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = -\frac{1}{s} + \frac{1}{s'}$$
$$H_{i\Sigma} = \frac{R_i d}{n\left(R_1 - R_2\right)}, \quad \delta \equiv d + H_{2\Sigma} - H_{1\Sigma} = \frac{n-1}{n}d$$

**Система из двух тонких линз**

$$H_{i\Sigma} = (-1)^i f_i \ell_{12}/\Delta$$
$$F_{i\Sigma} = (-1)^i f_i\left(1 + f_i/\Delta\right),$$
$$f_{i\Sigma} = (-1)^i f_1 f_2/\Delta.$$
$$\Delta = \ell_{12} - f_1 - f_2$$
$$\delta = \ell_{12} + H_{2\Sigma} - H_{1\Sigma} = \ell_{12}^2/\Delta.$$

Для оптической силы системы из формул (1.29), (1.38), (1.39) получаем выражения

$$\Phi_\Sigma = -\frac{1}{f_{2\Sigma}} = -\frac{\Delta}{f_1 f_2} = -\frac{\ell_{12}}{f_1 f_2} + \frac{1}{f_1} + \frac{1}{f_2}.$$

**Лупа**

$$N_L \equiv \frac{\varphi'}{\varphi} = \frac{y'}{y}\frac{L}{a - X'} = \frac{L}{f}\left(\frac{n'}{n} + \frac{aX}{f^2}\right)$$

$n = n' = 1$,

$$N_L = \frac{L/f}{1 + aX/f^2}.$$

$$N_L = \frac{y'}{y} = \frac{f}{X} = \frac{L - a}{f}.$$

**Микроскоп**

(картинку добавлю)

$$f_M = -\frac{f_1 f_2}{\Delta}, \quad f'_M = \frac{f'_1 f'_2}{\Delta},$$

$\Delta$ - оптический интервал, величина которого зависит от устройства и взаимного расположения объектива и окуляра.

Для увеличения безиммерсионого микроскопа ($n = n' = 1$), для которого $f_1 = -f'_1 > 0, f_2 = -f'_2 > 0$, из формул (1.44), (1.45) получаем выражения

$$N_M = \frac{L}{f_M} = -\frac{\Delta}{f_1}\frac{L}{f_2}, \quad \Delta = \ell_{12} - f_1 - f_2.$$

Величину $N_M$ можно представить в виде

$$N_M = N_1 N_2, \quad N_1 = -\frac{\Delta}{f_1}, \quad N_2 = \frac{L}{f_2},$$

$N_1$ - увеличение объектива, а $N_2$ - дополнительное увеличение, даваемое окуляром как лупой.

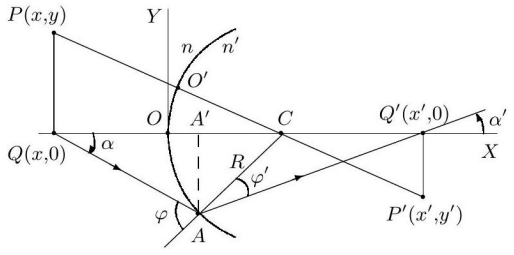
**Зеркальные трубы Кеплера и Галилея**

(картинки добавлю)

**Geometrical optics from labnik**

**1.**

**1.1.**




1.1 $\sin\xi \approx \xi \cos\xi \approx 1$
$\xi \ll 1$

$$QA \approx QA' \approx QO = -x$$
$$AQ' \approx A'Q' \approx OQ' = x' \quad A'C \approx OC = R \tag{1.1}$$

$$n\varphi = n'\varphi' \quad x\alpha = x'\alpha' = R(\varphi + \alpha) = R\left(\varphi' + \alpha'\right) \tag{1.2}$$

$$-\frac{n}{x} + \frac{n'}{x'} = \frac{n' - n}{R} \tag{1.3}$$

$$x\alpha = x'\alpha' \tag{1.4}$$

$$\frac{y'}{y} = \frac{x' - R}{x - R} \tag{1.5}$$

$$\frac{1}{x} + \frac{1}{x'} = \pm\frac{2}{R} \tag{1.6}$$

$$x = \tilde{x} - \tilde{x}_0 \quad x' = \tilde{x}' - \tilde{x}'_0 \quad y = \tilde{y} \quad y' = \tilde{y}' \tag{1.7}$$

$$\frac{x' - F'}{H' - F'} = \frac{H - F}{x - F} = \frac{y'}{y} = \frac{n\alpha}{n'\alpha'} \tag{1.8}$$

$$(x - H)\alpha = \left(x' - H'\right)\alpha' \tag{1.9}$$

$$F = x_0 - \frac{nR}{n' - n} \quad F' = x'_0 + \frac{n'R}{n' - n} \quad H = x_0 \quad H' = x'_0 \tag{1.10}$$

**1.2.**

$$x - H = 0 \quad x' - H' = 0 \tag{1.11}$$

$$x_H = H \quad x'_H = H' \tag{1.12}$$

$$x - F = 0 \quad x' - F' = 0 \tag{1.13}$$

$$x_F = F \quad x'_F = F' \tag{1.14}$$

$$f \equiv H - F \quad f' \equiv H' - F' \tag{1.15}$$

$$f = -F = \frac{nR}{n' - n} \quad f' = -F' = -\frac{n'R}{n' - n} \tag{1.16}$$

$$ny\alpha = n'y'\alpha' \tag{1.17}$$

$$\frac{n}{f} = -\frac{n'}{f'} \equiv \Phi \tag{1.18}$$

$$x_K = K \equiv H - (f + f') \quad x'_K = K' \equiv H' - (f + f') \tag{1.19}$$

$$\frac{\delta x'}{\delta x} = \frac{\alpha}{\alpha'}\frac{y'}{y} = \frac{n'}{n}\left(\frac{y'}{y}\right)^2 \tag{1.20}$$

1. $s \equiv x - Hs' \equiv x' - H'$.

$$\frac{s'+f'}{f'} = \frac{f}{s+f} = \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} \tag{1.21}$$

$$\frac{f}{s} + \frac{f'}{s'} = -1 \tag{1.22}$$

2. $X \equiv x - FX' \equiv x' - F'$.

$$\frac{X'}{f'} = \frac{f}{X} = \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} \tag{1.23}$$

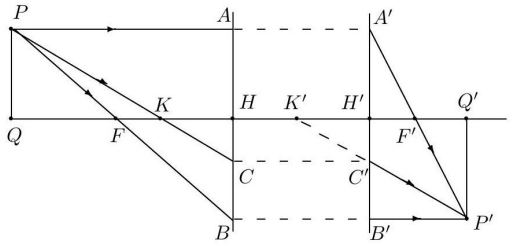
$n = n'$

$$-\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \quad XX' = -f^2 \tag{1.24}$$

$n = -n'f = f'$

**1.3.**

$(y'/y = -1)$ $(\alpha'/\alpha = -1)$

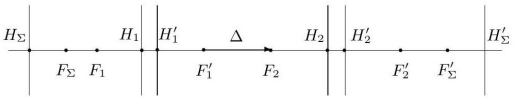



**Fermat's principle**

(this is a more philosophical one)

**2.**

$\Delta \equiv F_2 - F_1'$
$\Delta > 0$
$\Delta < 0$.




$F_1$ $F_2'$ $x = x_1 y = y_1$

$$\frac{x_1'}{f_1'} = \frac{f_1}{x} = \frac{y_1'}{y} \tag{1.25}$$

$f_1 = H_1 - F_1 f_1' = H_1' - F_1'$
$x_2 = x_1' - \Delta$
$y_2 = y_1'$.
$F_2' x_2' = x' y_2' = y'$

$$\frac{x'}{f_2'} = \frac{f_2}{x_1' - \Delta} = \frac{y'}{y_1'} \tag{1.26}$$

$f_2 = H_2 - F_2 f_2' = H_2' - F_2'$ -
$x_1' y_1'$

$$\frac{x' + f_2 f_2'/\Delta}{f_1' f_2'/\Delta} = \frac{-f_1 f_2/\Delta}{x - f_1 f_1'/\Delta} = \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} \tag{1.27}$$

$$\begin{array}{lll} f_\Sigma = -\frac{f_1 f_2}{\Delta} & H_\Sigma = \frac{f_1(f_1' - f_2)}{\Delta} & F_\Sigma = \frac{f_1 f_1'}{\Delta} \\ f_\Sigma' = \frac{f_1' f_2'}{\Delta} & H_\Sigma' = \frac{f_2'(f_1' - f_2)}{\Delta} & F_\Sigma' = -\frac{f_2 f_2'}{\Delta} \end{array} \tag{1.28}$$

$$\Phi_\Sigma = \frac{n}{f_\Sigma} = -\frac{n\Delta}{f_1 f_2} \tag{1.29}$$

**3.**

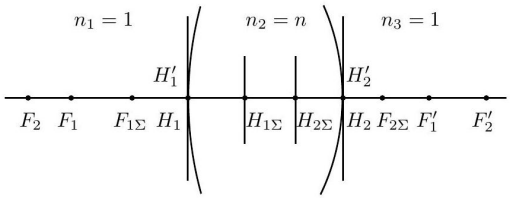
**3.1.**

$\Delta \to 0$.

$$\frac{x'}{x} = \frac{\delta x'}{\delta x} = \frac{f_2 f_2'}{f_1 f_1'} \quad \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} = \frac{f_2}{f_1'} \tag{1.30}$$

**3.2.**

$n_2 = n$.
$n_1 = n_3 = 1$.




$$\begin{array}{ll} f_1 = -F_1 = \frac{R_1}{n-1} & f_1' = -F_1' = -\frac{nR_1}{n-1} \\ f_2 = -F_2 = -\frac{nR_2}{n-1} & f_2' = -F_2' = \frac{R_2}{n-1} \end{array} \tag{1.31}$$

$$\Delta = d + f_1' - f_2 = d - \frac{n}{n-1}(R_1 - R_2) \tag{1.32}$$

$$\begin{array}{l} f_{1\Sigma} = -f_{2\Sigma} = \frac{1}{\Phi} = \frac{nR_1R_2}{(n-1)^2\Delta} \\ H_{i\Sigma} = -\frac{R_i d}{(n-1)\Delta} \quad F_{i\Sigma} = H_{i\Sigma} + f_{i\Sigma} \end{array} \tag{1.33}$$

$$d \ll \frac{n}{n-1}|R_1 - R_2|$$

$$\Phi_\Sigma = \frac{1}{f_{1\Sigma}} = -\frac{1}{f_{2\Sigma}} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = -\frac{1}{s} + \frac{1}{s'} \tag{1.34}$$

$$H_{i\Sigma} = \frac{R_i d}{n(R_1 - R_2)} \quad \delta \equiv d + H_{2\Sigma} - H_{1\Sigma} = \frac{n-1}{n} d \tag{1.35}$$

**3.3.**

$n = n'f = -f'$

$$\frac{f - a'}{f} = \frac{f}{a + f} = -\frac{X'}{f} = \frac{f}{X} = \frac{y'}{y} = \frac{\alpha}{\alpha'} \tag{1.36}$$

$$-\frac{1}{a} + \frac{1}{a'} = \frac{1}{f} \quad XX' = -f^2 \tag{1.37}$$

**3.4.**

$f'_{12} = -f_{12}$.

$$\begin{aligned} H_{i\Sigma} &= (-1)^i f_i \ell_{12}/\Delta \\ F_{i\Sigma} &= (-1)^i f_i \left(1 + f_i/\Delta\right) \\ f_{i\Sigma} &= (-1)^i f_1 f_2/\Delta \end{aligned} \tag{1.38}$$

$$\begin{aligned} \Delta &= \ell_{12} - f_1 - f_2 \\ \delta &= \ell_{12} + H_{2\Sigma} - H_{1\Sigma} = \ell_{12}^2/\Delta \end{aligned} \tag{1.39}$$

$$\Phi_\Sigma = -\frac{1}{f_{2\Sigma}} = -\frac{\Delta}{f_1 f_2} = -\frac{\ell_{12}}{f_1 f_2} + \frac{1}{f_1} + \frac{1}{f_2} \tag{1.40}$$

$\Delta = 0$
$\ell_{12} = 0$

**4.**

**4.1.**

$y'\ \varphi' = y'/\left(a - X'\right)$.
$\varphi = y/L$.

$$N_L \equiv \frac{\varphi'}{\varphi} = \frac{y'}{y}\frac{L}{a - X'} = \frac{L}{f}\left(\frac{n'}{n} + \frac{aX}{f^2}\right)^{-1} \tag{1.41}$$

$n = n' = 1$

$$N_L = \frac{L/f}{1 + aX/f^2} \tag{1.42}$$

$a = 0\ N_L = L/f$.
$X = 0$.
$a - X' = L$

$$N_L = \frac{y'}{y} = \frac{f}{X} = \frac{L - a}{f} \tag{1.43}$$

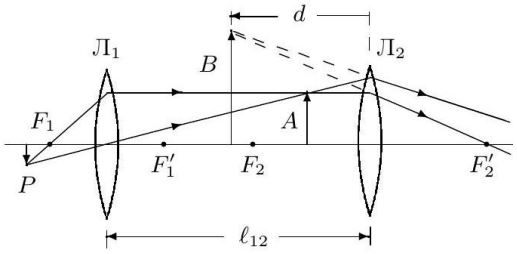
$a = -f$
$N_L = (L/f) + 1$.
$|X|/f \ll 1$

$|a|/f \leqslant 1$
$L/f \gg 1.$
$n = n' = 1$

$$N_L = \frac{L}{f} \tag{1.44}$$

$N_L = L/f$
$n = n' = 1$

**4.2.**



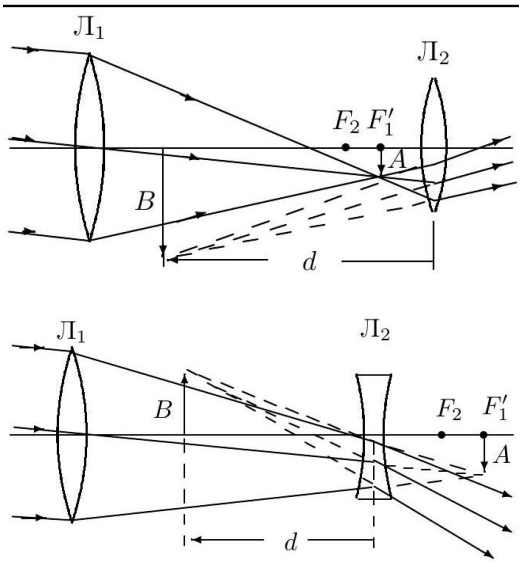

$$f_M = -\frac{f_1 f_2}{\Delta} \quad f'_M = \frac{f'_1 f'_2}{\Delta} \tag{1.45}$$

$(n = n' = 1)$
$f_1 = -f'_1 > 0 f_2 = -f'_2 > 0$

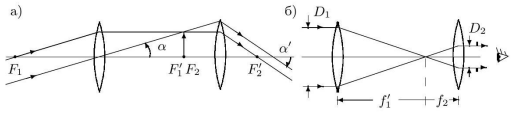
$$N_M = \frac{L}{f_M} = -\frac{\Delta}{f_1}\frac{L}{f_2} \quad \Delta = \ell_{12} - f_1 - f_2 \tag{1.46}$$

$$N_M = N_1 N_2 \quad N_1 = -\frac{\Delta}{f_1} \quad N_2 = \frac{L}{f_2} \tag{1.47}$$

**4.3.**



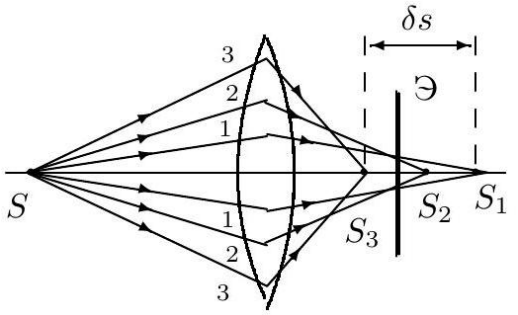

$$N_T \equiv \frac{\alpha'}{\alpha} = \frac{y}{y'} = \frac{f'_1}{f_2} = -\frac{f_1}{f_2} \tag{1.48}$$

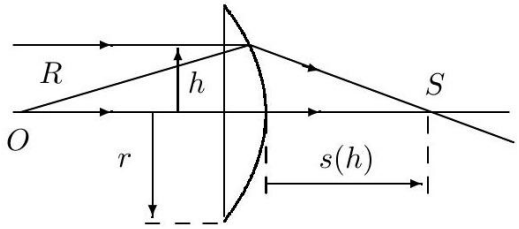



$$N_T = \mp\frac{D_1}{D_2} \tag{1.49}$$

## 5. Aberration

### 5.1. Spherical aberration




$$s(h) = \frac{R}{n-1}\left(1 - \frac{n^2h^2}{2R^2}\right) \tag{1.50}$$

$h \ll R$

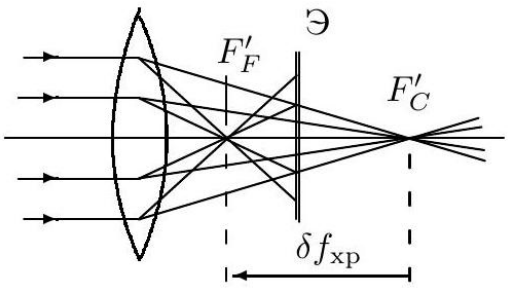
$$s(h)_{h\to 0}\frac{R}{n-1} \tag{1.51}$$

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{1.52}$$

$R_1 \to \infty$
$R_2 < 0.$

$$\delta s(h) = s(h) - s(0) = -\frac{1}{2}\frac{n^2h^2}{(n-1)R} = -\frac{1}{2}\left(\frac{n}{n-1}\right)^2\left(\frac{h}{f}\right)^2 f \tag{1.53}$$

### 5.2. Chromatic aberration




$\lambda_F = 4861$
$\lambda_C = 6563$

$$\delta f_{\text{xp}} = f_F - f_C \tag{1.54}$$

$$\nu = \frac{n_D - 1}{n_F - n_C} \tag{1.55}$$

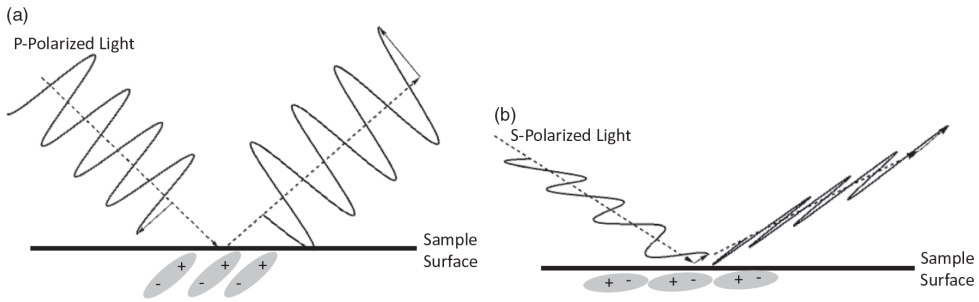
$\lambda_D = 5893$

$$\delta f_{\text{xp}} = -\frac{(n_D - 1)(n_D - 1)}{(n_F - 1)(n_C - 1)} \frac{f_D}{\nu} \approx -\frac{f_D}{\nu}. \tag{1.56}$$

$n_F - n_C \sim 10^{-2}$
$n_D$ $n_F$ $n_C$ 2%.
$f_F = f_C$

## 1.1.3 Reflection and transmission of light

**Fresnel equations**




for s-polarization:

$$r_\perp = -\frac{\sin(\theta - \theta'')}{\sin(\theta + \theta'')}, d_\perp = \frac{2\sin\theta''\cos\theta}{\sin(\theta + \theta'')}.$$

$n_{21} = n_2/n_1$

$$r_\perp = \frac{\cos\theta - \sqrt{n_{21}^2 - \sin^2\theta}}{\cos\theta + \sqrt{n_{21}^2 - \sin^2\theta}}, d_\perp = \frac{2\cos\theta}{\cos\theta + \sqrt{n_{21}^2 - \sin^2\theta}}$$

for p-polarization:

$$r_\parallel = -\frac{\operatorname{tg}(\theta - \theta'')}{\operatorname{tg}(\theta + \theta'')}, d_\parallel = \frac{2\sin\theta''\cos\theta}{\sin(\theta + \theta'')\cos(\theta + \theta'')}.$$

$$r_{\parallel 1} = -\frac{n_{21}^2 \cos\theta - \sqrt{n_{21}^2 - \sin^2\theta}}{n_{21}^2 \cos\theta + \sqrt{n_{21}^2 - \sin^2\theta}}, d_{\parallel} = \frac{2n_{21}\cos\theta}{n_{21}^2 \cos\theta + \sqrt{n_{21}^2 - \sin^2\theta}}.$$

(???)

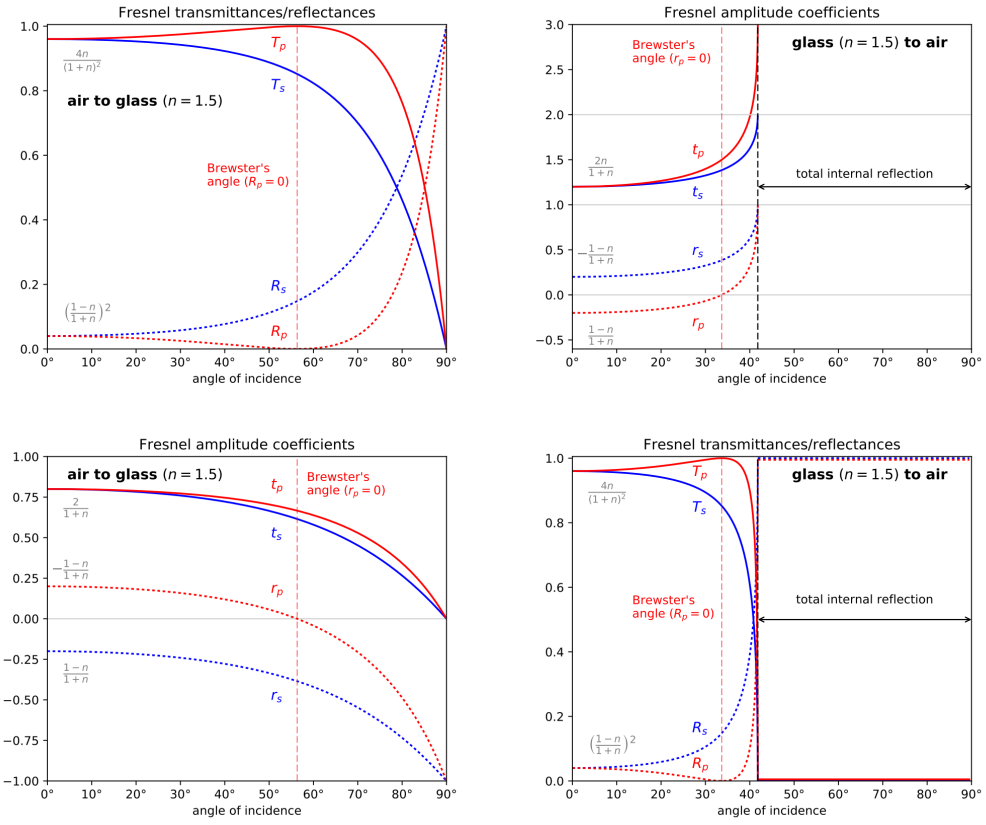
$$t_\mathrm{s} = \frac{2n_1\cos\theta_1}{n_1\cos\theta_1 + n_2\cos\theta_2} \qquad t_\mathrm{p} = \frac{2n_1\cos\theta_1}{n_1\cos\theta_2 + n_2\cos\theta_1}$$

$$r_\mathrm{s} = \frac{n_1\cos\theta_1 - n_2\cos\theta_2}{n_1\cos\theta_1 + n_2\cos\theta_2} \qquad r_\mathrm{p} = \frac{n_1\cos\theta_2 - n_2\cos\theta_1}{n_1\cos\theta_2 + n_2\cos\theta_1}$$

$$R_\mathrm{s} = \left|\frac{n_1\cos\theta_\mathrm{i} - n_2\cos\theta_\mathrm{t}}{n_1\cos\theta_\mathrm{i} + n_2\cos\theta_\mathrm{t}}\right|^2 = \left|\frac{n_1\cos\theta_\mathrm{i} - n_2\sqrt{1-\left(\frac{n_1}{n_2}\sin\theta_\mathrm{i}\right)^2}}{n_1\cos\theta_\mathrm{i} + n_2\sqrt{1-\left(\frac{n_1}{n_2}\sin\theta_\mathrm{i}\right)^2}}\right|^2$$

$$R_\mathrm{p} = \left|\frac{n_1\cos\theta_\mathrm{t} - n_2\cos\theta_\mathrm{i}}{n_1\cos\theta_\mathrm{t} + n_2\cos\theta_\mathrm{i}}\right|^2 = \left|\frac{n_1\sqrt{1-\left(\frac{n_1}{n_2}\sin\theta_\mathrm{i}\right)^2} - n_2\cos\theta_\mathrm{i}}{n_1\sqrt{1-\left(\frac{n_1}{n_2}\sin\theta_\mathrm{i}\right)^2} + n_2\cos\theta_\mathrm{i}}\right|.$$

Plots for glass-air transition (see wiki)




**задачах про отражение**

Проекции на нормаль вектора Пойтинга:

$$S_n^{(e)} + S_n^{(r)} = S_n^{(d)},$$

$e$ - падающий пучек, $r$ - отраженный, $d$ - преломленный пучки. Оно же:

$$I^{(e)}\cos\varphi - I^{(r)}\cos\varphi = I^{(d)}\cos\psi$$

**Закон Снеллиуса**

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

как кст доказать его?

**Полное внутреннее отражение**

есть такое и я хз.

**Особенности задач про элементарное прохождение света через вещество**

**Элементарные отражения (!?!?)**

Закон Снелиуса:
(также может добавляться половина длины волны. (???) также нужно указать, почему это так?)

### 1.1.4 Dispersion, Doppler effect

**дисперсии**

**Doppler effect**

$$f = \left(1 + \frac{\Delta v}{c}\right) f_0$$

(add from field theory about it!)

**Формула Лорентц-Лоренца**

$$\frac{n^2 - 1}{n^2 + 2} = \frac{4\pi}{3} N\alpha$$

(add from electrodynamics about it)

### 1.1.5 Interference, coherence

**Interference from two monochromatic waves**

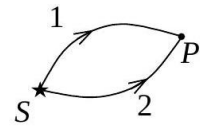
$$I := \overline{E^2} = \frac{1}{\Delta t}\int\limits_{\Delta t} E^2(t)dt \sim \frac{1}{\Delta t}\int\limits_{\Delta t} a^2(t)dt = \overline{a^2(t)}$$

если работаем с комплексной амплитудой $A(t)$ сигнала, записанном в комплексной форме $V(t) = a(t)e^{-i[\omega_0 t - \varphi(t)]} = A(t)e^{-i\omega_0 t}$, то переход такой:

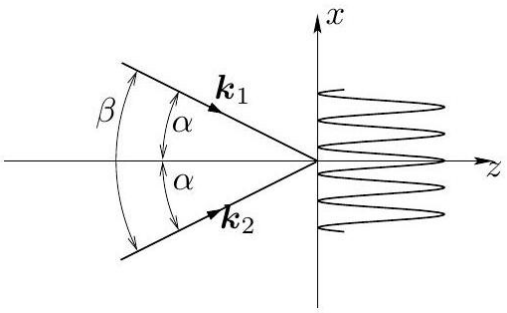
$$I = \overline{A(t)A^*(t)} = \overline{a^2(t)}$$

$$V := \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$ видность

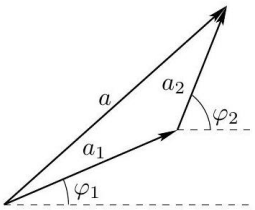
### 2.2.2. Plane waves



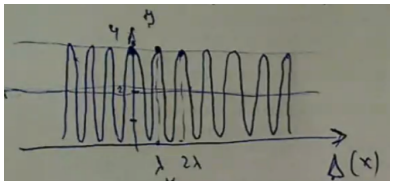

$\tau = \Delta/c$




$$E_1 = a_0 \cos(\omega t - kx \sin\alpha);$$
$$E_2 = a_0 \cos(\omega t + kx \sin\alpha),$$




$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Delta\varphi; \qquad I = 2I_0(1+\cos\Delta\varphi) \tag{2.18}$$



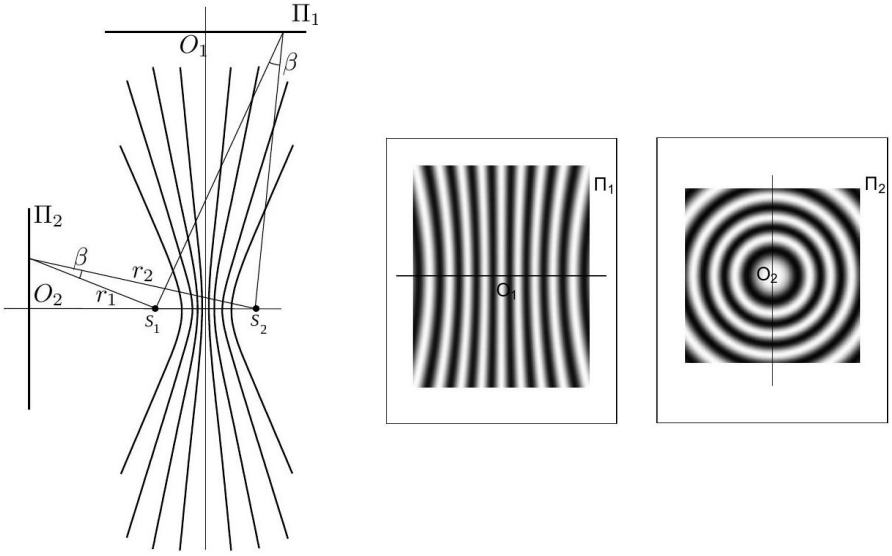

$I_{\max} = 4I_0; \quad I_{\min} = 0$

$$\varphi_2 - \varphi_1 := \frac{\omega}{c}\underbrace{(n_2 r_2 - n_1 r_1)}_{:=\Delta} = \frac{\omega}{c}\cdot\Delta \qquad k : \frac{\omega}{c} = \frac{2\pi}{\lambda} \tag{2.19}$$

$$I = 2I_0\big[1+\cos(\underbrace{2kx\sin\alpha}_{\text{step: } 2k\Delta x\sin\alpha = 2\pi})\big] \tag{2.23}$$

$$\Delta x = l = \frac{\lambda}{2\sin\frac{\beta}{2}} \overset{\sin\frac{\beta}{2}\approx\frac{\beta}{2}}{\approx} \frac{\lambda}{\beta} \qquad \text{interferece width} \tag{2.24}$$

### 2.2.1. Spherical waves




$\Delta\varphi = k\Delta$
$k = \frac{\omega}{c}$

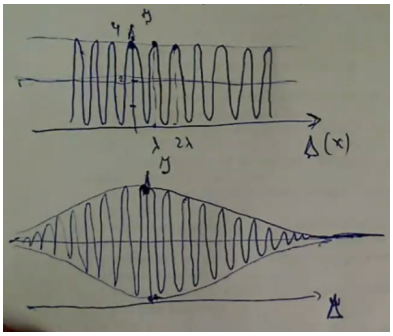
| | | |
|---|---|---|
| max | $\Delta = m\lambda$ | $\Delta\varphi = 2\pi m$ |
| min | $\Delta = m\lambda + \dfrac{\lambda}{2}$ | $\Delta\varphi = (2m+1)\pi$ |

$r_2 - r_1 = \text{const}$ (hyperbolas)
$r_2 - r_1 = m\lambda$

## Basics of time coherence

### Main properties of time coherence

The problem is that there is no one frequency in the wave from the source.




$$l_{\text{coh}} = N\lambda_{\text{coh}}$$

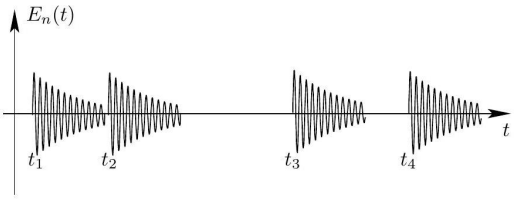
Our eye averages over $\sim 0.1$ s, typical light has coherence over $\sim 10^{-8}$ s, thus we don't see interference, because all light comes non-coherent.

We also assume that polarization of waves is not perpendicular, otherwise there will be no interference at all anyway.

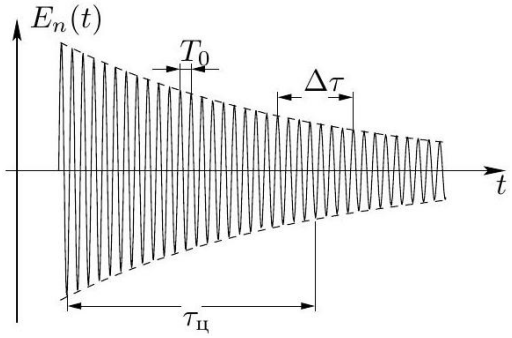
(??? what are the conditions when all is coherent temporally??)

### 3.1. Reasons for time coherence

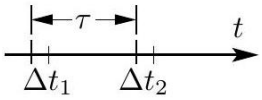
For a chaotic gases with a lot of spectra lines, there is a concept of pulses:




$$E(t) = \sum_n E_n(t) \qquad T_0 = \frac{2\pi}{\omega_0} \simeq 10^{-15}\text{c} \qquad \tau_\text{ц} \approx 10^{-8}$$




$$T_0 \ll \Delta\tau_\text{harmonic} \ll \tau_\text{ц}$$
$$E(t) = a(t)\cos\left[\omega_0 t - \varphi(t)\right]$$




The time $\tau_0$ during which the values of the amplitude $a\left(t+\tau_0\right)$ and the phase $\varphi\left(t+\tau_0\right)$ "forget" the previous values of $a(t)$ and $\varphi(t)$, is called coherence time (or the statistical memory time of the process).

$\Delta t \gg \tau_\text{ц}$ $\qquad \overline{e^{i\Delta\varphi(t)}} = \overline{\cos\Delta\varphi(t)} + i\overline{\sin\Delta\varphi(t)} = 0 \qquad \Delta\varphi(t) := \varphi(t+\tau) - \varphi(t)$

$\tau \ll \tau_\text{ц}$ $\qquad a(t) = a(t+\tau) \quad \varphi(t) = \varphi(t+\tau)$

$\Delta\omega \cdot t \sim \pi \qquad \Rightarrow \qquad$ no coherence

Mechanism of increasing of width:

$$\Delta\omega_\text{ц} \sim \frac{2\pi}{\tau_\text{ц}}. \tag{2.30}$$

many atoms $\Rightarrow$ collisions of atoms

temperature $\Rightarrow$ Doppler effect, $\omega = \omega_0\left(1 + \frac{v_n}{c}\cos\theta_n\right), \quad \omega_\text{Dopl} = v_\text{therm}/c$

atoms are not like oscillators $\Rightarrow$ very different lifetime of levels in atom

induced radiation in lasers $\Rightarrow$ time coherence can be up to $10^7$ times more than for the case of random radiation of atomic ensembles

### 3.4. Impact of coherence on interference

First, we introduce parameters for interference and description of coherence:

$$\bar{S} := \frac{c}{4\pi}\overline{E^2} \qquad I := \overline{E^2}$$

$$I = \frac{1}{\Delta t}\int\limits_{\Delta t} E^2(t)dt \sim \frac{1}{\Delta t}\int\limits_{\Delta t} a^2(t)dt = \overline{a^2(t)} \tag{2.31}$$

$$E(t) \equiv \operatorname{Re} V(t), \qquad A(t) = a(t)e^{i\varphi(t)}, \qquad V(t) = a(t)e^{-i[\omega_0 t - \varphi(t)]} = A(t)e^{-i\omega_0 t} \tag{2.32}$$

$$I = \overline{A(t)A^*(t)} = \overline{a^2(t)} \tag{2.33}$$

Function of time coherence (autocorrelation function):

$$\Gamma(\tau) := \overline{V(t) \cdot V^*(t+\tau)} = \frac{1}{\Delta t} \int_t^{t+\Delta t} V(t) \cdot V^*(t+\tau) d\tau \tag{2.34}$$

$\Delta t_{\text{human}} \sim 0.1$ s
$\Delta t_{\text{photo det}} = 10^{-3} - 10^{-5}$ s
always $\Delta t_{\text{photo det}} \gg T_0$

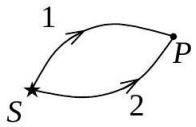
$$\Gamma(\tau) = \overline{A(t)A^*(t+\tau)}e^{i\omega_0\tau} = \overline{a(t)a(t+\tau)e^{i[\varphi(t)-\varphi(t+\tau)]}}e^{i\omega_0\tau} \tag{2.35}$$

degree of temporal coherence:

$$\gamma(\tau) := \frac{\Gamma(\tau)}{I_0} \tag{2.36}$$

$$\Gamma(\tau : T_0 \ll \tau \ll \tau_0) = \overline{a^2(t)}e^{i\omega_0\tau} = I_0 e^{i\omega_0\tau} \qquad \gamma(\tau) := e^{i\omega_0\tau}; \qquad |\gamma(\tau)| = 1$$

Now, we write the total intensity for the case:

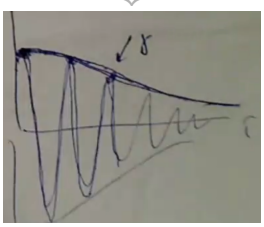



$\tau = \Delta/c$

$$I(\tau) = \overline{|V(t) + V(t+\tau)|^2} = \overline{|V(t)|^2} + \overline{|V(t+\tau)|^2} + 2\operatorname{Re}\overline{V(t)V^*(t+\tau)} \tag{2.38}$$

$I_0 := \overline{|V(t)|^2} = \overline{|V(t+\tau)|^2}$

$$I(\tau) = 2I_0\big(1 + \underbrace{|\gamma(\tau)|\cos\omega_0\tau}\big) \tag{2.39}$$



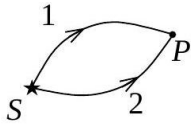
$$\text{visibility} \qquad \mathcal{V} := \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = |\gamma(\tau)| \tag{2.40}$$

Visibility (defined from the plow of intensity) is the same as time-coherence function (defined from averaging of the field). We can see it in the plot by subtracting and adding intensities and we will get the enveloping curve.

$\tau \ll \tau_0 \quad \Rightarrow \quad \gamma(\tau) \approx e^{i\omega_0\tau} \quad \Rightarrow \quad |\gamma(\tau)| = 1 \quad \Rightarrow \quad I(\tau) = 2I_0\,(1 + \cos\omega_0\tau)$
$\tau \geq \tau_0 \quad \Rightarrow \quad |\gamma(\tau)| = 0 \quad \Rightarrow \quad I(\tau) = 2I_0$
coherence length:

$$\Delta_{\max} \simeq c\tau_0 \tag{2.41}$$

### 3.5. Spectral properties



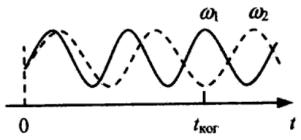

non-coherent

$I = \overline{|a_1 e^{i\omega t} + a_2 e^{i(\omega+\Delta\omega)t}|^2} \quad \Rightarrow \quad I = I_1 + I_2 + 2\sqrt{I_1 I_2}\overline{\cos\Delta\omega t} \quad \overset{\Delta t \gg 2\pi/\Delta\omega}{\to} \quad I = I_1 + I_2$

two-ray interference in terms of spectral intensity $J(\omega)$:

$$I_0 \equiv \int_0^{+\infty} J(\omega)d\omega \qquad \text{(Pythagor theorem, } a \sim \cos(\omega t - kx)\text{)}$$

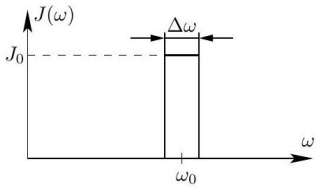
$$dI \overset{\text{from each}}{=} 2J(\omega)d\omega(1+\cos\omega\tau); \qquad I(\tau) = 2\int_0^{+\infty} J(\omega)(1+\cos\omega\tau)d\omega;$$

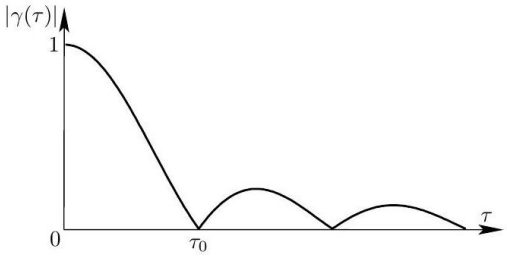



and for two waves the picture will be later restored. But it will not be restored for many lines. This is bad explained by "wave pulses".

Example 1:

$$J(\omega) = \begin{cases} J_0 & |\omega - \omega_0| < \frac{\Delta\omega}{2} \\ 0 & |\omega - \omega_0| > \frac{\Delta\omega}{2} \end{cases}$$




$$I(\tau) = 2\int_{\omega_0-\frac{\Delta\omega}{2}}^{\omega_0+\frac{\Delta\omega}{2}} J_0(1+\cos\omega\tau)d\omega = 2I_0\left[1 + \frac{\sin\frac{\Delta\omega\tau}{2}}{\frac{\Delta\omega\tau}{2}}\cos\omega_0\tau\right] \tag{2.44}$$

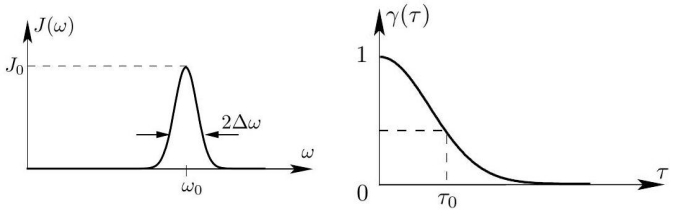
$I_0 = J_0\Delta\omega$

$$\mathcal{V} = |\gamma(\tau)| = \left|\frac{\sin\frac{\Delta\omega\tau}{2}}{\frac{\Delta\omega\tau}{2}}\right| \tag{2.45}$$

$$\tau_0: \qquad \frac{\Delta\omega\tau_0}{2} = \pi; \qquad \tau_0\Delta\omega = 2\pi \tag{2.46}$$

Example 2:

$$J(\omega) = J_0 e^{-\frac{(\omega-\omega_0)^2}{(\Delta\omega)^2}}$$




$$I(\tau) = 2I_0\left(1 + e^{-\left(\frac{\Delta\omega\tau}{2}\right)^2}\cos\omega_0\tau\right); \qquad \mathcal{V} = |\gamma(\tau)| = e^{-\left(\frac{\Delta\omega\tau}{2}\right)^2}$$

$$\frac{\tau_0\Delta\omega}{2} = 1 \tag{2.49}$$

### 3.6. Wiener–Khinchin theorem

$$\Gamma(\tau) = \int_{-\infty}^{+\infty} J(\omega)e^{i\omega\tau}d\omega \cdot \frac{1}{2\pi}$$

*Proof.*

$$I(\tau) = 2I_0 + 2\,\mathrm{Re}\,\overline{V(t)V^*(t+\tau)} = 2I_0 + 2\,\mathrm{Re}\,\Gamma(\tau)$$
$$\equiv 2I_0 + 2\,\mathrm{Re}\int_{-\infty}^{+\infty} J(\omega)e^{i\omega\tau}d\omega$$

Another proof is by plugging Fourier transform into the definition of $\Gamma(\tau)$, it is 3 lines long. □

$$\tau_0\Delta\omega \simeq 2\pi$$

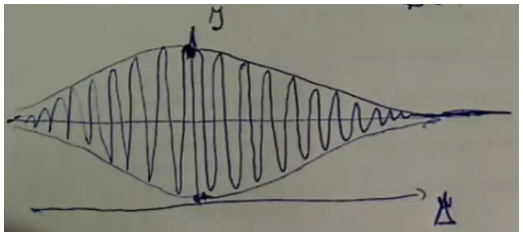
(since $\Gamma(\tau)$ is nonzero only for interval $\tau_0$ and $J(\omega)$ is nonzero only for $\Delta\omega$).

In another words, there is a field $f(t)$ and there is its spectrum $F(\omega)$ (by definition of Fourier). If we take Fourier from "spectrum" $|F(\omega)|^2$ (by definition of intensity), we will get not a field, but a correlation function $\Gamma(\tau)$.

(??? applications???)

#### Order in interference with time coherence

Count how many peaks are seen in



For waves with many wavelengths:

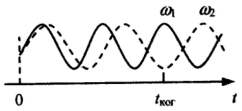
$$\Delta_{\max} = c\tau_0 = \frac{2\pi c}{\Delta\omega}; \qquad \tau_0 = \frac{2\pi}{\Delta\omega} \qquad \text{(uncert)}$$

$$\omega = 2\pi c/\lambda \qquad \Rightarrow \qquad \Delta\omega = \frac{2\pi c}{\lambda^2}\Delta\lambda \qquad \Rightarrow \qquad \Delta_{\max} = \frac{\lambda^2}{\Delta\lambda}$$

$$m_{\max} := \frac{\Delta_{\max}}{\lambda} = \frac{\lambda}{\Delta\lambda} \quad \text{(order of interference)}$$

For waves with two wavelengths:

$$\Delta = N\lambda_1 \equiv \left(N + \frac{1}{2}\right)\lambda_2 \qquad \Rightarrow \qquad 2N = \frac{\lambda}{\Delta\lambda}$$




And here we have two parameters that are the same: number of lines $N$, monochromaticity $N := \frac{\lambda}{\Delta\lambda} = \frac{\omega}{\Delta\omega} = \frac{V}{\Delta\nu}$. We can't determine exactly both, since for the first one, we can't say where exactly the spectrum vanishes, and for second, we don't know there is $\Delta\omega$ in the spectral density.

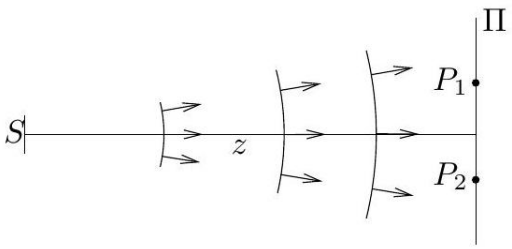
Also, there is a relation between temporal coherence and the width of the spectra:

$$t_{\text{coh}} = \frac{l_{\text{coh}}}{c} = \frac{\lambda N}{c}; \qquad t_{\text{coh}}\Delta\nu \sim 1.$$

Instead of defining $N$, we can work with visibility $V$ (which is also a time-coherence function). Then, it is connected to the spectrum density by Fourier transform (Wiener–Khinchin theorem).

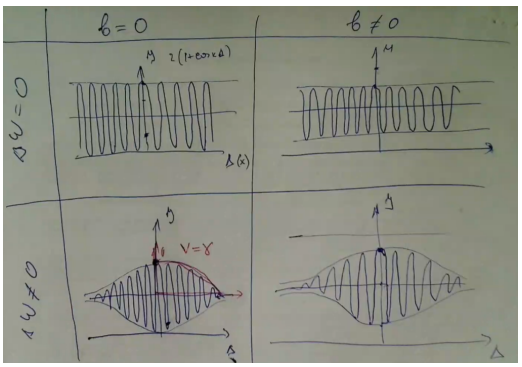
## Basics of spatial coherence

### Main properties of spatial coherence

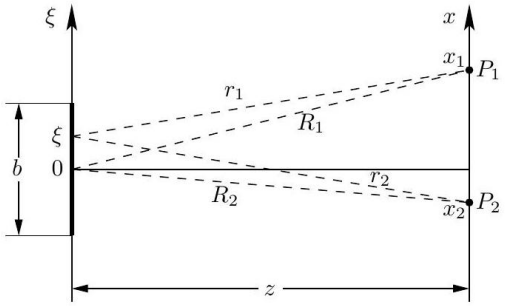



radius of spt coh $\rho_0$: $\frac{kb}{2z}\rho_0 = \pi$, $\qquad \rho_0 = \frac{\lambda}{b}z = \frac{\lambda}{\psi}$ $\qquad \psi := \frac{b}{z}$

(??? what are the conditions when all is coherent spatially??)



**Description of spatial coherence**




$$\rho := |x_1 - x_2|$$

$$V_1(t) = A_1(\boldsymbol{r}_1 t)\, e^{i\omega t} \quad V_2(t) = A_2(\boldsymbol{r}_2 t)\, e^{i\omega t}$$

We check, if the phase difference $\varphi_1'(t) - \varphi_2'(t) = \Delta\varphi'$ unchanged(or almost unchanged). Function of spatial coherence:

$$\Gamma_{12}(P_1 - P_2) = \overline{V_1(t)V_2^*(t)} = \overline{A_1(t)A_2^*(t)} \tag{2.55}$$

Degree of spatial coherence:

$$\gamma_{12}(P_1 - P_2) = \frac{\overline{V_1(t)V_2^*(t)}}{\sqrt{I_1 I_2}} \tag{2.56}$$

4.2. van Cittert–Zernike theorem:

$$\gamma(\rho) = \frac{\int J(\xi) e^{-i\frac{k}{z}\rho\xi} d\xi}{\int J(\xi) d\xi}.$$

*Proof.*

$$A_1 = \sum_i a_i e^{i\varphi_i} \frac{e^{ikr_{1i}}}{r_{1i}}; \qquad A_2 = \sum_j a_j e^{i\varphi_j} \frac{e^{ikr_{2j}}}{r_{2j}}$$

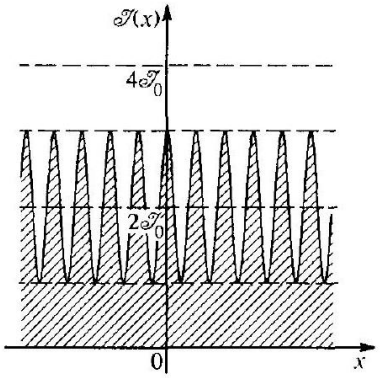
$$A_1 A_2^* = \sum_i \sum_j a_i a_j e^{i(\varphi_i - \varphi_j)} \frac{e^{ik(r_{1i} - r_{2j})}}{r_{1i} r_{2j}}; \qquad \overline{A_1 A_2^*} = \sum_i \overline{a_i^2} \frac{e^{ik(r_{1i} - r_{2i})}}{r_1 r_2}.$$

$$\overline{a^2(\xi)} = J(\xi)\Delta\xi, \qquad r_{12} = \sqrt{z^2 + (x_{12} - \xi)^2} \approx z + \frac{(x_{12} - \xi)^2}{2z}; \qquad r_1 - r_2 \approx \frac{x_1^2 - x_2^2}{2z} - \frac{\rho\xi}{z}$$

$$\Gamma(\rho) = \overline{V_1(t)V_2^*(t)} = \overline{A_1(t)A_2^*(t)} = \frac{e^{i\frac{k}{2z}\left(x_1^2 - x_2^2\right)}}{R_1 R_2} \sum_i J(\xi_i)\, e^{-i\frac{k}{z}\rho\xi_i} \Delta\xi_i \overset{\Delta\xi_i \to 0}{\to} \frac{e^{i\frac{k}{z}\rho x}}{R_1 R_2} \int J(\xi) e^{-i\frac{k}{z}\rho\xi} d\xi.$$

$x := (x_1 + x_2)/2$.

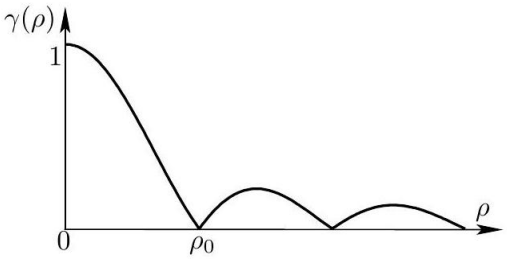
□



(????)

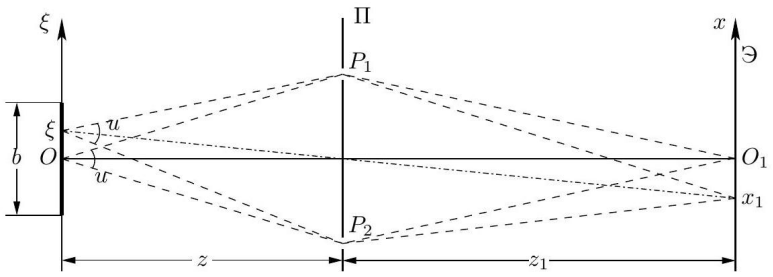
4.3. Example: homogeneous source

$$J(\xi) = \begin{cases} J_0 & |\xi| \leqslant \frac{b}{2} \\ 0 & |\xi| > \frac{b}{2} \end{cases} \overset{\text{van Cittert–Zernike}}{\Rightarrow} \gamma(\rho) = \frac{\sin \frac{kb}{2z}\rho}{\frac{kb}{2z}\rho},$$




radius of spt coh $\rho_0$: $\frac{kb}{2z}\rho_0 = \pi, \qquad \rho_0 = \frac{\lambda}{b}z = \frac{\lambda}{\psi} \qquad \psi := \frac{b}{z}$

Derivation of condition of interference with in the case of spatially extended sources in terms of aperture:




$\rho$ - difference between slits.

$$dI_0(d\xi) = 2J_0 d\xi \left[1 + \cos\frac{2\pi}{l}x\right]; \qquad l := \frac{\lambda}{\rho}z_1, \qquad x_1 = \xi\frac{z_1}{z}.$$

$$dI = 2J_0 d\xi \left[1 + \cos\frac{2\pi}{l}\left(x - \frac{z_1}{z}\xi\right)\right] \tag{2.63}$$

$$I(x) = 2I_0 \left[1 + \left(\frac{\sin \pi\frac{\rho}{\rho_0}}{\pi\frac{\rho}{\rho_0}}\right)\cos\frac{2\pi}{l}x\right] \tag{2.64}$$

$$\mathcal{V} = |\gamma(\rho)| = \left|\frac{\sin \pi\frac{\rho}{\rho_0}}{\pi\frac{\rho}{\rho_0}}\right| \qquad \rho_0 := \frac{\lambda}{b}z \tag{2.65}$$

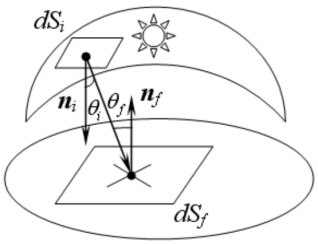
$\frac{\rho}{z} \leqslant \frac{\lambda}{b}$.

aperture: $u = \frac{\rho}{z}u \lesssim \frac{\lambda}{b}$

condition for interference: $u \lesssim \frac{\lambda}{b}$

Maximal size of the source: $b \lesssim \frac{\lambda}{u}$

**Intensity from wide sources**




Световой поток с площадки $dS_i$:

$$d^2\Phi := LdS_i\ d\Omega\ \cos\theta_i, \qquad d\Omega := \frac{dS_f}{r^2\left(\theta_f\right)}\cos\theta_f.$$

Освещенность площадки $dS_f$ потоком с площадки $dS_i$:

$$dE := \frac{d^2\Phi}{dS_f} = LdS_i\ \frac{d\Omega}{dS_f}\ \cos\theta_i = \frac{L\left(\theta_f\right)}{r^2\left(\theta_f\right)}\cos\theta_f\ \cos\left(\theta_i\left(\theta_f\right)\right)\ \frac{dS_i}{d\theta_f}d\theta_f.$$

$$E = \int\limits_{S_f} L\left(\theta_f\right) r^{-2}\left(\theta_f\right)\cos\theta_f \cdot \cos\left(\theta_i\left(\theta_f\right)\right)\frac{dS_i}{d\theta_f}d\theta_f$$

Example 1. освещение площадки, параллельной бесконечной плоскости:

$$L\left(\theta_f\right) = L; \quad \theta_i\left(\theta_f\right) = \theta_f; \quad r\left(\theta_f\right) = h/\cos\theta_f;$$

$$S_i = \pi h^2\,\mathrm{tg}^2\,\theta_f; \qquad \frac{dS_i}{d\theta_f} = 2\pi h^2\frac{\sin\theta_f}{\cos^3\theta_f};$$

$$E = \frac{L}{h^2}\int\limits_0^{\pi/2}\cos^4\theta_f 2\pi h^2\frac{\sin\theta_f}{\cos^3\theta_f}d\theta_f = \pi L\int\limits_0^{\pi/2}\sin 2\theta_f d\theta_f = \pi L$$

Example 2. Освещенность полусферой с ламбертовской яркостью (освещаемая площадка в центре полусферы).

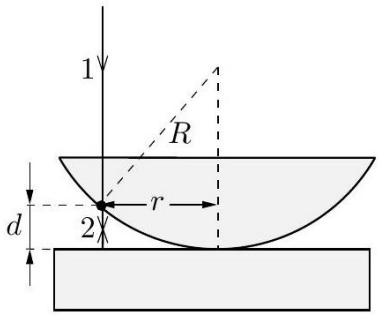
$$L\left(\theta_f\right) = L; \quad \theta_i\left(\theta_f\right) = 0; \quad r\left(\theta_f\right) = R; \quad dS_i = 2\pi R\sin\theta_f \cdot Rd\theta_f$$

$$E = \frac{L}{R^2}\int\limits_0^{\pi/2}\cos\theta_f \cdot 2\pi R^2\sin\theta_f d\theta_f = \pi L\int\limits_0^{\pi/2}\sin 2\theta_f d\theta_f = \pi L$$

(добавлю потом картинку, нарисую её аккуратно)

**Typical interferometers**

**5.4. Newton's rings**



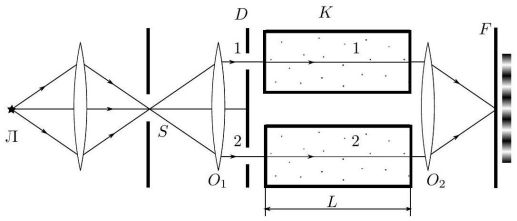

$d = r^2/2R.$

$$\Delta = 2d + \frac{\lambda}{2} = \frac{r^2}{R} + \frac{\lambda}{2}$$

$\Delta = (2m+1)\frac{\lambda}{2}$

$$r_m = \sqrt{m\lambda R}; \qquad r'_m = \sqrt{(2m-1)m\lambda R/2}.$$

**5.1. Rayleigh interferometer**



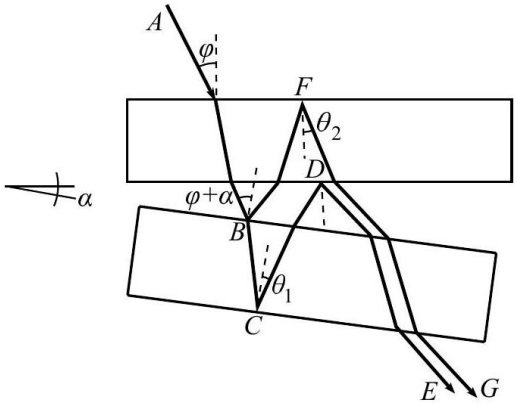

$$\varepsilon = n^2 = 1 + 4\pi N\alpha$$

$$n - 1 \approx \frac{\alpha}{2k_{\mathrm{B}}T}P$$

$\delta n = n_2 - n_1$

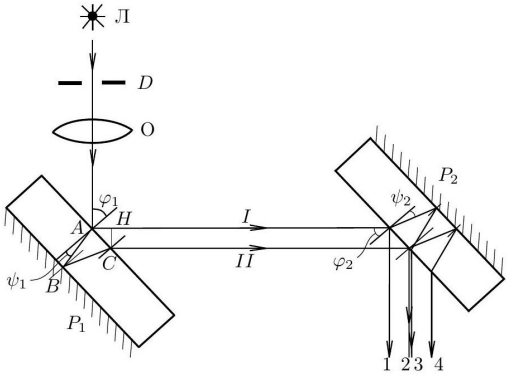
$$\delta n = \frac{\alpha}{2k_{\mathrm{B}}T}\delta P \tag{2.67}$$

### 5.2. Jamin interferometer



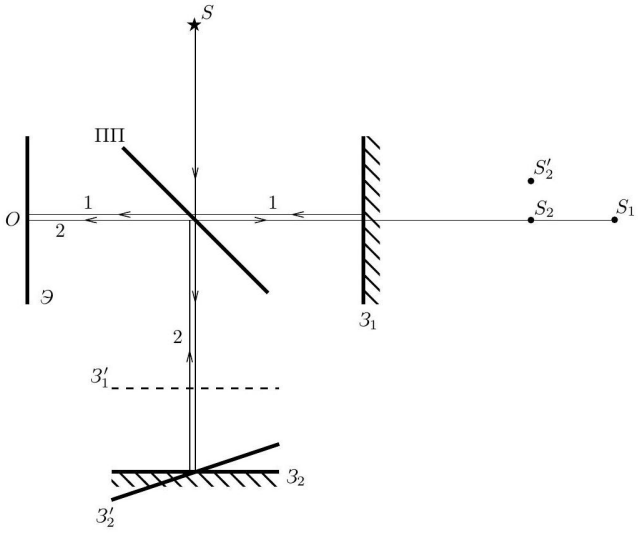

$2nd\,(\cos\theta_1 - \cos\theta_2)$

$$2nd\,(\cos\theta_1 - \cos\theta_2) \approx 2dn\left(\frac{\varphi^2}{2n^2} - \frac{(\varphi+\alpha)^2}{2n^2}\right) = \frac{2\alpha d}{n}\left(\varphi + \frac{\alpha}{2}\right) = m\lambda \qquad (2.68)$$




$$\delta\varphi = \varphi_{m+1} - \varphi_m = \frac{\lambda n}{2\alpha d}$$

### 5.3. Michelson interferometer



$$\sqrt{L^2+r_n^2}-\sqrt{(L-a)^2+r_n^2}\approx a+\frac{r^2a}{2L(L-a)}=\text{ const}$$

$m_0=\frac{a}{\lambda}$.

$$r_n\approx\sqrt{\frac{2nL(L-a)}{m_0}}\qquad \Delta r\approx\sqrt{\frac{L(L-a)}{2nm_0}}$$

$N=2\frac{l}{\lambda}$

$$v=\frac{\lambda}{2}\frac{N}{T} \tag{2.70}$$

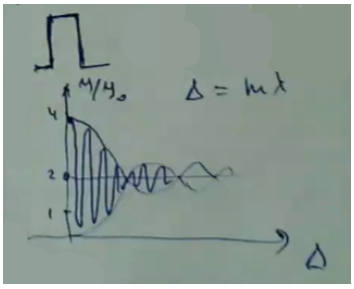
$\omega_1=\omega_0\sqrt{\frac{c+v}{c-v}}$
$\omega_2=\omega_1\sqrt{\frac{c+v}{c-v}}=\omega_0\frac{c+v}{c-v}$.
$\Delta\omega=\frac{2v}{c-v}\omega_0$.

$$N=\frac{\Delta\omega t}{2\pi}=2\frac{vt}{\lambda}\frac{1}{1-\frac{v}{c}}$$

**Other important about interference**

**Maximum of intensity when $\Delta=\left(m+\frac{1}{2}\right)\lambda$**

We take a very sharp pulse, so we will have visibility equals sin x/x. And there is a place where in $I(\Delta)=A(1+f(\Delta)\cos k\Delta)$ there is $f(\Delta)<0$. At this place for maximum there should be $\Delta=\left(m+\frac{1}{2}\right)\lambda$ (hald integer number of wavelengths).




So, for this special case of non-timecoherent sources. For real sources there are no sharp edges in the spectrum.

**Influence of polarization on interference**

(???)
$I_{\max}=(\sqrt{I_1}+\sqrt{I_2})^2,\quad I_{\min}=(\sqrt{I_1}-\sqrt{I_2})^2$.
$I=2I_0(1+\cos\Delta\varphi)$

$$2kx\sin\alpha=2\pi\quad\rightarrow\quad\Delta x=\frac{2\pi}{2k\sin(\alpha)}=\frac{\lambda}{2\sin(\alpha)}=\frac{\lambda}{\alpha_{total}}$$

т.о. ширина интерференционной полосы $l=\frac{\lambda}{2\alpha}$

## 1.1.6 Diffraction

### Main formulas of diffraction

$$p = \frac{\sqrt{\lambda z}}{b} \tag{3.1}$$

$p \ll 1 \quad \Rightarrow \quad$ геометрической оптики.

$p \gg 1 \quad \Rightarrow \quad$ называется областью дифракции Фраунгофера (или дальней волновой зоной).

$p \sim 1$ - областью дифракции Френеля (или ближней волновой зоной).

Здесь $z$ - расстояние от источника до поверхности наблюдения.

Граница $n$-ой зоны Френеля (Шустера): $r_n \approx \sqrt{n\lambda z}$, если размер отверстия $b$, то $n \approx b^2/\lambda z$.

В зависимости от $b^2/\lambda z$ мы имеем разные режимы:

**Дифракция Фраунгофера** - открыто $n \ll 1$ зоны, $b^2/\lambda z \ll 1$ или $p \gg 1$ - *дальняя волновая зона*

**Дифракция Френеля** - открыто $n \ll 1$ зоны, $b^2/\lambda z \approx 1$ или $p \gg 1$ - *ближняя волновая зона*

**Геометрическая оптика** - открыто $n \gg 1$ зоны, $b^2/\lambda z \gg 1$ или $p \ll 1$ - *самое первое пространство*

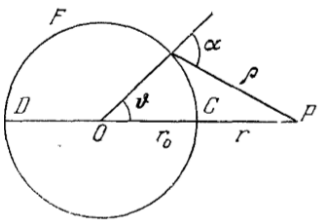
$$g_{\text{Fresnel}}(x, y) = (??)$$

$$g_{\text{Fraunghofer}}(x, y) = \frac{e^{ikR_0}}{i\lambda R_0} \iint f_0(\xi, \eta) e^{-i\left(\frac{kx}{R_0}\xi + \frac{ky}{R_0}\eta\right)} d\xi d\eta$$

В случае дифракции на круглом отверстии:




### Принцип Гюйгенса-Френеля



point $d\eta d\xi$

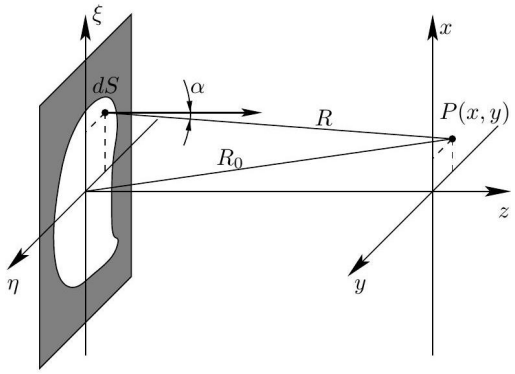
sp wave f point: $g(x,y,z) = \frac{1}{R}e^{ikR}d\eta d\xi;$

add angle: $g(x,y,z) = C\frac{1}{R}e^{ikR}\cos\alpha d\eta d\xi;$

$$E = \int \frac{adF}{\rho}e^{f(\omega t - k\rho)}$$

$$a = K(\alpha)e^{-ikr_0/r_0}$$

Геометрическая оптика работает на расстояниях от щели $z \ll b^2/\lambda$,




$$f_0(\xi\eta)\frac{e^{ikR}}{R}\cos\alpha \cdot d\xi d\eta$$

$$g(x,y) = \underbrace{K_0}_{\frac{1}{i\lambda}}\iint_S f_0(\xi\eta)\frac{e^{ikR}}{R}\cos\alpha d\xi d\eta \approx \frac{1}{i\lambda R_0}\iint f_0(\xi\eta)e^{ikR}d\xi d\eta$$

$$R = \sqrt{z^2 + (x-\xi)^2 + (y-\eta)^2} = z\sqrt{1 + \frac{(y-\xi)^2}{z} + \frac{(y-\eta)^2}{z}} \tag{3.11}$$

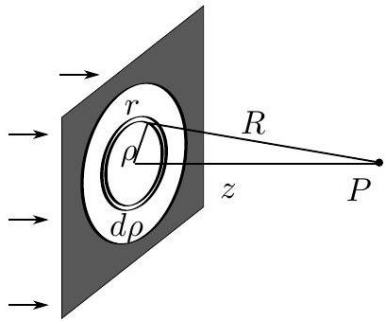
$$R \approx z + \frac{(x-\xi)^2}{2z} + \frac{(y-\eta)^2}{2z} \qquad \text{Fresnel appr}$$

$$g_{\text{Fresnel}}(x,y) = \frac{e^{ikz}}{i\lambda z}\iint f_0(\xi\eta)e^{i\frac{k}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]}d\xi d\eta \tag{3.13}$$

$$g_{\text{Fresnel}}(0,0|f_0(\xi,\eta) \equiv A_0) = A_0\frac{e^{ikz}}{i\lambda z}\iint e^{i\frac{k}{2z}(\xi^2+\eta^2)}d\xi d\eta \tag{3.14}$$

## 1.4. Fresnel diffraction

### 1.4.1. Fresnel diffraction on a round hole




$$g = A_0 \frac{e^{ikz}}{i\lambda z} \iint_{\rho \leqslant r} e^{i\frac{k}{2z}\rho^2} d\xi d\eta \tag{3.15}$$

$\rho = \sqrt{\xi^2 + \eta^2}$ - $d\xi d\eta$ $r$

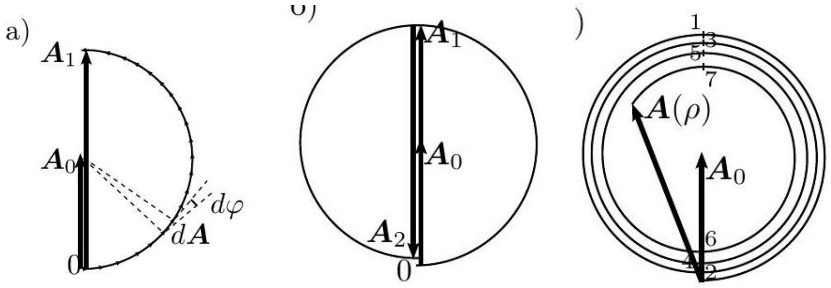
$$g = A_0 \frac{e^{ikz}}{i\lambda z} \int_0^r e^{i\frac{k}{2z}\rho^2} 2\pi\rho d\rho = A_0 e^{ikz} \left(1 - e^{-i\frac{k}{2z}r^2}\right) \tag{3.16}$$

$$I = |g|^2 = 2I_0 \left(1 - \cos\left(\frac{k}{2z}r^2\right)\right)$$

$r_m = \sqrt{m\lambda z} \quad m = 1, 2, 3 \ldots$

$I_{\max} = 4I_0$

$I_{\min} = 0$

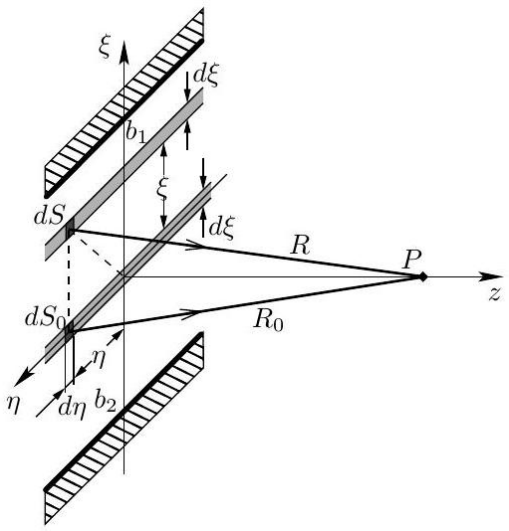



radius of the first Fresnel zone:

$$r_1 = \sqrt{\lambda z}$$

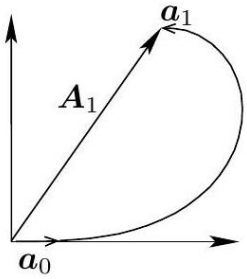
$$A = \left| \sum_{i=1}^{n} \boldsymbol{A}_i \right| = A_1 - A_2 + A_3 - \ldots$$

### 1.4.2. Fresnel diffraction on a slit

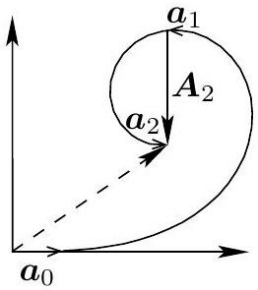



$$g = A_0 \int_{b_1}^{b_2} e^{\frac{ik}{2z}\xi^2} d\xi \tag{3.19}$$

$$\varphi = \frac{k}{2z}\xi^2, \qquad d\varphi = \frac{k}{z}\xi d\xi$$

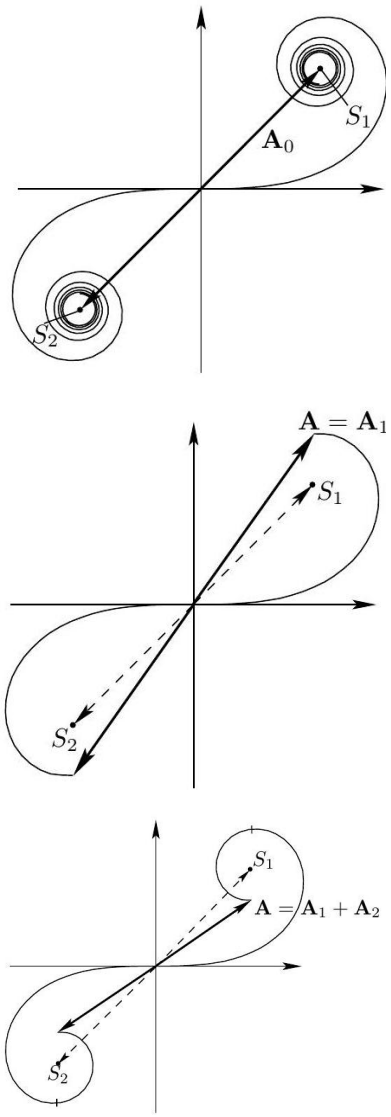



$\xi_1 = \sqrt{2\pi z/k} = \sqrt{\lambda z}$
$\frac{k}{2z}\xi_2^2 = 2\pi$
$\xi_2 = \sqrt{2\lambda z}.$
$\varphi = \frac{k}{2z}\xi^2$




$$\xi_m = \sqrt{m\lambda z} \quad m = 123\ldots \tag{3.20}$$

$\eta$. $(S_2)$ $\eta$ $(\xi < 0)$. $(b_1 \to \infty b_2 \to -\infty$ 3.9) $S_1$ $S_2$. $\boldsymbol{A}_0$




### 1.4.3. Conditions for Fresnel diffraction

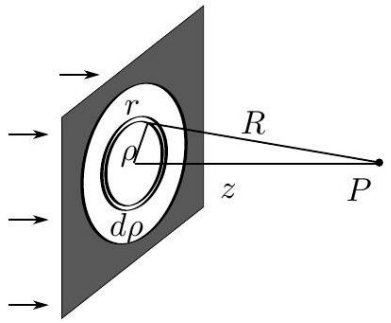
(write the idea form formulas above)

$$m = \frac{b^2}{\lambda z} \equiv \frac{1}{p^2}$$

$$p \sim 1$$

## 1.5. Fraunhofer diffraction

### 1.5.1. Fraunhofer diffraction on a round hole for 3D case



$$R = \sqrt{z^2 + (x-\xi)^2 + (y-\eta)^2} = \sqrt{R_0 - (2x\xi + 2y\eta) + (\xi^2 + \eta^2)}$$

$R_0 = \sqrt{z^2 + x^2 + y^2}$

$$R \approx R_0 - \frac{x\xi + y\eta}{R_0} + \frac{\xi^2 + \eta^2}{2R_0} \tag{3.21}$$

$\xi^2 + \eta^2 \leqslant b^2$.

$$\frac{b^2}{R_0} \ll \lambda \tag{3.22}$$

$$R \approx R_0 - \frac{x\xi}{R_0} - \frac{y\eta}{R_0}$$

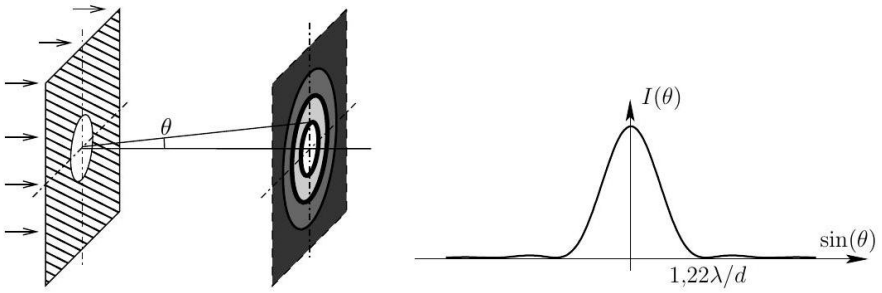
$$g(x,y) = \frac{e^{ikR_0}}{i\lambda R_0} \iint f_0(\xi\eta) e^{-i\left(\frac{kx}{R_0}\xi + \frac{ky}{R_0}\eta\right)} d\xi d\eta$$

$b^2/(\lambda z) \ll 1$.

$$u = \frac{kx}{R_0} \quad v = \frac{ky}{R_0}$$

$$g(u,v) = \frac{e^{ikR_0}}{i\lambda R_0} \iint f_0(\xi\eta) e^{-i(u\xi + v\eta)} d\xi d\eta \tag{3.23}$$

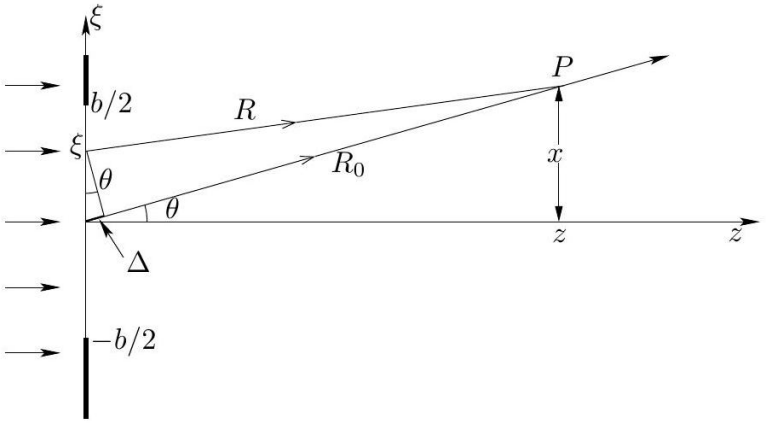
$$\boxed{\sin\theta = 1{,}22\frac{\lambda}{D}.} \qquad (3.37$$




$\sin\theta = 1,22\frac{\lambda}{D}$.

(see end of labnik for comments)

**Fraunhofer diffraction on a round hole in 1D case**



$$g(u) \propto \int_{-\infty}^{+\infty} f_0(\xi) e^{-iu\xi} d\xi \tag{3.24}$$

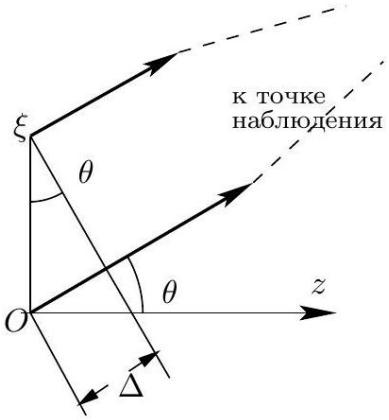
$$\Delta = \xi \sin\theta$$

$\varphi = -k\xi \sin\theta$

$$k \sin\theta = \frac{kx}{R_0} = u$$

$\sin\theta = \frac{x}{R_0}$

$$g(u) \propto \int_{-\infty}^{+\infty} f_0(\xi) e^{-ik\sin\theta\cdot\xi} d\xi$$




$$C(\omega) = \int f_0(t) e^{-i\omega t} dt$$

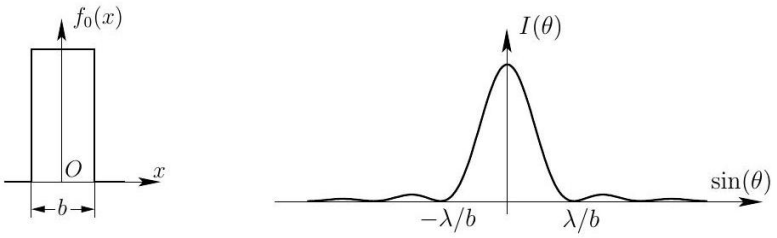
$u = k \sin\theta$ $\omega$ $f_0(t)$.
$b \ll \sqrt{\lambda z}$
$\sin\theta = x/R_0 = \text{const}$
$b \ll \sqrt{\lambda z}$

$$I(\theta) \propto \left| \int f_0(\xi) e^{-ik\sin\theta\cdot\xi} d\xi \right|^2 \tag{3.26}$$

$z \gg b^2/\lambda$

### 1.5.2. Fraunhofer diffraction on a slit

$$g(\theta) \propto \int_{-b/2}^{+b/2} e^{ikx\sin\theta} dx \propto \frac{\sin\left(\frac{kb}{2}\sin\theta\right)}{\frac{kb}{2}\sin\theta} \tag{3.27}$$



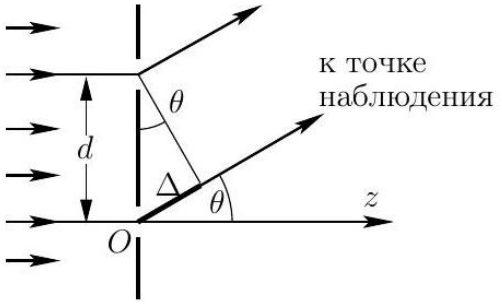

$\frac{kb}{2}\sin\theta = \pm\pi$

$$\sin\theta = \pm\frac{\lambda}{b} \tag{3.28}$$

$$|\sin\theta| \leqslant \frac{\lambda}{b}$$

### 1.5.3. Fraunhofer diffraction on two slits

$\Delta = d\sin\theta$

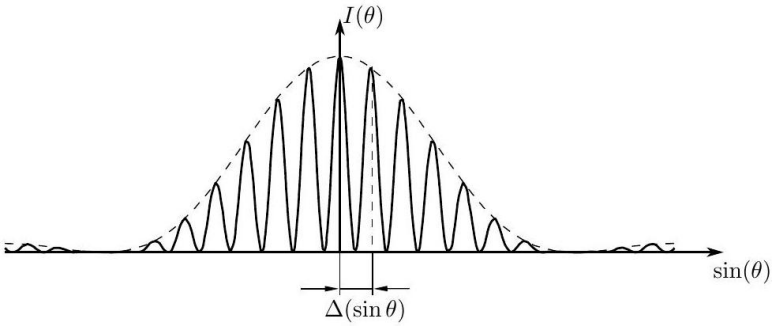



$$\alpha = -k\Delta = -kd\sin\theta$$

$g(\theta)e^{i\alpha}$.
. $g(\theta) + g(\theta)e^{i\alpha}$.
5

$$I(\theta) = |g(\theta)|^2 \cdot [1 + \cos(kd\sin\theta)]^2 \tag{3.29}$$

$kd\sin\theta = 2\pi$
$\Delta(\sin\theta) = \frac{\lambda}{d}$.

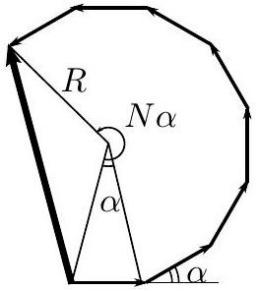


### 1.5.4. Fraunhofer diffraction on a grating

$\alpha_m = m\alpha = -mkd\sin\theta$

$$g_N(\theta) = g(\theta)\sum_{m=0}^{N-1} e^{im\alpha} \tag{3.30}$$

$R = \frac{1}{2|\sin\alpha/2|}$
$2R|\sin(N\alpha/2)|$.




$$|g_N(\theta)| = |g(\theta)| \cdot \left|\frac{\sin\frac{Nkd\sin\theta}{2}}{\sin\frac{kd\sin\theta}{2}}\right| \tag{3.31}$$

$I(\theta) = |g_N(\theta)|^2$

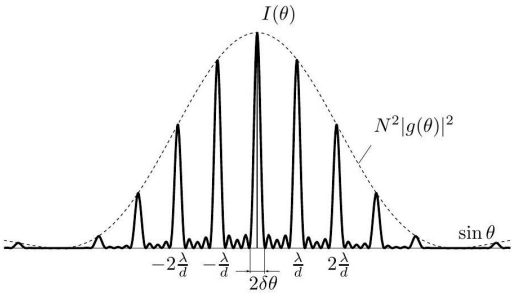
$$I(\theta) = |g(\theta)|^2 \left|\frac{\sin\frac{Nkd\sin\theta}{2}}{\sin\frac{kd\sin\theta}{2}}\right|^2 \tag{3.32}$$

$$d\sin\theta_m = m\lambda \quad m = 0 \pm 1 \pm 2\ldots \tag{3.33}$$

$-\alpha_m = kd\sin\theta_m = 2\pi m$.

$$g_N(\theta) = g(\theta)\frac{e^{iN\alpha} - 1}{e^{i\alpha} - 1}$$

$$g_N(\theta) = g(\theta)e^{i(N-1)\alpha/2}\frac{\sin\frac{Nkd\sin\theta}{2}}{\sin\frac{kd\sin\theta}{2}}$$




$I(\theta) = N^2|g(\theta)|^2$.

$$\lim_{\alpha\to 2\pi m}\frac{\sin(N\alpha/2)}{\sin(\alpha/2)} = N$$

$\theta = \theta_m + \delta\theta$

$$\delta(kd\sin\theta) = \frac{2\pi}{N} \qquad \delta(\sin\theta) = \frac{\lambda}{Nd}. \tag{3.34}$$

$$\delta\theta \approx \frac{\lambda}{Nd} \tag{3.35}$$

$m_{\max} = \frac{d}{\lambda}$.

$$m_{\max} \leqslant \frac{d}{b} \tag{3.36}$$

**Typical examples of diffraction**

**О дифракции на полупространстве**

хз, поизучаю - подумаю.

**Дифракция Френеля на круглом отверстии**

statement (???)

*Proof.* (????)

□

$$g(0,0) = A_0\frac{e^{ikz}}{i\lambda z}\iint e^{i\frac{k}{2z}(\xi^2+\eta^2)}d\xi d\eta$$

(тут нужно порешать на самом деле задачу эту, типа к вопросу, как нахрен мы интегрируем.)

переходим в обще к одному параметру $\rho$, через который считаем окружностями площадь.

в итоге получаем:

$$g = A_0\frac{e^{ikz}}{i\lambda z}\int\limits_0^r e^{i\frac{k}{2z}\rho^2}2\pi\rho d\rho = A_0e^{ikz}\left(1 - e^{-i\frac{k}{2z}r^2}\right)$$

Отсюда интенсивность в точке P равна

$$I = |g|^2 = 2I_0\left(1 - \cos\left(\frac{k}{2z}r^2\right)\right)$$

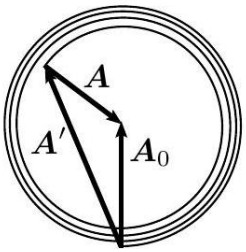
ну и дальше идет речь про векторные диаграммы и про всякие нюансы, позже доделаю.

**дифракция на диске**

(below formulas from labnik, where there is a Fresnel diffraction)
$\varphi = -\frac{k}{2z}\rho^2$.

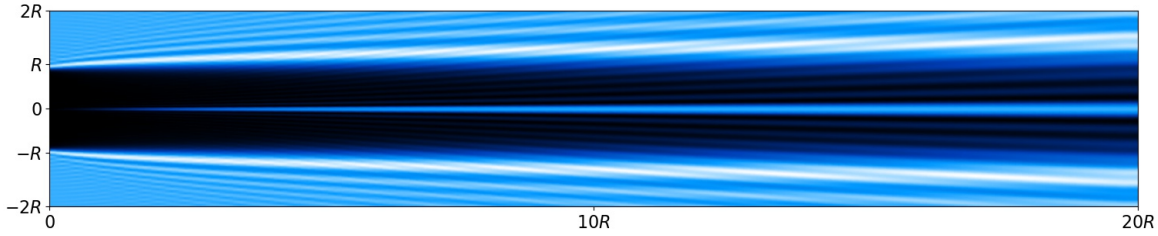
$$t(\rho) = e^{-i\frac{k}{2z}\rho^2}$$

$$\boldsymbol{A} = \boldsymbol{A}_0 - \boldsymbol{A}'$$




**пятно Пуассона-Араго**

пятно Пуассона видно при $\frac{d^2}{l\lambda} \geq 1$




**Дифракция на щели**

$$g = A_0 \int_{b_1}^{b_2} e^{\frac{ik}{2z}} \xi^2 d\xi$$

$$d\varphi = \frac{k}{z} \xi d\xi$$

**Other Useful methods and properties for diffraction**

**Transmission functions**

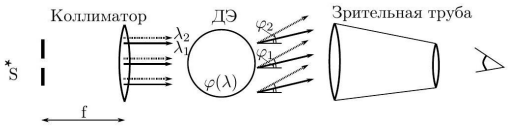
$$t(x, y) = a(x, y) e^{i\varphi(x,y)}$$

Например:
Фазовая синусоидальная решётка $t(x) = e^{im \cos \Omega x}$
Тонкая линза: $t(x, y) = e^{-i\frac{k}{2f}\left(x^2+y^2\right)}$
Дифракционная решётка - когда просто есть период и ширина щелей.
Амплитудная синусоидальная решётка: $t(x) = \beta(1 + m \cos \Omega x)$

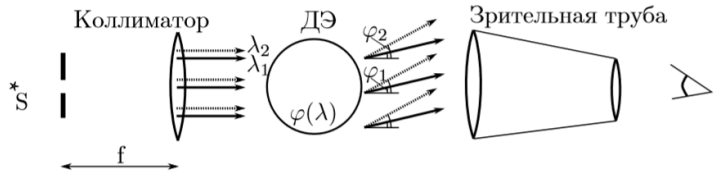
$$b > \lambda, z \gg \lambda \qquad \text{условия Кирхгофа}$$

## 1.1.7 Typical spectral devices and their properties

### 1. General ideas about spectral devices



$R = \frac{\lambda}{\delta\lambda}$ $\lambda$ $\lambda + \delta\lambda$.
$D = \frac{d\varphi}{d\lambda}$
$\delta\varphi \approx D\delta\lambda$.




*Коллиматор* - прибор, который из света от источника формирует пучок света, близкий к параллельному.

*Диспергирующий элемент* (ДЭ) - общее название для спектральных приборов.

**Цвета тонких пластинок**

**Клин**

(добавлю суть модели и картинку)

$$\Delta = (BD + DA)n - \left(AC - \frac{\lambda}{2}\right), \quad n(BD + DA) = \frac{2hn}{\cos r},$$
$$AC = 2h \operatorname{tg} r \sin i, \quad \frac{\sin i}{\sin r} = n$$

$h = ED$ - толщина клина;

$$\Delta = 2hn \cos r + \frac{\lambda}{2}.$$

(основные свойства напишу)

**Кольца Ньютона**

(коротко про постановку задачи)

$$\delta_m = \frac{r_m^2}{2R}$$

$$\Delta_m = 2\delta_m + \frac{\lambda}{2} = (2m+1)\frac{\lambda}{2} \qquad m \text{ dark ring},$$
$$\delta_m = \frac{m\lambda}{2},$$
$$r_m = \sqrt{m\lambda R} \qquad m \in \mathbb{Z},$$

Интерференционная картина будет отчетливой при малом $\delta$ (тонкая прослойка). Это не препятствует, однако, получению колец заметного радиуса, ибо $r_m = \sqrt{2R\delta}$, а $R$ - радиус кривизны линзы - может быть взят значительным (обычно 100-200 см).

Условие, облегчающее наблюдение колец Ньютона, состоит в очень малом наклоне поверхности линзы к поверхности пластинки. Подобный прием был много лет спустя применен в опытах Винера. Как уже упоминалось в §23, в одном из опытов, особенно

отчетливо определяющих положение пучностей и узлов по отношению к поверхности пластинки, Винер, пользуясь расположением, данным Ньютоном, получил стоячие волны в пространстве между линзой и пластинкой и наблюдал следы пучностей в виде концентрических колец, подобных кольцам Ньютона.

**Амплитудная решетка**

$$d \sin \varphi = m\lambda$$

Пусть на решётку, состоящую из $N$ штрихов, падает параллельный пучок света перпендикулярно её поверхности. Если $N = 2$, то две волны погасят друг друга, если между ними возникнет разность хода $\lambda/2$, если $N = 3$, то $\lambda/3$. В общем случае $N$ штрихов для полуширины линии $\delta\varphi$ получаем уравнение, решение которого совместно с уравнением (4.1) при $\delta\varphi \ll 1$ имеет вид

$$d \sin(\varphi_m + \delta\varphi) = m\lambda + \frac{\lambda}{N},$$
$$\delta\varphi = \frac{\lambda}{Nd \cos \varphi_m},$$

здесь $Nd \cos \varphi_m$ - видимый под углом $\varphi_m$ размер решётки.

$$R = Nm,$$

$N$ - число одновременно именно работающих штрихов решётки, не полное число штрихов освещённого участка решётки. Число штрихов $N$ определяется качеством реплики, размером источника света и т. д. Например, если источник является однородной линией шириной $b$ на расстоянии $z$ от решётки, то размер области решётки, засвеченной когерентно, определяется формулой $\rho \sim \lambda/\psi$, $\psi = b/z$ - угловой размер источника.

Дисперсия амплитудной решётки

$$D = \frac{d\varphi}{d\lambda} = \frac{m}{d \cos \varphi} = \frac{m}{\sqrt{d^2 - m^2\lambda^2}}.$$

Она возрастает с увеличением порядка спектра, для малых углов дифракции $\varphi \ll 1$:

$$D \approx m/d.$$

**фазовой решетке**

**установке Юнга**

(тут свойства)

**призме**

Дисперсия призмы:

$$D = \frac{d\varphi}{d\lambda} = \frac{a}{h}\frac{dn}{d\lambda}.$$

$$\delta\varphi = \frac{a}{h}\left|\frac{\partial n}{\partial \lambda}\right| \delta\lambda = \frac{\lambda}{h}$$

**эшелетте**

Дисперсия эшелетта

$$D = \frac{d\varphi}{d\lambda} = \frac{m}{d\cos\varphi} = \frac{m}{\sqrt{d^2 - (m\lambda + d\sin\psi)^2}}.$$

Тут углы дифракции нельзя считать малыми! В опытах углы падения $\psi \approx 30 - 60°$.

**бипризме (!?)**

(часто может в задачах встречаться, так что тут её суть)

**билинзе (!?)**

(часто может в задачах встречаться, так что тут её суть)

**Интерференция в плоскопараллельных пластинках и полосы равного наклона**

fringes of equal inclination, Haidinger fringes, равного наклона - это полосы, локализованные на бесконечности.

fringes of equal thickness - on the surface of interference setup.

**Интерференционные приборы и применения интерференции по Ландсбергу**

(спишу потом когда-то, по ка не до этого.)

**Интерферометр Жамена**

(120). § 29.

(пока он ни разу не нужен был)

**Интерферометр Майкельсона**

(123). § 30.

**Призмы**

**Простейшие линзы**

**Отражение и прохождение через пластин**

пар 2 овч

**О формулах Френеля**

(Тонга скорее всего вставлю потом.)

**Решетки**

(!!! в основах обязательно раздел про это!!!)

**Амплитудная дифракционная решетка**

(из лабника напишу, там сойдет написано.)

**Основные формулы**

Дифракционные максимумы определяются выражением:

$$d\sin\varphi_m = m\lambda$$

**энергия и максимумы**

$$dW = I_0\frac{b^2}{\lambda}\left[\frac{\sin\left(\frac{\pi}{\lambda}b\sin\varphi\right)}{\frac{\pi}{\lambda}b\sin\varphi}\right]^2\left[\frac{\sin\left(\frac{\pi}{\lambda}Nd\sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda}d\sin\varphi\right)}\right]^2 d\varphi$$

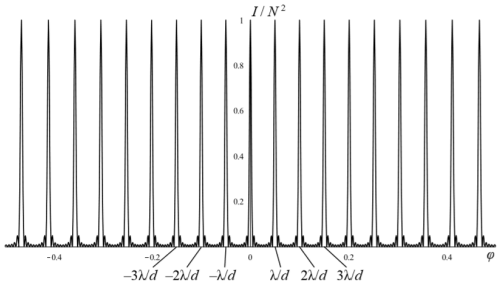
Дифракция на одной щели:

$$\left[\frac{\sin\left(\frac{\pi}{\lambda}b\sin\varphi\right)}{\frac{\pi}{\lambda}b\sin\varphi}\right]^2$$

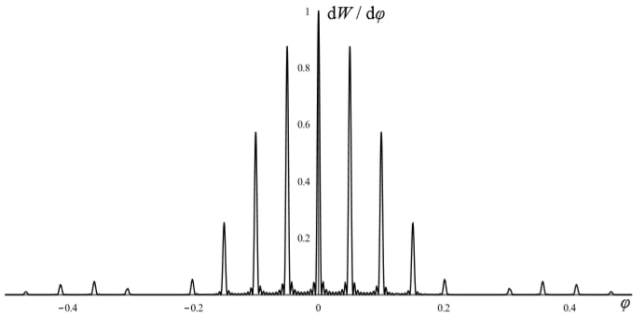



Ширина этого пика находится из приравнивания аргумента синуса к $2\pi$: $\Delta\phi = \frac{2\lambda}{b}$
Сложение от $N$ штрихов:

$$\left[\frac{\sin\left(\frac{\pi}{\lambda}Nd\sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda}d\sin\varphi\right)}\right]^2$$




Ширина этого пика аналогично: $\delta\phi = \frac{2\lambda}{Nd}$, поэтому, чтобы делать пики как можно более узкими создают $d$ как можно меньше.
а расстояние между максимами (вроде бы???) $\frac{\lambda}{d}$
Теперь видно, что общая картина имеет вид:




$d\sin\varphi_m = m\lambda$,

$$W_m \approx I_0\frac{b^2}{\lambda}N\left[\frac{\sin\pi\frac{b}{d}m}{\pi\frac{b}{d}m}\right]^2$$

$$I_0\frac{b^2}{d}N\sum_{-d/\lambda}^{d/\lambda}\left[\frac{\sin\pi\frac{b}{d}m}{\pi\frac{b}{d}m}\right]^2 \approx I_0\frac{b^2}{d}N\int\limits_{-d/\lambda}^{d/\lambda}\left(\frac{\sin\pi\frac{b}{d}x}{\pi\frac{b}{d}x}\right)^2 dx \approx I_0Nb$$

$$d\sin\theta_d = \left(m + \frac{1}{N}\right)\lambda$$

$$I_{\max} = I_0 \lim_{x\to\pi m}\left(\frac{\sin^2 Nx}{\sin^2 x}\right) = N^2 I_0$$

$$W_m = I_0\frac{b^2}{\lambda}\left[\frac{\sin\left(\frac{\pi}{\lambda}b\sin\varphi_m\right)}{\left(\frac{\pi}{\lambda}b\sin\varphi_m\right)}\right]_{\varphi_m-\frac{\lambda}{2d}}^{2\varphi_m+\frac{\lambda}{2d}}\left[\frac{\sin\left(\frac{\pi}{\lambda}Nd\sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda}d\sin\varphi\right)}\right]^2 d\varphi \approx I_0\frac{b^2}{\lambda}N\left[\frac{\sin\pi\frac{b}{d}m}{\pi\frac{b}{d}m}\right]^2$$

$$I_0\frac{b^2}{d}N\sum_{-d/\lambda}^{d/\lambda}\left[\frac{\sin\pi\frac{b}{d}m}{\pi\frac{b}{d}m}\right]^2 \approx I_0\frac{b^2}{d}N\int_{-d/\lambda}^{d/\lambda}\left(\frac{\sin\pi\frac{b}{d}x}{\pi\frac{b}{d}x}\right)^2 dx \approx I_0 Nb$$

$$\frac{I_0 b^2 N/d}{I_0 Nd}\left[\frac{\sin\pi\frac{b}{d}m_p}{\pi\frac{b}{d}m_p}\right]^2 = \left(\frac{b}{d}\right)^2\left[\frac{\sin\pi\frac{b}{d}m_p}{\pi\frac{b}{d}m_p}\right]^2.$$

**характеристики**

Для разрешающей способности по критерию Релея необходимо приравнять условия максимума волны с $\lambda$:

$$d\sin\theta_d = \left(m + \frac{1}{N}\right)\lambda$$

к минимуму с $\lambda + \Delta\lambda$:

$$d\sin\theta_d = m(\lambda + \Delta\lambda)$$

в этом случае мы имеем максимальный подрядок интерференции $m$ (да?), именно в нем

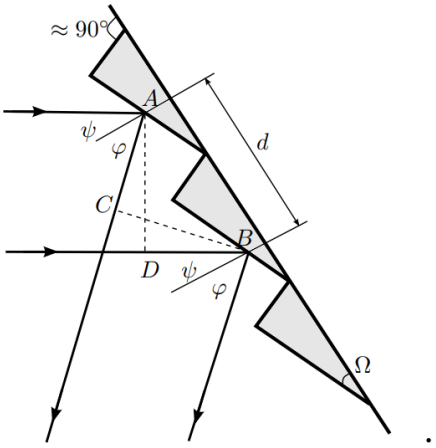
$$R = \frac{\lambda}{\Delta\lambda} = mN$$

- область дисперсии?

**фазовая решетка**

(какая-то хрень, которую я не хочу трогать)

Отражательная решётка (рис. 4.5), в которой угол $\Omega$ между рабочей гранью и плоскостью решётки (угол скоса) не превышает 20°, называется эшелеттом (фр. echelle - лестница). Для него, варьируя угол скоса и шаг решётки, получают рабочий порядок $m_p \leqslant 10$. Если угол скоса лежит в пределах $20 - 60°$, то решётка называется эшелле, её рабочий порядок $m_p = 5 - 500$.




Эшелетт можно представить как зеркало в форме лестницы. Найдём разность хода между лучами (рис. 4.5). Пусть на эшелетт падает плоская монохроматическая волна с длиной волны $\lambda$ угол падения, отсчитанный от нормали к решётке, равен $\psi$, угол дифракции $-\varphi$. Условие возникновения спектра порядка $m$:

$$AC - BD = d\,(\sin\varphi_m - \sin\psi) = m\lambda.$$

$$\varphi_6 = \psi + 2\Omega$$

Для эшелетта рабочим порядком спектра $m_p$ будет то целое число, которое соответствует минимальной ошибке решения уравнения

$$d\,(\sin\varphi_\sigma - \sin\psi) - \lambda m_p = 0.$$

В паспорте эшелетта указаны рабочий порядок спектра $m_p$ и рабочая длина волны $\lambda_p$. Чтобы устранить произвол в выборе угла падения света при расчёте угла блеска (4.4), считаем, что эшелетт работает в автоколлимационном режиме. В этом режиме свет падает перпендикулярно рабочей грани решётки ($\psi = -\Omega$) и отражается в обратном направлении ($\varphi = \varphi_m = \varphi_p = \Omega$), условие (4.3) возникновения рабочего спектра даёт

$$2d\sin\Omega = m_p\lambda_p.$$

В автоколлимационном режиме дифракция на одной ступенькезеркальце описывается так же, как и дифракция на отдельной щели амплитудной решётки, рис. 4.4а, с максимумом вблизи $\varphi \approx 0$. В отличие от амплитудной решётки, при углах $\varphi \approx 0$ располагается не нулевой порядок спектра (так называемая ахроматическая полоса), а порядок, определяемый соотношением (4.5), т. е. нумерацию порядков, приведённую на рис. 4.4 для амплитудной решётки, следует сместить на величину $m_p$. Доля мощности, попадающей в рабочий порядок, составляет $(b/d)^2$ от мощности светового потока, падающего на эшелетт ($b$ - ширина отдельной ступеньки, чуть меньшая шага решётки $d$).

**Линзы**

есть в бил.
- идеальная линза
- что там про фокусировку света мы знаем?
(с з.п.
Френеля некотрая связь)

**Интерферометр Фабри–Перо (!)**

**Основные свойства интерферометра Фабри-Перо**

Дисперсия интерферометра Фабри-Перо
Для малых углов $\theta_m \ll 1$ и больших порядков спектров имеем $m \approx M = 2L/\lambda$,

$$-2L\sin\theta_m d\theta = -2L\theta_m d\theta = md\lambda \approx \frac{2L}{\lambda}d\lambda,$$

и угловая дисперсия

$$D = \frac{d\theta}{d\lambda} = -\frac{m}{2L\sin\theta_m} \approx -\frac{1}{\lambda\theta_m}.$$

**Многолучевая интерференция**

**Функция пропускания интерферометра**

**Интерферометр как оптический резонатор**

**Энергия, запасаемая в интерферометре**

**описание**

Формула распределения интенсивности в проходящем ($I_{\text{прох}}$) и отраженном ($I_{\text{отр}}$) свете при многократной интерференции на плоскопараллельной пластинке, полагая, что коэффициент поглощения $A = 0,47$. Принимая интенсивность падающего пучка за 1, вывести формулу распределения интенсивности в проходящем ($I_{\text{прох}}$) и отраженном ($I_{\text{отр}}$) свете при многократной интерференции на плоскопараллельной пластинке, полагая, что коэффициент поглощения $A = 0$, так что $T + R = 1$. Ответ.

$$I_{\text{прох}} = \frac{T^2}{(1-R)^2}\frac{1}{1+\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}} = \frac{1}{1+\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}}, \qquad I_{\text{отр}} = \frac{\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}}{1+\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}}.$$

У к а з а н и е. Полагая коэффициент отражения для амплитуды равным $\rho$, а коэффициент пропускания - $\tau$ (коэффициент поглощения предполагается равным нулю, $\alpha = 0$), так что $R = \rho^2$ и $T = \tau^2$ и $R + T = 1$, найдем амплитуды проходящих $(0, 1, 2, 3, \ldots)$ и отраженных $(0', 1', 2', 3', \ldots)$ лучей (рис. 15). В соответствии с этим световое колебание в проходящем пучке дается выражением $\tau^2\rho^{2k}\exp[i(\omega t - k\cdot 2\pi m)]$, $k$ номер пучка ($k = 0, 1, 2, \ldots$), а $m$ - порядок интерференции. Для отраженных пучков имеем $\tau^2\rho^{2k-1}\exp[i(\omega t - 2\pi m)]$,
$k = 1, 2, 3, \ldots$ Для нулевого отраженного пучка ($k = 0$) имеем $\rho\exp[i(\omega t + \pi)] = -\rho\exp(i\omega t)$ (учет потери полуволны по сравнению с проходящим пучком вследствие разницы в условиях отражения). Итак, результирующее колебание: для проходящего пучка

$$A = \sum_{k=0}^{k=\infty}\tau^2\rho^{2k}\exp[i(\omega t - k\cdot 2nm)] = \frac{T}{1-R\exp(-i\cdot 2\pi m)}\exp i\omega t$$

$$B = -\exp(i\omega t)\left\{\rho - \tau^2\sum_{k=1}^{k=\infty}\rho^{2k-1}\exp[-ik\cdot 2\pi m]\right\} =$$

$$= -\exp(i\omega t)\rho\frac{1-(T+R)\exp(-i\cdot 2\pi m)}{1-R\exp(-i\cdot 2\pi m)} = -\exp(i\omega t)\rho\frac{1-\exp(-i\cdot 2\pi m)}{1-\mathrm{Re}\exp(-i\cdot 2\pi m)}$$

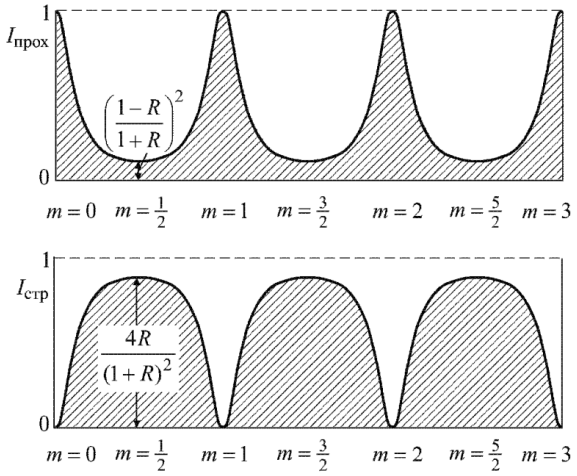
$I_{\text{прох}} = AA^*$, $I_{\text{отр}} = BB^*$,

$$I_{\text{прох}} = \frac{T^2}{1+R^2-2R\cos 2\pi m} = \frac{T^2}{(1-R)^2}\frac{1}{1+\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}} = \frac{1}{1+\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}},$$

$$I_{\text{отр}} = R\frac{2[1-\cos 2\pi m]}{1-2R\cos 2\pi m+R^2} = \frac{\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}}{1+\frac{4R}{(1-R)^2}\sin^2\frac{\psi}{2}},$$

$\psi = 2\pi m$.

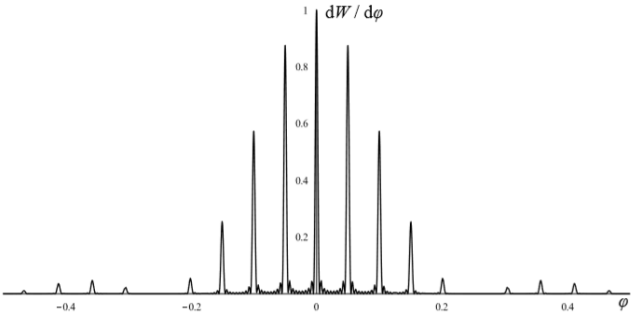
$I_{\text{прох}} + I_{\text{отр}} = 1$




**амплитуды волны при прохождении интерферометра Фабри-Перо**

Обсудим изменение амплитуды.

Так как интенсивность при отражении делится коэффициентами $r, t$, при этом $r+t=1$, амплитуды делятся так, как на рисунке **??**. Фазы считаются очевидным образом, главное не затупить, что разность фаз считается на уровне нормали к волнам.




$$\Delta = 2L\left(\frac{1}{\cos\theta} - \operatorname{tg}\theta\sin\theta\right) = 2L\cos\theta$$

$$2L\cos\theta_m = m\lambda$$

$$A = \frac{A_0 t}{1-re^{ik\Delta}}$$

$$I = |A|^2 = \frac{|A_0|^2 t^2}{1+r^2-2r\cos\left[\frac{4\pi}{\lambda}L\cos\theta\right]} = \frac{I_0}{1+\frac{4r}{(1-r)^2}\sin^2\left(\frac{2\pi}{\lambda}L\cos\theta\right)}$$

**Характеристики**

**Разрешающая способность**

без вывода, $\rho$ - коэффициент пропускания, пока:

$$R = \frac{2\pi L}{\lambda} \cdot \frac{\sqrt{\rho}}{1-\rho}$$

**дисперсионная область**

$$m(\lambda + \Delta\lambda) = (m+1)\lambda$$

**энергия и добротность**

**Формулы**

просто там какой-то нюанс есть, который я не понимаю вообще, что там с энергией? добротность хз почему, но определяется и ищется так:

$$Q = 2\pi \frac{E_0}{\Delta E}$$

$$Q = \frac{\omega}{\delta\omega}, \quad R = \frac{\lambda}{|\delta\lambda|} = \frac{\omega}{|\delta\omega|}$$

$$E_0 = 2I\tau(L) = 2I\frac{L}{c}$$

$$Q = 2\pi \frac{L}{\lambda} \cdot \frac{1}{1-\rho}$$

**Дисперсионная формула для ИФП**

При большой ширине спектра спектры различных порядков могут накладываться друг на друга. На рис. 4.3 отмечено наложение «красной» границы спектра 5-го порядка на «фиолетовую» границу спектра 6-го порядка. Предельная ширина спектрального интервала $\Delta\lambda$, при которой спектры соседних порядков перекрываются своими границами, называется дисперсионной областью. При этом $m(\lambda + \Delta\lambda) = (m+1)\lambda$ и дисперсионная область

$$\Delta\lambda = \frac{\lambda}{m}.$$

Ограничений на оптические спектры не возникает при использовании призмы и дифракционных решёток, работающих в первом порядке спектра, для них $\Delta\lambda \approx \lambda$.

Возможная для исследования ширина спектра источника уменьшается при увеличении рабочего порядка спектра дифракционной решётки. Особые требования к источнику возникают при использовании интерферометра Фабри-Перо, так как его рабочий порядок спектра весьма высок $(2L/\lambda)$ и дисперсионная область

$$\Delta\lambda = \frac{\lambda^2}{2L}$$

очень мала. Для порядка спектра $2L/\lambda \approx 5 \cdot 10^3$ дисперсионная область $\Delta\lambda \approx 1$ Å. Интерферометр Фабри-Перо обладает очень высокой разрешающей способностью, но при этом может использоваться только для анализа узкополосных спектров. Обычно перед интерферометром располагают светофильтры или другие спектральные приборы, вырезающие вблизи $\lambda$ полосу шириной менее $\Delta\lambda$.

**Классические примеры дифракции Фраунгофера**

**1D дифракция фраунгофера**

потом поучу.

$$g(u) \propto \int_{-\infty}^{+\infty} f_0(\xi) e^{-iu\xi} d\xi$$

**Дифракция на щели**

(начало задачи должно содержать слова, что мы именно Фраунгофера решаем, так как я удалил из главы это слово, чтобы было удобнее)

$$g(\theta) \propto \int_{-b/2}^{+b/2} e^{ikx\sin\theta} dx \propto \frac{\sin\left(\frac{kb}{2}\sin\theta\right)}{\frac{kb}{2}\sin\theta}$$

**Дифракция на двух щелях**

(начало задачи должно содержать слова, что мы именно Фраунгофера решаем, так как я удалил из главы это слово, чтобы было удобнее)

$$I(\theta) = |g(\theta)|^2 \cdot [1 + \cos(kd\sin\theta)]^2$$

**Дифракция на решётке**

(вставка теории из последней главы.

**Дифракция на круглом отверстии**

**поле в фокальной плоскости линзы**

**На синусоидальной решетке (??)**

Рассмотрим применения метода Рэлея на простейших задачах дифракции плоской волны на периодических структурах. Ограничимся сейчас случаем, когда волновые векторы всех волн лежат в плоскости $(x, z)$. 1. Амплитудной синусоидальной решёткой называется объект с функцией пропускания

$$D(x) = D_0(1 + m\cos\Omega x).$$

Период этой решётки равен

$$d = 2\pi/\Omega.$$

Имеется естественное ограничение на значения параметров $D_0$ и $m$:

$$D_0(1+m) \le 1,$$

определяемое тем, что прохождение тонкого экрана не приводит к локальному увеличению амплитуды поля.

Пусть на решётку (8.2.1) слева падает плоская волна под углом $Q_0$ (рис. 8.2.1):

$$A = A_0 e^{i\mathbf{k}_0\mathbf{r}}, u_0 = k\sin\theta_0, w_0 = k\cos\theta_0. \quad (8.2.4)$$

Для нахождения поля в области $z > 0$ воспользуемся граничным условием Кирхгофа:

$$A_S = D(x)A|_{z=-0} = A_0 D(x) e^{iu_0 x}. \quad (8.2.5)$$

Имея в виду выражение (8.2.1) для функции пропускания, получаем пространственное разложение Фурье функции $A_S$ :

$$A_S = A_0 D_0 e^{iu_0 x} + \frac{m}{2} A_0 D_0 e^{i(u_0+\Omega)x} + \frac{m}{2} A_0 D_0 e^{i(u_0-\Omega)x}.$$

$$\begin{aligned}
A(x,z) &= A_0(x,z) + A_{+1}(x,z) + A_{-1}(x,z)\\
A_0(x,z) &= A_0 D_0 \exp\left[iu^{(0)}x + iw^{(0)}z\right)]\\
A_{+1}(x,z) &= \frac{m}{2} A_0 D_0 \exp\left[iu^{(+1)}x + iw^{(+1)}z\right]\\
A_{-1}(x,z) &= \frac{m}{2} A_0 D_0 \exp\left[iu^{(-1)}x + iw^{(-1)}z\right]
\end{aligned}$$

$$\begin{aligned}
u^{(0)} &= u_0, w^{(0)} = \sqrt{k^2 - u_0^2},\\
u^{(+1)} &= u_0 + \Omega, w^{(+1)} = \sqrt{k^2 - (u_0+\Omega)^2},\\
u^{(-1)} &= u_0 - \Omega, w^{(-1)} = \sqrt{k^2 - (u_0-\Omega)^2}.
\end{aligned}$$

Пространственные частоты $u^{(0)}, u^{(+1)}, u^{(-1)}$ определяют направления распространения волн справа от решётки ($z > 0$). Полагая для этих волн $u = k\sin\theta$, записываем уравнения для отределения их направлений: $k\sin\theta^{(0)} = k\sin\theta_0, k\sin\theta^{(+1)} = k\sin\theta_0 + \Omega, k\sin\theta^{(-1)} = k\sin\theta_0 - \Omega$.

Вспоминая равенства $k = 2\pi/\lambda, \Omega = 2\pi/d$, получаем

$$d\left(\sin\theta - \sin\theta_0\right) = n\lambda, n = 0, \pm 1.$$

Мы пришли к известной формуле, указывающей направления на главные максимумы при дифракции на решётке. Заметим, что в рассматриваемом случае имеется всего три максимума. Это связано прежде всего с тем, что период решётки $d$ и эффективная ширина щели связаны соотношением $b_{\text{eff}} = d/2$. 2. Фазовой синусоидальной решёткой называется объект с функцией пропускания

$$D = e^{im\cos\Omega x}, \Omega = 2\pi/d.$$

Этот объект, не меняя амплитуды падающей на него волны, изменяет её фазу. Пусть на решётку (8.2.10) падает слева плоская волна (рис. 8.2.1):

$$A = A_0 e^{i\mathbf{k}_0\mathbf{r}}.$$

Найдём поле справа от решётки. Используем граничное условие

$$A|_{z=+0} = D(x)A|_{z=-0}\,,$$

или

$$A|_{z=+0} = A_0 e^{iu_0 x + im\cos\Omega x}, u_0 = k\sin\theta_0.$$

Обычно рассматривают фазовые решётки с малыми модуляциями фазы: $m \ll 1$. В этом случае получаем

$$A|_{z=+0} = A_0 e^{iu_0 x}(1 + im\cos\Omega x) =$$
$$= A_0 e^{iu_0 x} + \frac{im}{2} A_0 e^{i(u_0+\Omega)x} + \frac{im}{2} A_0 e^{i(u_0-\Omega)x}$$

Следовательно, от плоскости $z = +0$ исходят три волны, которым отвечают пространственные частоты

$$u = u_0 + n\Omega, n = 0, \pm 1.$$

Полагая

$$u = k\sin\theta, u_0 = k\sin\theta_0, k = 2\pi/\lambda,$$

находим уравнение, определяющее направления волн:

$$d\,(\sin\theta - \sin\theta_0) = n\lambda, n = 0, \pm 1.$$

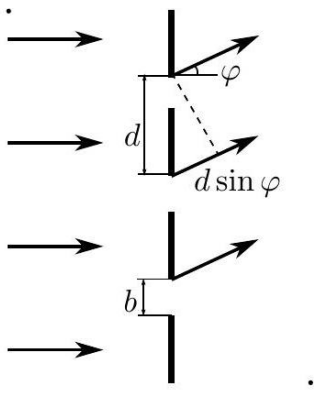
Таким образом, в пределе $m \ll 1$ остаются, как и в случае амплитудной синусоидальной решётки, три волны. Разница состоит в том, что в случае фазовой решётки фазы волн, отвечающих главным максимумам $n = \pm 1$, имеют дополнительный сдвиг на $\pi/2$ (благодаря наличию множителя $i$ во втором и третьем слагаемых в (8.2.13)).

Заметим, что если провести разложение (8.2.12) до более высоких степеней параметра $m$, то получим волны, распространяющиеся под углами, отвечающими значениям $n = \pm 2, \pm 3, \ldots$ Ограничение на величину $n$ при этом даётся, как и в случае обычной дифракционной решётки, условием $n \le d/\lambda$.

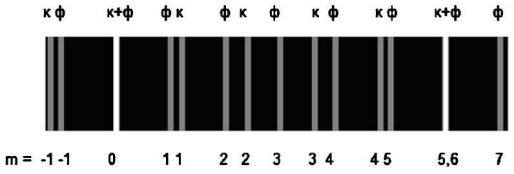
**2.**

**2.1.**

$$d\sin\varphi_m = m\lambda$$




$\varphi_{m+1} - \varphi_m \approx \lambda/d$

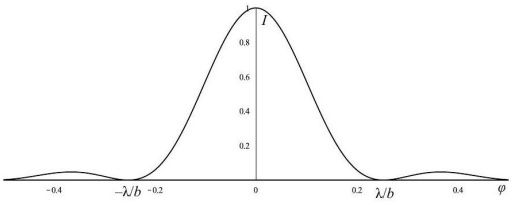


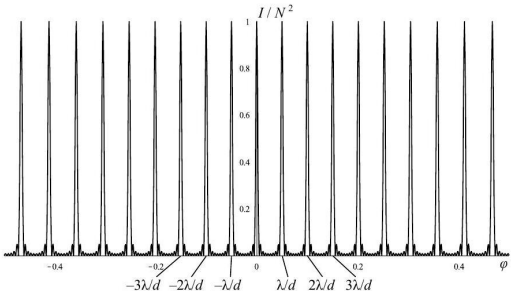

$$dW = I_0 \frac{b^2}{\lambda} \left[ \frac{\sin\left(\frac{\pi}{\lambda} b \sin\varphi\right)}{\frac{\pi}{\lambda} b \sin\varphi} \right]^2 \left[ \frac{\sin\left(\frac{\pi}{\lambda} N d \sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\varphi\right)} \right]^2 d\varphi \tag{4.2}$$

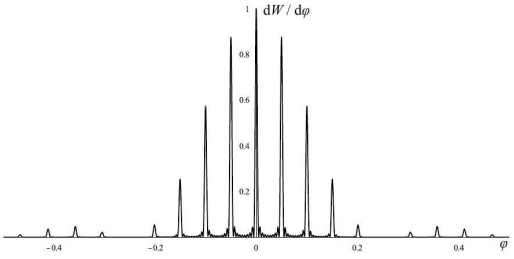
$$\left[ \frac{\sin\left(\frac{\pi}{\lambda} b \sin\varphi\right)}{\frac{\pi}{\lambda} b \sin\varphi} \right]^2$$

$b \gg \lambda$
$|\varphi| \leqslant \lambda/b$

$$\left[ \frac{\sin\left(\frac{\pi}{\lambda} N d \sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\varphi\right)} \right]^2$$



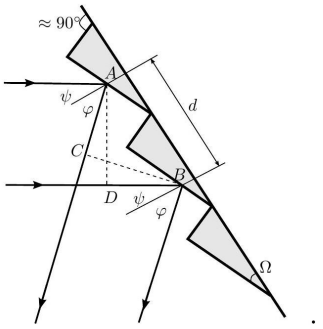

$$W_m = I_0 \frac{b^2}{\lambda} \left[ \frac{\sin\left(\frac{\pi}{\lambda} b \sin\varphi_m\right)}{\left(\frac{\pi}{\lambda} b \sin\varphi_m\right)} \right]_{\varphi_m - \frac{\lambda}{2d}}^{2\varphi_m + \frac{\lambda}{2d}} \left[ \frac{\sin\left(\frac{\pi}{\lambda} N d \sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\varphi\right)} \right]^2 d\varphi \approx I_0 \frac{b^2}{\lambda} N \left[ \frac{\sin \pi \frac{b}{d} m}{\pi \frac{b}{d} m} \right]^2$$

$$I_0 \frac{b^2}{d} N \sum_{-d/\lambda}^{d/\lambda} \left[ \frac{\sin \pi \frac{b}{d} m}{\pi \frac{b}{d} m} \right]^2 \approx I_0 \frac{b^2}{d} N \int_{-d/\lambda}^{d/\lambda} \left( \frac{\sin \pi \frac{b}{d} x}{\pi \frac{b}{d} x} \right)^2 dx \approx I_0 N b$$

$$\frac{I_0 b^2 N/d}{I_0 N d} \left[ \frac{\sin \pi \frac{b}{d} m_p}{\pi \frac{b}{d} m_p} \right]^2 = \left( \frac{b}{d} \right)^2 \left[ \frac{\sin \pi \frac{b}{d} m_p}{\pi \frac{b}{d} m_p} \right]^2$$

$d/b = 4$ $m_p = 2$ 3% .

**2.2.**




.

$$AC - BD = d\,(\sin\varphi_m - \sin\psi) = m\lambda$$

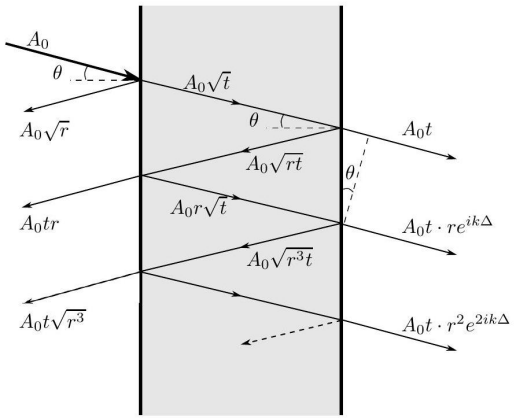
$$\varphi_\sigma = \psi + 2\Omega \tag{4.4}$$

$$d\,(\sin\varphi_\sigma - \sin\psi) - \lambda m_p = 0$$

$(\varphi = \varphi_m = \varphi_p = \Omega)$

$$2d\sin\Omega = m_p\lambda_p \tag{4.5}$$

**2.3. -**

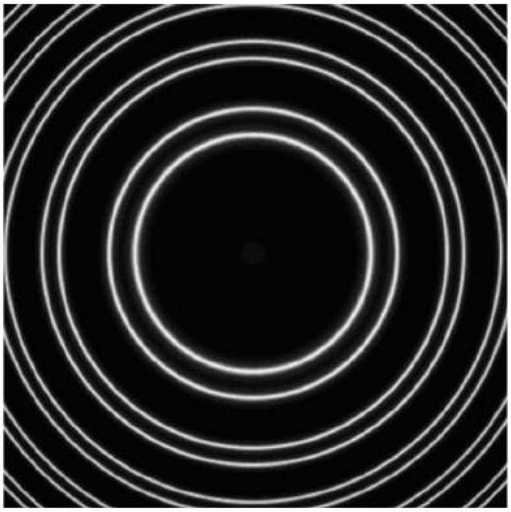


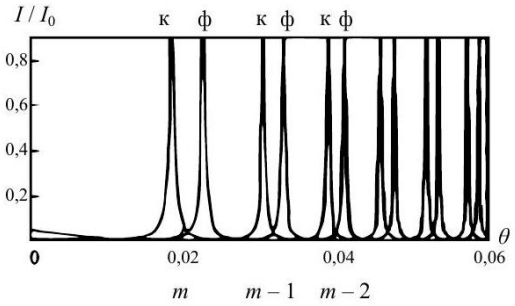

$$\Delta = 2L\left(\frac{1}{\cos\theta} - \operatorname{tg}\theta\sin\theta\right) = 2L\cos\theta \tag{4.6}$$

$$2L\cos\theta_m = m\lambda \tag{4.7}$$

$$A = \frac{A_0t}{1 - re^{ik\Delta}}$$

$$I = |A|^2 = \frac{|A_0|^2\,t^2}{1 + r^2 - 2r\cos\left[\frac{4\pi}{\lambda}L\cos\theta\right]} = \frac{I_0}{1 + \frac{4r}{(1-r)^2}\sin^2\left(\frac{2\pi}{\lambda}L\cos\theta\right)} \tag{4.8}$$

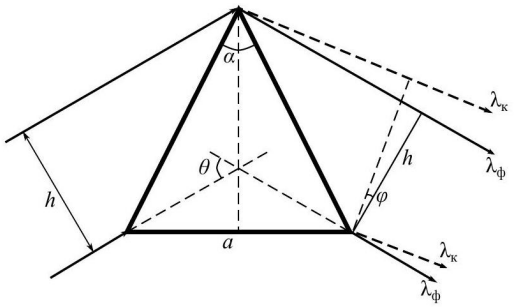




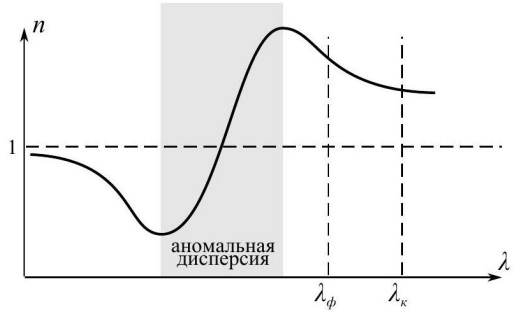

**2.4.**

$\lambda_1 = 4880\text{Å}$ $n_1 = 1661$ $\lambda_2 = 6327\text{Å}$ $n_2 = 1644$.

$$\frac{dn}{d\lambda} \approx \frac{n_1 - n_2}{\lambda_1 - \lambda_2} = -117 \cdot 10^3 \text{ cm}^{-1}$$




$$\Delta = a[n(\lambda + \delta\lambda) - n(\lambda)] \approx a\frac{dn}{d\lambda}\delta\lambda = h\delta\varphi \tag{4.9}$$



**3. Angular dispersion**

$$D(\lambda) = \frac{d\varphi}{d\lambda} \tag{4.10}$$

$$D = \frac{d\varphi}{d\lambda} = \frac{m}{d\cos\varphi} = \frac{m}{\sqrt{d^2 - m^2\lambda^2}} \tag{4.11}$$

$\varphi \ll 1$ $D \approx m/d$.

$$D = \frac{d\varphi}{d\lambda} = \frac{m}{d\cos\varphi} = \frac{m}{\sqrt{d^2 - (m\lambda + d\sin\psi)^2}} \tag{4.12}$$

$\psi \approx 30 - 60°$.
$\theta_m \ll 1$
$m \approx M = 2L/\lambda$

$$-2L\sin\theta_m d\theta = -2L\theta_m d\theta = md\lambda \approx \frac{2L}{\lambda}d\lambda$$

$$D = \frac{d\theta}{d\lambda} = -\frac{m}{2L\sin\theta_m} \approx -\frac{1}{\lambda\theta_m} \tag{4.13}$$

$$D = \frac{d\varphi}{d\lambda} = \frac{a}{h}\frac{dn}{d\lambda} \tag{4.14}$$

**4. Angular resolution**

$$R = \frac{\lambda}{\delta\lambda} \tag{4.15}$$

**4.1.**

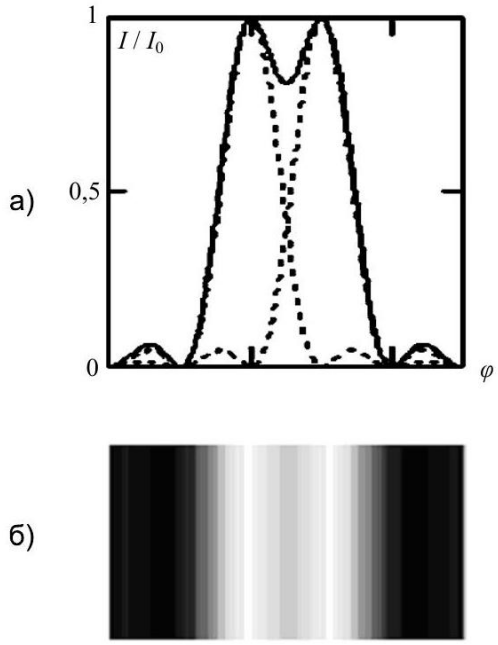
$$d\sin(\varphi_m + \delta\varphi) = m\lambda + \frac{\lambda}{N}$$

$$\delta\varphi = \frac{\lambda}{Nd\cos\varphi_m} \tag{4.16}$$

$$\Delta\varphi \approx D\delta\lambda = \frac{m}{d\cos\varphi_m}\delta\lambda \tag{4.17}$$

$$R = Nm \tag{4.18}$$

$\rho \sim \lambda/\psi$
$\psi = b/z$




$2\left(\frac{\sin \pi/2}{\pi/2}\right)^2 \approx 081$

$$\delta\varphi = \frac{a}{h}\left|\frac{\partial n}{\partial \lambda}\right| \delta\lambda = \frac{\lambda}{h}$$

$$R = \frac{\lambda}{\delta\lambda} = a\left|\frac{dn}{d\lambda}\right| \tag{4.19}$$

**4.2. -**

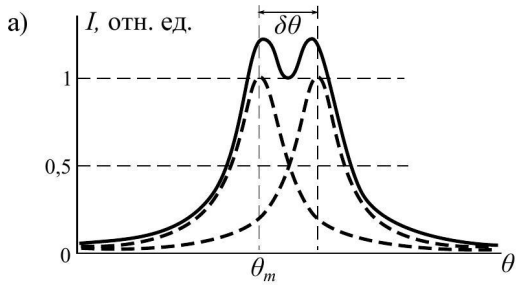
$\theta_m + \delta\theta/2$

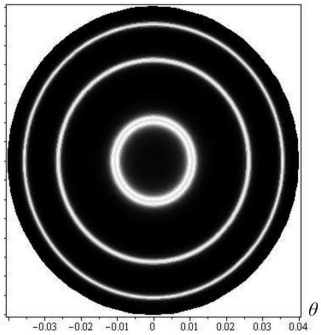
$$\frac{4r}{(1-r)^2}\sin^2\left[\frac{\pi}{\lambda}2L\cos\left(\theta_m + \delta\theta/2\right)\right] = 1$$

$\left(\left|\frac{\pi}{\lambda}\theta_m\delta\theta\right| \ll 1\right)$

$$\sin^2\left[\frac{\pi}{\lambda}2L\cos\left(\theta_m + \delta\theta/2\right)\right] =$$
$$= \sin^2\left(\frac{2\pi L}{\lambda}\cos\theta_m - \frac{\pi L}{\lambda}\sin\theta_m \cdot \delta\theta\right) \approx \left[\frac{\pi L}{\lambda}\theta_m\delta\theta\right]^2$$

$$\delta\theta = \frac{1-r}{2\pi\sqrt{r}}\frac{\lambda}{2L}\frac{1}{\theta_m} \approx D\delta\lambda/2 = \frac{m}{4L\theta_m}\delta\lambda$$






$r = 085\lambda = 57600\text{Å}$

$$\delta\lambda = 009\text{Å} L = 1\text{m})$$

$m \approx 2L/\lambda$

$$R = \frac{\lambda}{\delta\lambda} = \frac{\pi\sqrt{r}}{1-r}m \tag{4.20}$$

**5. Range of dispersion**

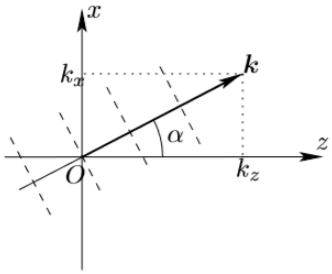
$m(\lambda + \Delta\lambda) = (m+1)\lambda$

$$\Delta\lambda = \frac{\lambda}{m} \tag{4.21}$$

$$\Delta\lambda = \frac{\lambda^2}{2L} \tag{4.22}$$

## 1.1.8 Fourier optics: basics

**Принципы фурье оптика**




В методе Рэлея волновое поле представляется в виде суперпозиции плоских волн, каждой из которых отвечает определённое направление волнового вектора. Иными словами, рассматривается пространственное преобразование Фурье. Если осуществить пространственное разделение этих компонент, то можно осуществлять преобразование поля, изменяя амплитуду или фазу отдельных фурье-гармоник. Разделение фурье-компонент производится, например, линзой в её фокальной плоскости, поскольку плоские волны, отличающиеся направлением, собираются в несовпадающие точки в фокальной плоскости (рис. 8.4.1):

$x = f\theta \approx f\frac{k_x}{k}$, или $k_x = \frac{kx}{f}$.

Поскольку поле $A_S$, падающее на линзу, есть суперпозиция волн разных направлений:

$$A_S(x) = \int a\left(k_x\right) e^{ik_x x} \frac{dk_x}{2\pi}, a\left(k_x\right) = \int A_S(x) e^{-ik_x x} dx,$$

то поле в фокальной плоскости $A_f$ даёт представление о спектре плоских волн, образующих изучаемое поле:

$$A_f(x) \sim a(kx/f)$$

В этой связи фокальную плоскость называют фурье-плоскостью. 8.4.2. Теория Аббе формирования изображений По теории, предложенной немецким физиком-оптиком Эрнестом Аббе (1904), формирование изображения с помощью одиночной линзы происходит в два этапа (рис. 8.4.2).

На первом этапе свет распространяется от плоскости $P_1$ до фурьеплоскости $F$, где формируется пространственный спектр предметной волны (первая дифракиия): линза Л осуществляет первое пространственное преобразование Фурье и создаёт в плоскости $F$ первичное изображение.

$$f(x,z) = ae^{i(k_x x + k_z z + \varphi)}$$

$$f(x,z) = ce^{i(ux + \sqrt{k^2-u^2}\cdot z)}$$

$$f(x,z) = f(x,0) \cdot e^{i\sqrt{k^2-u^2}z}$$

$$f_0(x) = \sum c_n e^{iu_n x}$$

$$f(x,z) = \sum_n c_n e^{i\left(u_n x + \sqrt{k^2-u_n^2}z\right)}$$

$$f(x,z) = \int C_0(u) e^{i\sqrt{k^2-u^2}z} e^{iux} du = \int C(u,z) e^{iux} du$$

**передаточная функция**

*передаточная функция*:

$$H(u) = e^{i\sqrt{k^2-u^2}z}$$

## 2. Fourier optics fundamentals

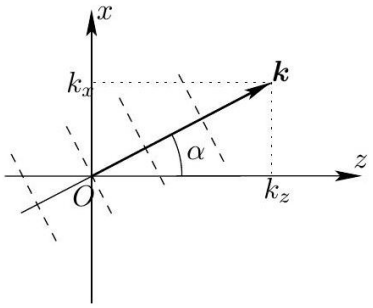
### 2.1.

$\omega_n f(t) = \sum c_n e^{i\omega_n t}$.

$$f(xz) = ae^{i(k_x x + k_z z + \varphi)} \tag{3.38}$$

$a$ - $\varphi$ - $k_x = k\sin\alpha$ $k_z = k\cos\alpha$ - $\boldsymbol{k}(k = |\boldsymbol{k}|)$ $x$ $z$ $(k_y = 0)$.
$c = ae^{i\varphi}$
$k_x^2 + k_z^2 = k^2$ . . $k_z = \sqrt{k^2 - u^2}$ (3.38)

$$f(xz) = ce^{i\left(ux + \sqrt{k^2-u^2}\cdot z\right)} \tag{3.39}$$




$$f(x0) = ce^{iux} \tag{3.40}$$

$$f(xz) = f(x0) \cdot e^{i\sqrt{k^2-u^2}z} \tag{3.41}$$

$H(u) = e^{i\sqrt{k^2-u^2}\cdot z}$
$f(t) = ce^{i\omega t}$
$u = k\sin\alpha$

$$f_0(x) = \sum c_n e^{iu_n x} \tag{3.42}$$

$u_n = k\sin\alpha_n$.
$k = |\boldsymbol{k}| = \omega/v$

$$f_0(x) = \frac{1}{2\pi}\int C_0(u)e^{iux}du \tag{3.43}$$

$$C_0(u) = \int\limits_{-\infty}^{\infty} f_0(x)e^{-iux}dx \tag{3.44}$$

$$f_0(x) = \sum_n c_n e^{inu_0 x} \tag{3.45}$$

$$c_n = \frac{1}{d}\int\limits_{-d/2}^{d/2} f_0(x)e^{-inu_0 x}dx \tag{3.46}$$

$d = \frac{2\pi}{u_0}$ - $f_0(x)$.

$$f_n(xz) = c_n e^{i\left(u_n x + \sqrt{k^2-u_n^2}z\right)}$$

$$f_n(x0) = c_n e^{iu_n x}$$

$$f(xz) = \sum_n c_n e^{i\left(u_n x + \sqrt{k^2-u_n^2}z\right)} \tag{3.47}$$

$$f(xz) = \int C_0(u)e^{i\left(ux+\sqrt{k^2-u^2}z\right)}du \tag{3.48}$$

$\varphi_n = \sqrt{k^2-u_n^2}\cdot z$

$$f(xz) = \int C_0(u)e^{i\sqrt{k^2-u^2}z}e^{iux}du = \int C(uz)e^{iux}du$$

$$C(uz) = C_0(u)e^{i\sqrt{k^2-u^2}z} \tag{3.49}$$

$$H(u) = e^{i\sqrt{k^2-u^2}z} \tag{3.50}$$

$$H(u) \approx e^{ikz}e^{-i\frac{z}{2k}u^2}$$

$$\sqrt{k^2-u^2}z \approx kz - \frac{z}{2k}u^2 \tag{3.51}$$

$$C(u,z) \approx C_0(u)e^{-i\frac{z}{2k}u^2} \tag{3.52}$$

**2.2.**

**2.2.1.**

$d = 2\pi/\Omega$ (. (3.4))

$$t(x) = 1 + m\cos\Omega x \quad (m < 1)$$

$a$ . . $f_s(x) = ae^{ikz}$. $z = 0$
$f_s = a$.
(3.2)

$$f_0(x) = a(1 + m\cos\Omega x)$$

$f_0(x)$ (3.46).

$$\cos\Omega x = \frac{1}{2}e^{i\Omega x} + \frac{1}{2}e^{-i\Omega x}$$

$$f_0(x) = a + \frac{am}{2}e^{i\Omega x} + \frac{am}{2}e^{-i\Omega x} \tag{3.53}$$

$z$
(3.53) ((3.42) (3.47))

$$f(xz) = ae^{ikz} + \frac{am}{2}e^{i\left(\Omega x+\sqrt{k^2-\Omega^2}z\right)} + \frac{am}{2}e^{i\left(-\Omega x+\sqrt{k^2-\Omega^2}z\right)} \tag{3.54}$$

. . $f_0(x)$ $z > 0$. - $a$ $z$. - $\frac{am}{2}$ $\pm\Omega$. $\sin\alpha = \pm\Omega/k = \pm\lambda/d$.
$x = 0$ $z = 0_+$
$d = \frac{2\pi}{\Omega}$ $\lambda$
$k\Omega \ll k$.
(3.51)

$$f(xz) = ae^{ikz} + \frac{am}{2}e^{ikz}\cdot e^{i\left(\Omega x-\frac{z}{2k}\Omega^2\right)} + \frac{am}{2}e^{ikz}\cdot e^{i\left(-\Omega x-\frac{z}{2k}\Omega^2\right)}$$

$$f(xz) = ae^{ikz}\left[1 + me^{-i\frac{z}{2k}\Omega^2}\cdot\cos\Omega x\right]$$

$z$ .

1. $z_n$ $\frac{z_n}{2k}\Omega^2 = 2\pi n$ . . $z_n = \frac{2d^2}{\lambda}n$ $e^{-i\frac{z_n}{2k}\Omega^2} = 1$ $f(xz) = ae^{ikz_n}(1 + m\cos\Omega x) = e^{ikz_n}f_0(x)$ .. $f_0(x)$.
   $I\,(xz_n) = I_0(x)$. $\mathcal{V} = \frac{I_{\max}-I_{\min}}{I_{\max}+I_{\min}}$
   $m$ $m^2$. $I\,(xz_n) = |f\,(xz_n)|^2 \approx a^2(1 + 2m\cos\Omega x)$ $\mathcal{V} \approx 2m$.
2. $z_n$ $e^{-i\frac{z_n}{2k}\Omega^2} = i$ . . $\frac{z_n}{2k}\Omega^2 = \frac{\pi}{2} + 2\pi n$ $f(xz) = a(1 + im\cos\Omega x)$. $I\,(xz_n) \approx a^2 = \text{const}$. $m^2$ $x$ $\mathcal{V} \approx 0$.

z - $z > 0$ $z$
$\sin\alpha = \pm\Omega/k = \pm\lambda/d$.

**2.2.2.**

$$t(x) = e^{im\cos\Omega x}$$

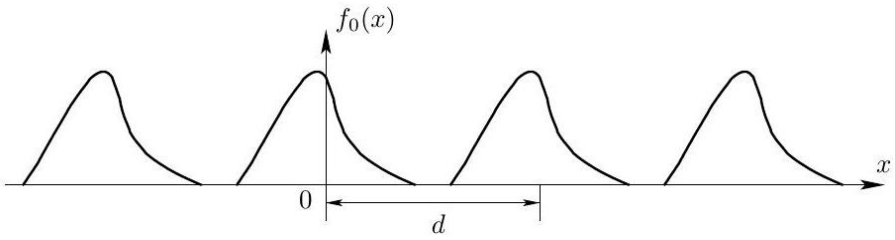
$m \ll 1$.

$$t(x) \approx 1 + im\cos\Omega x = 1 + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x}$$

$$f(xz) = ae^{ikz} + \frac{iam}{2}e^{i\left(\Omega x+\sqrt{k^2-\Omega^2}z\right)} + \frac{iam}{2}e^{i\left(-\Omega x+\sqrt{k^2-\Omega^2}z\right)}$$

$i = e^{i\pi/2}$

**2.3.** ()




$$f_0(x) = \sum c_n e^{in\frac{2\pi}{d}x} \tag{3.55}$$

(3.55) $u_n = n\frac{2\pi}{d}$ $\boldsymbol{k}_n$ $(u_n = k\sin\alpha_n)$
$\sin\alpha_n = n\frac{\lambda}{d}$.

$$f(xz) = \sum c_n e^{i\left(u_n x+\sqrt{k^2-u_n^2}z\right)}$$

$n$- $z = \text{const}$ $\varphi_n = \sqrt{k^2 - u_n^2}z \approx kz - \frac{zu_n^2}{2k}$ ((3.51)). $n$- $\varphi_0$ $z\varphi_0 = kz$.

$$\Delta\varphi_n = \varphi_0 - \varphi_n = \frac{z}{2k}\left(\frac{2\pi}{d}\right)^2 n^2 = \pi\frac{\lambda z}{d^2}n^2 \tag{3.56}$$

$$z_1 = \frac{2d^2}{\lambda} \tag{3.57}$$

( $p = \frac{\sqrt{\lambda z_1}}{d}$ . . ) $\Delta\varphi_n = 2\pi n^2$ .. $2\pi$.

$(n_1 2\pi/d$
$n_2 2\pi/d)$ $2\pi$ $(n_1^2 - n_2^2)$ $2\pi$.
$2\pi$
((3.55) $f_0(x)$) (3.57). ( $2\pi$)
. . $z_1$ $f_0(x)$ $e^{ikz}$

$$f\,(xz_1) = e^{ikz_1} f_0(x)$$

$$z_m = m\frac{2d^2}{\lambda} \quad (m = 12\ldots) \tag{3.58}$$

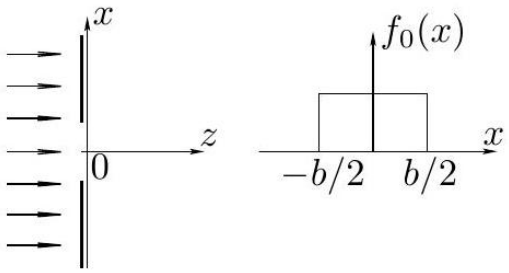
**2.4.**

$$f_0(x) = \begin{cases} 1 & |x| \leqslant b \\ 0 & |x| > b \end{cases}$$

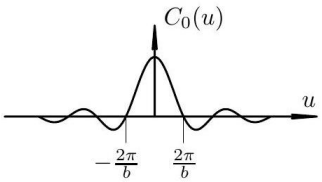
$$C_0(u) = b\frac{\sin\frac{b}{2}u}{\frac{b}{2}u} \tag{3.59}$$

$u = k\sin\theta$

$$|\Delta u| \lesssim \frac{2\pi}{b} \tag{3.60}$$




. 3.27.




. 3.28.
$\Delta x$ $\Delta u$ [8]

$$\Delta x \cdot \Delta u \sim 2\pi \tag{3.61}$$

- $b$ em

$$\Delta u = k\Delta(\sin\alpha)$$

$\Delta(\sin\alpha) \approx \lambda/b$

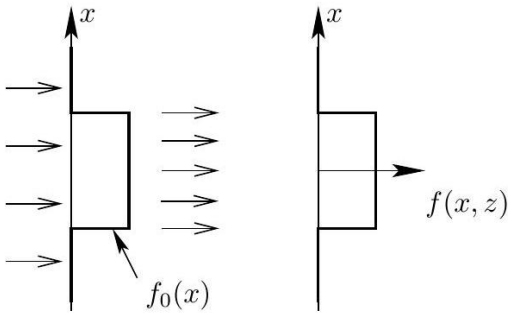
$$\Delta\alpha \approx \frac{\lambda}{b}$$

$b$.

[0]8 $\tau$ $b$ $\omega$ - $u$.

**2.5.**

**2.5.1.**




. 3.29.
(
)
$f(xz)$ $f_0(x)$ -
3.29.
() .
(b) $a$. $f_0(x)$
3.27 $C(u)$ (- $f_0(x)$) 3.28
$\Delta u$
$|\Delta u| < \frac{2\pi}{b}$.
.
$z(u = 0)$ $kz$. $u$ $\varphi(u) = \sqrt{k^2 - u^2} \cdot z$ (. (3.51)) $\varphi(u) \approx kz - \frac{z}{2k}u^2$.

$$\Delta\varphi < \frac{z}{2k}\left(\frac{2\pi}{b}\right)^2 = \pi\frac{\lambda z}{b^2} \tag{3.62}$$

$\pi$
$|u| < \frac{2\pi}{b}$ $kz$.
$f_0(x)$ $z = \text{const} > 0$
($e^{ikz}$) $f_0(x)$

$$f(xz) = e^{ikz} f_0(x)$$

$I(xz)$ $I(xz) = I_0(x) = |f_0(x)|^2$ . .
(3.62) $\Delta\varphi \ll \pi$

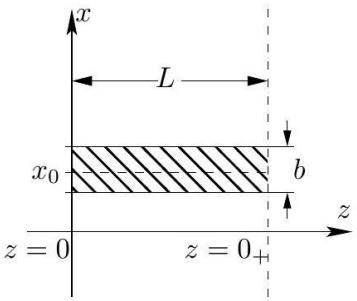
$$p = \frac{\sqrt{\lambda z}}{b} \ll 1 \tag{3.63}$$

$z$ $b$ $\lambda \to 0$ $\lambda$ (3.63) $z$ () $b$.

**2.5.2.**

(3.3). ( $z = 0$ $z = 0_+$ ) $L$. $b$ (. 3.30).
((3.3))
- -

$\Delta t$ $\Delta\omega$ $\Delta t\Delta\omega \sim 2\pi$.




. 3.30.
(3.63)

$$b \ll \sqrt{\lambda L} \tag{3.64}$$

-
$t(x) = a(x)e^{i\varphi(x)}$ - $a(x)$ - $\varphi(x) = k\,\mathrm{Ln}(x)$ - ($x = x_0 - b/2$ $x = x_0 + b/2$ $z = 0_+$ 3.30) $\Delta a \approx \frac{da}{dx}b$ $\Delta\varphi \approx \frac{d\varphi}{dx}b = kLb\frac{dn}{dx}$

$$\Delta a \approx \frac{da}{dx}b \ll a \quad \Delta\varphi \approx kLb\frac{dn}{dx} \ll \pi \tag{3.65}$$

(3.64) (3.65)

$$\sqrt{\lambda L} \ll b \ll \frac{\lambda}{2L\frac{dn}{dx}}$$

$L$ $\frac{dn}{dx}$

$$\frac{dn}{dx} \ll \frac{1}{L}\sqrt{\frac{\lambda}{L}} \tag{3.66}$$

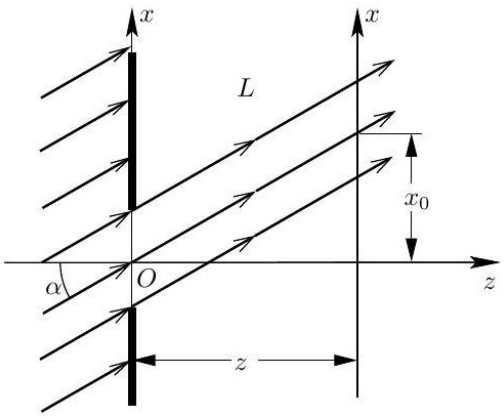
**2.5.3.**

$f_s(x) = ae^{ik\sin\alpha\cdot x}$ $u_0 = k\sin\alpha$ - $b$ (. 3.31). $z = 0_+$

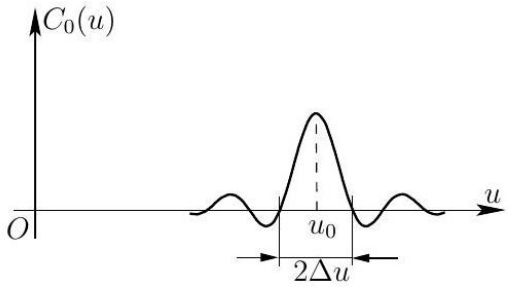
$$f_0(x) = \begin{cases} ae^{ik\sin\alpha\cdot x} & |x| \leqslant \frac{b}{2} \\ 0 & |x| > \frac{b}{2} \end{cases}$$

$z > 0$ ( $-\infty < x < \infty$)
($\alpha$) - $|x| \leqslant \frac{b}{2}$.
- $u_0$ -

$$C_0(u) = \frac{\sin\left[\frac{b}{2}(u - u_0)\right]}{\frac{b}{2}(u - u_0)} \tag{3.67}$$




. 3.31.




. 3.32.
($C_0(u)$ - - $f_0(x)$ . 3.32).
$b$ $u_0$ $|\Delta u| \leqslant \frac{2\pi}{b}$.
$z = 0$ $x_0$

$$C_1(u) = C_0(u)e^{-i(u-u_0)x_0} \tag{3.68}$$

($f_0(x)$ $x$ $x_0$ $e^{-i(u-u_0)x_0}$.)
$C_0(u)$ $z > 0$? $z$ $u_0$ $\varphi_0 \approx kz - \frac{z}{2k}u_0^2$. $u_0 + \Delta u$ ( $|\Delta u| < \frac{2\pi}{b}$)

$$\varphi = kz - \frac{z}{2k}(u_0 + \Delta u)^2 = kz - \frac{z}{2k}u_0^2 - \frac{z}{k}u_0 \cdot \Delta u - \frac{z}{2k}(\Delta u)^2.$$

$$\Delta\varphi = \varphi_0 - \varphi = \frac{z}{k}u_0 \cdot \Delta u + \frac{z}{2k}(\Delta u)^2$$

$$\Delta\varphi = z\sin\alpha \cdot \Delta u + \pi\frac{\lambda z}{b^2}$$

(3.63)

$$\Delta\varphi = z \cdot \sin\alpha\,(u - u_0)$$

$z$

$$C(u) \sim C_0(u)e^{-i\frac{z}{2k}u_0^2}e^{-iz\sin\alpha(u-u_0)}$$

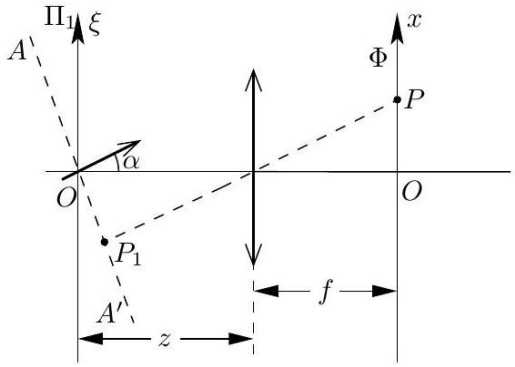
. .
$z$ $\alpha$ ( $u_0$) ($x$) $x_0 = z\sin\alpha$ ((3.68).
(3.63)
- .
$z \gg \frac{b^2}{\lambda}$ $\Delta x$
$|\Delta x| = \frac{\lambda}{b}z \gg b$.
$|x| < b$ $z \ll \frac{b^2}{\lambda}$.

**2.6.**

$\alpha$ (. 3.33) . . $u = k\sin\alpha$

$$x = f\,\mathrm{tg}\,\alpha \approx f\sin\alpha = \frac{fu}{k} \tag{3.69}$$

$\Pi_1$ $z$ $f_0(x)$. $\alpha_n$ . .
(3.42). $c_n e^{iu_n x}$
3.33 $\alpha = \alpha_n$ $\left(\sin\alpha_n = \frac{u_n}{k}\right)$ $c_n = a_n e^{i\varphi_n}$ $a_n$ $\varphi_n$ ($\xi = 0$ $\Pi_1$; $P_1$ $AA'$).




. 3.33.
$P$ $a_n$
$P$ $P_1$ ( $\xi = 0$) $P_1P$ $P_1P$ -
($P_1P$ ) $P_1P$ $\psi_n$. $P$ $\varphi_n + \psi_n$ $f(x_n) \sim a_n e^{i(\varphi_n+\psi_n)} = c_n e^{i\psi_n}$.
$c_0 e^{iu_0 x}$ ($u_0 = 0$) ($\alpha = 0$). $x = 0$ $\Phi$. $c_0 = a_0 e^{i\varphi_0}$ $a_0$ $\varphi_0$ . . $O$ $\Pi_1$. $\psi_0$ $x = 0$ $\Phi$
$OO$. $f(0) \sim a_0 e^{i(\varphi_0+\psi_0)} = c_0 e^{i\psi_0}$.
$z = f$ . . $\Pi_1$ -
($\cos\alpha \approx 1 - \alpha^2/2$) $P_1P = z\cos\alpha + \frac{f}{\cos\alpha} + \Delta_0$ $OO = z + f + \Delta_0$ ($\Delta_0$ ) $z = f$ .. $\psi_n = \psi_0$ (n-
(3.42)).
$\psi_n$
. .
$e^{i\psi_0}$ ($u_n$) $f(x_n) = c_n = C\left(\frac{kx_n}{f}\right)$. $x_n$
..
$n$

$$f(x) = C\left(\frac{kx}{f}\right) = \int f_0(\xi) e^{-i\frac{kx}{f}\xi} d\xi \tag{3.70}$$

$C(u)$ - $f_0(x)$.
$f(x)$ $f_0(\xi)$
$z \neq f$

$$f(x) = e^{i\frac{k}{2f}\left(1-\frac{z}{f}\right)x^2} \cdot C\left(\frac{kx}{f}\right)$$

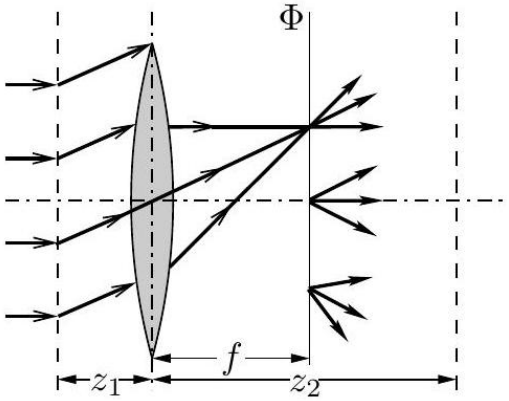
..
$\varphi(u) = \sqrt{k^2 - u^2}\Delta z$
$\Delta z = z - f$.

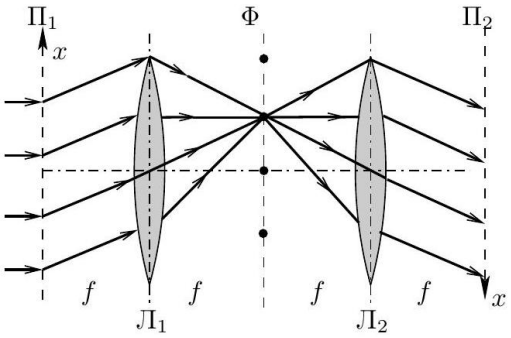
$$I(x) = \left|C\left(\frac{kx}{f}\right)\right|^2$$

$z$ (. .
).

**2.7.**

$u = k \sin \alpha$.




**2.8.**

**2.8.1.** ()

$$f_0(x) = \frac{1}{2\pi} \int C_0(u) e^{iux} dx$$

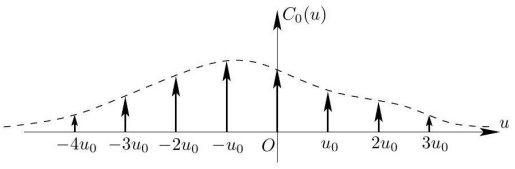
$$u_n = k \sin \alpha_n = k \frac{x_n}{f} = \frac{knd}{f} = 2\pi n \frac{d}{\lambda f}$$

$u_n = n u_0$
$u_0 = 2\pi \frac{d}{\lambda f}$

$$f(x) \sim \sum C_0 (n u_0) e^{i n u_0 x}$$

$$C_0 (n u_0) = \int\limits_{-\frac{a}{2}}^{\frac{a}{2}} f_0(x) e^{-i n u_0 x} dx \qquad (3.71)$$

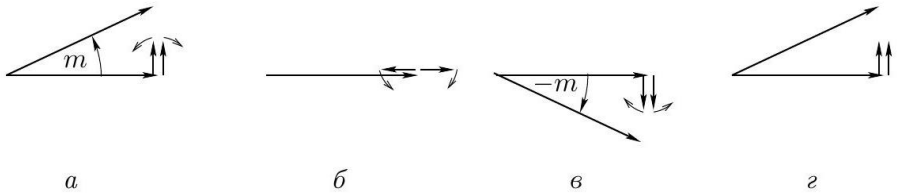
$(a \ f_0(x) \equiv 0 \ |x| > \frac{a}{2})$.

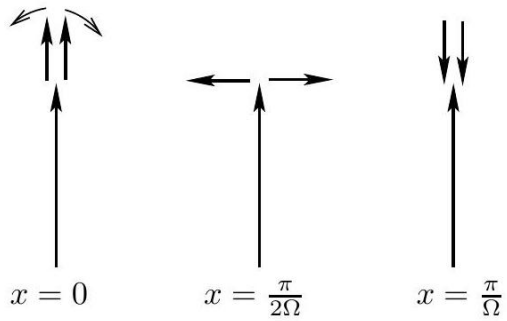



$d_0 = \frac{\lambda f}{d}$ $f_0(x)$.
$f(x) = \sum f_0 (x - n d_0)$

**2.8.2.**

$$f_0(x) = e^{im\cos\Omega x} \approx 1 + im\cos\Omega x = 1 + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x} \tag{3.72}$$




$$f(x) = e^{i\frac{\pi}{2}} + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x}$$

$$f(x) = i\left(1 + \frac{m}{2}e^{i\Omega x} + \frac{m}{2}e^{-i\Omega x}\right) = i(1 + m\cos\Omega x)$$

3.38.
$\pi/2$ $x = 0$
$x$
.

$$I(x) = |f(x)|^2 \approx (1 + m\cos\Omega x)^2 \approx 1 + 2m\cos\Omega x$$

$(m \ll 1$ $m^2)$. $f_0(x)$
- (. ).
. $\pi/2$ - -

$$f(x) = \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x} = im\cos\Omega x$$

-

$$I(x) = m^2\cos^2\Omega x$$

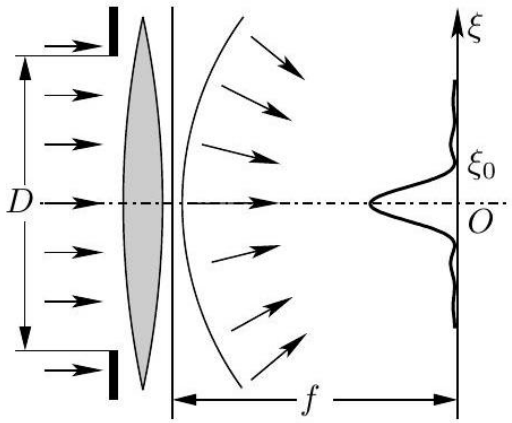
.

**2.9.**

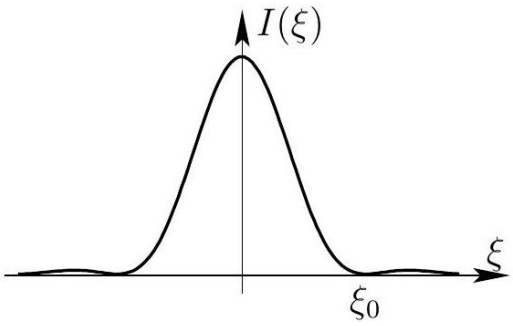
$D$ (3.8). ((3.7)) $D$

**2.9.1.**

. $D$ (. 3.39)




. 3.39.



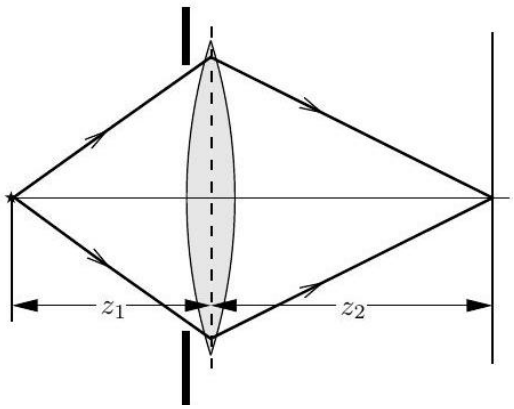

. 3.40.
() 3.39 $D$.
(..).
3.40 - (. (3.37))

$$\xi_0 = 122\frac{\lambda}{D}f \tag{3.73}$$

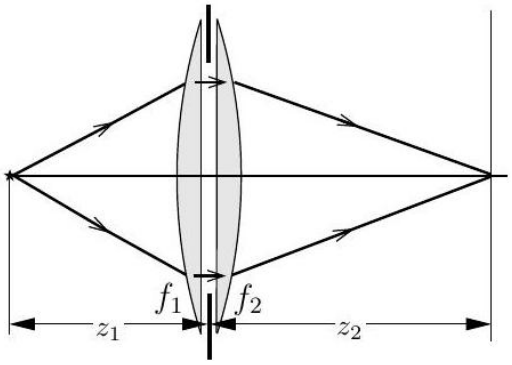
**2.9.2.**

$z_1$
$z_2$ $z_1$ $z_2$ $\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}$.

$$\xi_0 = 122\frac{\lambda}{D}z_2 \tag{3.74}$$

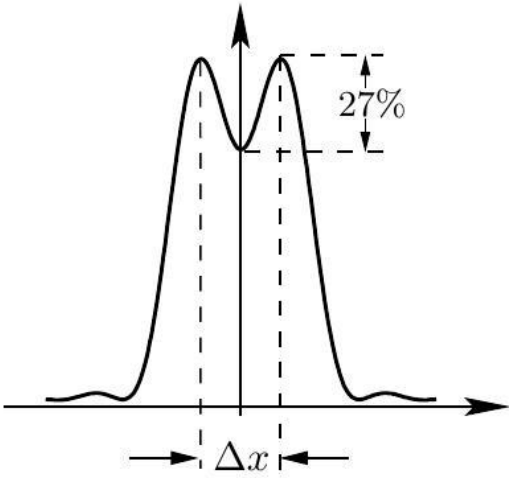


. 3.41.




. 3.42.
3.41 3.42 $f$ $_1$ $_2$ $f_1 = z_1$ $f_2 = z_2$. (. I) $\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{z_1} + \frac{1}{z_2}$
3.41 3.42 3.41. $_1$
$D$ $_2$.
($z_2$) - (3.74).

**2.9.3.**




$\Delta x = -\frac{z_2}{z_1}\ell$

$$\Delta x = \xi_0 = 122\frac{\lambda}{D}z_2$$

$l = \frac{z_1}{z_2}\Delta x$.

$$\ell_{\min} \approx 122\frac{\lambda}{D}z_1 \tag{3.75}$$

$\alpha = \frac{\ell}{z_1}$.

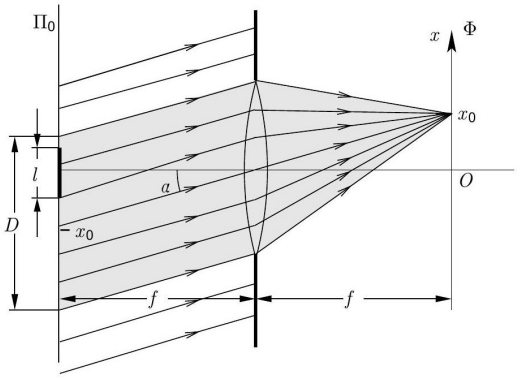
$$\alpha_{\min} \approx 122\frac{\lambda}{D} \tag{3.76}$$

**2.10.**

$u = k\sin\alpha$

$$f_0(x) = \frac{1}{2\pi}\int C_0(u)e^{iux}du \tag{3.77}$$

$u = k\sin\alpha$
$x_0 = f\,\mathrm{tg}\,\alpha \approx f\sin\alpha = f\frac{u}{k}$




$$-x_0 - \frac{D}{2} \leqslant x \leqslant -x_0 + \frac{D}{2} \qquad |x + x_0| \leqslant \frac{D}{2}$$

$|x| \leqslant \frac{D}{2}$
$u_0 = k\sin\alpha \approx k\,\mathrm{tg}\,\alpha = \frac{kx_0}{f}$ $\Delta u = \frac{2\pi}{D}$ (. 3.32). $D$ $\alpha$
$\frac{\lambda f}{D^2} \ll 1$ ( $D = 1$ $f = 2$ $\frac{\lambda f}{D^2} \simeq 10^{-4}$).
- $\Phi$ $C_0(u)e^{iux}$
$C_0(u)P_D\,(x + x_0)\,e^{iux}$ $x_0 = \frac{fu}{k}$ $P_D(\xi)$ -

$$P_D(\xi) = \begin{cases} 1 & |\xi| \leqslant \frac{D}{2} \\ 0 & |\xi| \geqslant \frac{D}{2} \end{cases}$$

$f_0(x)$ (3.77)

$$f(x) = \frac{1}{2\pi}\int C_0(u)P_D\left(x + \frac{fu}{k}\right)e^{iux}du \tag{3.78}$$

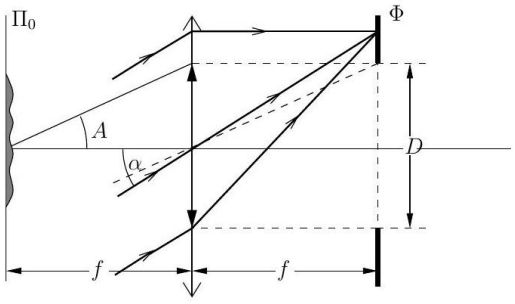
$|x| \leqslant \frac{l}{2}$ (3.77) (3.78)
$|x| > \frac{l}{2}$.
$C(u) = C_0(u)P_{\mathrm{D}}\left(x + \frac{fu}{k}\right)$. $l \ll D$ ( )
$u$ $P_D\left(\frac{fu}{k}\right)$. $\Delta u$ ($P_D\left(\frac{fu}{k}\right) = 1$)

$$C(u) = C_0(u)P_D\left(\frac{fu}{k}\right) \tag{3.79}$$

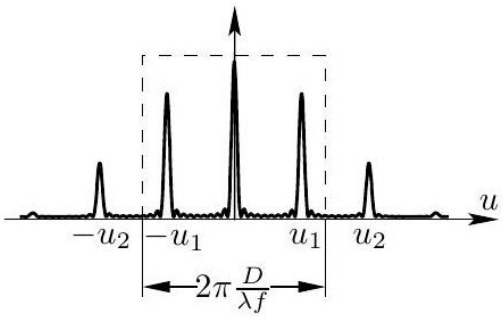
$$f(x) = \frac{1}{2\pi}\int C_0(u)P_D\left(\frac{fu}{k}\right)e^{iux}du \tag{3.80}$$

$$P_D\left(\frac{fu}{k}\right) = \begin{cases} 1 & |u| \leqslant \frac{kD}{2f} \\ 0 & |u| > \frac{kD}{2f} \end{cases} \tag{3.81}$$

$D$ $|u| \leqslant \frac{kD}{2f}$. $u = k\sin\alpha$
(
).




. 3.45. -




. 3.46.
(3.81)
$D$ -
$|\sin\alpha| > \frac{D}{2f}$ (. 3.45).
- $m\frac{\lambda}{d}f \leqslant \frac{D}{2}$ ($|u_m| \leqslant \pi\frac{D}{\lambda f}$). $P_D\left(\frac{fu}{k}\right)$ $|u| < \pi\frac{D}{\lambda f}$.

$$d_{\min} \geqslant \frac{2\lambda f}{D} \approx \frac{\lambda}{\sin A} \tag{3.82}$$

**фурье плоскость и теория аббе (???)**

от каждой гармоники после прохождения линзы в фокальной плоскости возникает пространственная картина спектра.

фокальную плоскость называют *фурье плоскостью.*

и говорят, что в фурье плоскости появляется первичное изображение.

в процессе распространения света от предмета до фурье-плоскости осуществляется преобразование Фурье светового поля (по терминологии Аббе — *первая дифракция*).

Формирование изображения по Аббе называется принципом двойной дифракции

**пространственная фильтрация**

(звучит интересно.
пространство фильтрует свет?)

Особая роль фурье-плоскости обусловлена тем, что именно в этой плоскости возможно избирательное воздействие на разные пространственные гармоники: установив в любой точке x фурье-плоскости маленькую пластинку, вносящую определённое поглощение и (или) фазовую задержку, мы изменим амплитуду и (или) фазу плоской волны с пространственной частотой u = kx f, не изменяя амплитуд и фаз других плоских волн.

В фурье плоскости можно что-то удалять, что-то добавлять в спектр.

Этим путём можно решать самые разнообразные задачи: улучшение качества изображений, разрешающей способности оптических систем, визуализация фазовых объектов, выполнение самых разнообразных преобразований пространственной структуры световых полей и т.д., т.е. решать широкий круг задач оптической обработки информации.

**наблюдение фазовых структур**

**метод фазового контраста и темного поля!!!**

(обещает быть очень интересным, даже я помню, что что-то такое было)
но на самом деле я ничего не помню, я очень рад буду дойти до этой темы.

**метод релея**

у участок свободного пространства можно рассматривать как простейший линейный пространственный фильтр.
(??? пока упущено.)

**Примеры расчёта дифракции спектральным методом**

то есть как я понял, можно по Г-Ф считать дифракцию, а можно по Фурье.
да?
скоро я узнаю точно, в чем же разница между этими методами.

**преобразование Фурье**

мощнейшая связка с матаном!!! обязательно уделю этому день.
(мб я уже написал это, вынесу потом в эту главу или дополню как-то, типа тут чисто математические методы будут, потому что там скорее всего много секретов.
а так - уже все сказано в первой подглаве тут, разве что я ничего не отработал.)
(в билетах написано, что это связано с полем в ф пл линзы)
стр 188 лабника!!!

**поле в фокусе плоской линзы**

?????
действительно, отличная ведь задача на фурье.

**Соотношение неопределенностей**

(допишу потом, очень редко это нужно)

**Другое о лазерах**

## 1.1.9 Radiation of light: basics

(see QM or QFT about it)

**Методы решения задач на излучение (!!!!!!)**

**процесса излучения (!?!?!!)**

**параметры излучения**

про интенсивность и все такое
какие параметры есть у нас?
все это.

### 1.1.10 Scattering of a light: basics

(see field theory about it and etc.)

**моделях рассеяния**

(основные модели тут. думаю, рассеяние важнее всего в оптике)

## 1.2 Mathematics for optics

**Fourier methods**

(write formulas for fourier transform, convolution, etc)
Parseval's identity:

$$\|f\|^2_{L^2(-\pi,\pi)} = \frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)|^2dx = \sum_{n=-\infty}^{\infty}|\hat{f}(n)|^2, \qquad \hat{f}(n) := \frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)e^{-inx}dx$$

$$\int_{-\infty}^{\infty}|\hat{f}(\xi)|^2d\xi = \int_{-\infty}^{\infty}|f(x)|^2dx \qquad f \in L^2(\mathbb{R})$$

**Useful Integrals**

# 2 Special parts of optics in a nutshell

## 2.1 Quantum optics fundamentals in a nutshell (!??!)

(see K Zhou's great notes)
(мб больше потом раздел сделаю, пока так, ибо не особо это нужно)

### 2.1.1 Laser optics: basics (?!)

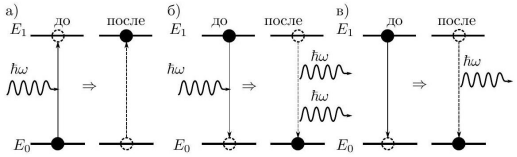
**Принцип работы лазеров**

**1. Laser radiation from labnik**



**1.1.**

$\hbar = h/2\pi$
$E_1 - E_0 = \hbar\omega.$




$\hbar\omega = E_1 - E_0.$

$$W_{0\to1} = W_{1\to0} = B\rho_\omega \tag{5.1}$$

**1.2.**

$I(x)dI = -\alpha I dx$

$$I(x) = I_0 \exp(-\alpha x)$$

$\alpha < 0$

$$I(x) = I_0 \exp(\gamma x) \tag{5.2}$$

$$\frac{dN_0}{N_0} = W_{0\to1}dt = B\rho_\omega dt$$

$$dN_\Phi^- = -dN_0 = -B\rho_\omega N_0 dt$$

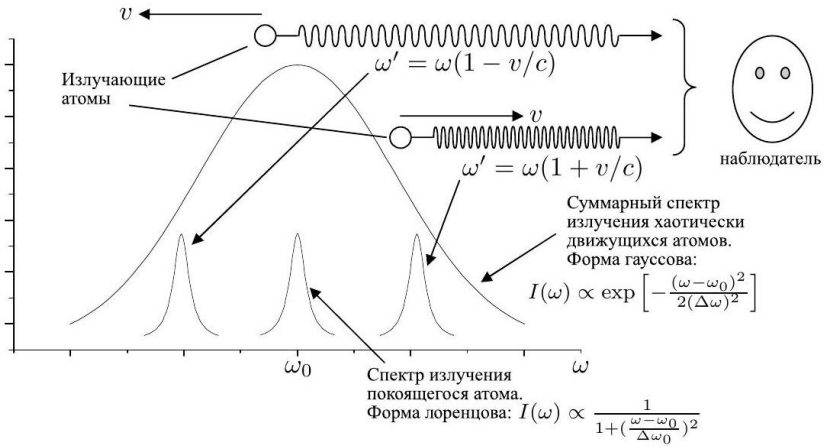
$$dN_\Phi^+ = B\rho_\omega N_1 dt$$

$$\frac{dI}{I} = \frac{dN_\Phi^+ + dN_\Phi^-}{\rho_\omega/\hbar\omega} = B\frac{\hbar\omega}{v}\left(N_1 - N_0\right)dx$$

$$\gamma = B\frac{\hbar\omega}{v}\Delta N \tag{5.3}$$

$\Delta N = N_1 - N_0.$ $(\gamma > 0)$
$N_1 > N_0.$

**1.3.**



$$\frac{\Delta\omega}{\omega} \sim \frac{v}{c}\cos\alpha$$

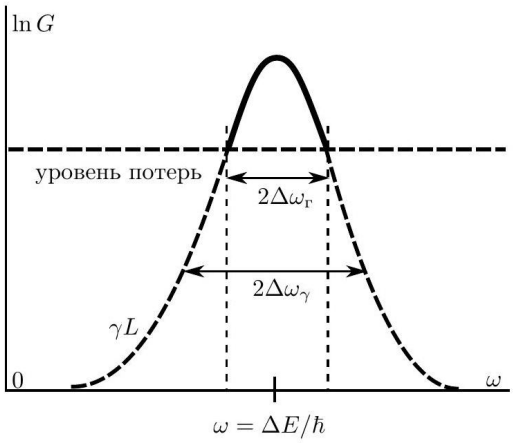
$$\Delta\omega \sim \omega\frac{v_T}{c} = \frac{\omega}{c}\sqrt{\frac{8k_{\mathrm{B}}T}{\pi m}} \tag{5.4}$$

**1.4.**

$$e^{\gamma L} \geqslant \frac{1}{\sqrt{Tr_1r_2}} \tag{5.5}$$

$$2\gamma L \geqslant -\ln T - \ln r_1r_2$$

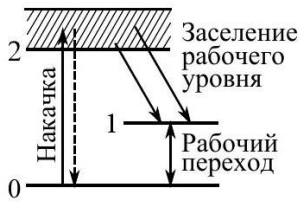
$I \propto \exp\left[\frac{(\omega-\omega_0)^2}{2\Delta\omega^2}\right]$
$\Delta\omega = \frac{\omega}{c}\sqrt{\frac{kT}{m}}$.




$G = 1/\left(Tr_1r_2\right)$.

**1.5. .**

$\frac{N_1}{N_0} = \exp\left(-\frac{E_1-E_0}{k_5T}\right)$.
$N_1 > N_0$




**2.**

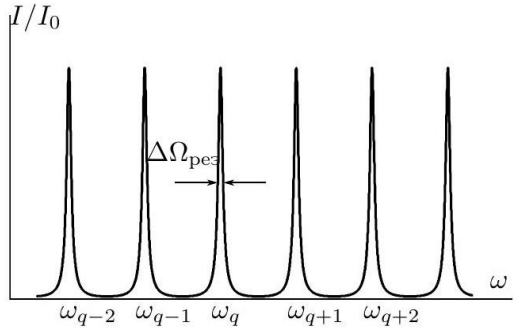
**2.1.**

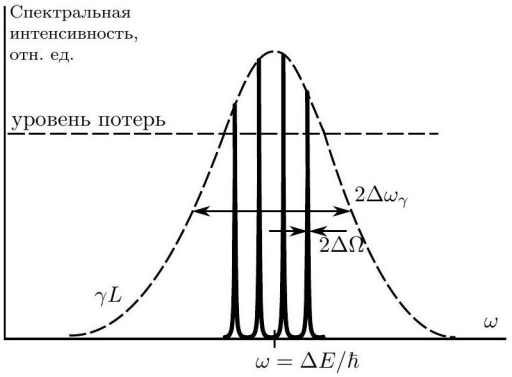
$$L = q\frac{\lambda}{2} \quad q = 123\ldots \tag{5.6}$$

$$\omega_q = q\frac{\pi c}{L} \quad \left(\nu_q = q\frac{c}{2L}\right) \quad q = 123\ldots \tag{5.7}$$

**2.2.**

$$Q \sim \frac{2\pi L}{\lambda}\frac{1}{1-r_1 r_2} \quad \Delta\Omega \sim \frac{\omega_q}{Q}$$



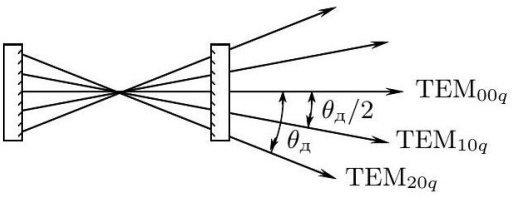




**2.3.**

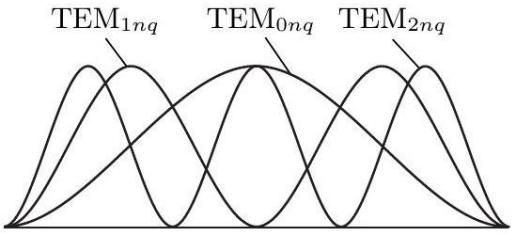
$$\frac{\Delta\Omega}{\omega} \sim \frac{\left(\Delta\Omega_{\text{рез}}\right)^2}{P}\cdot\frac{N_2}{\Delta N}$$

**2.4. ***

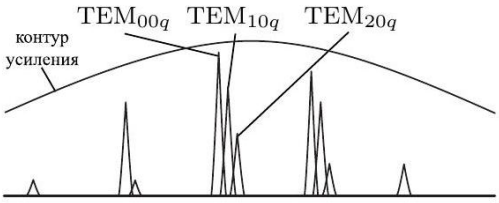
$$\theta = \pm\frac{\lambda}{2D}\sqrt{m^2+p^2} = \frac{\theta}{2}\sqrt{m^2+p^2} \tag{5.8}$$



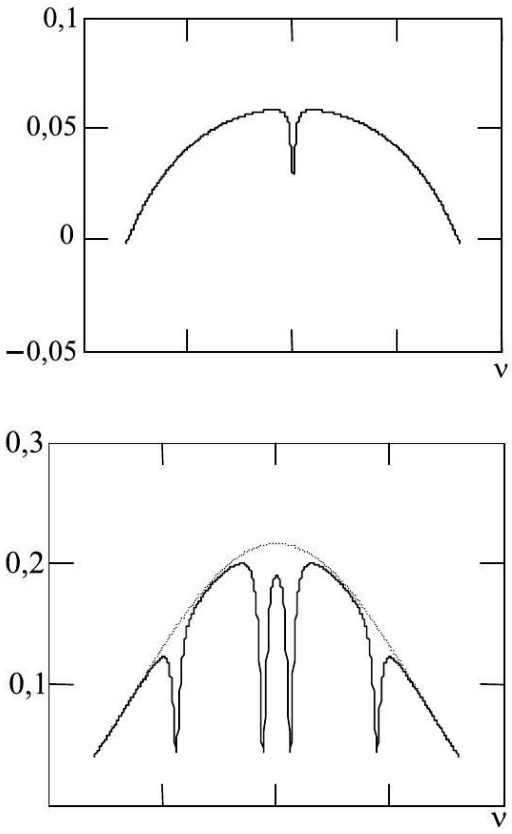

$$\omega_{mpq} \simeq \omega_{00q}\left[1+\frac{\left(m^2+p^2\right)\lambda^2}{8D^2}\right] \tag{5.9}$$




$$\frac{\omega_{mpq} - \omega_{00q}}{\omega_{00q+1} - \omega_{00q}} \simeq \frac{\lambda L\,(m^2 + p^2)}{4D^2} \tag{5.10}$$



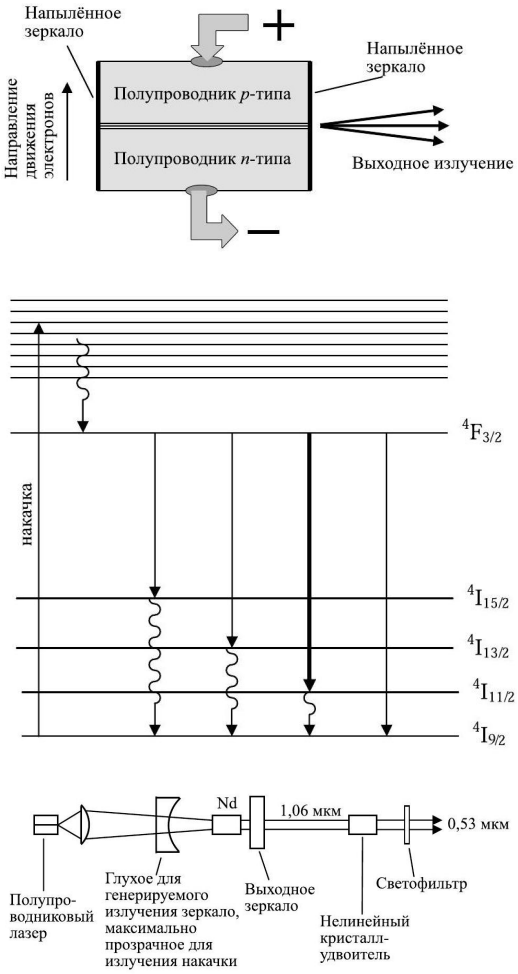

**2.5. ***




**3. Laser types ***

$10^9$ CD DVD Blue-Ray

.




### 2.1.2 Quantum States of Light

**Phase Space Distributions**

**Squeezed States**

**Other Nonclassical States**

### 2.1.3 Radiative Transitions, Jaynes–Cummings model, cavities

**Rabi Oscillations**

**The Jaynes–Cummings Model**

**Atoms in Cavities**

**когерентных состояниях**

### 2.1.4 Open optical systems, decoherences

### 2.1.5 Main formalism and equations of open systems

(see note on q optics for now)

### 2.1.6 Decoherence

(see note on q optics for now)

## 2.2 Some well-known optical setups

### 2.2.1 Optical fibers in a nutshell

(???)

# Part II
# Fundamentals of Optics

## 3 Fundamentals of Optics

Типичная оптическая конструкция распространения света - конструкция электромагнитных волны, которыми свет по сути и является. Приведем полную теорию и применения электромагнитных волн.

### 3.1 Philosophy of optics (?)

(все с нее и начинается, потом напишу очень глубоко ее.)

#### 3.1.1 The present moment

(конечно же, это то, что первое, что нужно обсудить)

#### 3.1.2 Beauty of the world

(конечно же, это то, что первое, что нужно обсудить)

#### 3.1.3 What is the nature of light?

Обсудим, как лучше подойти к описанию света.

**Свет как волны**

**Свет как квантовые частицы**

**Идея описания света интегралом по траекториям (?)**

**Идея описания света квантовой теорией поля (?)**

### 3.2 Electromagnetic waves for Optics

#### 3.2.1 Fundamentals of field theory for optics

(тут многое удалил, потому что проще заново написать. пока это в теории поля напишу, потом в оптику буду переписывать это для оптики.)

**Интенсивность по Ландау**

Таким образом, в геометрической оптике световую волну можно рассматривать как пучок лучей.

Лучи, однако, сами по себе определяют лишь направление распространения света в каждой точке; остается вопрос о распределении интенсивности света в пространстве.

Выделим на какой-либо из волновых поверхностей рассматриваемого пучка бесконечно малый элемент.

Из дифференциальной геометрии известно, что всякая поверхность имеет в каждой своей точке два, вообще говоря, различных главных радиуса кривизны.

Пусть ac и bd (рис. 7) - элементы главных кругов кривизны, проведенные на данном элементе волновой поверхности.

Тогда лучи, проходящие через точКи $a$ и $c$, пересекутся друг с другом в соответствующем центре кривизны $O_1$, а лучи, проходящие через $b$ и $d$, пересекутся в другом Центре кривИзны $O_2$.

При данных углах раствора лучей, исходящих из $O_1$ и $O_2$, длины отрезков $ac$ и $bd$ пропорЦиональны соответствующим радиусам кривизны $R_1$ и $R_2$ (т.е. длинам $O_1O$ и $O_2O$); площадь элемента поверхности пропорциональна произведению длин $ac$ и $bd$, т.е. пропорциональна $R_1R_2$. Другими словами, если рассматривать элемент волновой поверхности, ограниченный определенным рядом лучей, то при движениИ вдоль них площадь этого элемента будет меняться пропорцИонально $R_1R_2$.

С другой стороны, интенсивность, т.е. плотность потока энергии обратно пропорциональна площади поверхности, через которую проходит данное количество световой энергии.

Таким образом, мы приходим к выводу, что интенсивность

$$I = \frac{\text{const}}{R_1R_2}$$

эту формулу надо понимать следующим образом.

на каждом данном луче ($ab$ на рис. 7) существуют определенные точки $o_1$ и $o_2$, являющиеся центрами кривизны всех волновых поверхностей, пересекающих данный луч.

расстояния $oo_1$ и $oo_2$ от точки $o$ пересечения волновой поверхности с лучом до точек $o_1$ и $o_2$ являются радиусами кривизны $r_1$ и $r_2$ волновой поверхности в точке о.

таким образом, формула (54.1) определяет интенсивность света в точке о на данном луче как функцию от расстояний до определенных точек на этом луче.

подчеркнем, что эта формула непригодна для сравнения интенсивностей в разных точках одной и той же волновой поверхности.

Поскольку интенсивность определяется квадратом модуля поля, то длЯ изменения самого поля вдоль луча мы можем написать:

$$f = \frac{\text{const}}{\sqrt{R_1R_2}}e^{ikR}$$

ГДе в фазовом множителе $e^{ikR}$ под $R$ может подразумеваться как $R_1$, ТаК и $R_2$; величины $e^{ikR_1}$ и $e^{ikR_2}$ отличаются друг от друга тольКо Постоянным (длЯ данного луча) множителем, поскольку разность $R_1 - R_2$, расстояние между обоими центрами кривизны, постоянна.

Если оба радиуса кривизны волновой поверхности совпадают, то (54.1) и (54.2) имеют виД

$$I = \frac{\text{const}}{R^2}, \quad f = \frac{\text{const}}{R}e^{ikR}$$

Это имеет место, в частности, всегда в тех случаях, когда свет испус- кается точечным источником (волновые поверхности явлЯются тогДа Концентрическими сферами, а $R$ - расстоянием до источника света).

Из (54.1) мы видим, что интенсивность обращается в бесконечность в точках $R_1 = 0, R_2 = 0$, т.е. В центрах кривизны волновых поверхностей.

Применяя это ко всем лучам в пучке, находИм, что интенсивность света в данном пучке обращается в бесконечность, вообще говоря, на двух поверхностях - геометрическом месте всех центров кривизны волновых поверхностей.

Эти поверхности носят название каустик.

В частном случае пучка лучей со сферическими волновыми поверхностями обе каустики сливаются в одну точку (фокус).

Отметим, что, согласно известным из дифференциальной геометрии свойствам геометрического места центров кривизны семейства поверхностей, лучи касаются каустиК.

Надо иметь в виду, что (при выпуКлых волновых поверхностях) Центры кривизны волновых Поверхностей могут оказаться лежащими не на самиХ лучах, а на их продолженияХ за оптическую систему, от которой они исходят.

В таких случаях говорят о мнимьих каустиках (Или мнимых фокусах).

Интенсивность света при этом нигде не обраЩается в бесконечность.

Что касается обращения интенсивности в бесконечность, то в дейстВИтельности, разумеется, Интенсивность в точках каустиКи делается большой, но остается конечной (см. задачу к \$59).

Формальное обращение в бесконечность означает, что приближение геометрической оптики становится во всяКом случае неприменимым вблизи каустик.

С этим же обстоятельством связано и то, что изменение фазы вдоль луча может определяться формулой (54.2) только на участках луча, не включающих в себя точек его касания с каустиками.

Ниже (в §59) будет показано, что в действительности при прохождении мимо Каустики фаза поля уменьшается на $\pi/2$. Это значит, что если на участке луча до его касания первой каустики поле пропорционально множителю $e^{ikx}$($x$ - координата вдоль луча), то после прохождения мимо каустики поле будет пропорционально $e^{i(kx-\pi/2)}$. То же самое произойдет вблизи точки касания второй каустики, и за этой точКой поле будет пропорЦИонально $e^{i(kx-\pi)1)}$

**уравнения поля в среде**

**Суть**

!!!!!! разобраться наконец-то, что такое показатель преломления среды!!! давно путаница в этом.

$$\begin{aligned} \operatorname{rot}\vec{E} &= -\frac{1}{c}\frac{\partial\vec{B}}{\partial t}, \quad \operatorname{div}\vec{D} = 0, \\ \operatorname{rot}\vec{H} &= \frac{1}{c}\frac{\partial\vec{D}}{\partial t}, \qquad \operatorname{div}\vec{B} = 0. \end{aligned} \tag{3.1}$$

(электричество прямо очень нужно!) !!! чертова путаница с этими всеми темами!!! пока под ковер сметаю, потом в электричестве - это первая тема для разбора будет!!!

также есть материальные уравнения:

$$\boldsymbol{D = 4 + 4\pi P},\ \boldsymbol{B = H + 4\pi I}$$

(хороший вопрос, что тут что значит - потом отвечу)

В изотропных средах:

$$D = \upsilon E,\ B = \mu H$$

так как $P = \alpha E,\ I = kH$.

и также есть такие-то граничные условия.

## 3.2.2 Polarization of waves

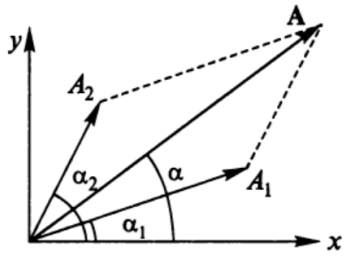
Обсуди подробно поляризацию.

**обзор задач поляризации**

Чаще всего нам придется иметь дело с задачами типа "какой-то свет шел через некоторый объект, который как-то влияет на поляризацию волны, что-то там раскладывается и потом собирается, и нужно определить, что со светом по итогу будет."

**описание вращающихся векторов**

Пусть есть два вектора под известными углами к заданной оси, которые в общем случае могут вращаться с $\omega$, требуется найти угол вектора, полученного в результате их суммы.




Из геометрии видно (????? мб тут написать лучше можно гораздо, пока не акутально рыться), что угол имеет вид

$$\operatorname{tg}\alpha = (A_1 \sin\alpha_1 + A_2 \sin\alpha_2) \,/\, (A_1 \cos\alpha_1 + A_2 \cos\alpha_2)$$

Также следует знать теорию эллипсов и вращения векторов с разными фазами и амплитудами. Она тут много применяется и есть очень нетривиальные ходы.

(потом напишу про нее тут)

Например, разберемся, что будет, если координаты точки вдоль осей двигаются по закону: $x = a\cos\omega t, \quad y = b\cos(\omega t + \delta)$?

Исключая время, находим траекторию движения:

$$\frac{x^2}{a^2} - \frac{2xy}{ab}\cos\delta + \frac{y^2}{b^2} = \sin^2\delta$$

Таким образом, фигура получается эллипсом.

(что еще известно?)

**векторы Герца (???)**

такое вообще нужно?

пироженко писала просто про это.

**векторную волну можно разложить на скалярные**

про это подробно напишу, говорят, что да.

пока не разорался еще.

**обзор типов поляризации**

Пусть на какой-то поляризационный объект падаем монохроматический свет (другие вроде случаи мы не рассматриваем). У него в общем случае $\mathbf{E}$ вращается по эллипсу (кст, нужно потом еще раз посмотреть это) вокруг $\mathbf{k}$.

Описание того, в какую сторону направлен в данный момент времени и как вращается **E**, производим в плоскости перпендикулярной вектору $k$.

Заметим, что **E** не перпендикулярен **k** в веществе, так что тут речь именно о выходе из вещества!

Начинаем описание с введения в нашей плоскости осей, на которые проецируем в данный момент вектор **E**.

Компоненты **E** имеют вид:

$$\begin{cases} E_x = E_{x0}\cos(\omega t + \varphi_x) \\ E_y = E_{y0}\cos(\omega t + \varphi_y) \end{cases}$$

То есть они гармонически колеблются.

Анализ поляризации удобно проводить в плоскости, которая вращается с $\omega$.

В ней наши проекции неподвижны, так что видно, как разность фаз $\varphi_1 - \varphi_2$ определяет тип поляризации.

Например, если амплитуды одинаковые и разность фаз $\pm\pi/2$, то это соответствует *круговой поляризации.* В данном случае в неподвижной плоскости **E** описывает круг, а в вращающейся - векторы-проекции перпендикулярны друг другу.

Также если $\varphi_1 - \varphi_2 = n\pi(n \in Z)$, то есть в подвижной плоскости векторы стоят вдоль линии,

относительно неподвижной - их фаза всегда будет совпадать, и перемещение будет синхронным, а, значит, **E** будет колебаться вдоль прямой.

В данном случае говорят о *линейной поляризации.*

Для нее вводят *Плоскость поляризации* - плоскость, построенную на векторах **E** и **k** (то есть параллельная направлению распространения,

но в то же время содержащая вектор напряженности волны).

Ее можно было бы ввести и для других поляризаций, но только в случае линейной она не меняется со временем, так что от нее будет польза.

(кст, какая?)

Вращение может быть в разные стороны, поэтому говорят о *левой и правой поляризации* по вращению относительно *наблюдателя, смотрящего навстречу лучу.*

(тут какие-то еще нюансы есть и задачи, типа как их на практике отличать.)

**способы получения линейно поляризованного света.**

Поляризацию о можно получить непосредственно от источника.

К таким явлениям относятся излучение анизотропных материалов (люминисценция и тепловоеизлучение при высоких температурах),

излучение веществ во внешних магнитных и электрических полях (эффекты Зеемана и Штарка), излучение частицы, движущейся быстрее скорости света в веществе (излучение Вавилова—Черенкова), тормозное излучение релятивистских электронов и, при некоторых условиях, излучение лазеров.

### 3.2.3 Refractive index (???)

(очень это не тривиальный параметр, про него много есть вопросов)

**Со многими отражениями фотон в сумме идет с меньшей скоростью, чем скорость света?**

(хочется это доказать, достаточно интересный факт, что такое действительно есть.)

### 3.2.4 Other about polarization

(все это в теории поля написано, тут лишь указываю.)

#### поляризация при отражении

есть *угол Брюстера.*
мда.
зависимость коэф отражения от угла:
...
я просто совершенно не усвоил суть это й формулы

#### о получении поляризованного света

самое простое - преобразовывать естественный свет. устройства для этого - *поляризаторы* они предназначенное для получения полностью или частично поляризованного (?) оптического излучения из излучения с произвольным состоянием поляризации.

Например, если поставить фильтр, который будет в одном направлении поглощать более, чем во втором - это будет называться *линейный дихроизм.* Вообще, дихроизм — явление, состоящее в различном поглощении веществом света в зависимости от его поляризации.

например, *эллиптический дихроизм* — состоит в различном поглощении веществом света с правой и левой эллиптической поляризацией.

Если при этом поляризация у света круговая - он *круговой.*

Интересно то, что вещества, обладающие этим свойством - при разных углах наблюдения имеют разный (всего 2) цвет!! (почему так?)

Другие способы разберем позже.

#### поляроиды

Рассмотрим *полароид* - поляризатор, с помощью которого получают линейно-поляризованный свет.

У поляроида есть *главная плоскость*, вдоль которой он пропускает свет в одном направлении, а интенсивность прошедшего поляризованного света определяется *законом Малюса*:

$$I = I_0 \cos^2 \varphi$$

При прохождении поляроида степень поляризации можно находить так: (хз, потом посмотрю, это в [1])

$$P = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}$$

Здесь $I_i$ соответственно максимальное и минимальное значения интенсивности света, прошедшего через анализатор при различных углах.

Разберем разные случаи прохождения поляроидов.

Что будет, если пройде через него круговой поляризованный свет?

Но также и в природе есть источники поляризованного света.

источники: (потом поизучаю, почему это так??)

- относятся излучение анизотропных (?

  каких и почему?) материалов

- излучение Вавилова—Черенкова (надеюсь, изучу его)

- при некоторых условиях, излучение лазеров.
- тормозное излучение релятивистских электронов

**нелинейные поляризационные среды**

**частично поляризованный свет**

(пар 50 лл2, и тут есть интересные темы про плоские волны, я просто сейчас не этим занимаюсь.)

**главные формулы**

В жизни часто мы имеем дело с *частично поляризованным светом* - это смесь линейно поляризованного и естественного. Для него вводят понятие *степени поляризации*:

$$P = \frac{I_{pol}}{I_{pol} + I_{natural}}$$

**Другое о поляризации**

**поляризация в КТП**

Для общего кругозора приведем, что есть поляризация в КТП.
хз, так и не разобрался, мб это связано, мб нет, не знаю.

**Поляризация при преломлении**

хз

## 3.3 Photometric concepts and units (by Landsberg)

### 3.3.1 Basic concepts

(???)

### 3.3.2 Transition from energy quantities to luminous quantities

(???)

### 3.3.3 Units for luminous measurements

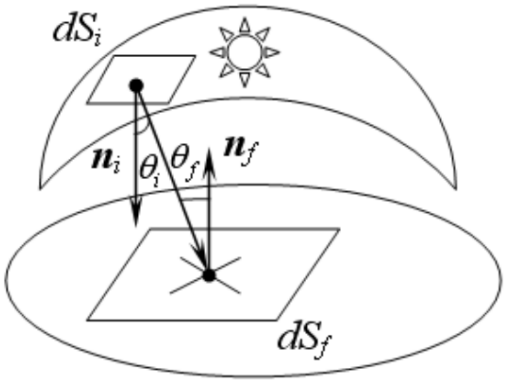
(???)

### 3.3.4 Light measurements (photometry)

(???)

**задача об освещенности протяженным источником (!?)**

(разберу потом)




Световой поток с площадки $dS_i$ : $d^2\Phi = LdS_i \cdot d\Omega \cdot \cos\theta_i$. Освещенность площадки $dS_f$ потоком с площадки $dS_i$:

$$dE = \frac{d^2\Phi}{dS_f} = LdS_i \cdot \frac{d\Omega}{dS_f} \cdot \cos\theta_i = \ldots \left[ d\Omega = \frac{dS_f}{r^2\left(\theta_f\right)} \cos\theta_f \right] \ldots = LdS_i \cdot r^{-2}\left(\theta_f\right) \cos\theta_f \cdot \cos\theta_i =$$
$$= L\left(\theta_f\right) r^{-2}\left(\theta_f\right) \cos\theta_f \cdot \cos\left(\theta_i\left(\theta_f\right)\right) \left(\frac{dS_i}{d\theta_f}\right) d\theta_f$$

Полная освещенность площадки $dS_f$:

$$E = \int\limits_{S_f} L\left(\theta_f\right) r^{-2}\left(\theta_f\right) \cos\theta_f \cdot \cos\left(\theta_i\left(\theta_f\right)\right) \left(\frac{dS_i}{d\theta_f}\right) d\theta_f$$

## 3.4 Other Fundamentals of Optics

### 3.4.1 Path Integral for Optics

(мб сделаю раздел большим, пока знаю слишком мало, чтобы этим заниматься)

### 3.4.2 Limits of Applicability of Wave Optics

**Теория по Ландау**

По определению плоской монохроматической волны ее амплитуда везде и всегда одинакова.

Такая волна бесконечна по всем направлениЯм в пространстве и существует на протяжении всего времени от - $\infty$ до + о.

Всякая же волна с не везде и не всегда постоянной амплитудой

может быть лишь более или менее монохроматической.

Мы займемся теперь выяснением вопроса о степени немонохроматичности волн.

Рассмотрим электромагнитную волну с амплитудой, являющейся в каждой точке пространства функцией времени.

Пусть $\omega_0$ есть неко-

Конечно, само монохроматическим, можно, однако, разложить на монохроматические компоненты, т.е. в интеграл Фурье.

Амплитуда компоненты этого разложения с частотой $\omega$ пропорциональна интегралу

$$A \sim \int_{-\infty}^{+\infty} \mathbf{E}_0(t) e^{i(\omega-\omega_0)t} dt$$

Множитель $e^{i(\omega-\omega_0)t}$ является периодической функцией, среднее значение которой равно нулю.

Если бы $\mathrm{E}_0$ было вообще постоянным, то интеграл был бы в точности равен нулю при всех $\omega \neq \omega_0$.

Если же $\mathbf{E}_0(t)$ переменно, но почти не меняется на протяжении промежутков времени порядка $1/|\omega - \omega_0|$, то интеграл почти равен нулю, тем точнее, чем медленнее меняется $\mathbf{E}_0$.

Для того чтобы интеграл был заметно отличен от нуля, необходимо, чтобы $\mathbf{E}_0(t)$ заметно менялось на протяжении промежутка времени порядка $1/|\omega - \omega_0|$.

Обозначим посредством $\Delta t$ порядок величины промежутка времени, в течение которого амплитуда волны в данной точке пространства заметно меняется.

Из приведенных соображений следует теперь, что наиболее отличающиеся от $\omega_0$ частоты, входящие в спектральное

разложение этой волны с заметными интенсивностями, определяются из условия

$$1/|\omega - \omega_0| \sim \Delta t.$$

Если обозначить посредством $\Delta\omega$ интервал частот (вокруг средней частоты $\omega_0$) в спектральном разложении, то, следовательно, имеет место соотношение

$$\Delta\omega \Delta t \sim 1$$

Мы видим, что действительно волна тем более монохроматична (т.е. $\Delta\omega$ тем меньше), чем больше $\Delta t$, т.е. чем медленнее меняется в каждой точке пространства ее амплитуда.

Соотношения, аналогичные (58.1), легко вывести и длЯ волнового вектора.

Пусть $\Delta x, \Delta y, \Delta z$ - порядки величин расстояний вдоль осей $x, y, z$, на которых заметно меняется амплитуда волны.

В данный момент времени поле волны как функция от координат имеет вид

$$\mathbf{E}_0(\mathbf{r}) e^{i\mathbf{k}_0\mathbf{r}}$$

где $\mathrm{k}_0$ - некоторое среднее значение волнового вектора.

Совершенно аналогично выводу (58.1) можно найти интервал $\Delta$ к значений, имеющихся в разложении рассматриваемой волны в интеграл Фурье:

$$\Delta k_x \Delta x \sim 1, \quad \Delta k_y \Delta y \sim 1, \quad \Delta k_z \Delta z \sim 1$$

**волна, излучавшаяся в течение конечного времени**

Рассмотрим, в частности, волну, излучавшуюся в течение некоторого конечного интервала времени.

Обозначим посредством $\Delta t$ порядок величины этого интервала.

Амплитуда в данной точке пространства во всяком случае заметно изменяется за время $\Delta t$, в течение которого волна успеет целиком пройти через эту точку.

На основании соотношения (58.1) мы можем теперь сказать, что «степень немонохроматичности» такой волны $\Delta\omega$ во всяком случае не может быть меньше, чем $1/\Delta t$ (но может, конечно, быть и больше):

$$\Delta\omega \geq \frac{1}{\Delta t}$$

Аналогично, если $\Delta x, \Delta y, \Delta z$ - порядки величины размеров волны в пространстве, то для интервалов значений компонент волнового вектора, входящих в разложение волны, находим:

$$\Delta k_x \geq \frac{1}{\Delta x}, \quad \Delta k_y \geq \frac{1}{\Delta y}, \quad \Delta k_z \geq \frac{1}{\Delta z}$$

Из этих формул следует, что если мы имеем пучок света конечной ширины, то направление распространения света в таком пучке не может быть строго постоянным.

Направляя ось $x$ по направлению (среднему) света в пучке, мы получаем:

$$\theta_y \geq \frac{1}{k\Delta y} \sim \frac{\lambda}{\Delta y}$$

где $\theta_y$ - порядок величины отклонения пучка от среднего направления в плоскости $xy, a\lambda$ - длина волны.

С другой стороны, формула (58.5) дает ответ на вопрос о предельной резкости оптических изображений.

Пучок света, все лучи которого согласно геометрической оптике должны были бы пересечься в одной точке, в действительности дает изображение не в виде точки, а в виде некоторого пятна.

Для ширины $\Delta$ этого пятна имеем согласно (58.5)

$$\Delta \sim \frac{1}{k\theta} \sim \frac{\lambda}{\theta}$$

где $\theta$ - угол раствора пучка.

Эту формулу можно применить не только К изображению, но и к предмету.

Именно, можно утверждать, что при наблюдении исходящего из светящейся точки пучка света эту точку нельзя отличить от тела размера $\lambda/\theta$. Соответственно этому формула (58.6) определяет предельную разрешающую силу микроскопа.

Минимальное значение $\Delta$, достигающееся при $\theta \sim 1$, есть $\lambda$, в полном согласии с тем, что пределы геометрической оптики определяются длиной волны света.

(??????)

**порядок величины наименьшей ширины светового пучка, получающегося от параллельного пучка света на расстоянии $l$ от диафрагмы**

(что????)

Найти порядок величины наименьшей ширины светового пучка, получающегося от параллельного пучка света на расстоянии $l$ от диафрагмы.

Обозначив размер отверстия диафрагмы через $d$, имеем из (58.5) для угла отклонения лучей («угла дифракции») значение $\sim \lambda/d$, откуда ширина пучка порядка $d + \frac{\lambda}{d}$ l.

Наименьшее значение этой величины $\sim \sqrt{\lambda l}$.

## 3.4.3 Change of Phase While Propagation

(круто понимать, но не такая приоритетная тема, чтобы ставить выше.)

**изменение фазы при прохождении волны через фокус**

????

[2] пар 27 п. 8.??? и стр 268

#### изменение фазы при прохождении через каустику

тоже такое есть, я хз, почему

#### изменение фазы при отражении

хз почему, но там Френель что-то насчитал.

# 4 Geometrical Optics

## 4.1 Aberration of the Optical System by Landsberg

(here, there is a good theory about it, but it wasn't translated to English. I'll write it later, if I need it)

#### Caustic Surface. The Nature of Its Symmetry

(277).

#### Aberrations Caused by Wide Beams of Rays

(277).§83.

#### Aberrations Caused by Thin Off-Axis Inclined Beams

(280). §84.

#### Astigmatism Caused by System Asymmetry

(283). § 85.

#### Aplanatism. The Sine Condition

(284). § 86.

#### Aberrations Caused by the Dependence of the Refractive Index on the Wavelength (Chromatic Aberrations)

(286).

## 4.2 Optical Instruments According to Landsberg

(here, there is a good theory about it, but it wasn't translated to English. I'll write it later, if I need it)

## 4.3 General Laws of Geometric Optics

### 4.3.1 Transition to Geometric Optics from Wave Optics

#### Конструкция перехода

(потом напишу)

**О случаях применения перехода в геометрическую оптику (?!!)**

(применять её можно много где, тут укажу, где конкретно. пока не проработал)

**волновые поверхности**

Однако часто электромагнитные волны, не являющиеся плоскими, тем не менее таковы, что их можно рассматривать как плоские в каждом небольшом участке пространства.

Для этого необходимо, чтобы амплитуда и направление волны почти не менялись на протяжении расстояний порядка длины волны.

Если выполнено это условие, то можно ввести так называемые волновые поверхности, во всех точках которых фаза волны в данный момент времени одинакова (для плоской волны это — плоскости, перпендикулярные к направлению ее распространения).

В каждом небольшом участке пространства можно говорить о направлении распространения волны, нормальном к волновой поверхности.

**определение лучей**

При этом можно ввести понятие лучей — линий, касательная к которым в каждой точке совпадает с направлением распространения волны. Изучение законов распространения волн в этом случае состав- ляет предмет геометрической оптики.

Другими словами, геометрическая оптика соответствует предельному случаю малых длин волн, $\lambda \to 0$. Займемся теперь выводом основного уравнения геометрической оПтиКи - уравнения, определяющего направление лучей.

Пусть $f$ есть любая величина, описывающая поле волны (любая из компонент Е Или Н).

В плоской монохроматической волне $f$ имеет вид

$$f = ae^{i(\mathrm{kr}-\omega t+\alpha)} = a\exp\left[i\left(-k_i x^i + \alpha\right)\right]$$

(мы опускаем знак Re; везде подразумевается вещественная часть).

Напишем выражение для поля в виде

$$f = ae^{i\psi}$$

В случае, когда волна не плоская, но геометрическая оптика применима, амплитуда $a$ является, вообще говоря, функцией координат

Оптика соответствует предельному случаю малых длин волн, $\lambda \to 0$. Займемся теперь выводом основного уравнения геометрической оптики - уравнения, определяющего направление лучей.

## 4.3.2 The Eikonal Equation: Basics

Обсудим подробно уравнение эйконала.

**Обзор эйконала**

**определение эйконала**

Пусть $f$ есть любая величина, описывающая поле волны (любая из компонент Е Или Н).

В плоской монохроматической волне $f$ имеет вид

$$f = ae^{i(\mathrm{kr}-\omega t+\alpha)} = a\exp\left[i\left(-k_i x^i + \alpha\right)\right]$$

(мы опускаем знак Re; везде подразумевается вещественная часть).

Напишем выражение для поля в виде

$$f = ae^{i\psi}$$

В случае, когда волна не плоская, но геометрическая оптика применима, амплитуда $a$ является, вообще говоря, функцией координат и времени, а фаза $\psi$, называемая также эйконалом, не имеет простого Вида, как в (53.1).

Существенно, однако, что эйконал $\psi$ является большой величиной.

Это видно уже из того, что он меняется на $2\pi$ на протяжении длины волны, а геометрическая оптика соответствует пределу $\lambda \to 0$.

В малых участках пространства и интервалах времени эйконал $\psi$ можно разложить в ряд; с точностью до членов первого порядка имеем:

$$\psi = \psi_0 + \mathbf{r}\frac{\partial\psi}{\partial\mathbf{r}} + t\frac{\partial\psi}{\partial t}$$

(начало координат и начало отсчета времени выбраны в рассматриваемом участке пространства и интервале времени; значения производных берутся в начале координат).

Сравнивая это выражение с (53.1), мы можем написать:

$$\mathbf{k} = \frac{\partial\psi}{\partial\mathbf{r}} \equiv \operatorname{grad}\psi, \quad \omega = -\frac{\partial\psi}{\partial t}$$

В соответствии с тем, что в каждом небольшом участке пространства (и в небольших интервалах времени) волну можно рассматривать как плоскую.

В четырехмерном виде соотношения (53.3) напишутся как

$$k_i = -\frac{\partial\psi}{\partial x^i}$$

где $k_i$ - волновой 4-вектор.

**обзор уравнения эйконала**

Компоненты 4-вектора $k^i$ связаны соотношением $k_i k^i = 0$. Подставляя сюда (53.4), находим уравнение:

$$\frac{\partial\psi}{\partial x_i}\frac{\partial\psi}{\partial x^i} = 0$$

и времени, а фаза $\psi$, называемая также эйконалом, не имеет простого Вида, как в (53.1).

Существенно, однако, что эйконал $\psi$ является большой величиной.

Это видно уже из того, что он меняется на $2\pi$ на протяжении длины волны, а геометрическая оптика соответствует пределу $\lambda \to 0$.

В малых участках пространства и интервалах времени эйконал $\psi$ можно разложить в ряд; с точностью до членов первого порядка имеем:

$$\psi = \psi_0 + \mathbf{r}\frac{\partial\psi}{\partial\mathbf{r}} + t\frac{\partial\psi}{\partial t}$$

(начало координат и начало отсчета времени выбраны в рассматриваемом участке пространства и интервале времени; значения производных берутся в начале координат).

Сравнивая это выражение с (53.1), мы можем напИсать:

$$\mathbf{k} = \frac{\partial \psi}{\partial \mathbf{r}} \equiv \operatorname{grad} \psi, \quad \omega = -\frac{\partial \psi}{\partial t}$$

В соответствии с тем, что в каждом небольшом участке пространства (и в небольших интервалах времени) волну можно рассматривать как плоскую.

В четырехмерном виде соотношения (53.3) напишутся как

$$k_i = -\frac{\partial \psi}{\partial x^i}$$

где $k_i$ - волновой 4-вектор.

Мы видели в §48, что компоненты 4-вектора $k^i$ связаны соотношением $k_i k^i = 0$. Подставляя сюда (53.4), находим уравнение:

$$\frac{\partial \psi}{\partial x_i} \frac{\partial \psi}{\partial x^i} = 0$$

Из формы уравнения эйконала вытекает замечательная аналогия между геометрической оптикой и механикой материальных частиц.

Движение материальной частицы определяется уравнением Гамильтона-Якоби (16.11).

Это уравнение, как и уравнение эйконала, является уравнением в частных производных первого порядка и второй степени.

Как известно, действие $S$ связано с импульсом p и фунКЦией Гамильтона $\mathscr{H}$ частицы соотношениями

$$\mathbf{p} = \frac{\partial S}{\partial \mathbf{r}}, \quad \mathscr{H} = -\frac{\partial S}{\partial t}$$

Сравнивая эти формулы с формулами (53.3), мы виДим, что волновой вектор волны играет в геометрической оптике роль импульса частицы В механике, а частота - роль функЦии Гамильтона, т.е. энергиИ частиЦы.

Абсолютная величина $k$ волнового вектора связана с частотой посредством формулы $k = \omega/$ с.

Это соотношение аналогично соотношенИЮ $p = \mathscr{E}/c$ между импульсом и энергией частицы с массой, равной нулю, и скоростью, равной скорости света.

Для частиц имеют место уравнения Гамильтона

$$\dot{\mathbf{p}} = -\frac{\partial \mathscr{H}}{\partial \mathbf{r}}, \quad \mathbf{v} = \dot{\mathbf{r}} = \frac{\partial \mathscr{H}}{\partial \mathbf{p}}$$

Ввиду указанной аналогии мы можем непосредственно написать подобные уравнения для лучей:

$$\dot{\mathrm{k}} = -\frac{\partial \omega}{\partial \mathrm{r}}, \quad \dot{\mathrm{r}} = \frac{\partial \omega}{\partial \mathrm{k}}$$

В пустоте $\omega = ck$, так что $\dot{\mathrm{k}} = 0, \mathrm{v} = c\mathbf{n}$(n - единичный вектор вдоль направления распространения), т. е., как и следовало, в пустоте лучи ЯвлЯются прямыми линиями, вдоль которых свет распространяется со скоростьЮ с.

АналогиЯ между волновым вектором волны и импульсом частиЦы В особенности ясно проявляется в следующем обстоятельстве.

Рассмотрим волну, представляющую собой наложение монохроматических волН с частотами, лежащими в некотором небольшом интервале и занимающую некоторую конечную область пространства (так называемый волновой пакет).

Вычислим 4-импульс поля этой волны, воспользовавшись формулой (32.6) с тензором энергии-импульса (48.15) (для каждой монохроматической компоненты).

Заменяя в этой формуле $k^i$ некоторым его средним значением, получим выражение вида

$$P^i = Ak^i$$

где коэффициент пропорциональности $A$ между двумя 4-векторами $P^i$ и $k^i$ есть некоторый скаляр.

В трехмерном виде это соотношение дает:

$$\mathbf{P} = A\mathbf{k}, \quad \mathscr{E} = A\omega$$

Таким образом, мы видим, что импульс и энергия волнового пакета преобразуются от одной системы отсчета к другой соответственно как волновой вектор и частота.

Продолжая аналогию, можно установить для геометрической оптики принцип, аналогичный принципу наименьшего действия в механике.

Однако его при этом нельзя будет написать в гамильтоновой форме, лучей видна, впрочем, и непосредственно из указанного выше обстоятельства, что распространение лучей аналогично движению частиц с массой, равной нулю.

Если волна обладает определенной постоянной частотой $\omega$, то зависимость ее поля от времени определяется множителем вида $e^{-i\omega t}$. Поэтому для эйконала такой волны мы можем написать:

$$\psi = -\omega t + \psi_0(x, y, z)$$

где $\psi_0$ - функция только от координат.

Уравнение эйконала (53.5) принимает теперь вид

$$(\operatorname{grad}\psi_0)^2 = \frac{\omega^2}{c^2}$$

Волновые поверхности являются поверхностями постоянного эйконала, т.е. семейством поверхностей вида $\psi_0(x, y, z) = \text{const}$.

Лучи же В каждой точке нормальны к соответствующей волновой поверхности; их направление определяется градиентом $\nabla\psi_0$.

**Принцип наименьшего действия**

Как известно, в случае, когда энергия постоянна, принцип наименьшего действия частицы можно напИсать также и в виде так называемого принципа МопертюИ:

$$\delta S = \delta \int \mathbf{p} d\mathbf{l} = 0$$

заданными ее положениями.

Импульс предполагается при этом выражЕнным как функция от энергии и координат частицы.

Аналогичный принцип для лучей называется принципом Ферма.

В этом случае мы можем написать по аналогии:

$$\delta\psi = \delta \int \mathbf{k} d\mathbf{l} = 0$$

**эйконал и интеграл по траекториям**

(Зинн-Жюстен тут, но сперва квантмех сделаю)

#### О применениях эйконала

(обсудим, где она нужна?)
Мб в детектировании разностей фаз? пока хз.

#### О появлении

#### О местах применения

Обсудим, где эйконал нужен.
(важно! пока слабо вижу)

#### Связь эйконала с волновой функцией

я хз, мб потом пропишу, скорее её нет.

### переход в геометрическую оптику

#### явное применение эйконального приближения !!!!

(самое важное!!)
как можно скорее это пропишу.

#### описание приближения

Рассмотрим произвольную волну, у которой каждой точки ее фронта свое направление распространения и своя амплитуда.

Для описания такой волны можно ввести так называемые волновые поверхности, во всех точках которых фаза волны в данный момент времени одинакова.

Лучи можно ввести как линий, касательная к которым в каждой точке совпадает с направлением распространения волны в данный момент времени.

Если такие допущения верны, то говорят, что мы занимаемся изучением *геометрической оптики.* и в ней не затрагивается вопрос о природе света.

#### основные уравнения геометрической оптики

к каким итогам мы придем и что тут важно? что вообще известно?
в конце изучения сюда все впишу
формула про эйконал
формула про интенсивность
формула тонкой линзы и школа

### Фазы и эйконал

#### обзор

В геометрической оптике фазы волны связаны так называемым уравнением Эйконала. Ниже разберемся, что это и как применять.

Как известно, в простейшем случае волна имеет вид $f = ae^{i\psi}$.

Фаза $\psi$ называется *эйконалом* и в общем случае может быть сложной и зависеть от разных точек пространства.

А если ввести *четырехмерный волновой вектор* (мб отдельно все про 4d соберу вместе): $k^i = \left(\frac{\omega}{c}, \mathbf{k}\right)$, у которого из определения $\mathbf{k}$: $k^t k_l = 0$, то эйконал иметь будет вид:

$$k_i x^i = \omega t - \mathbf{kr}$$

(пара свойств про тензор энергии-импульса и эффект доплера, очень это мотивирует создать главу про энергию волны)

Эйконал это большая величина, так как он должен на при малом приращении в $\lambda$ меняться на $2\pi$.

также заметим, что теперь поверхности постоянной фазы можно называть *поверхностями постоянного эйконала.*

таким образом, подставляя наше решение в уравнения Максвелла, они принимают вид: (нужно кстати его запомнить, типа как чисто уравнения для плоской волны)

(нужно эти рассуждения вообще в теории поля делать и там эту базу наработать, в этом нет оптики! в оптике должно быть только переопределение через новые оптические термины и все.)

$$k \times E = \frac{\omega}{c} B, \; kD = 0$$

**Свойства плоской электромагнитной волны**

Изучим плоскую волну подробно.

- $k, E, H$ взаимно ортогональны.

  (вроде из уравнений максвелла, когда мы уже подставляем решение.)

- Наличие принципа суперпозиции.

  (что кстати интересно, почему, как там волны, закругляются что ли, они ведь в то же время перпендикулярны друг другу еще)

данная формула работает даже иногда тогда, когда волна не плоская (и как же?)

окей, думаем про общий вид формулы, но мы знаем также в каком-то приближении ее частный вид. поэтому попробуем связать эйконал с тем, что уже известно в волновом уравнении. раскладываем эйконал:

$$\psi = \psi_0 + \mathbf{r}\frac{\partial \psi}{\partial \mathbf{r}} + t\frac{\partial \psi}{\partial t}$$

(позже допишу в формулу, понятно, что производные в начале координат берутся.)

теперь видим:

$$\mathbf{k} = \frac{\partial \psi}{\partial r} \equiv \mathrm{grad}\psi, \quad \omega = -\frac{\partial \psi}{\partial t}. \quad \Leftrightarrow \quad k_i = -\frac{\partial \psi}{\partial x^i}$$

то есть тут уже сразу используется теория поля, сразу метрика играет роль, все через тензоры пишется.

а если что-то не понятно - см про теорпол листок.

Известно, что все волновые 4-вектора удовлетворяют соотношению $k_i k^i = 0$, поэтому получается уравнение для $\psi$, называемое уравнением эйконала:

$$\frac{\partial \psi}{\partial x_i}\frac{\partial \psi}{\partial x^i} = 0$$

Посмотрим на различные частные случаи этого уравнения. У плоской полны с одной частотой $\omega$ зависимость от времени определяется посредством множителя $e^{-i\omega t}$, поэтому эйконал имеет вид:

$$\psi = -\omega t + \psi_0(x, y, z).$$

Тогда уравнение эйконала примет вид:

$$(\mathrm{grad}\psi_0)^2 = \frac{\omega^2}{c^2}$$

) потом буду разбираться с тем, что выше.)
описывается волна с помощью введения *комплексной амплитуды.*

**уравнение эйконала действительно основное уравнение геометрической оптики**

(почему так? хз, значит, другие все уравнения из него должны следовать, мб уравнение линзы? хз, как)

### Примеры применений уравнения эйконала (???!!)

Обсудим применения уравнения эйконала.

### Принцип наименьшего действия в оптике

#### Основная конструкция принципа наименьшего действия в оптике

Можно также установить для геометрической оптики принцип, аналогичный принципу наименьшего действия в механике. Однако если в механике этот принцип был в лагранжевом формализме, то в оптике необходимо писать его в гамильтоновом,

потому что невозможно ввести функцию, аналогичную функции Лагранжа для оптических явлений.

Действительно, функция Лагранжа находится как $L = \mathbf{p}\frac{\partial H}{\partial \mathbf{p}} - H$, для оптики она запишется, заменяя $H$ на $\omega$ и $\mathbf{p}$ на $\mathbf{k}$, таким образом, $L = \mathbf{k}\frac{\partial \omega}{\partial \mathbf{k}} - \omega$.

Но это выражение равно нулю, потому что $\omega = ck$.

Поэтому напишем принцип наименьшего действия в виде принципа Мопетюри:

$$\delta S = \delta \int \mathbf{p} d\mathbf{l} = 0$$

где интегрирование производится по траектории частицы между начальным и конечным положениями. Так что заменив импульс на волновой вектор, мы получаем аналогичный принцип наименьшего действия, называемый принципом Ферма:

$$\delta\psi = \delta \int \mathbf{k} d\mathbf{l} = 0$$

Для пустоты $\mathbf{k} = \omega\mathbf{n}/c$, а так как $\mathbf{n}d\mathbf{l} = dl$, то получаем

$$\delta \int dl = 0$$

что соответствует прямолинейному распространению лучей

!!!! место нетривиального примера!!!

**уравнение эйконала в формализме интеграла по траекториям**

**уравнение Эйконала в случае искривленных пространств**

(вроде ничего сложного, но мб и нет)
Из того, что квадрат волнового 4-вектора равен нулю

$$k_i k^i = 0$$

Подставляя сюда $\partial\psi/\partial x^i$ вместо $k_i$($\psi$ - эйконал), находим уравнение эйконала в гравитационном поле:

$$g^{ik}\frac{\partial\psi}{\partial x^i}\frac{\partial\psi}{\partial x^k} = 0$$

В предельном случае малых скоростей релятивистские уравнения движения частицы в гравитационном поле должны перейти в соответствующие нерелятивистские уравнения.
(многие детали тут добавлю позже)

### 4.3.3 The Eikonal Equation: Special Features and Applications

### 4.3.4 Limits of Geometric Optics

(пар 58 ЛЛ2 и это важно. пока в теории поля пропишу)

**аналогия геометрической оптики и механики частиц (??)**

(пока хз, что за она.)

## 4.4 Laws of Geometric Optics for Typical Setups

### 4.4.1 Formula of a Thin Lens

(that well-known basic theory)

# 5 Other Methods and Ideas of Geometrical Optics

(??????? I'll add it from Bork Wolf later)

## 5.1 Geometric Theory and Modeling of Optical Systems

При изложении материала данного раздела предполагается знакомство читателей с элементарными понятиями и законами геометрической оптики. Вопрос применимости приближения геометрической оптики к реальным системам рассматривается в разделе III, посвящённом явлению дифракции. Более полные сведения о затронутых вопросах можно найти в рекомендованной к данному разделу литературе.

## 5.2 1. General Consideration of Centered Optical Systems

Оптические приборы, предназначенные для формирования изображений, в большинстве случаев представляют собой так называемые центрированные оптические системы - однородные преломляющие или отражающие среды, отделённые одна от другой сферическими поверхностями, центры кривизны которых лежат на одной прямой, называемой главной оптической осью.

Получаемые с помощью оптических систем действительные или мнимые изображения образуются световыми пучками, которые называются гомоцентрическими, если выйдя из одной точки и пройдя оптическую систему, пучки или их продолжения снова сходятся в одной точке.

Идеальной оптической системой называют систему, в которой имеет место гомоцентричность пучков и изображение подобно предмету. Идеальная оптическая система может быть реализована с достаточным приближением в виде центрированной оптической системы, если при формировании изображения ограничиться параксиальными пучками, то есть пучками, идущими вблизи оси симметрии. Строгий критерий параксиальности световых пучков будет сформулирован ниже.

## 5.2.1 1.1. An Elementary Optical Cell

Рассмотрим ключевую для геометрической оптики задачу о формировании изображения элементарной ячейкой центрированной оптической системы - сферической границей раздела двух оптически однородных сред с различными показателями преломления, представленной на рис. 1.1.

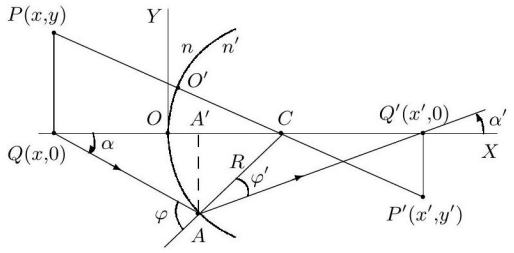



Рис. 1.1. Формирование изображения сферической преломляющей поверхностью

На этом рисунке прямая $QQ'$ проходит через центр $C$ кривизны сферической поверхности радиуса $R$, разграничивающей среды с показателями преломления $n$ и $n'$. Эта прямая принимается за главную оптическую ось системы. Направим вдоль $QQ'$ координатную ось $X$ декартовой прямоугольной системы координат $(x, y, z)$ с началом отсчёта в точке $O$ пересечения её со сферической поверхностью. Абсцисса $x$ считается положительной, если направление отсчёта совпадает с направлением распространения света вдоль оптической оси; в противоположном случае она считается отрицательной. Если падающий свет распространяется слева направо, то принятое правило знаков совпадает с правилом знаков в аналитической геометрии. Радиусы кривизны сферических поверхностей отсчитываются от поверхности к центру кривизны, так что на рис. 1.1$R > 0$.

Координатная ось $Y$ лежит в плоскости рисунка. Ордината $y$ считается положительной, если соответствующая точка лежит выше оптической оси, и отрицательной, если она лежит ниже. Отсчёт углов в плоскости рисунка (плоскости $XY$) производится в направлении от главной оптической оси к лучу, причём положительным считается угол, растущий в направлении против вращения часовой стрелки. Ввиду симметрии вращения относительно главной оптической оси достаточно ограничиться рассмотрением задачи в координатной плоскости $XY$.

Координаты предмета и его изображения здесь и ниже будем обозначать одинаковыми символами, отмечая штрихами символы, относящиеся к изображениям.

Пусть в точке $Q(x, 0)$ находится точечный источник света. Произвольный луч $QA$ после преломления на сферической поверхности, пройдя по пути $AQ$, попадёт в точку $Q'(x', 0)$. В принятой системе знаков $x < 0, x' > 0$.

Найдём связь между $x$ и $x'$ в предположении, что все отмеченные на рис. 1.1 углы столь малы, что можно пользоваться приближениями $\sin\xi \approx \xi, \cos\xi \approx 1$, где $\xi \ll 1$ - любой из рассматриваемых малых углов в нашем или аналогичном построении. В этом приближении, которое и называется параксиальным, выполняются соотношения

$$\begin{gathered} QA \approx QA' \approx QO = -x \\ AQ' \approx A'Q' \approx OQ' = x', \quad A'C \approx OC = R \end{gathered} \tag{1.1}$$

Теперь, используя закон преломления и формулы тригонометрии, из рис. 1.1 с учётом сформулированных выше правил знаков получаем систему уравнений

$$n\varphi = n'\varphi', \quad x\alpha = x'\alpha' = R(\varphi + \alpha) = R\left(\varphi' + \alpha'\right) \tag{1.2}$$

из которой находим связи между абсциссами точек $Q$ и $Q'$ и углами с главной осью произвольного параксиального луча, связывающего эти точки:

$$-\frac{n}{x} + \frac{n'}{x'} = \frac{n' - n}{R} \tag{1.3}$$

$$x\alpha = x'\alpha' \tag{1.4}$$

Из формулы (1.3) видно, что в рассматриваемом приближении положение точки $Q'$ не зависит от угла $\alpha$. Следовательно, все параксиальные лучи, выходящие из одной точки главной оптической оси, после преломления на сферической поверхности пересекутся примерно в одной точке также на главной оптической оси.

Пусть теперь точечный источник не лежит на главной оптической оси, а находится в точке $P(x, y)$ над точкой $Q(x, 0)$, см. рис. 1.1. Прямую $PCP'$, проходящую через центр $C$ кривизны сферической поверхности и заканчивающуюся в точке $P'\left(x', y'\right)$ под точкой $Q'\left(x', 0\right)$, можно рассматривать как оптическую ось. Тогда полученные выше результаты относятся и к точкам $P, P'$, а также к любой другой аналогичной паре точек приосевой области системы. Из подобия треугольников $PCQ$ и $P'CQ'$ получаем связь между ординатами точек $P$ и $P'$:

$$\frac{y'}{y} = \frac{x' - R}{x - R} \tag{1.5}$$

Подчеркнём, что сказанное выше справедливо и для мнимого источника, когда источником света служит точка схождения продолжений падающих лучей, а также и для отражающей сферической поверхности, для которой надо формально изменить знак у показателя преломления. При изменении знака кривизны сферической поверхности по отношению к падающему лучу нужно изменить знак перед $R$. Так, полагая в (1.3) $n = -n'$, получим формулу для выпуклого (+) или вогнутого (-) сферического зеркала:

$$\frac{1}{x} + \frac{1}{x'} = \pm\frac{2}{R} \tag{1.6}$$

Отметим, что при $n = n'$ система утрачивает преломляющие свойства и не представляет интереса.

Уравнения (1.3) - (1.5) описывают формирование изображения элементарной оптической ячейкой - сферической поверхностью, разделяющей среды с разными показателями преломления. Преобразуем эти уравнения к более общему виду, позволяющему установить некоторые важные свойства центрированных оптических систем в параксиальных лучах. Для этого введём в пространствах предметов и

изображений разные системы координат, получающиеся из использованной при выводе (1.3) - (1.5) единой системы с центром в точке $O$ параллельным переносом вдоль главной оптической оси:

$$x = \tilde{x} - \tilde{x}_0, \quad x' = \tilde{x}' - \tilde{x}'_0, \quad y = \tilde{y}, \quad y' = \tilde{y}' \tag{1.7}$$

Здесь $\tilde{x}_0$ и $\tilde{x}'_0$ - координаты точки $O$ (вершины преломляющей поверхности) в новых системах координат. Тильды над новыми переменными далее будем для упрощения опускать. Проведя замену переменных (1.7) в формулах (1.3) - (1.5), после несложных преобразований получим систему уравнений

$$\frac{x' - F'}{H' - F'} = \frac{H - F}{x - F} = \frac{y'}{y} = \frac{n\alpha}{n'\alpha'} \tag{1.8}$$

$$(x - H)\alpha = (x' - H')\,\alpha' \tag{1.9}$$

где введены постоянные величины

$$F = x_0 - \frac{nR}{n' - n}, \quad F' = x'_0 + \frac{n'R}{n' - n}, \quad H = x_0, \quad H' = x'_0 \tag{1.10}$$

определяемые свойствами преломляющей поверхности и выбором систем координат.

Первые два равенства системы (1.8) устанавливают соответствие между точками пространства предметов (в системе координат $x, y$) и
точками пространства изображений (в системе координат $x', y'$). Уравнения (1.8) определяют также связь между поперечным $y'/y$ и угловым $\alpha'/\alpha$ увеличениями рассмотренной оптической ячейки и показывают, как эти увеличения зависят от координат. Уравнение (1.9) определяет соотношение между углами, которые образует с главной оптической осью произвольный параксиальный луч, связывающий сопряжённые точки этой оси.

## 5.2.2 1.2. Centered Optical System

Две или несколько элементарных оптических ячеек, рассмотренных выше, расположенные последовательно на одной оптической оси, образуют центрированную оптическую систему. Применяя преобразования координат по формулам (1.8) к каждой из преломляющих поверхностей, принимая за предмет для очередной поверхности изображение, созданное предыдущей поверхностью, можно найти положение окончательного изображения, даваемого всей системой. Более того, можно доказать [1], что несколько последовательно выполненных преобразований вида (1.8) эквивалентны одному такому преобразованию с некоторыми параметрами $F, F', H, H'$.

Согласно формулам (1.8) четыре параметра $F, F', H$ и $H'$, имеющие размерность длины, определяют четыре особые точки на главной оптической оси системы и соответственно четыре особые плоскости, проходящие через эти точки перпендикулярно главной оси. Выясним их физический смысл.

Из первого и второго равенств системы (1.8) следует, что сопряжённые плоскости

$$x - H = 0, \quad x' - H' = 0 \tag{1.11}$$

отображаются друг в друга с поперечным увеличением $y'/y = +1$. Такие плоскости называются главными плоскостями оптической системы. Первая плоскость называется главной плоскостью пространства предметов, или передней главной плоскостью, вторая - главной плоскостью пространства изображений, или задней главной плоскостью.

Точки пересечения главных плоскостей с главной оптической осью

$$x_H = H, \quad x'_H = H' \tag{1.12}$$

называются главными точками центрированной системы.

Плоскости, определяемые уравнениями

$$x - F = 0, \quad x' - F' = 0 \tag{1.13}$$

называются соответственно фокальной плоскостью пространства предметов (передней фокальной плоскостью) и фокальной плоскостью пространства изображений (задней фокальной плоскостью). Как видно из первого равенства системы (1.8), каждая точка первой из этих плоскостей изображается бесконечно удалённой точкой: все лучи, вышедшие из одной и той же точки этой плоскости, после прохождения через оптическую систему становятся параллельными. Видно также, что в одной из точек второй плоскости (1.13) пересекаются после прохождения через оптическую систему параллельные лучи, вышедшие из бесконечно удалённой точки пространства предметов.

Точки пересечения фокальных плоскостей с главной оптической осью

$$x_F = F, \quad x'_F = F' \tag{1.14}$$

называются фокальными точками, или главными фокусами оптической системы. Расстояния от фокальных точек до соответствующих главных точек называются главными фокусными расстояниями системы:

$$f \equiv H - F, \quad f' \equiv H' - F' \tag{1.15}$$

Здесь $f$ - фокусное расстояние в пространстве предметов, а $f'$ - фокусное расстояние в пространстве изображений. В соответствии с принятым выше правилом знаков фокусное расстояние считается положительным, если падающий луч идёт в направлении от главного фокуса к соответствующей главной плоскости.

Отметим, что для исходной элементарной ячейки - сферической поверхности (рис. 1.1) - главные точки совпадают с вершиной $O$, а фокусные расстояния и положения главных фокусов (относительно точки $O$) задаются формулами

$$f = -F = \frac{nR}{n' - n}, \quad f' = -F' = -\frac{n'R}{n' - n} \tag{1.16}$$

Из второго равенства системы (1.8) вытекает, что поперечное увеличение $y'/y$, называемое также просто увеличением системы, не зависит от $y$. С учётом осевой симметрии, это означает, что изображение плоского предмета, перпендикулярного главной оптической оси, подобно самому предмету. Если $y'/y > 0$, то изображение прямое, если же $y'/y < 0$, то изображение обратное. Аналогичные утверждения в случае осевой симметрии системы относятся и к увеличению $z'/z$. Для краткости ниже мы не будем упоминать о координатах $z$.

Из третьего равенства системы (1.8) следует соотношение, называемое теоремой Лагранжа-Гельмгольца:

$$ny\alpha = n'y'\alpha' \tag{1.17}$$

означающее, что величина пу не изменяется при прохождении параксиального луча через иентрированную оптическуюо систему.

[01] См., напр., п. 2 «Сложение центрированных оптических систем».

Используя определения (1.15), из уравнений (1.8), (1.9) нетрудно получить важную формулу

$$\frac{n}{f} = -\frac{n'}{f'} \equiv \Phi \tag{1.18}$$

устанавливающую связь между главными фокусными расстояниями центрированной оптической системы. Определяемая этой формулой величина Φ называется оптической силой системы.

Обратимся теперь к угловому увеличению системы, представляющему собой отношение $\alpha'/\alpha$ углов, которые образуют с главной оптической осью произвольные падающий и проходящий параксиальные лучи, связывающие две сопряжённые точки на главной оси. Если $\alpha'/\alpha = +1$, то падающий и проходящий лучи параллельны. Согласно (1.8) с угловым увеличением $+1$ отображаются друг в друга точки главной оптической оси с координатами, которые с учётом формул (1.15), (1.18) представляются в виде

$$x_K = K \equiv H - (f + f'), \quad x'_K = K' \equiv H' - (f + f') \tag{1.19}$$

Эти точки называются узловыми. Всякий луч, проходящий через узловую точку $K$, после прохождения через оптическую систему остаётся параллельным своему исходному направлению и проходит через вторую узловую точку системы $K'$. Как видно из (1.19), узловые точки смещены относительно главных точек на одинаковые отрезки, равные сумме фокусных расстояний системы. Плоскости, проходящие через узловые точки перпендикулярно главной оптической оси, называются узловыми плоскостями. При равенстве показателей преломления пространств предметов и изображений согласно (1.18), (1.19) узловые точки и плоскости совпадают соответственно с главными точками и плоскостями центрированной системы.

Ещё одно важное свойство узловых точек проявляется при рассмотрении осевого, или продольного, увеличения оптической системы, определяемого как отношение длины $\delta x'$ изображения бесконечно малого отрезка, параллельного главной оптической оси, к длине $\delta x$ самого отрезка. Из уравнений (1.8) находим выражения для продольного увеличения

$$\frac{\delta x'}{\delta x} = \frac{\alpha}{\alpha'}\frac{y'}{y} = \frac{n'}{n}\left(\frac{y'}{y}\right)^2 \tag{1.20}$$

связывающее его с угловым и поперечным увеличениями системы. Эти соотношения показывают, что в общем случае продольное увеличение
не равно поперечному увеличению, и, значит, изображение бесконечно малого обдёмного предмета в центрированной оптической системе, вообще говоря, не подобно самому предмету. Для сохранения подобия бесконечно малого объёмного предмета и его изображения необходимо и достаточно, чтобы продольное увеличение по модулю было равно поперечному увеличению. Как видно из формулы (1.20), это будет при условии $\alpha'/\alpha = \pm 1$. Таким образом, если малый обдёмный предмет помещён в одной из узловых ($\alpha'/\alpha = +1$) или так называемых обратных узловых ($\alpha'/\alpha = -1$) точек, то его изображение будет подобно предмету.

Формулы (1.20) указывают на особенности центрированных оптических систем, отличающихся знаком отношения $n'/n$. Если $n'/n > 0$ (такая система называется диоптрческой), то направления распространения падающего и прошедшего света одинаковы. Если же $n'/n < 0$ (такая система называется катоптрической), то направления распространения падающего и прошедшего света противоположны. Из формул (1.20) для продольного увеличения при этом следует общее для всех оптических систем правило: если предмет перемещается вдоль оптической оси в направлении

распространения падающего света, то его изображение перемещается в направлении распространения прошедшего света, и наоборот.

Некоторые частные случаи:

1. Начала координатных систем помещены в главные точки. В новых переменных $s \equiv x - H, s' \equiv x' - H'$. При сохранении прежних обозначений для поперечных координат уравнения (1.8) принимают вид

$$\frac{s' + f'}{f'} = \frac{f}{s + f} = \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} \tag{1.21}$$

Первое уравнение этой системы можно представить в виде, подобном формуле тонкой линзы

$$\frac{f}{s} + \frac{f'}{s'} = -1 \tag{1.22}$$

с разными фокусными расстояниями, связанными соотношением (1.18).

2. Начала координатных систем помещенъ в главнъе фокусы. В новых переменных $X \equiv x - F, X' \equiv x' - F'$. При сохранении прежних обозначений для поперечных координат уравнения (1.8) принимают простой вид

$$\frac{X'}{f'} = \frac{f}{X} = \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} \tag{1.23}$$

Первое соотношение этой системы называют формулой Ньютона.

3. Среда по обе стороны оптической системы одна и та же. В этом наиболее распространённом случае формулы, описывающие оптическую систему, значительно упрощаются. Возможны два варианта.

В диоптрической системе $n = n'$, и по формуле (1.18) $f = -f'$. При этом, как видно из (1.23) и (1.19), угловое увеличение есть величина, обратная поперечному увеличению, и узловые точки системы совпадают с главными точками. Вместо (1.22) и (1.23) имеем

$$-\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}, \quad XX' = -f^2 \tag{1.24}$$

В катоптрической системе $n = -n', f = f'$, а формулы (1.23) и (1.22) переходят в формулу зеркала.

### 5.2.3 1.3. Building an Image in a Centered System

Главные, фокальные и узловые точки центрированной оптической системы называются её кардинальными точками. Иногда к ним также относят обратные главные $(y'/y = -1)$ и обратные узловые $(\alpha'/\alpha = -1)$ точки, координаты которых могут быть получены из уравнений (1.8). Положение двух главных и двух фокалъных кардиналъных точек на главной оптической оси системы полностъю характеризует чентрированную оптическую систему: зная положение этих точек, можно найти положение изображения любого предмета, даваемого системой. Пара узловых точек также может быть использована для нахождения положения изображения.

Главные точки (плоскости) могут лежать как внутри, так и вне системы и при этом могут располагаться несимметрично относительно поверхностей, ограничивающих

оптическую систему. Задние кардинальные точки (плоскости) системы могут лежать впереди передних. Типичный пример - рассеивающая линза, у которой главный фокус пространства изображений (задний фокус) лежит впереди главного фокуса пространства предметов (переднего фокуса). Ясно, что часто используемые в названиях кардинальных точек прилагательные «передний» и «задний» являются не совсем удачными, в связи с чем ниже они будут применяться только в тех случаях, когда это не вызовет недоразумений.

Процедуру построения изображения в центрированной оптической системе по известным положениям на главной оси главных $H, H'$ и фокальных $F, F'$ кардинальных точек иллюстрирует рис. 1.2. На этом рисунке оси $x, x'$ независимых в общем случае систем координат пространств предметов и изображений проходят через точки $Q, Q'$, а оси $y, y'$ лежат в плоскости рисунка. Для полноты картины показаны и узловые точки $K, K'$, лежащие в данном случае согласно (1.19) левее
соответствующих главных точек $H, H'$ на величину отрезка, равного алгебраической сумме фокусных расстояний $f + f'\,(f > 0, f' < 0)$.

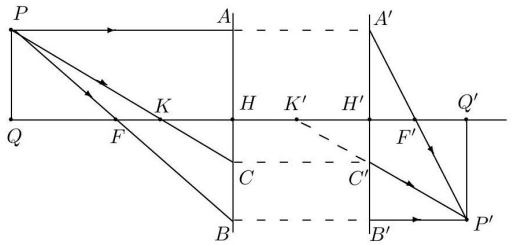



Рис. 1.2. Построение изображения в центрированной оптической системе

Построение изображения отрезка $PQ$, перпендикулярного главной оптической оси, начнём направив луч $PA$ параллельно главной оси. Он (или его продолжение в обратном направлении) встретит главные плоскости в сопряжённых точках $A$ и $A'$, лежащих на одинаковом расстоянии от главной оси $(y'/y = 1)$. Далее этот луч пройдёт через задний фокус $F'$. Двумя точками $A'$ и $F'$ положение луча в пространстве изображений определяется полностью. Второй луч из точки $P$ проведём через передний фокус $F$. Он (или его продолжение) встретится с передней главной плоскостью в точке $B$. Проведём прямую, параллельную главной оптической оси, через точку $B$ и сопряжённую с ней точку $B'$ на задней главной плоскости до пересечения с лучом $A'F'$ в точке $P'$, которая и будет изображением точки $P$. При этом проекция $Q$ точки $P$ на главную оптическую ось изобразится проекцией $Q'$ точки $P'$ на ту же ось. Отрезок $PQ$ изображается отрезком $P'Q'$. Попутно мы получили правило построения изображения точки и в том случае, когда она лежит на главной оптической оси.

На рис. 1.2 также показаны лучи $PC, C'P'$, проходящие (сами или их продолжения) через узловые точки $K$ и $K'$. Эти лучи параллельны $(\alpha'/\alpha = 1)$ и встречаются с главными плоскостями в сопряжённых точках $C$ и $C'$ на одинаковых расстояниях от главной оптической оси. Как видим, лучи, проходящие через узловые точки, положение которых на главной оси определяется положением главных и фокальных точек, тоже могут использоваться для построения изображения в центрированной системе. Напомним, что в том случае, когда по обе стороны системы находится одна и та же среда, узловые точки и плоскости сливаются с главными точками и плоскостями.

Таким образом, построение изображения предмета в центрированной оптической системе по известным положениям главных и фокалъных точек может быть выполнено с использованием двух лучей из трёх, представленных на рис. 1.2, или их продолжжений в обратном направлении.

Отдельно подчеркнём, что представленная на рис. 1.2 картина фактически является эквивалентной схемой реальной оптической системы и реального хода лучей,

преломляющихся на сферических поверхностях, разграничивающих среды с различными оптическими свойствами.

## 5.3 2. The Addition of Centered Optical Systems

Пусть две центрированные системы имеют общую главную оптическую ось. Если известны параметры каждой системы, а также их взаимное расположение, то аналитическим расчётом или геометрическим построением можно определить положение всех кардинальных точек сложной оптической системы, состоящей из этих двух систем.

Рассматриваемая система схематически изображена на рис. 1.3. Кардинальные точки первой и второй систем отмечены соответствующими нижними индексами. Штрихами выделены кардинальные точки пространства изображений первой системы и аналогичные точки пространства предметов второй системы. Величина $\Delta \equiv F_2 - F_1'$ представляет расстояние от заднего фокуса первой системы до переднего фокуса второй системы и называется оптическим интервалом двух систем. В соответствии с принятым правилом знаков $\Delta > 0$, если падающий свет идёт от фокуса $F_1'$ к фокусу $F_2$, как отмечено стрелкой на рис. 1.3, в противоположном случае $\Delta < 0$. Заданием оптического интервала полностью определяется взаимное расположение складываемых систем.

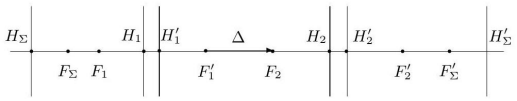



Рис. 1.3. Сложение центрированных систем

Поместим начала координат исходных систем в их фокальные точки, а фокусы $F_1$ и $F_2'$ примем соответственно за начала координат в пространстве предметов и пространстве изображений для сложной системы в целом. Координаты предмета $x = x_1, y = y_1$ и его изображения
$x_1', y_1'$ в первой системе согласно (1.23) связаны уравнениями

$$\frac{x_1'}{f_1'} = \frac{f_1}{x} = \frac{y_1'}{y} \tag{1.25}$$

где $f_1 = H_1 - F_1, f_1' = H_1' - F_1'$ - фокусные расстояния первой системы. Примем изображение, даваемое первой системой, за предмет для второй системы. Его координаты в системе с началом отсчёта в точке $F_2$ будут $x_2 = x_1' - \Delta$, $y_2 = y_1'$. При этом координаты изображения, даваемого второй системой (и, значит, всей сложной системой) относительно начала отсчёта $F_2' : x_2' = x', y_2' = y'$, определятся уравнениями

$$\frac{x'}{f_2'} = \frac{f_2}{x_1' - \Delta} = \frac{y'}{y_1'} \tag{1.26}$$

где $f_2 = H_2 - F_2, f_2' = H_2' - F_2'$ - фокусные расстояния второй системы. Исключая промежуточные координаты $x_1', y_1'$ из уравнений (1.25), (1.26), получаем формулы преобразования в виде, аналогичном (1.8):

$$\frac{x' + f_2 f_2'/\Delta}{f_1' f_2'/\Delta} = \frac{-f_1 f_2/\Delta}{x - f_1 f_1'/\Delta} = \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} \tag{1.27}$$

где $n$ и $n'$ - показатели преломления крайних сред сложной системы. Сравнивая эти уравнения с (1.8), находим фокусные расстояния и координаты фокальных и главных точек сложной системы:

$$\begin{aligned} f_\Sigma &= -\tfrac{f_1 f_2}{\Delta}, \quad H_\Sigma = \tfrac{f_1(f_1'-f_2)}{\Delta}, \quad F_\Sigma = \tfrac{f_1 f_1'}{\Delta} \\ f_\Sigma' &= \tfrac{f_1' f_2'}{\Delta}, \quad\quad H_\Sigma' = \tfrac{f_2'(f_1'-f_2)}{\Delta}, \quad F_\Sigma' = -\tfrac{f_2 f_2'}{\Delta} \end{aligned} \tag{1.28}$$

Напомним, что продольные координаты в пространствах предметов и изображений сложной системы отсчитываются в (1.27), (1.28) соответственно от фокусов $F_1$ и $F_2'$ складываемых систем.

Оптическая сила рассматриваемой сложной системы согласно (1.18), (1.28) даётся формулой

$$\Phi_\Sigma = \frac{n}{f_\Sigma} = -\frac{n\Delta}{f_1 f_2} \tag{1.29}$$

На рис. 1.3 показаны положения главных и фокальных точек сложной системы, рассчитанные по формулам (1.28), исходя из характеристик и расположения двух заданных на рисунке центрированных систем. Отмечены также главные плоскости сложной системы. Из взаимного расположения главных и фокальных точек видно, что в результате
сложения двух собирающих систем образовалась рассеивающая сложная система.

Таким образом, последовательное применение коллинеарных преобразований в форме (1.8) снова приводит к коллинеарному соотношению того же вида между координатами предмета и его изображения в сложной системе. Полученные выше формулы могут быть использованы как для построения изображения в сложной системе, так и для её аналитического исследования. Ниже аналитический метод использован при рассмотрении в качестве примеров телескопической системы и системы из двух тонких линз, а также при изучении основных оптических приборов.

## 5.4 3. Examples of Centered Optical Systems

### 5.4.1 3.1. The Telescopic System

Оптическая система может не иметь фокальных плоскостей. Такая система называется афокальной или телескопической. Она является предельным случаем обычной системы, у которой фокальные плоскости сдвинуты в бесконечность. Как видно из формул (1.28), телескопической становится система из двух обычных систем, если их оптический интервал $\Delta \to 0$. Выполнив этот предельный переход в уравнениях (1.27), получим формулы для преобразования координат и коэффициенты увеличений телескопической системы:

$$\frac{x'}{x} = \frac{\delta x'}{\delta x} = \frac{f_2 f_2'}{f_1 f_1'}, \quad \frac{y'}{y} = \frac{n}{n'}\frac{\alpha}{\alpha'} = \frac{f_2}{f_1'} \tag{1.30}$$

Из выражений (1.30) следует, что в телескопической системе: 1) всякий параллельный пучок света после прохождения через систему остаётся параллельным; 2) продолъное, поперечное и угловое увеличения постоянны, то есть не зависят от положения предмета.

Оптическая сила телескопической системы, как видно из (1.29), равна нулю.

### 5.4.2 3.2. Thick Lens

В качестве ещё одного примера применения выражений (1.28) для аналитического определения характеристик сложной системы решим задачу о толстой линзе, которая представляет собой две сферические поверхности с радиусами кривизны $R_1$ и $R_2$, ограничивающие объём оптического материала с показателем преломления $n_2 = n$.

Обычно показатели преломления крайних сред $n_1$ и $n_3$ одинаковы. Будем считать, что $n_1 = n_3 = 1$.

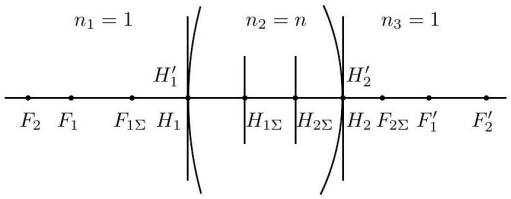



Рис. 1.4. Толстая линза как центрированная система

Рассчитаем параметры линзы, рассматривая её как сложную систему, состоящую из двух подсистем - сферических поверхностей на границах раздела сред, - представленных на рис. 1.4.

Начала координат каждой из подсистем поместим в точки $H_i = H_i' = 0 (i = 1, 2)$ пересечения преломляющих поверхностей и главных плоскостей каждой подсистемы с главной оптической осью (т.е. в вершинах поверхностей линзы). Формулы (1.16) для фокусных расстояний $f_i, f_i'$ и координат фокусов $F_i, F_i'$ при преломлении на передней ($i = 1$) и задней ($i = 2$) поверхностях линзы тогда дают

$$\begin{array}{ll} f_1 = -F_1 = \frac{R_1}{n-1}, & f_1' = -F_1' = -\frac{nR_1}{n-1} \\ f_2 = -F_2 = -\frac{nR_2}{n-1}, & f_2' = -F_2' = \frac{R_2}{n-1} \end{array} \tag{1.31}$$

Пусть $d$ - толщина линзы, при этом оптический интервал

$$\Delta = d + f_1' - f_2 = d - \frac{n}{n-1}(R_1 - R_2) \tag{1.32}$$

С учётом полученных выражений по формулам (1.28), (1.29) находим фокусные расстояния и оптическую силу толстой линзы как сложной системы, а также координаты её главных и фокальных точек:

$$\begin{aligned} f_{1\Sigma} = -f_{2\Sigma} = \frac{1}{\Phi} = \frac{nR_1R_2}{(n-1)^2\Delta} \\ H_{i\Sigma} = -\frac{R_i d}{(n-1)\Delta}, \quad F_{i\Sigma} = H_{i\Sigma} + f_{i\Sigma} \end{aligned} \tag{1.33}$$

При $i = 1$ эти формулы дают параметры системы, относящиеся к пространству предметов, а при $i = 2$ - к пространству изображений, причём в отличие от формул (1.28) отсчёт координат ведётся от вершин
соответствующих поверхностей линзы. Изменение систем координат по сравнению с общими формулами (1.28) обусловлено удобством использования реальных объектов в качестве тел отсчёта при проведении измерений.

Если радиусы $R_i$ различны и выполняется неравенство

$$d \ll \frac{n}{n-1}|R_1 - R_2|$$

то величиной $d$ в (1.32) можно пренебречь. Как следует из формул (1.33), (1.22), в этом приближении «не очень толстой» линзы

$$\Phi_\Sigma = \frac{1}{f_{1\Sigma}} = -\frac{1}{f_{2\Sigma}} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = -\frac{1}{s} + \frac{1}{s'} \tag{1.34}$$

$$H_{i\Sigma} = \frac{R_i d}{n(R_1 - R_2)}, \quad \delta \equiv d + H_{2\Sigma} - H_{1\Sigma} = \frac{n-1}{n} d \tag{1.35}$$

Выражения (1.34) показывают, что в этом случае для определения фокусного расстояния и преобразования координат, отсчитываемых от главных плоскостей, можно пользоваться формулами тонкой линзы. Формулы (1.35) позволяют при этом определить положения главных плоскостей, не совпадающих из-за малой, но конечной толщины линзы, и найти расстояние $\delta$ между ними.

### 5.4.3 3.3. Thin Lens

В модели тонкой линзы принимается, что вершины сферических поверхностей линзы и её главные точки совпадают и лежат в оптическом центре линзы - середине отрезка, соединяющего вершины линзы. В общем случае разных оптических сред в пространствах предметов и изображений при этом выполняются соотношения (1.21)-(1.23). При одинаковых средах по обе стороны линзы $n = n', f = -f'$, и формулы $(1.21) - (1.23)$ для тонкой линзы принимают вид

$$\frac{f - a'}{f} = \frac{f}{a + f} = -\frac{X'}{f} = \frac{f}{X} = \frac{y'}{y} = \frac{\alpha}{\alpha'} \tag{1.36}$$

Здесь и далее в модели тонкой линзы вместо координат $s$ и $s'$ (расстояний от главных точек оптической системы) используются продольные координаты предметов $a$ и изображений $a'$, отсчитываемые от оптического центра линзы. Преобразования продольных координат определяются при этом известными формулами тонкой линзы:

$$-\frac{1}{a} + \frac{1}{a'} = \frac{1}{f}, \quad XX' = -f^2 \tag{1.37}$$

### 5.4.4 3.4. A System of Two Thin Lenses

Центрированная система из двух тонких линз, расположенных на расстоянии $\ell_{12}$ в среде с показателем преломления $n = 1$, описывается формулами (1.28), в которых надо положить $f'_{1,2} = -f_{1,2}$. В результате для нашей сложной системы находим положения главных и фокальных точек относительно соответствующих линз и фокусные расстояния в виде

$$\begin{aligned} H_{i\Sigma} &= (-1)^i f_i \ell_{12}/\Delta \\ F_{i\Sigma} &= (-1)^i f_i \left(1 + f_i/\Delta\right) \\ f_{i\Sigma} &= (-1)^i f_1 f_2/\Delta \end{aligned} \tag{1.38}$$

При $i = 1$ эти формулы дают параметры системы, относящиеся к пространству предметов, а при $i = 2$ - к пространству изображений. Заметим, что согласно выбранному нами правилу знаков в этих системах отсчёта для действительного предмета $x < 0$, для действительного изображения $x' > 0$.

Оптический интервал $\Delta$, расстояние $\ell_{12}$ между линзами и расстояние $\delta$ между главными плоскостями системы связаны соотношениями

$$\begin{aligned} \Delta &= \ell_{12} - f_1 - f_2 \\ \delta &= \ell_{12} + H_{2\Sigma} - H_{1\Sigma} = \ell_{12}^2/\Delta \end{aligned} \tag{1.39}$$

Для оптической силы системы из формул (1.29), (1.38), (1.39) получаем выражения

$$\Phi_\Sigma = -\frac{1}{f_{2\Sigma}} = -\frac{\Delta}{f_1 f_2} = -\frac{\ell_{12}}{f_1 f_2} + \frac{1}{f_1} + \frac{1}{f_2} \tag{1.40}$$

Как видно, при $\Delta = 0$, когда совпадают задний фокус первой линзы и передний фокус второй линзы, система становится телескопической, и её оптическая сила обращается в нуль. При $\ell_{12} = 0$, когда тонкие линзы установлены вплотную, оптическая сила системы равна сумме оптических сил составляющих её линз.

## 5.5 4. Optical Devices

### 5.5.1 4.1. Magnifying Glass

Лупа, как и микроскоп, даёт увеличенное изображение малого предмета, находящегося на небольшом расстоянии от глаза. Увеличение лупы (и микроскопа) $N$ определяется как отношение угла $\varphi'$, под которым виден малый предмет через лупу (микроскоп), к углу $\varphi$, под которым
предмет был бы виден невооружённым глазом с расстояния наилучшего зрения $L = 25$ см. При увеличениях, не превосходящих 5, в качестве лупы обычно применяется простая собирающая линза, при больших увеличениях - система, состоящая из нескольких таких линз.

При нахождении общей для лупы и микроскопа формулы для увеличения воспользуемся уравнениями центрированной системы в форме (1.23). Предмет помещается между передним главным фокусом и передней главной плоскостью лупы, которая даёт его прямое мнимое изображение. Пусть $X, y$ - координаты точки предмета относительно переднего главного фокуса, $X', y'$ - координаты соответствующей точки изображения относительно заднего главного фокуса, $a$ - координата глаза относительно того же заднего главного фокуса. Глаз через лупу видит отрезок $y'$ под углом $\varphi' = y'/(a - X')$. Соответствующий отрезок предмета был бы виден невооружённым глазом с расстояния наилучшего зрения под углом $\varphi = y/L$. Отсюда для увеличения лупы и микроскопа с учётом (1.23) получаем формулы

$$N_L \equiv \frac{\varphi'}{\varphi} = \frac{y'}{y}\frac{L}{a - X'} = \frac{L}{f}\left(\frac{n'}{n} + \frac{aX}{f^2}\right)^{-1} \tag{1.41}$$

В простейшем случае, когда $n = n' = 1$, эта формула принимает вид

$$N_L = \frac{L/f}{1 + aX/f^2} \tag{1.42}$$

Если глаз находится в заднем главном фокусе лупы, то $a = 0$ и $N_L = L/f$. Увеличение будет тем же самым, если поместить предмет в передний главный фокус лупы, где $X = 0$. В этом случае глаз должен быть аккомодирован на бесконечность, куда уходит изображение предмета. Если же глаз аккомодирован на расстояние наилучиего зрения, то $a - X' = L$ и

$$N_L = \frac{y'}{y} = \frac{f}{X} = \frac{L - a}{f} \tag{1.43}$$

Как видим, в этом случае увеличение лупы равно её поперечному увеличению. Если лупой служит тонкая линза, вплотную придвинутая к глазу, то $a = -f$, так что $N_L = (L/f) + 1$.

Рассматриваемый предмет всегда помещается вблизи переднего главного фокуса, то есть $|X|/f \ll 1$, величина $|a|/f \leqslant 1$, отношение $L/f \gg 1$. Поэтому при $n = n' = 1$ увеличение лупы (и микроскопа) определяется формулой

$$N_L = \frac{L}{f} \tag{1.44}$$

Увеличение угла зрения в $N_L = L/f$ раз фактически означает, что при $n = n' = 1$ лупа (и микроскоп) в случае аккомодации глаза на расстояние наилучшего зрения $L$ позволяет разглядывать предмет на значительно меньшем расстоянии $f$ от глаза.

## 5.5.2 4.2. The Microscope

Микроскоп состоит из двух собирающих систем линз - объектива и окуляра, расположенных на расстоянии $\ell_{12}$ друг от друга в трубе, называемой тубусом. Длина тубуса у всех микроскопов, выпускаемых отечественной промышленностью, составляет 16 см, что значительно превышает фокусные расстояния объектива и окуляра. Предмет помещается на малом расстоянии перед передним фокусом объектива. Между предметом и объективом может находиться иммерсионная жидкость с показателем преломления $n$, тогда такой микроскоп называется иммерсионным. Ход лучей в микроскопе схематично показан на рис. 1.5.

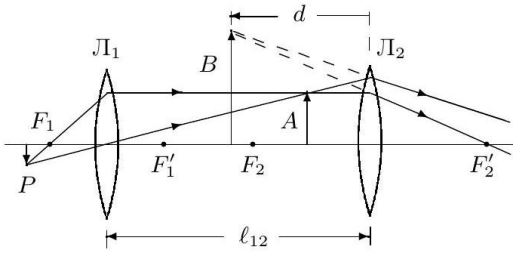



Рис. 1.5. Ход лучей в микроскопе

Объектив $_1$ даёт действительное перевёрнутое увеличенное изображение $A$ предмета $P$, которое рассматривается через окуляр $\Omega_2$, действующий как лупа. Мнимое изображение $B$, даваемое окуляром, располагается на некотором расстоянии $d$ от окуляра. Наводя микроскоп, как и любой другой оптический прибор, на резкость, наблюдатель автоматически устанавливает такое расстояние $d$, которое удобно для аккомодации глаза.

Поскольку глаз обладает значительной областью аккомодации (область аккомодации нормального человеческого глаза простирается с расстояния около 10 см до бесконечности), расстояние $d$ даже для одного и того же наблюдателя может существенно изменяться от опыта к опыту. При изменении аккомодации оптический прибор, вооружающий глаз, должен быть перефокусирован. В микроскопе это делается путём перемещения всей оптической системы относительно предмета.

Если $f_1$ и $f_1'$ - главные фокусные расстояния объектива в пространствах предметов и изображений, а $f_2$ и $f_2'$ - соответствующие характе ристики окуляра, то согласно (1.28) фокусные расстояния микроскопа как сложной системы даются формулами

$$f_M = -\frac{f_1 f_2}{\Delta}, \quad f_M' = \frac{f_1' f_2'}{\Delta} \tag{1.45}$$

где $\Delta$ - оптический интервал, величина которого зависит от устройства и взаимного расположения объектива и окуляра.

Для увеличения безиммерсионого микроскопа ($n = n' = 1$), для которого $f_1 = -f_1' > 0, f_2 = -f_2' > 0$, из формул (1.44), (1.45) получаем выражения

$$N_M = \frac{L}{f_M} = -\frac{\Delta}{f_1}\frac{L}{f_2}, \quad \Delta = \ell_{12} - f_1 - f_2 \tag{1.46}$$

Величину $N_M$ можно представить в виде

$$N_M = N_1 N_2, \quad N_1 = -\frac{\Delta}{f_1}, \quad N_2 = \frac{L}{f_2} \tag{1.47}$$

где $N_1$ - увеличение объектива, а $N_2$ - дополнительное увеличение, даваемое окуляром как лупой. Для иммерсионного микроскопа увеличение (1.46) согласно (1.41) надо умножить на $n$ - показатель преломления иммерсионной жидкости.

В микроскопе фокусные расстояния $f_1$ и $f_2$, а также оптический интервал $\Delta$ - положительны. Фокусное расстояние $f_M$ всей системы, а с ним и увеличение $N_M$ - отрицательны, так что изображение, получаемое в микроскопе, - перевёрнутое (обратное).

## 5.5.3 4.3. The Viewing Tubes

Зрительные трубы, основными элементами которых, как и в случае микроскопа, являются объектив и окуляр, предназначены для наблюдения удалённых предметов. Ход лучей в зрительных трубах схематически показан на рис. 1.6, 1.7. Уменьшенное обратное изображение $A$ удалённого предмета, даваемое объективом, находится практически в его фокальной плоскости. Мнимое изображение $B$, даваемое окуляром, располагается на расстоянии $d$ от окуляра, что отражено на рис. 1.6, 1.7. В теории зрительных труб для определённости считается, что глаз аккомодирован на бесконечность. При этом мнимое изображение $B$ должно располагаться в бесконечности, и, следовательно, промежуточное изображение $A$ должно находиться в фокальной плоскости окуляра, а задний фокус объектива должен быть совмещён с передним фокусом окуляра. В таком случае зрительные трубы представляют собой телескопические системы, рассмотренные выше в п. 3.1.

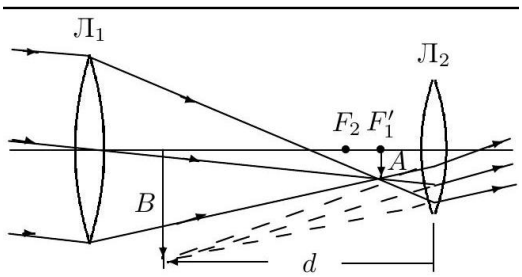



Рис. 1.6. Ход лучей в зрительной трубе Кеплера

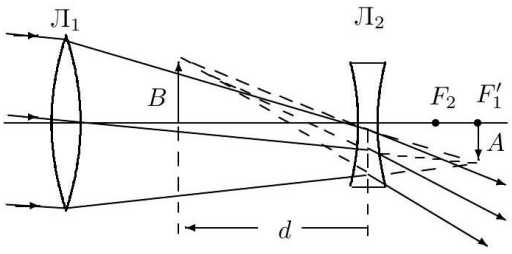



Рис. 1.7. Ход лучей в зрительной трубе Галилея

Для любых телескопических систем, у которых $n = n'$, согласно (1.30), (1.18) угловое увеличение

$$N_T \equiv \frac{\alpha'}{\alpha} = \frac{y}{y'} = \frac{f_1'}{f_2} = -\frac{f_1}{f_2} \tag{1.48}$$

обратно поперечному увеличению и по абсолютной величине равно отношению фокусных расстояний объектива и окуляра. Угловое увеличение $N_T$ в этом случае называют просто увеличением трубы, опуская прилагательное «угловое». Эта величина показывает, во сколько раз угол, под которым виден бесконечно удалённый малый предмет в трубу, больше угла, под которым он был бы виден невооружённым глазом (рис. 1.8а).

Отношение фокусных расстояний в равенствах (1.48) может иметь разный знак. Объективом зрительной трубы всегда является собирающая система, для которой переднее фокусное расстояние $f_1 > 0$. Окуляром трубы Кеплера (рис. 1.6, 1.8б) является собирающая система, переднее фокусное расстояние которой $f_2 > 0$, так что труба Кеплера даёт перевёрнутое изображение предмета. Окуляром трубы Галилея (рис. 1.7), напротив, является рассеивающая система, переднее фокус ное расстояние которой $f_2 < 0$, так что труба Галилея даёт прямое изображение. Поэтому бытовые зрительные трубы, бинокли и т. д. делаются по схеме Галилея.

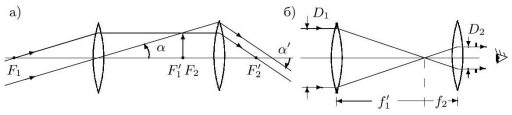



Рис. 1.8. Ход лучей в телескопе

Модуль отношения фокусов в равенствах (1.48), как видно из рис. 1.8б, на котором изображена модель телескопа Кеплера из двух тонких собирающих линз, равен отношению ширины падающего пучка лучей к ширине соответствующего выходящего пучка. Ширина пучка, прошедшего объектив, определяется диаметром $D_1$ его оправы; ширина пучка, выходящего из окуляра, - диаметром $D_2$ изображения оправы объектива, даваемого окуляром. Таким образом,

$$N_T = \mp\frac{D_1}{D_2} \tag{1.49}$$

где верхний знак относится к трубам Кеплера, а нижний - к трубам Галилея.

В том случае, когда диаметр $D_2$ пучка, выходящего из окуляра, равен диаметру $d_0$ зрачка наблюдателя ($d_0 \approx 5$ мм), увеличение телескопа называется нормальным. При этом диаметре выходного пучка освещённость изображения на сетчатке глаза достигает своего максимального значения. Уменьшение $D_2$ ведёт к квадратичному уменьшению освещённости изображения на сетчатке глаза, тогда как с ростом $D_2$ уменьшается увеличение телескопа при неизменной освещённости.

## 5.6 5. Basic Aberrations of Optical Systems

В идеальных оптических системах лучи, вышедшие из одной точки объекта, пересекаются в одной и той же точке изображения независимо от угла испускания и длины волны света. В реальных системах такие зависимости имеют место, из-за чего заметно ухудшается качество изображения. Основные погрешности при этом возникают из-за сферической и хроматической аберраций линз, входящих в состав оптических систем.

### 5.6.1 5.1. Spherical Aberration

Сферическая аберрация появляется при преломлении широких (непараксиальных) пучков света на сферических поверхностях линз. Рис. 1.9 поясняет возникновение сферической аберрации.

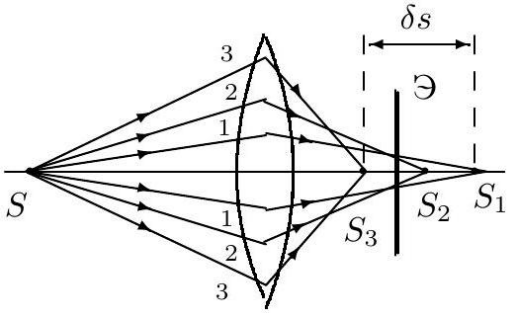



Рис. 1.9. Сферическая аберрация

Лучи 1, 2, 3, преломляющиеся в линзе на различных расстояниях от центра, пересекают оптическую ось в разных точках $S_1, S_2, S_3$. Поэтому преломлённые пучки лучей не гомоцентричны; на экране Э вместо точечного изображения получается расплывчатое пятно - кружок рассеяния. Сферическую аберрацию характеризуют с помощью так называемой продольной аберрации $\delta s$, равной расстоянию между точками пересечения крайних и центральных лучей с главной оптической осью. Продольная аберрация зависит от материала линзы и от кривизны её поверхностей.

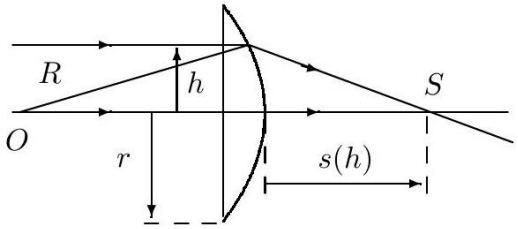



Рис. 1.10. Сферическая аберрация плосковыпуклой линзы

$$s(h) = \frac{R}{n-1}\left(1 - \frac{n^2h^2}{2R^2}\right) \tag{1.50}$$

где $n$ - показатель преломления стекла, а $R$ - радиус кривизны поверхности линзы. Для параксиальных лучей $h \ll R$ и

$$s(h)_{h\to 0}\frac{R}{n-1} \tag{1.51}$$

что согласуется с известным выражением для фокусного расстояния линзы, ограниченной сферическими поверхностями в воздухе,

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{1.52}$$

при радиусах $R_1 \to \infty$ и $R_2 < 0$.

Характеристической кривой сферической аберрации называют зависимость

$$\delta s(h) = s(h) - s(0) = -\frac{1}{2}\frac{n^2h^2}{(n-1)R} = -\frac{1}{2}\left(\frac{n}{n-1}\right)^2\left(\frac{h}{f}\right)^2 f \tag{1.53}$$

При $h = r$, где $r$ - радиус линзы, эта формула определяет продольную сферическую аберрацию линзы. Формула (1.53) показывает, что собирающая и рассеивающая линзы имеют сферические аберрации разных знаков, что даёт возможность составлять комбинации положительных и отрицательных линз с малой сферической аберрацией. По этой причине большинство современных объективов склеено из двух или большего числа

линз. Разумеется, сферическую аберрацию можно было бы устранить, делая поверхности линзы несферическими. Однако это представляет значительные технические трудности.

## 5.6.2 5.2. Chromatic Aberration

Хроматическая аберрация - зависимость фокусного расстояния линзы от длины волны - возникает вследствие дисперсии показателя преломления стёкол, то есть из-за того, что показатель преломления зависит от длины волны света $n = n(\lambda)$. Возникновение хроматической аберрации поясняет рис. 1.11.

В результате хроматической аберрации на экране Э возникает окрашенный кружок. Хроматическую аберрацию принято характеризовать разностью фо-

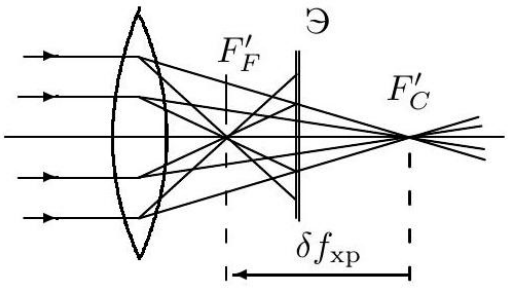



Рис. 1.11. Хроматическая аберрация кусных расстояний для двух характерных спектральных линий водорода, расположенных в крайних частях видимой области спектра: $\lambda_F = 486,1$ нм (голубая линия $F$ водорода), $\lambda_C = 656,3$ нм (красная линия $C$ водорода):

$$\delta f_{\text{хр}} = f_F - f_C \tag{1.54}$$

Для характеристики дисперсионных свойств стёкол часто пользуются так называемым коэффициентом дисперсии, или числом Аббе:

$$\nu = \frac{n_D - 1}{n_F - n_C} \tag{1.55}$$

где $n_F$ и $n_C$ - показатели преломления для линий $F$ и $C$ водорода, а $n_D$ - показатель преломления для жёлтой линии $D$ натрия $\lambda_D = 589,3$ нм (среднее значение длины волны жёлтого дублета натрия). Используя формулы (1.52), (1.54) и (1.55), для продольной хроматической аберрации получаем удобное для расчётов выражение

$$\delta f_{\text{хр}} = -\frac{(n_D - 1)(n_D - 1)}{(n_F - 1)(n_C - 1)} \frac{f_D}{\nu} \approx -\frac{f_D}{\nu}. \tag{1.56}$$

Использованное здесь приближение является довольно хорошим, поскольку согласно табличным данным для крайних частот видимого спектра $n_F - n_C \sim 10^{-2}$, так что в числителе формулы (1.56) $n_D$ можно заменить на $n_F$ и $n_C$ с относительной погрешностью результата около 2%.

Согласно табличным данным для тяжёлых стёкол с большим показателем преломления число Аббе существенно меньше, чем для лёгких стёкол. Поэтому хроматическая аберрация у линз, изготовленных из тяжёлых стёкол (тяжёлый флинт), в несколько раз больше, чем у линз из лёгких стёкол (крон, лёгкий крон). Заметим также, что собирающие и рассеивающие линзы имеют хроматические аберрации разных знаков. Это обстоятельство позволяет получать ахроматические объективы ($f_F = f_C$),

комбинируя короткофокусную положительную линзу из крона с длиннофокусной рассеивающей линзой из флинта. Чтобы уменьшить потери света на отражение, линзы обычно склеивают, поэтому радиусы вогнутой поверхности отрицательной линзы и выпуклой поверхности положительной линзы делают одинаковыми.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§9 – 16, 21 – 24.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. 12-14.

3* Дитчберн Р. Физическая оптика. - М.: Наука, 1965. Гл. 7, 14, 15.
4* Бори М., Вольф Э. Основы оптики. - М.: Наука, 1973. Гл. 2, 2.3; гл. 3-5.

# 6 Interference. Coherence

### 6.0.1 1.1. The Wave Equation. The Simplest Types of Waves

Волновое уравнение, описывающее поведение электромагнитных волн (в том числе волн света), - важнейшее следствие уравнений Максвелла, которым подчиняются все явления электромагнетизма. В явлениях, о которых идёт речь в этой книге (кроме раздела VII, посвящённого поляризации света), векторный характер электрического и магнитного полей в волне оказывается несущественным, поэтому волновой процесс можно описывать скалярной функцией координат и времени $E(x, y, z, t)$ (или, коротко, $E(\boldsymbol{r}, t)$) и понимать под $E$ какуюлибо компоненту напряжённости электрического (или магнитного) поля в электромагнитной волне.

Волновое уравнение для скалярного поля $E$ имеет вид

$$\nabla^2 E - \frac{1}{v^2}\frac{\partial^2 E}{\partial t^2} = 0 \tag{2.1}$$

где $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ - оператор Лапласа,

$$v = \frac{c}{n}$$

- скорость электромагнитной волны в среде, $c$ - скорость света в вакууме, $n$ - показатель преломления среды.

В наиболее простом случае функция $E$ зависит лишь от одной координаты $z$, т. е. в любой плоскости, перпендикулярной оси $z$ (плоскости $z = \text{const}$), величина $E$ одна и та же, независимо от координат $x, y$ в этой плоскости. Тогда уравнение (2.1) принимает вид

$$\frac{\partial^2 E}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2 E}{\partial t^2} = 0 \tag{2.2}$$

Легко проверить непосредственной подстановкой, что произвольная функция вида $E(z \mp vt)$ является решением уравнения (2.2). Общее решение уравнения (2.2) - произвольная линейная комбинация таких функций.

Зависимость вида $E(z, t) = E(z - vt)$ соответствует интуитивному представлению о волне, как о возмущении, которое распространяется в пространстве в направлении оси $z$ с конечной скоростью $v$. Действительно, значение, которое принимает величина $E$ в точке $z$ в момент времени $t$ повторяется в точке $z + \Delta z$ в момент времени $t + \Delta t$, где $\Delta z = v\Delta t > 0$ (поскольку аргумент функции $z - vt$ при этом не изменится: $z - vt = (z + \Delta z) - v(t + \Delta t)$). Волне, бегущей в обратном направлении (навстречу оси $z$), соответствует аргумент $z + vt$.

### 1.1.1. Плоская монохроматическая волна

Волна, описываемая функцией координат и времени вида

$$E(z,t) = a\cos(\omega t - kz) \tag{2.3}$$

где $\omega$ и $k$ - константы, называется плоской монохроматической волной. При фиксированном $z$ мы имеем гармонический колебательный процесс частоты $\omega$ и амплитуды $a$.

Можно убедиться непосредственной подстановкой, что функция (2.3) удовлетворяет волновому уравнению (2.2), если

$$k = \frac{\omega}{v} \tag{2.4}$$

Константа $k$ называется волновым числом. Константа $a$ в (2.3) амплитуда волны, а аргумент косинуса $\omega t - kz = \varphi$ - фаза волны. Очевидно, в любой фиксированной плоскости $z = \text{const}$ колебания величины $E$ происходят «в такт», синхронно (обычно говорят синфазно). Поверхности, на которых колебания возмущения $E$ происходят синфазно, называются волновыми поверхностями. В данном случае волновые поверхности - плоскости $z = \text{const}$ (перпендикулярные оси $z$). Поэтому волна (2.3) называется плоской. Поскольку колебания происходят по гармоническому закону, волна (2.3) также называется гармонической (в оптике принят термин монохроматическая, т.е. «одноцветная»).

Величина

$$v = \frac{\omega}{k}$$

называется фазовой скоростью. Это скорость, с которой перемещается поверхность постоянной фазы колебаний - волновая поверхность. Действительно, поверхности равных фаз определяются соотношением
$\omega t - kz = \text{const}$, откуда в результате дифференцирования по времени получим $\omega - k\frac{dz}{dt} = 0$, и $\frac{dz}{dt} = \omega/k$.

Ещё одно важное понятие - волновой вектор $k$. Это вектор, совпадающий по направлению с нормалью к волновой поверхности (и вектором скорости, с которой перемещается волновая поверхность), причём модуль вектора $\boldsymbol{k}$ равен волновому числу $|\boldsymbol{k}| = k = \omega/v$.

Разность фаз $\Delta\varphi$ колебаний на двух волновых поверхностях, разделённых расстоянием $\Delta z$, равна $\Delta\varphi = k\Delta z$. Расстояние между волновыми поверхностями, колебания на которых происходят с разностью фаз $\Delta\varphi = 2\pi$ (т.е. колебания на этих поверхностях происходят синхронно, в такт), называется длиной волны и обозначается буквой $\lambda$:

$$\lambda = \frac{2\pi}{k}. \tag{2.5}$$

Используя определение (2.4) волнового числа $k$, находим

$$\lambda = \frac{2\pi}{\omega}v = vT \tag{2.6}$$

где $T = 2\pi/\omega$ - период колебания. Из (2.6) ясно, что длина волны есть расстояние, на которое перемещается волновая поверхность за время, равное одному периоду колебания.

В общем случае плоская монохроматическая волна (распространяющаяся в произвольном направлении) имеет вид

$$E(\boldsymbol{r},t) = a\cos\left(\omega t - \boldsymbol{k}\cdot\boldsymbol{r} - \varphi_0\right) \tag{2.7}$$

где $\boldsymbol{k}\cdot\boldsymbol{r}$ - скалярное произведение волнового вектора $\boldsymbol{k}\,(k_x,k_y,k_z)$ и радиуса-вектора $\boldsymbol{r}$ точки наблюдения $\boldsymbol{r}(x,y,z) : \boldsymbol{k}\cdot\boldsymbol{r} = k_x x + k_y y + k_z z$. Направление распространения волны (направление, в котором перемещаются волновые поверхности) задаётся направлением вектора $\boldsymbol{k}$. Для общности в аргумент косинуса введена начальная фаза $\varphi_0$ (фаза колебаний в момент $t = 0$ в начале координат).

Уравнение $\boldsymbol{k}\cdot\boldsymbol{r} = \text{const}$ - это уравнение плоскости, нормаль к которой совпадает по направлению с вектором $\boldsymbol{k}$ и, как следует из (2.7), во всех точках $(x,y,z)$, принадлежащих плоскости $\boldsymbol{k}\cdot\boldsymbol{r} = \text{const}$, колебания происходят синфазно. Предоставляем читателю убедиться, что функция (2.7) удовлетворяет волновому уравнению (2.1), если

$$k^2 \equiv k_x^2 + k_y^2 + k_z^2 = \left(\frac{\omega}{\upsilon}\right)^2 \tag{2.8}$$

причём частота волны может быть любой. Направление распространения волны также произвольно: проекции волнового вектора $\boldsymbol{k}$ можно изменять как угодно, лишь бы сумма их квадратов удовлетворяла равенству (2.8). Таким образом, уравнение (2.7) описывает целую совокупность, множество решений волнового уравнения - множество плоских волн, отличающихся частотой колебаний и направлением распространения.

Плоские монохроматические волны (2.7) называют также бегущими плоскими волнами, подчёркивая таким образом, что волновые поверхности перемещаются в пространстве, и скорость их перемещения - это и есть фазовая скорость.

Плоская монохроматическая волна представляет собой математическую идеализацию: ведь волновые поверхности этой волны - плоскости неограниченной протяжённости, амплитуда колебаний отлична от нуля во всех точках волновых поверхностей (от минус до плюс бесконечности). Это - такая же идеализация, как бесконечно длящийся, неограниченный во времени гармонический колебательный процесс $\cos\omega t$.

### 1.1.2. Сферическая волна

Волна, описываемая выражением

$$E(r,t) = \frac{a}{r}\cos\left(\omega t - kr - \varphi_0\right) \tag{2.9}$$

называется сферической [1]. В уравнении (2.9) $r$ - расстояние от начала координат (где расположен источник волны) до точки наблюдения $(x,y,z)$ :

$$r = \sqrt{x^2 + y^2 + z^2}$$

Согласно (2.9), колебания $E$ происходят синфазно во всех точках пространства, отстоящих от источника на фиксированное расстояние $r = \text{const}$, т. е. волновые поверхности - это сферы, центр которых (источник волны) находится в начале координат.

Обратим внимание на различие уравнений (2.7) и (2.9). В уравнении плоской волны аргумент косинуса содержит скалярное произведение векторов $\boldsymbol{k}$ и $\boldsymbol{r}$. В уравнении сферической волны - произведение волнового числа $k$ на расстояние $r$ точки наблюдения от источника волны. Дифференцируя равенство $\omega t - kr = \text{const}$, получаем $\frac{dr}{dt} = \frac{\omega}{k} > 0$, т. е. волновые поверхности бегут от источника, и волну называют расходящейся.

[1] Заметим, что реальная (векторная) поперечная электромагнитная волна не может быть чисто сферической (элементарным источником является диполь, а не точка). Приближение сферической волны в оптике означает, что мы можем считать некоторый участок волнового фронта имеющим сферическую форму с точностью до малых поправок.

Напротив, выражение

$$E(r,t) = \frac{a}{r}\cos(\omega t + kr)$$

описывает сферическую волну, в которой волновые поверхности бегут к началу координат $\left(\frac{dr}{dt} = -\frac{\omega}{k} < 0\right)$. Это - сходящаяся сферическая волна.

Наконец, ещё одно существенное обстоятельство: в отличие от плоской волны амплитуда колебаний в сферической волне по мере увеличения расстояния от источника убывает по закону $A(r) = \frac{a}{r}$.

## 6.0.2 1.2. Monochromatic Waves. The Complex Amplitude. The Helmholtz Equation

В самом общем случае монохроматическая волна (не обязательно плоская или сферическая) имеет вид

$$E(\boldsymbol{r},t) = a(\boldsymbol{r})\cos(\omega t - \varphi(\boldsymbol{r})) \tag{2.10}$$

где $\boldsymbol{r}$ - радиус-вектор точки наблюдения.

Функции координат $a(\boldsymbol{r})$ и $\varphi(\boldsymbol{r})$ определяют пространственную структуру волны, в том числе и форму волновых поверхностей.

При изучении монохроматических волн, так же как и при изучении гармонических колебаний, удобно использовать комплексную форму записи. Наряду с волной (2.10) рассмотрим волновой процесс вида

$$E_1(\boldsymbol{r},t) = a(\boldsymbol{r})\sin(\omega t - \varphi(\boldsymbol{r})) \tag{2.11}$$

Ясно, что если функция (2.10) описывает некоторый волновой процесс (т. е. удовлетворяет волновому уравнению), то функция (2.11) также описывает волну. Ясно также, что линейная комбинация функций (2.10) и (2.11) вида

$$V(\boldsymbol{r},t) = E(\boldsymbol{r},t) - iE_1(\boldsymbol{r},t) \tag{2.12}$$

где $i$ - мнимая единица, также удовлетворяет волновому уравнению (линейная суперпозиция решений волнового уравнения также является решением).

Используя тождество $\cos\alpha - i\sin\alpha = e^{-i\alpha}$, перепишем (2.12) в виде

$$V(\boldsymbol{r},t) = a(\boldsymbol{r})e^{-i[\omega t - \varphi(\boldsymbol{r})]}$$

Функцию $V(\boldsymbol{r},t)$ можно записать в виде произведения двух функций, одна из которых

$$f(\boldsymbol{r}) = a(\boldsymbol{r})e^{i\varphi(\boldsymbol{r})} \tag{2.13}$$

зависит только от координат, а вторая $e^{-i\omega t}$ - только от времени:

$$V(\boldsymbol{r},t) = f(\boldsymbol{r})e^{-i\omega t} \tag{2.14}$$

Комплексная функция $f(\boldsymbol{r})$ полностью определяет пространственную структуру волны (т.е. распределение в пространстве амплитуд $a(\boldsymbol{r})$ и фаз $\varphi(\boldsymbol{r})$ колебаний) и называется комплексной амплитудой волны. Будем для краткости называть эту функцию волновым полем или просто полем. В частности, для плоской волны (2.7)$\varphi(\boldsymbol{r}) = \boldsymbol{k}\cdot\boldsymbol{r}+\varphi_0$ и $a(\boldsymbol{r}) = a =$ const, поэтому

$$f(x,y,z) = ae^{i(\boldsymbol{k}\cdot\boldsymbol{r}+\varphi_0)} = ae^{i(k_x x+k_y y+k_z z+\varphi_0)}$$

Комплексная амплитуда сферической волны (2.9) есть

$$f(r) = ae^{i\varphi_0}\frac{e^{ikr}}{r}$$

Итак, комплексная функция $V(\boldsymbol{r}, t)$, определяемая равенством (2.14), должна быть решением волнового уравнения

$$\nabla^2 V - \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2} = 0 \tag{2.15}$$

Подставляя (2.14) в (2.15), приходим (после сокращения на множитель $e^{-i\omega t}$) к уравнению

$$\nabla^2 f + k^2 f = 0 \tag{2.16}$$

Уравнение (2.16) для комплексных амплитуд называется уравнением Гельмгольца.

### 6.0.3 2.1. Two-Beam Interference. The Visibility of the Stripes

Пусть в пространстве распространяются две монохроматические волны одинаковой частоты с амплитудами $a_1$ и $a_2$, и пусть в некоторой точке наблюдения их фазы равны $\varphi_1$ и $\varphi_2$ соответственно.

Согласно принципу суперпозиции, результирующий колебательный процесс в точке наблюдения представляет собой сумму колебаний, создаваемых каждой из волн, т. е. гармоническое колебание той же частоты $\omega$. Интенсивность результирующего колебания (квадрат амплитуды $I = a^2$) можно найти используя правило сложения векторов (рис. 2.1):

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Delta\varphi \tag{2.17}$$

где $\Delta\varphi = \varphi_1 - \varphi_2$ - разность фаз, $I_1 = a_1^2$ и $I_2 = a_2^2$ - интенсивности слагаемых волн.

Напомним, что интенсивность - величина,

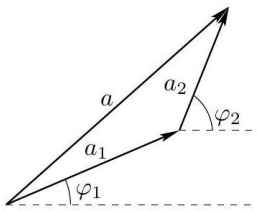



Рис. 2.1. Сложение колебаний пропорциональная плотности потока энергии в волне. Равенство (2.17) показывает, что плотность потока энергии в результирующей волне не равна в общем случае сумме потоков энергии в слагаемых волнах. В пространстве, где налагаются две волны, происходит перераспределение потоков энергий: в некоторых точках пространства результирующая интенсивность больше суммы интенсивностей слагаемых волн ($\cos\Delta\varphi > 0$), в других точках, наоборот, результирующий поток энергии меньше суммы потоков энергии в слагаемых волнах ($\cos\Delta\varphi < 0$). Это явление называется интерференцией. При наложении волн одинаковой интенсивности $I_1 = I_2 = I_0$ имеем из (2.17):

$$I = 2I_0(1 + \cos\Delta\varphi) \tag{2.18}$$

Чередующиеся максимумы $I_{\max}$ и минимумы $I_{\min}$ результирующей интенсивности образуют интерференционные полосы.

Двухлучевая интерференция. В любой двухлучевой интерференционной схеме свет от одного источника приходит в точку наблюдения по двум различным путям $r_1$ и $r_2$ (двум «плечам» интерференционной схемы) с разностью фаз $\Delta\varphi = k_2 r_2 - k_1 r_1$, где $k_1 = n_1\omega/c$ и $k_2 = n_2\omega/c$ (частота волны при переходе из одной среды в другую остаётся неизменной). Таким образом,

$$\varphi_2 - \varphi_1 = \frac{\omega}{c}\left(n_2 r_2 - n_1 r_1\right) = \frac{\omega}{c}\cdot\Delta \tag{2.19}$$

где $n_2$ и $n_1$ - показатели преломления среды вдоль путей $r_2$ и $r_1$ соответственно. Произведение показателя преломления на расстояние, пройденное светом, называют оптическим путём, а

$$\Delta = n_2 r_2 - n_1 r_1 \tag{2.20}$$

есть разность оптических путей двух плеч интерферометра (оптическая разностью хода).

Если амплитуды (и интенсивности) волн в точке наблюдения одинаковы, из (2.18) получаем

$$I = 2I_0\left(1 + \cos\frac{\omega}{c}\Delta\right) \tag{2.21}$$

Эта формула справедлива при интерференции любых монохроматических волн одинаковой частоты и интенсивности.

Видность. Контраст интерференционной картины принято характеризовать величиной видности $\mathcal{V}$, определяемой равенством

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{2.22}$$

Интенсивность максимальна и равна $I_{\max} = (a_1 + a_2)^2$ при $\Delta\varphi = 2\pi m$, где $m$ - целое число, называемое порядком интерференции. Геометрическое место точек, удовлетворяющих этим условиям, образует максимумы (светлые интерференционные полосы) $m$-го порядка. При $\Delta\varphi = (2m+1)\pi$ возникают минимумы (тёмные интерференционные полосы). Видность максимальна (и равна единице) при равных амплитудах волн, при этом $I_{\max} = 4I_0$ вдвое больше суммы интенсивностей слагаемых волн, а $I_{\min} = 0$.

## 6.0.4 2.2. Interference of Different Types of Waves. The Width of the Interference Fringes

### 2.2.1. Интерференция сферических волн

Пусть волны излучаются в пустоте ($n = 1$) двумя точечными источниками $S_1$ и $S_2$ (источники излучают синфазно с одинаковыми амплитудами), $r_1$ и $r_2$ - расстояние от источников до точки наблюдения. Разность фаз колебаний в точке наблюдения $\Delta\varphi = k\Delta$, где $\Delta = r_2 - r_1$ - разность хода волн, $k = \frac{\omega}{c}$ - волновое число.

При $\Delta = m\lambda (\Delta\varphi = 2\pi m)$ получаем интерференционный максимум (светлую полосу) $m$-го порядка. При $\Delta = m\lambda + \frac{\lambda}{2}$ - интерференционный минимум ($\Delta\varphi = (2m+1)\pi$). Если расстояния $r_1$ и $r_2$ до точки наблюдения много больше расстояния между источниками $d$, то амплитуды (и интенсивности) слагаемых волн примерно одинаковы $I_1 \approx I_2 = I_0$ и тогда справедливо равенство (2.21).

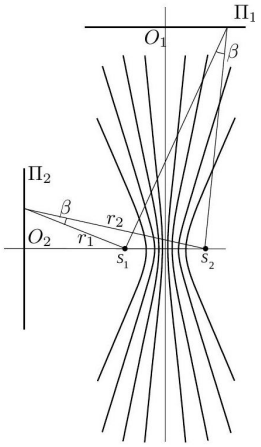



Рис. 2.2. Интерференция сферических волн

Геометрическое место точек, для которых $r_2 - r_1 = \text{const}$, - это гиперболоиды вращения. Таким образом, поверхности интерференционных максимумов - это гиперболоиды $r_2 - r_1 = m\lambda$,
сечения которых плоскостью, в которой лежат источники $S_1$ и $S_2$ (плоскость рисунка), показаны на рис. 2.2.

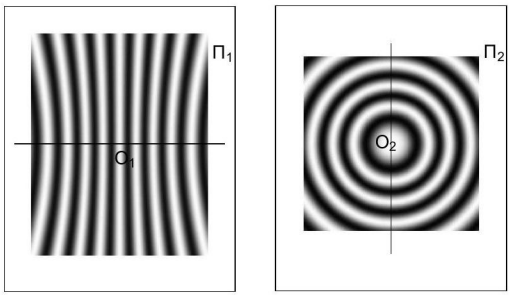



Рис. 2.3. Картина интерференции сферических волн в плоскостях $\Pi_1$ и $\Pi_2$

Форма интерференционных полос, наблюдаемых на каком-либо экране, зависит от его положения. Например, на экране $\Pi_1$, параллельном линии $S_1S_2$, соединяющей источники, полосы имеют вид почти параллельных прямых, если рассматривать малую область вблизи точки $O_1$ - центра экрана (рис. 2.3). На экране $\Pi_2$, плоскость которого перпендикулярна линии $S_1S_2$, полосы равной интенсивности концентрические кольца, центры которых лежат в точке $O_2$ экрана $\Pi_2$.

### 2.2.2. Интерференция плоских волн

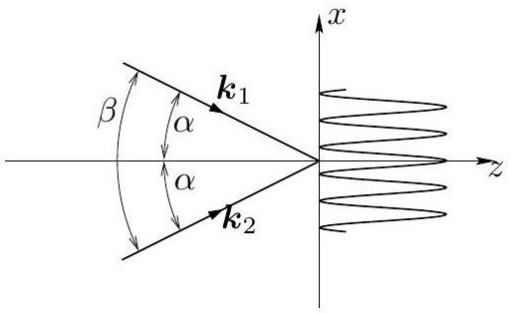



Рис. 2.4. Интерференция плоских волн

Рассмотрим результат интерференции двух плоских волн: $E_1 = a_0 \cos(\omega t - kx \sin\alpha)$ и $E_2 = a_0 \cos(\omega t + kx \sin\alpha)$ одинаковой частоты и амплитуды, падающих на экран (плоскость наблюдения $z = 0$) под углами $\alpha$ и $-\alpha$ (рис. 2.4). Используя (2.18), получаем

$$I = 2I_0[1 + \cos(2kx \sin\alpha)] \tag{2.23}$$

Переход от одной светлой полосы (интерференционного максимума) к другой соответствует изменению аргумента косинуса в (2.23) на $2\pi$:

$$2k\Delta x \sin\alpha = 2\pi$$

откуда

$$\Delta x = l = \frac{\lambda}{2\sin\frac{\beta}{2}} \tag{2.24}$$

($\beta = 2\alpha$ - угол схождения волн, рис. 2.4).
Для малых углов схождения $\sin\frac{\beta}{2} \approx \frac{\beta}{2}$ имеем простую оценку

$$l \approx \frac{\lambda}{\beta} \tag{2.25}$$

Расстояние $l$ между двумя соседними максимумами (или минимумами) интерференционной картины $I(x)$ называется шириной интерференционной полосы. В реальных условиях почти всегда можно использовать приближённую формулу (2.25) для ширины полосы, поскольку на малых участках плоскости наблюдения (порядка нескольких полос $l$) волновые поверхности интерферирующих волн можно считать плоскими. В частности, её можно использовать для оценки ширины полосы при интерференции сферических волн при условии, что $r_1$ и $r_2$ много больше расстояния между источниками (рис. 2.2).

## 6.1 3. The Statistical Nature of Radiation. Coherence

Волновые или колебательные процессы, протекающие согласованно во времени и пространстве, называют когерентными. Два гармонических колебания являются когерентными, если их разность фаз постоянна во времени. Стабильная интерференционная картина двух волн может наблюдаться, только если колебания в интерферирующих волнах остаются когерентными - то есть имеют практически неизменную разность фаз - в течение времени, достаточного для наблюдения сигнала детектором [2].

### 6.1.1 3.1. Statistical Model of Light Emission by a Source

Колебания электрического поля в волне, излучаемой любым реальным источником света, не являются строго гармоническими. В противном случае мы бы сталкивались с явлением интерференции буквально на каждом шагу. Мы не видим интерференционные полосы на стене комнаты, освещённой двумя электрическими лампочками. Это означает, что при описании излучения реального источника света следует отказаться от использования «гармонической» идеализации и более внимательно рассмотреть структуру излучения.

Опишем её в общих чертах, используя простейшую из возможных моделей процесса излучения света тепловым или газоразрядным источником, например, светящимся газом. Существуют три важных обстоятельства, определяющих структуру излучения. Во-первых, излучает огромное число возбуждённых атомов источника. Второе

[0]2 Иными словами, времени инерционности, в течение которого детектор производит усреднение попадающего на него излучения. Например, для человеческого глаза это величина $\sim 0,1$ с.

обстоятельство состоит в том, что процесс излучения атома длится конечное время, такое излучение принято называть «цугом» - «обрывком» (в отличие от бесконечно длящегося гармонического процесса). Третье обстоятельство связано со случайным, статистическим характером механизма возбуждения атомов: к возбуждению может приводить столкновение между атомами при тепловом, хаотическом движении, электронный удар, поглощение кванта света. Всё это приводит к тому, что моменты возникновения излучаемых цугов являются случайными, не связанными между собой величинами.

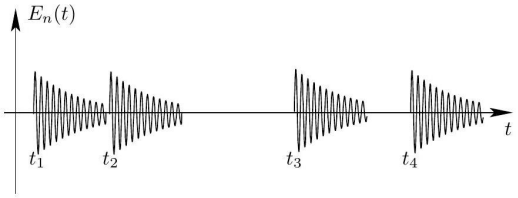



Рис. 2.5. Последовательность цугов, излучаемых атомом

На рис. 2.5 показана последовательность цугов излучения какоголибо $n$-го атома. Цуги возникают («зажигаются», «вспыхивают») в случайные моменты времени $t_1, t_2, t_3 \ldots$ Суммарное излучение источника

$$E(t) = \sum_n E_n(t) \tag{2.26}$$

содержит в каждый момент времени огромное число «живых» (отличных от нуля) слагаемых.

Если следовать классическим представлениям, излучающий атом это возбуждённый случайным толчком осциллятор (электрон в атоме). Будем далее полагать, что все осцилляторы среды имеют одну и ту же собственную частоту $\omega_0$ (период колебаний оптического диапазона излучения $T_0 = 2\pi/\omega_0 \simeq 10^{-15}$с). Колебания возбуждённого осциллятора затухают вследствие потери энергии на излучение (т. н. радиационное затухание), поэтому цуг излучения повторяет процесс затухающих колебаний классического осциллятора (рис. 2.6).

Из анализа классической модели следует, что время затухания (длительность цуга излучения) равно по порядку величины $\tau_ц \approx 10^{-8}$ с. Важно обратить внимание на соотношение величин $T_0$ и $\tau_ц$ : $T_0/\tau_ц \lesssim 10^{-7}$, т.е. цуг излучения атома содержит огромное число ($\sim 10^7$) периодов (добротность осциллятора $\sim 10^7$). Это означает, что на интервале времени $\Delta\tau$, малом по сравнению с $\tau_ц$ (но содержащем большое число периодов $T_0 : T_0 \ll \Delta\tau \ll \tau_ц$), колебание поля в цуге излучения практически неотличимо от гармонического (рис. 2.6).

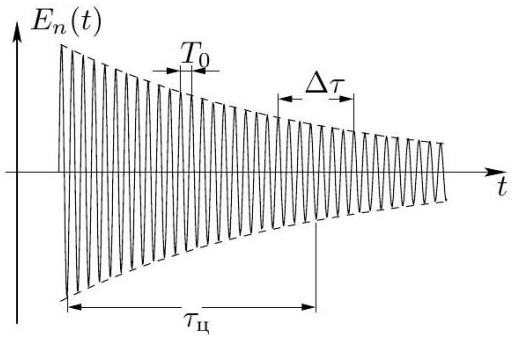



Рис. 2.6. Отдельный цуг излучения колебание затухающего осциллятора

Представим суммарный (2.26) процесс в виде хаотически-модулированного по амплитуде $a(t)$ и фазе $\varphi(t)$ колебания с несущей частотой $\omega_0$:

$$E(t) = a(t)\cos\left[\omega_0 t - \varphi(t)\right] \tag{2.27}$$

Легко видеть, что время, в течение которого происходит заметное изменение амплитуды $a(t)$ и фазы $\varphi(t)$, как раз и определяется длительностью цугов $\tau_{\text{ц.}}$ Действительно, цуги, составляющие суммарный процесс $E(t)$ на интервале времени $\Delta\tau_1$, не «доживают» до интервала $\Delta\tau_2$, если временной промежуток $\tau$ между ними превышает $\tau_{\text{ц.}}$ $:\geqslant \tau_{\text{ц}}$ (рис. 2.7).

За время $\tau_{\text{ц}}$ происходит полная смена излучаемых цугов новыми цугами. Складываясь с новыми фазовыми соотношениями (ведь моменты возникновения новых цугов совершенно не зависят от того, в какие моменты возникли цуги, выбывшие из игры), они образуют суммарный процесс, амплитуда и фаза которого совершенно не зависят от значений $a$ и $\varphi$ в процессе, образованном прежней совокупность цугов.

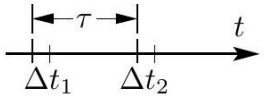



Рис. 2.7

Время $\tau_0$, за которое значения амплитуды $a\,(t+\tau_0)$ и фазы $\varphi\,(t+\tau_0)$ «забывают» прежние значения $a(t)$ и $\varphi(t)$, называют временем статистической памяти процесса (или временем когерентности).

В рассматриваемой нами модели цугов $\tau_0 = \tau_{\text{ц}}$. В частности, при $\tau > \tau_{\text{ц фаза}}^{(}\,(t+\tau)$ может с равной вероятностью принять любое значение в интервале $[0, 2\pi]$ независимо от значения фазы $\varphi(t)$ и, следовательно, равновероятно любое значение разности фаз $\Delta\varphi(t) = \varphi(t+\tau) - \varphi(t)$. Заметим, что при этом среднее за время $\Delta t \gg \tau_{\text{ц}}$ значение величины $e^{i\Delta\varphi(t)}$ равно нулю:

$$\overline{e^{i\Delta\varphi(t)}} = \overline{\cos\Delta\varphi(t)} + i\overline{\sin\Delta\varphi(t)} = 0 \tag{2.28}$$

Разумеется, «забывание» значений $a(t)$ и $\varphi(t)$ происходит постепенно: за время $\tau < \tau_{\text{ц}}$ «выбывает из игры» (заканчивается и заменяется новыми) лишь доля общего числа цугов, равная в среднем отношению $\tau/\tau_{\text{ц}}$. Часть же цугов, равная $1 - \tau/\tau_{\text{ц}}$, составляет суммарный процесс колебаний и на интервале $\Delta\tau_1$, и на интервале $\Delta\tau_2$, поэтому частично сохраняется корреляция (когерентность) колебаний $E(t)$ и $E(t+\tau)$. За промежуток времени $\tau \ll \tau_{\text{ц}}$ (сильное неравенство!) лишь ничтожная доля цугов в сумме (2.26), равная $\tau/\tau_{\text{ц}} \ll 1$, заменяется новыми. Подавляющая же доля составляет суммарный процесс колебания и на интервале $\Delta\tau_1$, и на интервале $\Delta\tau_2$; амплитуда и фаза суммарного колебания сохраняются на малом интервале $\tau \ll \tau_{\text{ц}}$ почти неизменными:

$$a(t) = a(t+\tau), \quad \varphi(t) = \varphi(t+\tau) \quad \text{при} \quad \tau \ll \tau_{\text{ц}} \tag{2.29}$$

Резюмируя, можно сказать, что когерентность постепенно уменьшается от максимального значения при $\tau \ll \tau_0$ до почти полной некогерентности при $\tau > \tau_0$.

## 6.1.2 3.2. The Width of the Line. Mechanisms of Broadening

Конечное время излучения атомов приводит к немонохроматичносmu: ширина $\Delta\omega_{\text{ц}}$ спектра излучения, состоящей из хаотичного набора цугов, может быть оценена из соотношения неопределённостей:

$$\Delta\omega_{\text{ц}} \sim \frac{2\pi}{\tau_{\text{ц}}}. \tag{2.30}$$

Столкновение между атомами при тепловом хаотическом движении может приводить не только к возбуждению осцилляторов (и изучаемых ими цугов), но и к уменьшению времени излучения уже возбуждённого атома (к уменьшению «времени жизни» возбуждённого состояния), а также к «сбою» фазы излучаемой волны (что эквивалентно появлению нового цуга). В сильно разреженном газе (при малой плотности) среднее время между столкновениями $\tau_{\text{ст}}$ оказывается больше времени «естественного затухания» цугов $\tau_{\text{ц}}$ - в этом случае столкновения практически не влияют на длительность цугов. Однако при увеличении плотности газа среднее время между столкновениями уменьшается. Например, при $\tau_{\text{ст}} \approx 10^{-9}$ с длительность цугов определяется в основном временем $\tau_{\text{ст}}$. Уменьшение времени жизни возбуждённого состояния
(длительности цугов излучения) приводит к уширению спектра излучения ($\Delta\omega_{\text{ст}} \sim 2\pi/\tau_{\text{ст}}$) - так называемое ударное уширение.

Если речь идёт об излучении нагретого газа, существенную роль может играть эффект Доплера. Частота излучения цуга, регистрируемого в точке приёма, зависит от скорости движения $v_n$ излучающего $n$-го атома:

$$\omega = \omega_0 \left(1 + \frac{v_n}{c} \cos\theta_n\right)$$

где $\theta_n$ - угол между направлением на атом-источник и вектором его скорости. Процесс колебаний $n$-го цуга в точке приёма можно записать в виде

$$E_n(t) = a_n \cos\left[(\omega_0 + \Delta\omega_n)\, t\right]$$

где $\Delta\omega_n = \omega_0 \frac{v_n}{c} \cos\theta_n$. Результирующее колебание в точке наблюдения $E(t) = \sum E_n(t)$ будет представлять собой сумму цугов с близкими частотами, случайно распределёнными вблизи $\omega_0$. Две гармоники близких частот $\omega_0$ и $\omega_0 + \Delta\omega_n$ остаются когерентными в течение времени $\tau_n \lesssim 2\pi/\Delta\omega_n$ (то есть пока разность фаз между ними $\Delta\varphi = (\omega + \Delta\omega_n)\, t - \omega t$ много меньше $2\pi$). Как известно, скорости молекул газа в тепловом равновесии подчиняются распределению Максвелла, причём характерная ширина этого распределения $\Delta v$ равна по порядку величины среднему значению тепловой скорости $v_T$. Следовательно, ширина спектра излучения газа $\Delta\omega_{\text{д}}$ может быть оценена как $\Delta\omega_{\text{д}} \sim \frac{v_T}{c}$, соответствующее время когерентности $\tau_{\text{д}} \sim 2\pi/\Delta\omega_{\text{д}}$. Это так называемое доплеровское уширение линии. При высокой температуре ($\sim$ 1000 K) для лёгких атомов $v_T/c \simeq 10^{-5}$ и $\tau_{\text{д}} \simeq 10^{-10}$ с. В этом случае время когерентности в суммарном колебании определяется именно эффектом Доплера ($\tau_{\text{ц}} > \tau_{\text{ст}} > \tau_{\text{д}}$).

Итак, суммарное колебание (2.26) в каждый момент времени может быть представлено как суперпозиция огромного числа «живых» цугов с некоторой средней длительностью $\tau_0$. Именно это время будем в дальнейшем называть временем когерентности. В описанных выше случаях $\tau_0$ определяется минимальным временем процесса: если доминируют столкновительные процессы ($\tau_{\text{ст}} \ll \tau_{\text{ц}}, \tau_{\text{Д}}$), то $\tau_0 \simeq \tau_{\text{ст}}$, и, аналогично, если ширина линии определяется эффектом Доплера ($\tau_{\text{д}} \ll \tau_{\text{ст}}, \tau_{\text{ц}}$), то $\tau_0 \simeq \tau_{\text{д}}$. Заметим, что наша модель справедлива, если период колебаний в любом случае много меньше длительности цугов: $T_0 \ll \tau_0$.

Оценка времени затухания цугов, приведённая выше ($\tau_{\text{ц}} \approx 10^{-8}$с), относится к процессу излучения изолированного классического осциллятора (отдельного атома, не взаимодействующего с другими атомами среды). Эта модель во многих случаях неприменима. Квантовомеханическое рассмотрение показывает возможность существования в некото рых ситуациях состояний с временем жизни, существенно превышающим $\tau_{\text{ц}}$, а также состояний с временем жизни, на много порядков (до $10^{-13}$ с)

меньшим $\tau_{\text{ц.}}$ Кроме того, представление, основанное на случайном характере моментов возникновения цугов излучения различных атомов среды (моментов $t_n$ в сумме (2.26)), неприменимо, если речь идёт об индуцированном излучении лазеров, в которых время когерентности $\tau_0$ результирующего излучения лазера может на много порядков (до $10^7$ раз) превышать длительность цугов излучения отдельных атоМОВ.

В дальнейшем мы будем говорить об излучении, для которого выполняется условие $T_0 \ll \Delta\tau \ll \tau_0$, в силу чего на интервале времени $\Delta\tau$ излучение практически не отличается от гармонического с частотой $\omega_0 = \frac{2\pi}{T_0}$ (квазимонохроматическое излучение).

### 6.1.3 3.3. The Time Coherence Function

В оптике измеряемой величиной является средний за время регистрации $\Delta t$ поток энергии

$$\bar{S} = \frac{c}{4\pi}\overline{E^2}$$

Величину $I = \overline{E^2}$, пропорциональную среднему потоку энергии, называют интенсивностью:

$$I = \frac{1}{\Delta t}\int\limits_{\Delta t} E^2(t)dt \sim \frac{1}{\Delta t}\int\limits_{\Delta t} a^2(t)dt = \overline{a^2(t)} \tag{2.31}$$

При изучении статистических характеристик излучения удобно использовать комплексную форму записи: реальному колебанию (2.27) ставят в соответствие комплексную функцию

$$V(t) = a(t)e^{-i[\omega_0 t - \varphi(t)]} = A(t)e^{-i\omega_0 t} \tag{2.32}$$

где $A(t) = a(t)e^{i\varphi(t)}$ - комплексная амплитуда; очевидно, что $E(t) = \operatorname{Re} V(t)$. Интенсивность, согласно (2.31), может быть выражена через комплексную амплитуду:

$$I = \overline{A(t)A^*(t)} = \overline{a^2(t)} \tag{2.33}$$

Введём количественную меру временной когерентности, которая, как мы увидим, соответствует приведённым выше качественным соображениям о степени когерентности. Функция

$$\Gamma(\tau) = \overline{V(t)\cdot V^*(t+\tau)} = \frac{1}{\Delta t}\int\limits_{t}^{t+\Delta t} V(t)\cdot V^*(t+\tau)d\tau \tag{2.34}$$

называется функцией временной когерентности случайного процесса $V(t)$. Черта в (2.34) означает усреднение величины $V(t)V^*(t+\tau)$ за время $\Delta t$, определяемое инерционностью регистрирующего устройства. В большинстве используемых устройств (видеокамера, фотоэлемент, глаз человека) величина $\Delta t$ на несколько порядков превышает время $\tau_0$ (например, инерционность зрения человека $\sim 0,1$ с, у некоторых типов фотоприёмников $\Delta t = 10^{-3} - 10^{-5}$ с). Однако существуют современные детекторы и соответствующие электронные схемы с очень малым временем разрешения $\Delta t < \tau_0$, при этом в любом случае $\Delta t \gg T_0$: ни один фотодетектор не успевает следить за изменениями светового поля, происходящими с частотой $\omega_0 \approx 10^{15}\text{с}^{-1}$. Мы будем предполагать инерчионный способ регистрации: $\Delta t \gg \tau_0$. Используя (2.34) и (2.32), имеем

$$\Gamma(\tau) = \overline{A(t)A^*(t+\tau)}e^{i\omega_0\tau} = \overline{a(t)a(t+\tau)e^{i[\varphi(t)-\varphi(t+\tau)]}}e^{i\omega_0\tau} \tag{2.35}$$

Исследуем функцию $\Gamma(\tau)$. При $\tau = 0$ получаем

$$\Gamma(0) = \overline{A(t)A^*(t)} = \overline{a^2(t)} = I_0$$

Значение функции когерентности при $\tau = 0$ равно интенсивности. Будем далее полагать процесс излучения стаиионарным, т. е. считать что ни интенсивность, ни функция когерентности не зависят от $t$, т. е. не зависят от положения интервала усреднения $\Delta t$ в (2.34) на оси времени. Легко показать, что $\Gamma(0) = I_0$ - максималъное значение функции когерентности. Нормированная функция

$$\gamma(\tau) = \frac{\Gamma(\tau)}{I_0} \tag{2.36}$$

называется степенью временной когерентности. При $\tau = 0$ имеем $\gamma(0) = 1$, это максимальное значение степени когерентности.

Из (2.29) ясно, что при $\tau \ll \tau_0$ (но $\tau \gg T_0$)

$$\Gamma(\tau) = \overline{a^2(t)}e^{i\omega_0\tau} = I_0 e^{i\omega_0\tau}$$

При этом

$$\gamma(\tau) = e^{i\omega_0\tau} \quad \text{и} \quad |\gamma(\tau)| = 1 \tag{2.37}$$

Наконец, при $\tau > \tau_0$, как следует из (2.28): $\overline{e^{i[\varphi(t)-\varphi(t+\tau)]}} \simeq 0$, поэтому $\Gamma(\tau) \simeq 0, \gamma(\tau) \simeq 0$.

Итак, мы видим, что поведение функции временной когерентности $\Gamma(\tau)$ (и степени когерентности $\gamma(\tau)$) соответствует нашим качественным представлениям о корреляции (когерентности) колебаний $V(t)$ и $V(t+\tau)$
при $\tau \ll \tau_0$ (при этом $\Gamma(\tau) \simeq \Gamma(0), \gamma(\tau) = 1$) и об отсутствии корреляции (когерентности) при $\tau > \tau_0$ (при этом $\Gamma(\tau) \simeq 0, \gamma(\tau) \simeq 0$).

## 6.1.4 3.4. Interference and Coherence

Один из наиболее эффективных способов получения информации о статистических характеристиках излучения (в частности, о степени

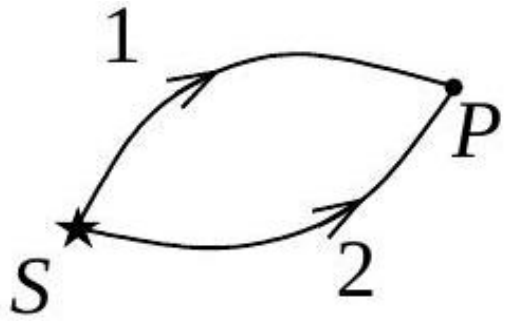



Рис. 2.8 временной когерентности) - измерение видности (контрастности) интерференционной картины в двухлучевой интерференционной схеме: излучение маленького («точечного») источника $S$ приходит в точку наблюдения $P$ по двум путям (двум «плечам») интерференционной схемы (рис. 2.8). Разность длин плеч $\Delta$ (разность путей) приводит к относительному запаздыванию процессов колебаний $V_1 = V(t)$ и

$V_2 = V(t+\tau)$ на время $\tau = \Delta/c$. Результат интерференции - интенсивность суммарного колебания - можно записать в виде

$$I(\tau) = \overline{|V(t) + V(t+\tau)|^2} = \overline{|V(t)|^2} + \overline{|V(t+\tau)|^2} + 2\,\mathrm{Re}\,\overline{V(t)V^*(t+\tau)} \tag{2.38}$$

Если $I_0 = \overline{|V(t)|^2} = \overline{|V(t+\tau)|^2}$ - равные интенсивности интерферирующих волн, то из (2.38) получаем

$$I(\tau) = 2I_0\,(1 + |\gamma(\tau)|\cos\omega_0\tau) \tag{2.39}$$

(полагаем, что фаза функции когерентности равна нулю). Используя определение видности $\mathcal{V}$(2.22), получаем

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = |\gamma(\tau)| \tag{2.40}$$

Измеряя видность интерференционных полос, отвечающих разным $\tau = \Delta/c$, мы измеряем степень временной когерентности $|\gamma(\tau)|$.

Если $\tau \ll \tau_0$, то $\gamma(\tau) \approx e^{i\omega_0\tau}$ и $|\gamma(\tau)| = 1$. Мы приходим к формуле $I(\tau) = 2I_0\,(1 + \cos\omega_0\tau)$, которая совпадает с формулой (2.18), определяющей результат интерференции строго монохроматических волн. Тем самым, получен ответ на вопрос о применимости синусоидальной идеализации: при сложении когерентных волн всё происходит так, как если бы волны были строго монохроматическими.

При $\tau \geqslant \tau_0$ волны становятся некогерентными: $|\gamma(\tau)| \simeq 0$ - интенсивность $I(\tau) = 2I_0$ равна сумме интенсивностей слагаемых волн. Отсюда следует оценка максимально допустимой для наблюдения интерференции оптической разности хода $\Delta_{\max}$ волн:

$$\Delta_{\max} \simeq c\tau_0 \tag{2.41}$$

Величину $c\tau_0$ называют длиной когерентности (это - длина цугов излучения источника).

## 6.1.5 3.5. Spectral Analysis

О свойствах квазимонохроматического излучения можно судить по спектральным измерениям, изучая с помощью какого-либо спектрального прибора распределение энергии излучения по частоте и получая тем самым величину $J(\omega)$, называемую спектральной интенсивностью. То есть величина $J(\omega)d\omega$ есть интенсивность излучения на частоте $\omega$ в бесконечно узкой полосе $d\omega$.

Заметим, что если излучение содержит две спектральные компоненты с частотами $\omega$ и $\omega + \Delta\omega$, то фотоприемник зарегистрирует величину $I = \overline{|a_1 e^{i\omega t} + a_2 e^{i(\omega+\Delta\omega)t}|^2}$ (интенсивности спектральных компонент

$I_1 = \overline{a_1^2}$ и $I_2 = \overline{a_2^2}\Big)$. Мы получаем $I = I_1 + I_2 + 2\sqrt{I_1 I_2}\overline{\cos\Delta\omega t}$. Если время регистрации $\Delta t \gg 2\pi/\Delta\omega$ (поскольку $\overline{\cos\Delta\omega t} = 0$), то $I = I_1 + I_2$, т.е. регистрируемая интенсивность равна сумме интенсивностей спектральных компонент.

Сложение интенсивностей означает некогерентность компонент разных частот. Таким образом, результат интерференции при использовании квазимонохроматического источника можно рассчитать следующим образом: нужно взять отдельную монохроматическую (гармоническую) компоненту частоты $\omega$, интенсивности $J(\omega)d\omega$ и найти результат двухлучевой интерференции, создаваемый этой компонентой:

$$dI = 2J(\omega)d\omega(1 + \cos\omega\tau) \tag{2.42}$$

Результирующая интерференционная картина представляет собой (в силу некогерентности волн разных частот!) сумму интенсивностей картин, созданных всеми спектральными компонентами, содержащимися в излучении источника:

$$I(\tau) = 2\int_0^{+\infty} J(\omega)(1+\cos\omega\tau)d\omega \tag{2.43}$$

Суммарная интенсивность $I_0$ равна сумме интенсивностей всех спектральных компонент:

$$I_0 = \int_0^{+\infty} J(\omega)d\omega$$

**Examples**

Рассмотрим два примера.

1) Спектральная интенсивность $J(\omega)$ имеет вид «столбика» шириной $\Delta\omega$ со средней частотой $\omega_0$ (рис. 2.9):

$$J(\omega) = \begin{cases} J_0 & \text{при } |\omega-\omega_0| < \frac{\Delta\omega}{2} \\ 0 & \text{при } |\omega-\omega_0| > \frac{\Delta\omega}{2} \end{cases}$$

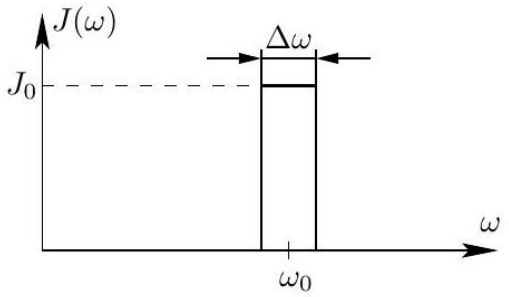



Рис. 2.9. Спектральная интенсивность

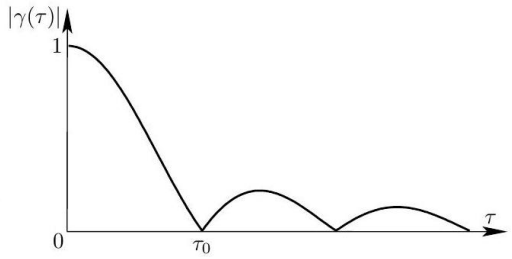



Рис. 2.10. Степень когерентности

Используя (2.43), находим

$$I(\tau) = 2\int_{\omega_0-\frac{\Delta\omega}{2}}^{\omega_0+\frac{\Delta\omega}{2}} J_0(1+\cos\omega\tau)d\omega = 2I_0\left[1+\frac{\sin\frac{\Delta\omega\tau}{2}}{\frac{\Delta\omega\tau}{2}}\cos\omega_0\tau\right] \tag{2.44}$$

где $I_0 = J_0\Delta\omega$ - интенсивность каждой из интерферирующих волн. Из (2.44) получаем видность картины:

$$\mathcal{V} = |\gamma(\tau)| = \left| \frac{\sin \frac{\Delta\omega\tau}{2}}{\frac{\Delta\omega\tau}{2}} \right| \tag{2.45}$$

Функция $|\gamma(\tau)|$ показана на рис. 2.10. В качестве оценки за величину $\tau_0$ можно принять первый ноль функции $|\gamma(\tau)|$:

$$\frac{\Delta\omega\tau_0}{2} = \pi, \quad \text{т. е.} \quad \tau_0\Delta\omega = 2\pi \tag{2.46}$$

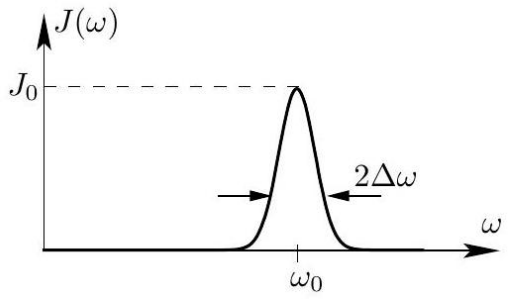



Рис. 2.11. Спектральная интенсивность

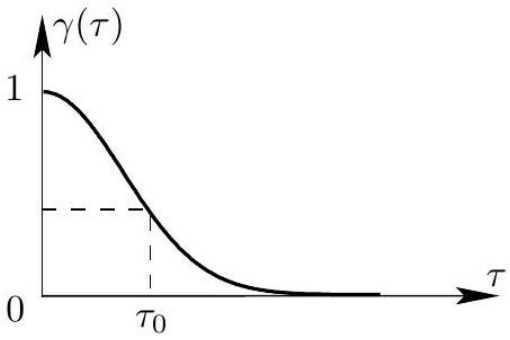



Рис. 2.12. Степень когерентности

2. Спектральная линия имеет гауссову форму (рис. 2.11):

$$J(\omega) = J_0 e^{-\frac{(\omega-\omega_0)^2}{(\Delta\omega)^2}} \tag{2.47}$$

Используя (2.43), находим

$$I(\tau) = 2I_0 \left(1 + e^{-\left(\frac{\Delta\omega\tau}{2}\right)^2} \cos\omega_0\tau\right) \tag{2.48}$$

Видность $\mathcal{V} = |\gamma(\tau)| = e^{-\left(\frac{\Delta\omega\tau}{2}\right)^2}$ показана на рис. 2.12. Время когерентности можно оценить из условия спадания видности в $e$ раз:

$$\frac{\tau_0\Delta\omega}{2} = 1 \tag{2.49}$$

Таким образом, видно, что в обоих приведённых примерах ширина спектральной линии оказывается порядка обратного времени когерентности.

### 6.1.6 3.6. The Wiener-Khinchin Theorem

Сравним формулы (2.38) и (2.43). В первой из них результат интерференции выражается через функцию когерентности. Перепишем её в виде

$$I(\tau) = 2I_0 + 2\,\mathrm{Re}\,\overline{V(t)V^*(t+\tau)} = 2I_0 + 2\,\mathrm{Re}\,\Gamma(\tau) \tag{2.50}$$

Во второй формуле тот же результат интерференции выражается через спектральную интенсивность $J(\omega)$. Перепишем её в виде

$$I(\tau) = 2I_0 + 2\,\mathrm{Re}\int\limits_{-\infty}^{+\infty} J(\omega)e^{i\omega\tau}d\omega \tag{2.51}$$

Сравнивая (2.50) и (2.51), приходим к фундаментальной связи между $\Gamma(\tau)$ и $J(\omega)$: функиия когерентности $\Gamma(\tau)$ является (обратным) преобразованием Фурье спектральной интенсивности $J(\omega)$:

$$\Gamma(\tau) = \int\limits_{-\infty}^{+\infty} J(\omega)e^{i\omega\tau}d\omega \tag{2.52}$$

Соотношение (2.52) называется теоремой Винера-Хинчина.

Из этой теоремы следует, что ширина спектра излучения $\Delta\omega$ (полоса частот, в которой заметно отлична от нуля спектральная интенсивность $J(\omega)$) связана со временем когерентности $\tau_0$ (интервалом времени $\tau \leqslant \tau_0$, в котором отлична от нуля функция когерентности $\Gamma(\tau)$) соотношением неопределённостей:

$$\tau_0\Delta\omega \simeq 2\pi \tag{2.53}$$

Иллюстрацией этого факта являются два рассмотренных в предыдущем параграфе примера.

Используя соотношение неопределённостей, можно выразить максимально допустимую для наблюдения интерференции разность хода $\Delta_{\max} = c\tau_0$ через ширину спектра излучения источника $\Delta\omega$ : $\Delta_{\max} = \frac{2\pi c}{\Delta\omega}$ или, т.к. $\omega = 2\pi c/\lambda$ и $\Delta\omega = \frac{2\pi c}{\lambda^2}\Delta\lambda$, получить полезную оценку максимально допустимой разности хода волн и максимального наблюдаемого порядка интерференции $m_{\max} = \Delta_{\max}/\lambda$:

$$\Delta_{\max} = \frac{\lambda^2}{\Delta\lambda}, \quad m_{\max} = \frac{\lambda}{\Delta\lambda} \tag{2.54}$$

## 6.2 4. Spatial Coherence

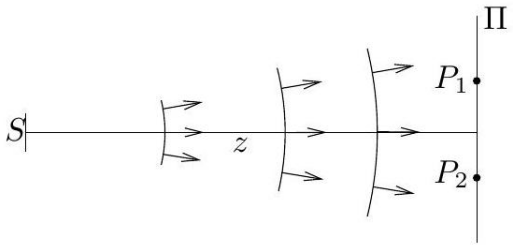
### 6.2.1 4.1. Degree of Spatial Coherence

Пусть источником светового поля $E(\boldsymbol{r}, t)$ в плоскости наблюдения П является протяжённый квазимонохроматический источник $S$, находящийся на расстоянии $z$ (рис. 2.13). Колебания поля в точках $P_1$ и $P_2$ плоскости наблюдения запишем, как и ранее, в комплексной форме

$$V_1(t) = A_1\left(\boldsymbol{r}_1, t\right)e^{i\omega t}, \quad V_2(t) = A_2\left(\boldsymbol{r}_2, t\right)e^{i\omega t}$$

($\boldsymbol{r}_1$ и $\boldsymbol{r}_2$ - координаты точек $P_1$ и $P_2$).

Рассмотрим вопрос о когерентности колебаний $V_1(t)$ и $V_2(t)$. Каждый из колебательных процессов характеризуется временем когерентности $\tau_0$, т.е. в течение времени $\tau \ll \tau_0$ (но много большем периода колебаний $T_0$) амплитуды $a_1(t)$ и




$a_2(t)$, а также фазы колебаний

Рис. 2.13 $\varphi_1(t)$ и $\varphi_2(t)$ сохраняются почти неизменными; сохраняется неизменной при этом и разность фаз $\Delta\varphi = \varphi_1(t) - \varphi_2(t)$. По прошествии времени $\tau_0$ амплитуды $a_1$ и $a_2$, а также фазы колебаний принимают новые значения (скажем, $\varphi_1'$ и $\varphi_2'$), которые некоррелированы со своими прежними значениями $\varphi_1$ и $\varphi_2$, т. е. $\varphi_1$ с равной вероятностью имеет любое значение в интервале $[0, 2\pi]$, независимо от значения $\varphi_1'$ спустя время $\tau_0$, то же самое касается и фазы $\varphi_2$. Однако возникает вопрос: сохраняется ли неизменной (или почти неизменной) разность фаз $\varphi_1'(t) - \varphi_2'(t) = \Delta\varphi'$, т. е. выполняется ли равенство $\Delta\varphi \simeq \Delta\varphi'$? Другими словами, изменяются ли фазы $\varphi_1$ и $\varphi_2$ согласованно? Если разность фаз, а также соотношения амплитуд $a_1(t)$ и $a_2(t)$ сохраняются неизменными (или почти неизменными) в течение всего времени регистрации $\Delta t\,(\Delta t \gg \tau_0)$, то тогда колебания в точках $P_1$ и $P_2$ будут когерентными (или почти когерентными). Если же через время $\tau_0$ разность фаз $\Delta\varphi'$ может принять с равной вероятностью любое значение, не зависящее от предшествующего значения $\Delta\varphi\,(\tau_0$ секунд назад), то за время регистрации $\Delta t \gg \tau_0$ разность фаз множество раз изменяется, принимая с равной вероятностью любое значение в интервале $[0, 2\pi]$. В этом случае колебания в точках $P_1$ и $P_2$ окажутся некогерентными.

Рассмотрение колебаний в различных точках пространства позволяет ввести понятие пространственной когерентности. Количественной мерой пространственной когерентности является функция пространственной когерентности:

$$\Gamma_{12} = \overline{V_1(t)V_2^*(t)} = \overline{A_1(t)A_2^*(t)} \tag{2.55}$$

которая зависит от расстояния между точками наблюдения. Нормированная функция

$$\gamma_{12} = \frac{\overline{V_1(t)V_2^*(t)}}{\sqrt{I_1 I_2}} \tag{2.56}$$

называется степенью пространственной когерентности.

Вопрос о пространственной когерентности возникает, если источник является протяжённым, и разные его участки $ds$ излучают некогерентно. Действительно, излучение разных участков - это излучение разных совокупностей атомов, моменты возникновения излучаемых ими цугов не связаны между собой, поэтому разность фаз излучаемых ими волн изменяется за время регистрации $\Delta t \gg \tau_0$ множество раз, принимая с равной вероятностью любое значение в интервале $[0, 2\pi]$.

## 6.2.2 4.2. The Van Zitter-Zernike Theorem

Найдём колебания поля, созданного протяжённым квазимонохроматическим источником $S$ в точках $P_1\,(x_1)$ и $P_2\,(x_2)$ плоскости наблюдения, находящейся на расстоянии $z$ от источника - излучающего отрезка ширины $b$ (рис. 2.14), расстояние между точками $P_1$ и $P_2$ равно $\rho = |x_1 - x_2|$. Колебания в каждой из точек $P_1$ и $P_2$ - это сумма колебаний, созданных всеми малыми отрезками $\Delta\xi$ источника. Амплитуда

$a_i = a(\xi_i)$ колебаний излучающей площадки $\Delta\xi_i$, имеющей координату $\xi_i$, и её фаза $\varphi_i = \varphi(\xi_i)$ сохраняются неизменными в течение времени, малом по сравнению со временем когерентности $\tau_0$. При распространении волны от площадки $\Delta\xi_i$ до точек $P_1$ и $P_2$ возникает набег фазы $kr_{1i}$ и $kr_{2i}$ соответственно. Поэтому суммарное колебание в точках $P_1$ и $P_2$ можно записать в виде

$$A_1 = \sum_i a_i e^{i\varphi_i} \frac{e^{ikr_{1i}}}{r_{1i}}, \quad A_2 = \sum_j a_j e^{i\varphi_j} \frac{e^{ikr_{2j}}}{r_{2j}}$$

$A_1$ и $A_2$ - комплексные амплитуды (малая площадка излучает сферическую волну, отсюда множители $1/r$). Мы получаем

$$A_1 A_2^* = \sum_i \sum_j a_i a_j e^{i(\varphi_i - \varphi_j)} \frac{e^{ik(r_{1i} - r_{2j})}}{r_{1i} r_{2j}} \tag{2.57}$$

При усреднении величины $A_1 A_2^*$ за время $\Delta t \gg \tau_0$ учтём некогерентность колебаний, созданных разными площадками $\Delta\xi_i$ и $\Delta\xi_j$ (при $i \neq j$) : $\overline{e^{i(\varphi_i - \varphi_j)}} = 0$. Т.е. отличны от нуля в (2.57) только слагаемые с $i = j$:

$$\overline{A_1 A_2^*} = \sum_i \overline{a_i^2} \frac{e^{ik(r_{1i} - r_{2i})}}{r_1 r_2} \tag{2.58}$$

Далее мы учтём, что среднее значение квадрата амплитуды излучающей площадки $\Delta\xi$ равно $\overline{a^2(\xi)} = J(\xi)\Delta\xi$, где $J(\xi)$ - интенсивность излучения единичной площадки, имеющей координату $\xi$ (интенсивность

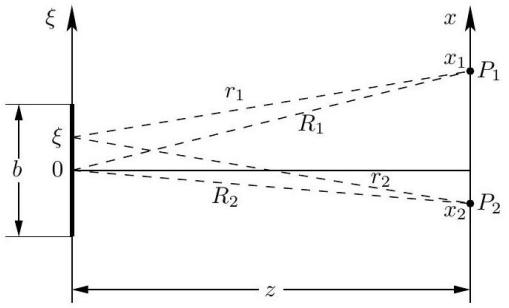



Рис. 2.14. К выводу функции пространственной когерентности

излучения площадки пропорциональна её размеру $\Delta\xi$ - она состоит из некогерентных точечных излучателей!). Примем также, что размеры источника $b$ и расстояние $\rho$ малы по сравнению с $z$. В этом приближении можно положить $\frac{1}{r_1 r_2} \approx \frac{1}{R_1 R_2}$ ($R_1$ и $R_2$ - расстояния от центра источника - точки $O$ - до точек $P_1$ и $P_2$). При вычислении фазового множителя $e^{ik(r_1 - r_2)}$ необходимо учесть дополнительный член в разложении $r_1, r_2$ в ряд Тейлора, поскольку ошибки при оценке $r_1$ и $r_2$ должны быть малы по сравнению с длиной волны $\lambda$. Тогда запишем

$$r_{1,2} = \sqrt{z^2 + (x_{1,2} - \xi)^2} \approx z + \frac{(x_{1,2} - \xi)^2}{2z}$$

при условии, что отброшенные члены разложения много меньше $\lambda$. Отсюда имеем

$$r_1 - r_2 \approx \frac{x_1^2 - x_2^2}{2z} - \frac{\rho\xi}{z}$$

Тогда из (2.58) находим

$$\Gamma(\rho) = \overline{V_1(t)V_2^*(t)} = \overline{A_1(t)A_2^*(t)} = \frac{e^{i\frac{k}{2z}\left(x_1^2 - x_2^2\right)}}{R_1 R_2} \sum_i J\left(\xi_i\right) e^{-i\frac{k}{z}\rho\xi_i} \Delta\xi_i$$

Переходя к непрерывному распределению в пределе $\Delta\xi \to 0$, получаем

$$\Gamma(\rho) = \frac{e^{i\frac{k}{z}\rho x}}{R_1 R_2} \int J(\xi) e^{-i\frac{k}{z}\rho\xi} d\xi \tag{2.59}$$

где $x = (x_1 + x_2)/2$. Нормированная функция (степень пространственной когерентности) имеет вид

$$\gamma(\rho) = \frac{\int J(\xi) e^{-i\frac{k}{z}\rho\xi} d\xi}{\int J(\xi) d\xi} \tag{2.60}$$

Формула (2.59) выражает замечательный результат: функиия пространственной когерентности $\Gamma(\rho)$ является (с точностью до несущественного множителя) преобразованием Фуръе распределения интенсивности излучения по источнику $J(\xi)$ (это - так называемая теорема Ван-Циттера-Цернике). В роли частоты $\omega$ здесь выступает пространственная частота $\frac{k\rho}{z}$.

### 6.2.3 4.3. Homogeneous Source

Для однородного источника (все точки которого излучают с одинаковой интенсивностью $J_0$)

$$J(\xi) = \begin{cases} J_0 & \text{при } |\xi| \leqslant \frac{b}{2} \\ 0 & \text{при } |\xi| > \frac{b}{2} \end{cases}$$

получаем из (2.60)

$$\gamma(\rho) = \frac{\sin\frac{kb}{2z}\rho}{\frac{kb}{2z}\rho} \tag{2.61}$$

Функция $|\gamma(\rho)|$ показана на рис. 2.15.

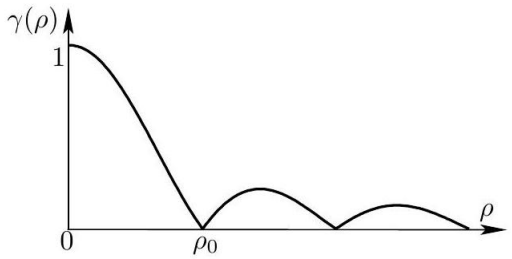



Рис. 2.15. Степень пространственной когерентности однородного источника в зависимости от расстояния между точками наблюдения Найдём радиус пространственной когерентности $\rho_0$ - максимальное расстояние между точками наблюдения, при котором степень когерентности не обращается в нуль. Его можно оценить по полуширине главного максимума функции $\gamma(\rho)$ : $\frac{kb}{2z}\rho_0 = \pi$, откуда

$$\rho_0 = \frac{\lambda}{b} z = \frac{\lambda}{\psi} \tag{2.62}$$

где $\psi = \frac{b}{z}$ - угловой размер источника.

Функция $|\gamma(\rho)|$ описывает постепенное уменьшение степени когерентности от единицы при $\rho = 0$ до нуля по мере увеличения расстояния между точками наблюдения $P_1$ и $P_2$, при $\rho = \rho_0$ степень когерентности становится равной нулю. Отметим однако, что наличие боковых лепестков функции $\gamma(\rho)$ (рис. 2.15) означает частичное (незначительное) восстановление степени когерентности колебаний в точках, расстояние между которыми превышает $\rho_0$.

## 6.2.4 4.4. Spatial Coherence and Interference Aperture

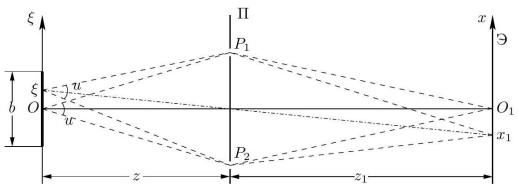



Рис. 2.16. Двухлучевая интерференционная схема с протяжённым источником

Измерить степень пространственной когерентности колебаний в точках $P_1$ и $P_2$ можно «сведя» эти колебания с помощью двухлучевой интерференционной схемы в какую-либо точку плоскости наблюдения, например, так, как показано на рис. 2.16: проделаем в точках $P_1$ и $P_2$ маленькие отверстия в непрозрачном экране П, расстояние между точками $P_1$ и $P_2$ равно $\rho$. Колебания поля в этих точках являются источниками волн справа от экрана П. Эти волны, интерферируя, создадут на экране Э картину интерференции, видность которой позволит нам судить о степени когерентности колебаний в точках $P_1$ и $P_2$ (мы описали схему опыта Юнга). Результирующую картину интерференции найдём сложив интерференционные картины, образованные излучением всех малых (некогерентных!) площадок $d\xi$ источника. Площадка $d\xi_0$, находящаяся в центре источника (в точке $O$), создаёт в плоскости наблюдения Э картину интерференции

$$dI_0 = 2J_0 d\xi \left[1 + \cos\frac{2\pi}{l}x\right]$$

где $J_0 d\xi$ - интенсивность одной из интерферирующих волн, $l = \frac{\lambda}{\rho}z_1$ - ширина интерференционной полосы. Ясно, что нулевая полоса (соответствующая нулевой разности хода) находится на оси симметрии в точке $O_1$. Элемент источника, имеющий координату $\xi$, создаёт картину, смещённую на расстояние $x_1 = \xi\frac{z_1}{z}$. (Полагаем, что расстояния $z_1$ и $z$ велики по сравнению с размером источника и расстоянием $\rho$ между точками $P_1$ и $P_2$):

$$dI = 2J_0 d\xi \left[1 + \cos\frac{2\pi}{l}\left(x - \frac{z_1}{z}\xi\right)\right] \tag{2.63}$$

Результирующую картину найдём суммированием (2.63) по всем площадкам $d\xi$ источника. Интегрируя (2.63), получаем

$$I(x) = 2I_0 \left[1 + \left(\frac{\sin\pi\frac{\rho}{\rho_0}}{\pi\frac{\rho}{\rho_0}}\right)\cos\frac{2\pi}{l}x\right] \tag{2.64}$$

Из (2.64) находим видность интерференционной картины, которая, как ясно из сравнения с (2.61), равна модулю пространственной степени когерентности:

$$\mathcal{V} = |\gamma(\rho)| = \left|\frac{\sin\pi\frac{\rho}{\rho_0}}{\pi\frac{\rho}{\rho_0}}\right| \tag{2.65}$$

Для оценки можно принять, что для наблюдения интерференции в опыте Юнга расстояние между отверстиями $P_1P_2 = \rho$ не должно превышать радиуса когерентности $\rho_0$. Из (2.62) находим $\frac{\rho}{z} \leqslant \frac{\lambda}{b}$.

гол $u = \frac{\rho}{z}$ между лучами, выходящими из какой-либо точки источника, которые затем сходятся в точке наблюдения (на экране, где наблюдается интерференция), называется апертурой интерференции. Итак, необходимое условие наблюдения интерференции (для любой интерференционной схемы):

$$u \lesssim \frac{\lambda}{b} \tag{2.66}$$

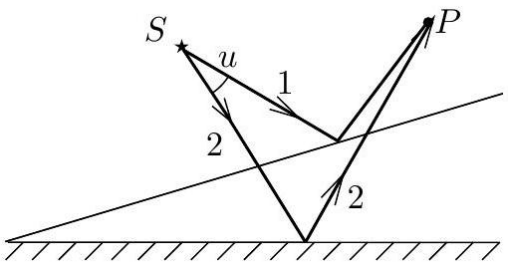
Чем меньше апертура интерференции, тем большего размера источник можно использовать в данной интерференционной схеме.

## 6.2.5 4.5. Localization of the Interference Pattern

Из-за немонохроматичности и протяжённости источника света интерференционная картина имеет хорошую контрастность (видность) только в определённых областях пространства, т. е. интерференционная картина локализована.

В качестве примера рассмотрим интерференцию на тонком воздушном клине, пренебрегая для простоты преломлением на поверхностях, образующих клин (рис. 2.17).

а)




б)

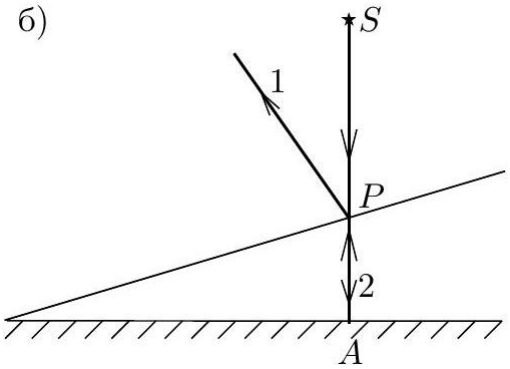



Рис. 2.17. Интерференция на клине (полосы равной толщины)

Для любых положений источника $S$ и точки наблюдения $P$ можно единственным образом построить два луча 1 и 2, исходящие из точки $S$, затем отражающиеся от верхней и нижней грани клина соответственно (рис. 2.17а) и сходящиеся в точке $P$ (если точка наблюдения находится с противоположной стороны клина, в ней будут сходиться не сами лучи, а их продолжения). Каждому положению точки наблюдения соответствует определённая разность хода приходящих в неё лучей 1 и 2. Следовательно, интерференционную картину от точечного источника можно наблюдать во всех точках пространства, а интерференционные полосы при этом будут не локализованы.

С другой стороны, если источник $S$ является протяжённым, то, как следует из изложенного в предыдущем параграфе, интерференцию можно будет наблюдать лишь в тех областях, в которых апертура интерференции $u$ (угол между интерферирующими лучами, исходящими из $S$) достаточно мала: согласно формуле (2.66), максимально допустимый размер источника равен $\lambda/u$.

Максимальная видность картины соответствует точкам наблюдения, для которых апертура интерференции равна нулю $u = 0$. Этому условию удовлетворяют точки, находящиеся на поверхности клина, как проиллюстрировано на рис. 2.17б. Интерференционные полосы, лока лизованные на поверхности клина (или, например, тонкой плёнки), называют полосами равной толщины. Такое название связано с тем, что разность хода интерферирующих лучей определяется оптической толщиной клина в точке наблюдения $P$ (на рис. 2.17 б она равна $2PA$ и постоянна вдоль ребра клина).

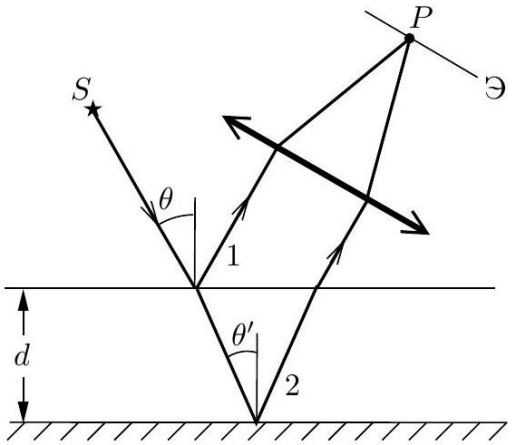



Рис. 2.18. Интерференция на пластине постоянной толщины (полосы равного наклона)

Плоскопараллельная пластина даёт возможность наблюдать интерференционные полосы, локализованные на бесконечности. На рис. 2.18 они наблюдаются в фокальной плоскости линзы. Такие полосы называют полосами равного наклона, поскольку каждой интерференционной полосе соответствует определённый угол наклона падающего луча $\theta$ и определённая разность хода $2d\cos\theta'$. Контраст картины для полос равного наклона не зависит от размеров источника, поскольку, как нетрудно видеть из рис. 2.18, апертура интерференции равна нулю.

## 6.3 5. Interferometers

Измерительные приборы, использующие явление интерференции, называют интерферометрами. Оптические интерферометры применяются в физических экспериментах для измерения длин волн спектральных линий, показателей преломления прозрачных сред, абсолютных и относительных длин, для контроля качества оптических деталей и их поверхностей. По числу интерферирующих пучков интерферометры разделяются на два класса - многолучевые и двухлучевые. В данном разделе мы рассматриваем двухлучевую интерференцию. Принцип работы всех двухлучевых интерферометров одинаков: свет от источника разделяется на два пучка, идущих по двум различным путям. Затем

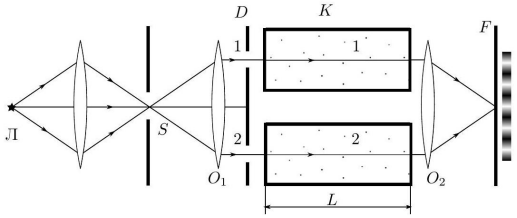



Рис. 2.19. Схема интерферометра Релея

эти пучки сводятся вместе, появляющаяся интерференционная картина исследуется с помощью регистрирующего прибора или визуально.

Вид интерференционной картины зависит от способа получения когерентных пучков, оптической разности хода, относительной интенсивности, размеров источника, спектрального состава света. Получают когерентные пучки двумя способами - делением волнового фронта и делением амплитуды.

В первом способе пучок делится, проходя через два близко расположенных отверстия (как, например, в опыте Юнга). Метод деления волнового фронта прост в реализации, его недостаток - большая апертура интерференции, и как следствие - небольшая интенсивность, поскольку источник должен иметь малые размеры.

Второй способ - деление амплитуды - реализуется, когда пучок делится на одной или нескольких частично отражающих поверхностей. Деление амплитуды может применяться при работе с протяжёнными источниками, что обеспечивает большую интенсивность картины (например, в интерферометрах Жамена и Майкельсона).

## 6.3.1 5.1. The Rayleigh Interferometer

В интерферометре Юнга, использующем метод деления волнового фронта, на экран попадает лишь малая доля светового потока от источника. Этот недостаток можно устранить, если поставить линзу после экрана с двойной щелью и наблюдение вести в точке, сопряжённой с источником света. Дальнейшие усовершенствования данной схемы привели к созданию получившего широкое распространение интерферометра Релея (рис. 2.19).

В этой схеме в качестве источника используют узкую щель $S$ (шириной $b$), освещаемую сколлимированным светом от лампочки Л. Щель расположена в фокальной плоскости объектива $O_1$. После экрана $D$ с двойной щелью два параллельных пучка проходят через кюветы 1 и 2 и попадают на объектив $O_2$, в фокальной плоскости которого рассматривается интерференционная картина. Для обеспечения когерентности пучков 1 и 2 ширина входной щели $S$ выбирается так, чтобы радиус когерентности на экране $D$ был больше расстояния между щелями 1 и 2. На экране $F$ рассматривается дифракционное изображение щели $S$, ширина изображения определяется шириной одной щели на экране $D$. Если интерференция рассматривается в белом свете, ширина щелей на экране $D$ выбирается в $2-3$ раза меньше расстояния между ними, так что в центральном максимуме видно $5-7$ интерференционных полос.

Для более точного измерения смещения полос из-за изменения коэффициента преломления в одном из плеч в интерферометре Релея используют две интерференционные картины. Верхняя часть пучков 1 и 2 проходит через кюветы $K$ с газом. Нижняя часть, проходящая под кюветами, образует в фокальной плоскости объектива $O_2$ неподвижную интерференционную картину. Для отдельного рассмотрения каждой системы полос (пучки света 1 и 2 параллельны и системы полос в фокальной плоскости совпадают) используют окуляр с цилиндрической линзой, которая пространственно разделяет эти системы полос. При заполнении кювет газами с одинаковыми коэффициентами преломления обе системы полос (верхняя и нижняя)

совпадают. Различие коэффициентов преломления газов в кюветах приводит к смещению верхней системы полос относительно нижней из-за возникновения между лучами 1 и 2 оптической разности хода $\Delta = L\,(n_2 - n_1)$ ($L$ - длина кювет). По числу полос между центрами обеих картин можно рассчитать изменение коэффициента преломления $n$.

Зависимость показателя преломления газа от давления и температуры. Воспользуемся известной формулой диэлектрической проницаемости $\varepsilon$ для газа невзаимодействующих диполей:

$$\varepsilon = n^2 = 1 + 4\pi N\alpha$$

где $N$ - концентрация молекул, $\alpha$ - поляризуемость молекулы (в ед. СГС). Эта формула справедлива для разреженных газов, и коэффициент преломления их мало отличается от единицы. Учитывая зависимость давления $P$ газа от температуры $P = Nk_{\text{в}}\,T$, где $k_{\text{в}}$ - константа Больцмана, получим соотношение

$$n - 1 \approx \frac{\alpha}{2k_{\text{B}}T}P$$

Тогда для разности показателей преломления $\delta n = n_2 - n_1$, измеряемой с помощью интерферометра Релея, и разности давлений $\delta P$, измеряемой с помощью манометра, имеем простое соотношение:

$$\delta n = \frac{\alpha}{2k_{\text{B}}T}\delta P \tag{2.67}$$

## 6.3.2 5.2. The Jamin Interferometer

Получить два плеча интерферометра можно делением амплитуды на толстой плоскопараллельной пластине. Большую разность хода можно компенсировать использованием второй такой же пластины (рис. 2.20). Впервые такую интерференцию наблюдал Брюстер. Эти полосы обычно называют полосами Брюстера. В быту их можно наблюдать на многих оконных стеклопакетах.

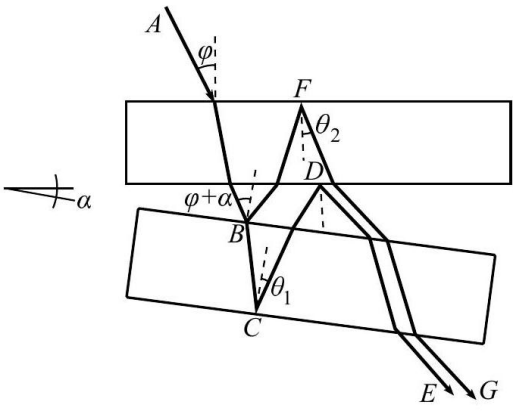



Рис. 2.20. Схема наблюдения полос Брюстера

Пластины располагаются под небольшим углом $\alpha$. Свет от источника $A$ последовательно проходит через две пластины, частично отражаясь. Интерферируют лучи $E$ и $G$ с минимальной разностью хода, показанные на рис. 2.20. Разность хода этих лучей (без учёта разности хода в воздухе) равна $2nd\,(\cos\theta_1 - \cos\theta_2)$, где $d$ - толщина пластин, $n$ - их показатель преломления. Обозначим угол падения луча на первую пластину $\varphi$, тогда на вторую луч падает под углом $\alpha + \varphi$. Для малых углов условие максимума

$$2nd\left(\cos\theta_1-\cos\theta_2\right)\approx 2dn\left(\frac{\varphi^2}{2n^2}-\frac{(\varphi+\alpha)^2}{2n^2}\right)=\frac{2\alpha d}{n}\left(\varphi+\frac{\alpha}{2}\right)=m\lambda \qquad (2.68)$$

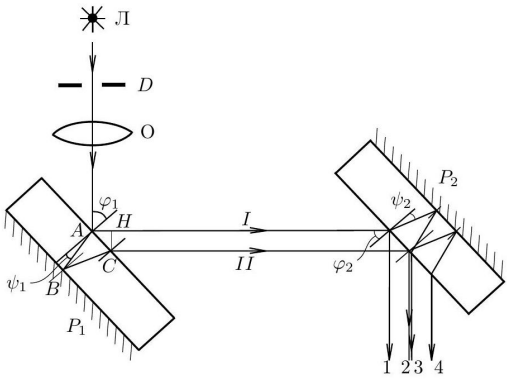



Рис. 2.21. Ход лучей в интерферометре Жамена

В данном случае мы наблюдаем полосы равного наклона. Из (2.68) можно выразить угловое расстояние между полосами:

$$\delta\varphi=\varphi_{m+1}-\varphi_m=\frac{\lambda n}{2\alpha d}$$

Подобная схема применяется в интерферометре Жамена, где для исключения разности хода в воздухе интерферирующие пучки между пластинами идут перпендикулярно бисекторной плоскости, а сами пластины располагаются под углом 45° к падающему свету (рис. 2.21). Полосы равного наклона будут наблюдаться в этом случае как полосы, параллельные ребру двугранного клина, образованного пластинами (разности хода при прохождении в воздухе не будет). Присутствие лучей 1 и 4 ухудшает чёткость интерференционной картины, поэтому их устраняют с помощью диафрагм.

Так как наблюдаются полосы равного наклона, протяжённый источник располагают в фокусе коллиматора, а полосы наблюдают в телескоп. Для идеальных плоскопараллельных пластин в трубе будет видно геометрическое изображение источника, а на нем будут видны полосы, ширина которых будет зависеть от угла $\alpha$ между пластинами.

При наблюдении в белом свете центральная полоса оказывается ахроматичной; она окружена двумя глубокими минимумами. Если труба наблюдения горизонтальна, нулевая полоса оказывается в центре поля зрения при горизонтальной ориентации клина. При повороте одной из пластин вокруг горизонтальной оси, параллельной оси клина, изменяется ширина интерференционных полос, как в рассмотренной системе полос Брюстера. Если повернуть зеркало вокруг вертикальной оси, меняется ориентация клина, нулевая полоса смещается вверх или вниз.

Расстояние между двумя плечами интерферометра Жамена будет зависеть от толщины пластин, и для достаточно толстых пластин в одно из плеч интерферометра можно поставить кювету с газом, коэффициент преломления которого можно измерить по смещению интерференционных полос. Допустимая клиновидность толстых пластин определится дифракционной расходимостью пучка света в каждом плече. Проверить клиновидность пластин можно, уменьшив угол между ними $\alpha$ до нуля. Если при этом в поле зрения будет видно не более двух полос равной толщины, пластины хорошие. При достаточно толстых пластинах (около 2-3 см) в каждое из двух плеч интерферометра можно поставить кюветы с газом (как в интерферометре Релея) для измерения малых изменений коэффициента преломления газов.

## 6.3.3 5.3. The Michelson Interferometer

Схема интерферометра приведена на рис. 2.22. В этом интерферометре когерентные волны получаются методом деления амплитуды при использовании полупрозрачного зеркала ПП. Интерференционная картина наблюдается в плоскости экрана Э. Источником света является точка, получающаяся за счёт фокусировки лазерного излучения с помощью дополнительной линзы. В данной схеме интерференции требуется большая временная когерентность, поэтому используется лазерный источник излучения.

Картина интерференции на экране Э может быть построена следующим образом. Отражаясь в зеркалах ПП, З $3_1$ и $3_2$, источник $S$ даёт два изображения $S_1$ и $S_2$ (рис. 2.22). Если зеркало $3_2$ не перпендикулярно падающему лучу $2, S_2$ переходит в $S_2'$. То есть интерференциию на экране можно рассматривать как интерференциию от двух точечных источников.

Если экран расположен перпендикулярно линии, соединяющей два источника, линиями равной разности фаз будут окружности

$$\sqrt{L^2 + r_n^2} - \sqrt{(L-a)^2 + r_n^2} \approx a + \frac{r^2 a}{2L(L-a)} = \text{ const}$$

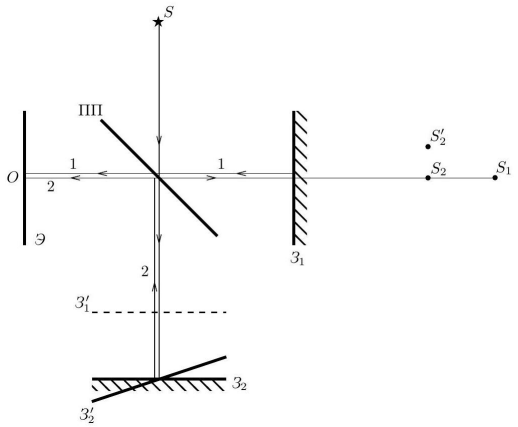



Рис. 2.22. Схема интерферометра Майкельсона

где $r_n$ - радиус окружности, $a$ - расстояние между изображениями источника $S$, а $L$ - расстояние от $S_1$ до экрана.

Интенсивность в центре картины ($r = 0$) определяется величиной $a$, равной разности хода между лучами 1 и 2. Порядок интерференции в центре равен $m_0 = \frac{a}{\lambda}$. Если в центре имеется максимум, тогда радиус $n$-го интерференционного кольца, отсчитанного от центра, будет определяться формулой

$$r_n \approx \sqrt{\frac{2nL(L-a)}{m_0}} \tag{2.69}$$

где $n = 0, 1, 2 \ldots$ При больших $n$ расстояние между полосами будет равно

$$\Delta r \approx \sqrt{\frac{L(L-a)}{2nm_0}}$$

При малом повороте зеркала $_2$ изображение источника $S_2$ переходит в $S_2'$. Если от центра экрана изображения источников будут видны под углом $\beta$, то в центре будут видны полосы, ширина которых согласно (2.25) приближенно равна $\lambda/\beta$.

По интерференционной картине можно определять длину волны источника света. При движении зеркала $3_1$ к экрану (зеркало $3_2$ установлено без наклона) разность хода со

временем уменьшается, интерференционные кольца стягиваются к центру, как бы исчезая в нём. При смещении зеркала на расстояние $l$ в центре исчезнет $N = 2\frac{l}{\lambda}$ колец. При равномерном перемещении зеркала, если за время $T$ зарегистрировано исчезновение $N$ колец, скорость перемещения зеркала равна

$$v = \frac{\lambda}{2}\frac{N}{T} \tag{2.70}$$

Для более точного рассмотрения необходимо воспользоваться релятивистскими формулами для эффекта Доплера. Учтём, что в системе координат, связанной с зеркалом $3_1$, движущимся со скоростью $v$, частота излучения источника $\omega_1$ отличается от исходной $\omega_0$, так что $\omega_1 = \omega_0\sqrt{\frac{c+v}{c-v}}$. Эта волна отражается от неподвижного зеркала и попадает на движущийся (в системе зеркала) приёмник. Частота приёма будет равна $\omega_2 = \omega_1\sqrt{\frac{c+v}{c-v}} = \omega_0\frac{c+v}{c-v}$. В центре экрана колебания $\omega_0$ и $\omega_2$ складываются, и регистрируемая интенсивность меняется с частотой $\Delta\omega = \frac{2v}{c-v}\omega_0$. Число периодов колебаний за время $t$ равно

$$N = \frac{\Delta\omega t}{2\pi} = 2\frac{vt}{\lambda}\frac{1}{1-\frac{v}{c}}$$

Для небольших скоростей эта формула совпадает с (2.70).

## 6.3.4 5.4. Newton's Rings

Этот классический опыт используется для определения радиуса кривизны сферических поверхностей линз. В этом опыте наблюдается интерференция волн, отражённых от границ тонкой воздушной прослойки, образованной сферической поверхностью линзы и плоской стеклянной пластиной. При нормальном падении света (рис. 2.23) интерференционные полосы локализованы на сферической поверхности и являются полосами равной толщины.

Геометрическая разность хода между интерферирующими лучами равна удвоенной толщине воздушного зазора $2d$ в данном месте. Для точки на сферической поверхности, находящейся на расстоянии $r$ от оси системы, имеем $r^2 = R^2 - (R-d)^2 = 2Rd - d^2$, где $R$ - радиус кривизны сферической поверхности (рис. 2.23).

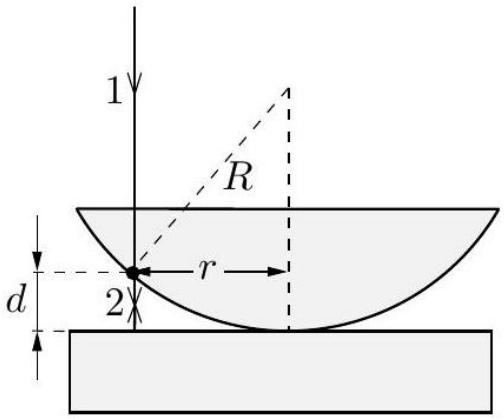



Рис. 2.23. Схема наблюдения колец Ньютона

При $R \gg d$ получим $d = r^2/2R$. С учётом изменения фазы на $\pi$ при отражении волны от оптически более плотной среды (на границе воздух-стекло) получим оптическую разность хода интерферирующих лучей:

$$\Delta = 2d + \frac{\lambda}{2} = \frac{r^2}{R} + \frac{\lambda}{2}$$

Условие интерференционного минимума $\Delta = (2m+1)\frac{\lambda}{2} (m = 0, 1, 2, \ldots)$, откуда получаем для радиусов тёмных колец

$$r_m = \sqrt{m\lambda R} \tag{2.71}$$

Аналогично для радиусов $r'_m$ светлых колец

$$r'_m = \sqrt{(2m-1)m\lambda R/2} \tag{2.72}$$

**Литература**


1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. III.
2. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 7.
3. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. IV, VI, VII.
4. Бутиков Е.И. Оптика. - М.: Высшая школа, 1986. Гл. 5.
5. Борн М., Волъф Э. Основы оптики. - М.: Наука, 1973. Гл. 7, 10.


# 7 Diffraction

Дифракция - отклонение волны за счет прохождение поверхностей.

В самом широком смысле слова дифракцией называют отклонения в распространении волн от законов геометрической оптики.

А если дифракция происходит на объектах порядка длины волны или где показатель преломления сильно меняется - говорят о *рассеянии* (?????)

Будем подразумевать, что распространение света происходит в трехмерном пространстве,

но иногда будем разбирать и двумерные случаи.

## 7.1 1. Diffraction

### 7.1.1 1.1. Setting the Task. The Wave Parameter

Дифракцией, в самом широком смысле слова, называют отклонения в распространении волн от законов геометрической оптики. Частный случай дифракции - огибание волной препятствия и её проникновение в область геометрической тени.

Пусть препятствие (непрозрачный экран с отверстием, расположенный в плоскости $z = 0$) освещается волной, созданной источниками, расположенными слева от экрана. Требуется определить волновое поле справа от экрана, в области $z > 0$ (рис. 3.1), в частности в плоскости наблюдения П, расположенной на расстоянии $z$ от экрана.

Для решения этой задачи необходимо найти ответ на два основных вопроса. Первый вопрос ставится следующим образом: известно поле $f_s(x, y)$, созданное в плоскости $z = 0$ (в отсутствие экрана) какими-либо «сторонними» источниками, расположенными в области $z < 0$. Затем в плоскость $z = 0$ помещается непрозрачный

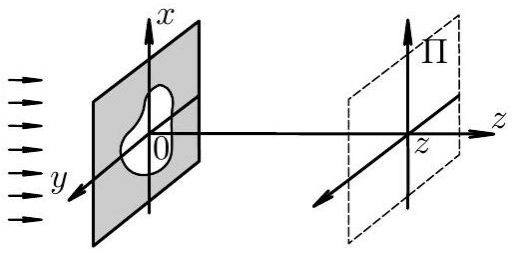



Рис. 3.1. К постановке задачи экран с отверстием. Требуется найти поле $f_0(x, y)$, возникающее в плоскости, непосредственно примыкающей к экрану справа от него (эту плоскость далее будем называть плоскостью $z = 0_+$). Это и есть граничное условие, необходимое для решения дифракционной задачи.

Второй вопрос: каким образом, зная граничное поле $f_0(x, y)$, определить искомое поле $f(x, y, z)$ в области $z > 0$ (т.е. в пространстве справа от экрана).

Волновой параметр. Основными параметрами, существенно определяющими характер дифракционных явлений, являются длина волны $\lambda$, размер отверстия $b$ и расстояние до плоскости наблюдения $z$. Как показывает дальнейший анализ, характер дифракционных явлений определяется значением волнового параметра

$$p = \frac{\sqrt{\lambda z}}{b} \tag{3.1}$$

Область значений волнового параметра $p \ll 1$ является областью применимости приближения геометрической оптики. Область значений $p \gg 1$ называется областью дифракции Фраунгофера (или дальней волновой зоной). Промежуточная область $p \sim 1$ - областью дифракции Френеля (или ближней волновой зоной).

## 7.1.2 1.2. Boundary Conditions. The Transparency Function

Определение граничных условий представляет собой исключительно трудную проблему. Действительно, пусть речь идёт об электромагнитной волне, падающей на проводящий экран с отверстием. Переменное электрическое поле волны порождает на проводящем экране переменные токи, которые в свою очередь индуцируют «вторичную» электромагнитную волну. Эта волна зависит как от физических свойств экрана, так и от геометрии отверстия. Суммарное волновое поле в плоскости $z = 0_+$представляет собой суперпозицию первичной волны (поля сторонних источников) и вторичной волны, индуцированной токами проводимости, причём электрическое и магнитное поля суммарной волны должны удовлетворять известным в электродинамике граничным условиям. Имеется лишь небольшое число задач, для которых найдено точное решение (например, задача дифракции плоской волны на идеально проводящем шаре).

Мы, следуя Кирхгофу, определим приближённые граничные условия следующим образом: в той части плоскости $z = 0_+$, которая затенена непрозрачным экраном, граничное поле полагается равным нулю. На открытой части волнового фронта (т.е. в области отверстия) граничное поле полагается равным полю сторонних источников. Другими словами, в этой области установка экрана с отверстием совершенно не влияет на падающую волну. Итак, граничное поле определяется следующим равенством:

$$f_0(x, y) = \begin{cases} f_s(x, y) \text{ в точках, принадлежащих отверстию,} \\ 0 \text{ в точках, затенённых экраном.} \end{cases} \tag{3.2}$$

Ещё раз подчеркнём, что граничные условия (3.2) являются приближёнными: наличие экрана в действительности приводит к тому, что поле на отверстии отлично от падающего поля $f_s(x, y)$, и особенно сильно оно искажено вблизи краев отверстия; да и в

затенённой области поле отлично от нуля вблизи краев. Однако, как показывает сравнение теоретических выводов, основанных на приближённых граничных условиях Кирхгофа, с экспериментом, согласие оказывается удовлетворительным при условии, что, во-первых, линейные размеры отверстия $b$ велики по сравнению с длиной волны: $b > \lambda$ и, во-вторых, плоскость наблюдения находится на расстоянии, много большем длины волны $z \gg \lambda$.

Равенству (3.2) можно придать другую форму, если ввести функцию пропускания непрозрачного экрана с отверстием $t(x, y)$ согласно формуле

$$t(x,y) = \begin{cases} 1, & \text{если точка } (x,y) \text{ принадлежит отверстию,} \\ 0, & \text{если точка } (x,y) \text{ лежит вне отверстия.} \end{cases}$$

Тогда

$$f_0(x,y) = f_s(x,y) \cdot t(x,y) \tag{3.3}$$

Равенство (3.3) можно обобщить на случай произволъного тонкого экрана. Представим себе тонкую пластинку-транспарант, амплитудная пропускаемость которой (прозрачность) меняется от точки к точке по закону $a(x, y)$, а оптическая толщина (произведение показателя преломления на толщину пластинки) - по закону $\Delta(x, y)$. Пусть такая пластинка освещается плоской нормально падающей волной. Что происходит с волной на выходе из пластинки в плоскости $z = 0_+$? Переменная поглощательная способность транспаранта приводит к тому, что амплитуда колебаний на выходе будет функцией координат $a(x, y)$. Переменная оптическая толщина $\Delta(x, y)$ приводит к различному в разных точках набегу фазы $\varphi(x, y) = k \cdot \Delta$, так что распределение фаз колебаний на выходе будет описываться функцией $\varphi(x, y)$.

Определим комплексную пропускаемость транспаранта согласно формуле

$$t(x,y) = a(x,y)e^{i\varphi(x,y)}$$

Тогда комплексная амплитуда волны на выходе из пластинки может быть найдена по той же формуле (3.3), которая и определяет граничное поле $f_0(x, y)$ для произвольного тонкого экрана.

Таким образом, изменение амплитуды волны и набег фазы в любой точке $(x, y)$ тонкой пластинки (транспаранта) рассчитывают вдоль тонкого луча, пронизывающего транспарант по нормали в точке $(x, y)$.

Другими словами, мы полагаем, что распространение волны непосредственно через тонкий экран подчиняется законам геометрической (лучевой) оптики (см. п. 2.5 - область геометрической оптики).

### 1.2.1. Примеры тонких экранов

Приведём несколько примеров тонких экранов (транспарантов), часто используемых в оптике в качестве дифракционных элементов.

1. Амплитудные экраны

Амплитудная синусоидальная решётка - тонкая пластинка, амплитудная прозрачность которой меняется от точки к точке по закону [1]

$$t(x) = \beta(1 + m\cos\Omega x) \tag{3.4}$$

Константа $m$ - глубина модуляции, $\Omega = \frac{2\pi}{d}$ - пространственная частота решётки с периодом $d$.

Дифракционная решётка, функция пропускания которой показана на рис. 3.2. Это непрозрачный экран с рядом параллельных равноотстоящих щелей, $d$ - период решётки, $b$ - ширина щелей.

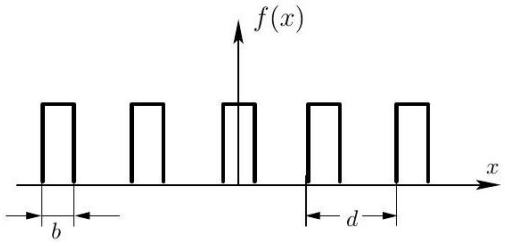



Рис. 3.2. Функция пропускания дифракционной решетки

2. Фазовые экраны

Если амплитудная прозрачность транспаранта не меняется от точки к точке, а изменяется только набег фазы $\varphi(x, y)$, то такой транспарант называется фазовым. Функция пропускания фазового экрана $t(x, y) = e^{i\varphi(x,y)}$.

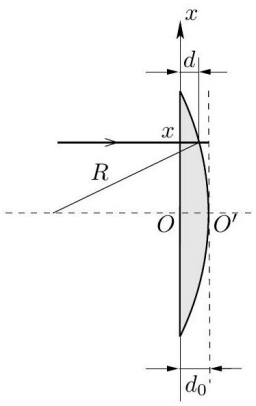
**Фазовая синусоидальная решётка**

$$t(x) = e^{im\cos\Omega x} \tag{3.5}$$

$m$ - глубина модуляции фазы, $\Omega = \frac{2\pi}{d}$ - пространственная частота решётки. При $m \ll 1$ имеем

$$t(x) \simeq 1 + im\cos\Omega x$$

Тонкая линза. Рассчитаем набег фазы вдоль луча, пронизывающего линзу на расстоянии $x$ от оптической оси (от плоскости, примыкающей к линзе слева, до плоскости, примыкающей к линзе справа). Он равен (рис. 3.3) $knd + k(d_0 - d)$. Интерес представляет лишь относительный набег фазы (по отношению к лучу, пронизывающему линзу вдоль оптической оси). Он равен $knd_0 = knd + kn(d_0 - d)$. Далее, поскольку $d_0 - d = R - \sqrt{R^2 - x^2}$ и $\sqrt{R^2 - x^2} \approx R - \frac{x^2}{2R}$, где $R$ - радиус кривизны поверхности линзы (последнее означает, что сферическая форма поверхности заменяется параболической), то имеем




[0]1 Значение $t(x)$ не может превосходить единицу (амплитуда волны на выходе из экрана не может превосходить амплитуду на входе), поэтому в формулу (3.4) введён множитель $\beta < 1$, который во многих случаях, когда это не может вызвать недоразумений, полагают равным единице.

Рис. 3.3. Тонкая линза

Введя обозначение $\frac{1}{f} = \frac{n-1}{R}$($f$ - фокусное расстояние линзы), получаем функцию пропускания тонкой линзы:

$$t(x) = e^{-i\frac{k}{2f}x^2} \tag{3.6}$$

В двумерном случае имеем

$$t(x,y) = e^{-i\frac{k}{2f}\left(x^2+y^2\right)} \tag{3.7}$$

Если обе поверхности линзы сферические с радиусами кривизны $R_1$ и $R_2$, то для фокусного расстояния тонкой линзы $f$ легко получить

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} + \frac{1}{R_2}\right)$$

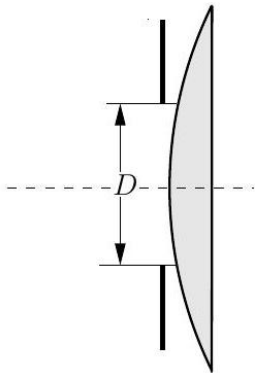



Рис. 3.4. Линза с диафрагмой

Любая реальная линза имеет конечные размеры. Можно учесть этот факт, накрыв «идеальную линзу» с функцией пропускания (3.7) непрозрачным экраном с отверстием (диафрагмой) диаметра $D$, центр которого совпадает с центром линзы (рис. 3.4). Очевидно, функция пропускания задиафрагмированной линзы (т. е. линзы диаметра $D$) имеет вид

$$t(x,y) = P_{\mathrm{D}}(x,y)e^{-i\frac{k}{2f}\left(x^2+y^2\right)} \tag{3.8}$$

где $P_D(x,y)$ - единично-нулевая функция:

$$P_D(x,y) = \begin{cases} 1 & \text{при } |x^2+y^2| \leqslant \left(\frac{D}{2}\right)^2 \\ 0 & \text{при } |x^2+y^2| > \left(\frac{D}{2}\right)^2 . \end{cases}$$

## 7.1.3 1.3. The Huygens-Fresnel Principle

Принцип Гюйгенса-Френеля является одним из широко используемых рецептов для расчёта дифракционных задач.

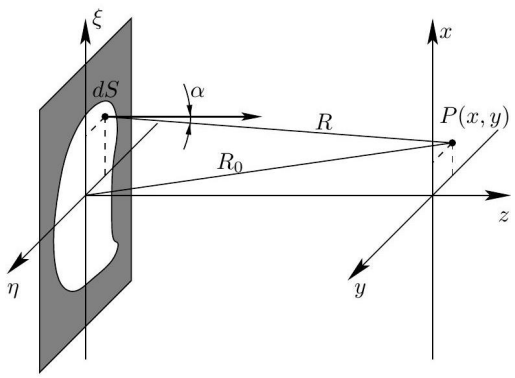



Рис. 3.5. К принципу Гюйгенса-Френеля

Пусть волна света, созданная источниками, расположенными в области $z < 0$, достигла плоскости $z = 0$ (рис. 3.5). Световое поле в этой плоскости нам известно. Пусть его комплексная амплитуда есть $f_0(x, y) = a_0(x, y)e^{i\varphi_0(x,y)}$, где функции $a_0(x, y)$ и $\varphi_0(x, y)$ описывают распределение амплитуд и фаз колебаний в плоскости $z = 0$. Что происходит с волной далее, по мере её распространения в области $z > 0$, как изменяется структура волны - распределение амплитуд и фаз колебаний, что собой представляет световое колебание в некоторой точке наблюдения $P$?

Согласно принципу Гюйгенса, каждую точку $(\xi, \eta)$ плоскости $z = 0$, куда пришла волна (каждую малую площадку $ds = d\xi d\eta$), можно рассматривать как источник вторичной волны. То есть можно представить себе, что волна возбуждает колебания некоторого фиктивного источника (осциллятора), который и переизлучает вторичную волну. Частота $\omega$ этой переизлучённой волны совпадает с частотой исходной монохроматической волны. Френель дополнил принцип Гюйгенса, предложив рассматривать световое колебание в любой точке наблюдения в области $z > 0$ как результат интерференциии этих вторичных волн.

Предполагается, что амплитуда излучения вторичного источника пропорциональна амплитуде $a_0(\xi, \eta)$ колебания, созданного реальной волной, пришедшей к площадке $ds$. Фаза колебания также задаётся фазой $\varphi_0(\xi, \eta)$ пришедшей к элементу $ds$ волны. Напомним, что согласно принятым граничным условиям (3.2) на открытой части волнового фронта, т.е. в области отверстия, волна не искажается препятствием, причём работают лишь вторичные источники, находящиеся на открытой части, не затенённой непрозрачным экраном. Затенённые площадки не переизлучают.

Далее предполагается, что маленькая площадка ds переизлучает, подобно точечному источнику, сферическую волну, т. е. для вычисления вклада, который даёт эта площадка в суммарное колебание в точке наблюдения $P$, нужно учесть ослабление амплитуды и набег фазы (так же, как в сферической волне $\frac{1}{R}e^{ikR}$, $R$ - расстояние от площадки $ds$ до точки наблюдения). Наконец, предполагается, что амплитуда колебания пропорциональна видимой из точки наблюдения площади элемента $ds$, т. е. пропорциональна $ds\cos\alpha$. Таким образом, вклад элемента $ds$ пропорционален величине

$$f_0(\xi, \eta)\frac{e^{ikR}}{R}\cos\alpha \cdot d\xi d\eta$$

Полное световое колебание $g(x, y)$ есть результат интерференции всех вторичных волн, посылаемых всеми площадками $ds$, расположенными в области отверстия:

$$g(x, y) = K_0 \iint_S f_0(\xi, \eta)\frac{e^{ikR}}{R}\cos\alpha d\xi d\eta$$

[0]2 Строгое обоснование данного принципа на основе рассмотрения волнового уравнения было дано Кирхгофом (см., напр., [1, §43]).

Можно показать [3], что нормировочный коэффициент пропорциональности равен $K_0 = \frac{1}{i\lambda}$.

Таким образом, количественная формулировка принципа ГюйгенсаФренеля принимает вид

$$g(x,y) = \frac{1}{i\lambda}\iint_S f_0(\xi,\eta)\frac{e^{ikR}}{R}\cos\alpha d\xi d\eta \tag{3.9}$$

Приближённые формулировки. При условии, что размер препятствия (например, размер отверстия в непрозрачном экране) мал по сравнению с расстоянием $R_0$ до точки наблюдения, амплитудный множитель $\frac{1}{R}$, учитывающий уменьшение амплитуды в сферической волне по мере удаления от вторичного источника $ds$, можно заменить постоянной величиной $\frac{1}{R_0}$ (рис. 3.5). Множитель наклона $\cos\alpha$ также считаем приблизительно одинаковым (и равным единице) для всех вторичных источников, расположенных в области отверстия. Тогда в этом приближении принцип Гюйгенса-Френеля приобретает следующий вид:

$$g(x,y) = \frac{1}{i\lambda R_0}\iint f_0(\xi,\eta)e^{ikR}d\xi d\eta \tag{3.10}$$

Точное выражение для расстояния $R$ от вторичного источника (малой площадки, расположенной в точке $(\xi,\eta)$) до точки наблюдения $P(x,y,z)$ равно

$$R = \sqrt{z^2+(x-\xi)^2+(y-\eta)^2} = z\sqrt{1+\frac{(y-\xi)^2}{z}+\frac{(y-\eta)^2}{z}} \tag{3.11}$$

Вычисляя величину $R$, входящую в фазовый множитель $e^{ikR}$, мы не можем довольствоваться грубой оценкой $R \approx R_0 = \sqrt{z^2+x^2+y^2}$, поскольку ошибка при вычислении фазы колебаний $kR$ должна быть мала по сравнению с $\pi$ и, следовательно, ошибка в вычислении $R$ мала по сравнению с $\frac{\lambda}{2}$. В противном случае результат интерференции волн, приходящих в точку наблюдения от различных вторичных источников, находящихся в области отверстия, будет найден неверно. Считая поправку к единице под знаком радикала в (3.11) малой, оставим только два члена разложения в ряд Тейлора, полагая, что отброшенные слагаемые вносят малую поправку по сравнению с $\lambda$. Это приближение называется френелевским:

$$R \approx z + \frac{(x-\xi)^2}{2z} + \frac{(y-\eta)^2}{2z} \tag{3.12}$$

С помощью (3.9) и (3.12) получаем выражение, которое используют обычно для расчёта френелевских дифракционных картин:

$$g(x,y) = \frac{e^{ikz}}{i\lambda z}\iint f_0(\xi,\eta)e^{i\frac{k}{2z}\left[(x-\xi)^2+(y-\eta)^2\right]}d\xi d\eta \tag{3.13}$$

Если отверстие освещается плоской волной амплитуды $A_0$: $f_0(\xi,\eta) \equiv A_0$, а точка наблюдения лежит на оси $z(x=0,y=0)$, то из (3.13) получаем

$$g(0,0) = A_0\frac{e^{ikz}}{i\lambda z}\iint e^{i\frac{k}{2z}\left(\xi^2+\eta^2\right)}d\xi d\eta \tag{3.14}$$

[03] См. литературу, рекомендованную к разделу.

**Теория по Ландау 2**

Дифракция Законы геометрической оптики строго точны лишь в идеальном слу- чае, когда длину волны можно рассматривать как бесконечно малую. Чем хуже выполнено это условие, тем сильнее проявляются отклонения от геометрической оптики.

Явления, наблюдающиеся в результате этих отклонений, носят название явлений дифракции. Явления дифракции можно наблюдать, например, если на пути распространения света 1) находятся препятствия — непрозрачные тела (будем называть их экранами) произвольной формы или, например, если свет проходит через отверстия в непрозрачных экранах.

Если бы законы геометрической оптики строго выполнялись, то за экранами находились бы области тени, резко отграниченные от областей, куда свет попадает.

Дифракция же приводит к тому, что вместо резкой границы между светом и тенью получается довольно сложная картина распределения интенсивности света.

Эти явления дифракции тем силь- нее выражены, чем меньше размеры экранов и отверстий в них или чем больше длина волны. Задача теории дифракции заключается в том, чтобы при данном расположении и форме тел (и расположении источников света) опре- делить распределение света, т.е.

электромагнитное поле во всем про- странстве.

Точное разрешение этой задачи возможно только путем решения волнового уравнения с соответствующими граничными усло- виями на поверхности тел, зависящими еще к тому же и от оптических свойств материала.

Такое решение обычно представляет большие мате- матические трудности. Однако во многих случаях оказывается достаточным приближенный метод решения задачи о распределении света вблизи границы между светом и тенью.

Этот метод применим в случаях слабого отклоне- ния от геометрической оптики.

Тем самым предполагается, во-первых, что все размеры велики по сравнению с длиной волны (это относится как к размерам экранов или отверстий в них, так и к расстояниям от тел до точек испускания и наблюдения света); во-вторых, рассмат- риваются лишь небольшие отклонения света от направления лучей, определяемых геометрической оптикой. Рассмотрим какой-нибудь экран с отверстием, через которое прохо- дит свет от данных источников.

Рис.

9 изображает этот экран в разрезе (жирная линия); свет идет слева направо.

Будем обозначать посред-

ством $u$ любую из компонент поля E или H.

При этом под $u$ мы будем подразумевать поле как функцию только от координат, т.е.

без множителя $e^{-i\omega t}$, определяющего зависимость от времени.

Нашей задачей является определение интенсивности света, т.е.

поля $u$ в любой точке наблюдения $P$ за экраном.

При приближенном решении этой задачи в случаях, когда отклонения от геометрической оптики малы, можно

счИтать, что в точках отверстиЯ поле таково, каким ОНО было бы при отсутствии вообще какого-либо ЭКрана.

Другими словами, значения поля здесь те, которые следуют из геометрической оптики.

Во всех же точках, находяЩихся непосредственно за экраном, поле можно положить равным нулю.

При этом, очевидНО, свойства самого экрана (материала, из которого он сделан) вообще не играют роли.

Очевидно также, что в рассматриваемых случаях для дифракЦИИ сушественна только форма края отверстиЯ и не существенна форма непрозрачного экрана.

Проведем какую-нибудь поверхность, закрываЮШу Ю отверстие в экране и ограниченную его краями (разрез такой поверхности на рис.

9 изображен

штриховой линией).

Эту поверхность разобьем на участки с площаДью $df$, малые по сравнению с размерами отверстия, но большие по сравнению с длиной волны света.

Мы можем тогда рассматривать каждый из этих участков, до которых дошла световая волна, так, как будто бы он сам делается источником световой волны, распространяющейся во все стороны от этого участка.

Поле в точке $P$ мы будем рассматривать как результат наложения полей, исходящиХ из всех участков $df$ поверхности, закрывающей отверстие (так называемый принцип Гюйгенса).

Поле, создаваемое участком $df$ в точке $P$, пропорционально значению $u$ поля в самом участке $df$ (напоминаем, что поле в $df$ мы предполагаем таким, каким оно было бы при отсутствии экрана).

Кроме того, оно пропорционально проекции $df_n$ площади $df$ на плоскость, перпендикулярную к направлению $\mathbf{n}$ луча, пришедшего из источника света в df.

Это следует из того, что какой бы формой ни обладал участок $df$, через него будут проходить одинаковые лучи, если только его проекция $df_n$ будет неизменной, а потому и его действие на поле в точке $P$ будет одинаковым.

Таким образом, поле, создаваемое в точке $P$ участком $df$, пропорЦионально $udf_n$. Далее, надо еще учесть изменение амплитуды и фазы волны при ее распространении от $df$ к точке $P$. Закон этого изменения определяется формулой (54.3).

Поэтому $udf_n$ надо умножить еще на $\frac{1}{R}e^{ikR}$( где $R$ - расстояние от $df$ до $P$, а $k$ - абсолютная величина

волнового вектора света), и мы находим, что искомое поле равно

$$au\frac{e^{ikR}}{R}df_n$$

где $a$ есть неизвестная пока постоянная.

Полное же поле в точке $P$, Являющееся результатом наложения полей, создаваемых всеми $df$, есть

$$u_P = a\int\frac{ue^{ikR}}{R}df_n$$

где интеграл распространен по поверхности, ограниченной краем отверстия.

Этот интеграл в рассматриваемом приближении не зависит, Конечно, от формы этой поверхности.

Формула (59.1) применима, очевидно, и к дифракции не от отверстия на экране, а от экрана, вокруг которого свет может свободно распространяться.

В этом случае поверхность интегрирования в (59.1) простирается во все стороны от края экрана.

Для определения постоянной $a$ рассмотрим плоскую волну, распространяющуюся вдоль оси $x$; волновые поверхности параллельны Плоскости $yz$. Пусть $u$ есть значение поля в плоскости $yz$. Тогда В точке $P$, которую мы выберем на оси $x$, поле равно $u_P = ue^{ikx}$. С другой стороны, поле в точке $P$ можно определить исходя из формуЛы (59.1), выбрав в качестве поверхности интегрирования, например, плоскость уз.

При этом ввиду малости угла дифракции в интеграле существенны только точки плоскости $yz$, близкие к началу координат, Т.

е.

точКи, в которых $y, z \ll x$($x$ - коорДината точки $P$).

Тогда

$$R = \sqrt{x^2 + y^2 + z^2} \approx x + \frac{y^2 + z^2}{2x}$$

и (59.1) даст:

$$u_P = au\frac{e^{ikx}}{x}\int\limits_{-\infty}^{+\infty} e^{ik\frac{y^2}{2x}}dy \cdot \int\limits_{-\infty}^{+\infty} e^{ik\frac{z^2}{2x}}dz$$

где $u$ - постоянная (поле в плоскости $yz$); в множителе $1/R$ можно положить $R \approx x =$ const.

Стоящие здесь интегралы подстановкой $y = \xi\sqrt{2x/k}$ приводятся к виду

$$\int\limits_{-\infty}^{+\infty} e^{i\xi^2}d\xi = \int\limits_{-\infty}^{+\infty} \cos\xi^2 d\xi + i\int\limits_{-\infty}^{+\infty} \sin\xi^2 d\xi = \sqrt{\frac{\pi}{2}}(1+i)$$

и мы получаем: $u_P = aue^{ikx}\frac{2i\pi}{k}$. С другой стороны, $u_P = ue^{ikx}$ и, следовательно, $a = \frac{k}{2\pi i}$. Подставляя это в (59.1), находим окончательное

решение поставленной задачи в виде

$$u_P = \int \frac{ku}{2\pi iR}e^{ikR}df_n$$

При выводе формулы (59.2) источник света предполагался, по суЩеству, точечным, а самый свет - строго монохроматическим.

Случай реального протяженного источника, испускающего немонохроматический свет, не нуждается, однако, в особом исследовании.

Вследствие полной независимости (некогерентности) света, испускаемого различными точками источника, и некогерентности различных спектральных компонент испускаемого света суммарный результат дифракции своДится просто к сумме распределений интенсивности, получающихся от дифракЦии каждой из независимых компонент света.

Применим формулу (59.2) для решения вопроса об изменении фазы при прохождении луча через точку его касания с каустикой (см.

конец \$54).

Выберем в качестве поверхности интегрирования в (59.2) какую-либо волновую поверхность и будем определять поле $u_P$ В точке $P$, лежащей на некотором данном луче на расстоянии $x$ от точки его пересечения с выбранной волновой поверхностью (эту точку выберем в качестве начала координат $O$, а в качестве плоскости $yz$ - плоскость, касательную к волновой поверхности в точке $O$).

При интегрировании В (59.2) существен только небольшой участок волновой поверхности вблизи точки О.

Если плоскости $xy$ и $xz$ выбраны совпадающими с главными плоскостями кривизны волновой поверхности в точке О, то вблизи этой точки уравнение поверхности есть

$$X = \frac{y^2}{2R_1} + \frac{z^2}{2R_2}$$

где $R_1$ и $R_2$ - радиусы кривизны.

Расстояние же $R$ от точки волновой Поверхности с координатами $X, y, z$ до точки $P$ с координатами $x, 0, 0$ есть

$$R = \sqrt{(x-X)^2 + y^2 + z^2} \approx x + \frac{y^2}{2}\left(\frac{1}{x} - \frac{1}{R_1}\right) + \frac{z^2}{2}\left(\frac{1}{x} - \frac{1}{R_2}\right)$$

Вдоль волновой поверхности поле $u$ можно считать постоянным; то же касается множителя $1/R$. Поскольку мы интересуемся только изменением фазы волны, то коэффициент опускаем и пишем просто:

$$u_P \sim \frac{1}{i}\int e^{ikR}df_n \approx$$

$$\approx \frac{e^{ikx}}{i}\int\limits_{-\infty}^{+\infty} e^{ik\frac{y^2}{2}\left(\frac{1}{x}-\frac{1}{R_1}\right)}dy \cdot \int\limits_{-\infty}^{+\infty} e^{ik\frac{z^2}{2}\left(\frac{1}{x}-\frac{1}{R_2}\right)}dz.$$

Центры кривизны волновой поверхности лежат на рассматриваемом луче в точках $x = R_1$ и $x = R_2$; это и есть точки касания

лучом обеих каустик.

Пусть $R_2 < R_1$. При $x < R_2$ коэффициенты при $i$ в показателях подынтегральных выражений в обоих интегралах положительны, и каждый из этих интегралов содержит множитель 1+.

Поэтому на участке луча до касания первой каустики имеем $u_P \sim e^{ikx}$. При $R_2 < x < R_1$, т.е.

на отрезке луча между двумя точками касания, интеграл по $dy$ содержит множитель $1 + i$, а интеграл по $dz$ - множитель $1 - i$, так что их произведение вовсе не содержит $i$. Таким образом имеем здесь: $u_P \sim -ie^{ikx} = e^{i(kx-\pi/2)}$, т.е.

при прохождении луча вблизи первой каустики фаза дополнительно меняется на - $\pi/2$. Наконец, при $x > R_1$ Имеем $u_P \sim -e^{ikx} = e^{i(kx-\pi)}$, т.

е.

при прохождении вблизи второй каустики фаза еще раз меняется на $-\pi/2$.

З а д а ч а Определить распределение интенсивности света вблизи точки касания луча с каустикой.

Р е ш е н и е.

Для решения задачи пользуемся формулой (59.2), производя в ней интегрирование по какой-либо волновой поверхности, достаточно удаленной от рассматриваемой точки касания луча с каустикой.

На рис.

10$ab$ есть сечение этой волновой поверхности, а $a'b'$ - сечение каустики; $a'b'$ есть эволюта кривой ab.

Мы интересуемся распределением интенсивности вблизи точки $O$ касания луча $QO$ с каустикой; длина $D$ отрезка $QO$ луча предполагается достаточно большой.

Посредством $x$ обозначим расстояние от точки $O$ вдоль нормали к каустике, причем будем считать положительными значения $x$ для точек, лежащих на нормали по направлению к центру кривизны.

Подынтегральное выражение в (59.2) есть функция от расстояния $R$ от произвольной точки $Q'$ на волновой поверхности до точки $P$. По известному свойству эволюты сумма длины отрезка $Q'O'$ касательной в точке $O'$, и длины дуги $OO'$ равна длине $QO$ касательной в точке $O$. Для близких друг к другу точек $O$ и $O'$ имеем $OO' = \theta\rho$($\rho$ - радиус кривизны каустики в точке $O$).

Поэтому длина $Q'O' = D - \theta\rho$. Расстояние же $Q'O$ (по прямой) равно приближенно (угол $\theta$ предполагается малым):

$$Q'O \approx Q'O' + \rho\sin\theta = D - \theta\rho + \rho\sin\theta \approx D - \frac{\rho\theta^3}{6}$$

Наконец, расстояние $R = Q'P$ равно $R \approx Q'O - x\sin\theta \approx Q'O - x\theta$, т.

e.

$$R \approx D - x\theta - \frac{1}{6}\rho\theta^3$$

Подставляя это выражение в (59.2), найдем:

$$u_P \sim \int_{-\infty}^{+\infty} e^{-ikx\theta - \frac{ik\rho}{6}\theta^3} d\theta = 2\int_0^{\infty} \cos\left(kx\theta + \frac{k\rho}{6}\theta^3\right) d\theta$$

(медленно меняющийся множитель $1/D$ в подынтегральном выражении не существен по сравнению с экспоненциальным множителем, и мы считаем его постоянным).

Вводя новую переменную интегрирования $\xi = (k\rho/2)^{1/3}\theta$, ПОЛу ЧИМ:

$$u_P \sim \Phi\left(x\sqrt[3]{\frac{2k^2}{\rho}}\right)$$

где $\Phi(t)$ есть так называемая функция Эйри [1]).

Для интенсивности $I \sim |u_P|^2$ напишем:

$$I = 2A\left(\frac{2k^2}{\rho}\right)^{1/6} \Phi^2\left(x\sqrt[3]{\frac{2k^2}{\rho}}\right)$$

(о выборе постоянного множителя см.

ниже).

При больших положительных значениях $x$ имеем отсюда асимптотическую формулу

$$I \approx \frac{A}{2\sqrt{x}} \exp\left(-\frac{4x^{3/2}}{3}\sqrt{\frac{2k^2}{\rho}}\right)$$

т.е.

интенсивность экспоненциально убывает (область «тени»).

При больших же по абсолютной величине отрицательных значениях $x$ имеем:

$$I \approx \frac{2A}{\sqrt{-x}} \sin^2\left[\frac{2(-x)^{3/2}}{3}\sqrt{\frac{2k^2}{\rho}} + \frac{\pi}{4}\right]$$

т.е.

интенсивность быстро осциллирует; усредненное по этим осцилляциям значение равно

$$\bar{I} = \frac{A}{\sqrt{-x}}$$

Отсюда выясняется смысл постоянной $A$ - она определяет интенсивность вдали от каустики, которая получилась бы из геометрической оптики без учета явлений дифракции.

Наибольшее значение, равное $0,949$, функция $\Phi(t)$ имеет при $t = -1,02$; соответственно этому наибольшая интенсивность достигается при $x\,(2k^2/\rho)^{1/3} = -1,02$ и равна

$$I = 2,03Ak^{1/3}\rho^{-1/6}$$

(в самой же точке касания луча с каустикой, $x = 0$, имеем $I = 0,89Ak^{1/3}\rho^{-1/6}$, так как $\Phi(0) = 0,629$). Таким образом, вблизи каустики интенсивность пропорциональна $k^{1/3}$, т.е. $\lambda^{-1/3}$($\lambda$ - длина волны).

При $\lambda \to 0$ интенсивность, как и следовало (ср.

\$54), стремится к бесконечности.

### Принцип Гюйгенса

Обсудим подробно принцип Гюйгенса

### определение принципа Гюйгенса

**Определение 7.1.** *Согласно принципу Гюйгенса, каждую точку $ds = \xi, \nu$ плоскости, куда пришла волна,*

*можно рассматривать как источник вторичной сферической волны.*

*)[3])*

*Заметим, что мы тут связываем площадь с точкой так: точка - это $ds$.*

*внутри этой точки уже нет других точек. (кстати, это важный момент, можно потом подумать, почему так, если бы мы так не делали, то наша теория бы провалилась. мб это как раз и главное, что заставляет ее работать?)*

Если так на самом деле происходит, то каждая точка пространства мгновенно (со скоростью света) чувствует волну от любой другой точки.

Но пока мы думаем, что это просто математическое описание, а почему это работает - наверное потом поймем, но так на самом деле скорее всего не происходит.

(вообще, место подумать.

в этой теории нет ни ур Максвелла, ни интерференции, как тут не спутать ее с чем-то, что потом добавляем?)

С помощью принципа Гюйгенса для описания света теперь мы можем рассматривать произвольную поверхность, включающую в себя источники,

и нам более не нужно смотреть за ними.

Например, в случае, если свет приходит из далека на отверстие, мы можем заменить неизвестный нам источник поверхностью на ней и исследовать именно ее.

Также из этих соображений мы можем строить волновые фронты, нужно только убедиться в том,

что волны позади фронта интерферируют и гасят друг друга.

)как?

что там у [2] про это?

пока просто ок, пусть так есть.)

Пока что мы умеем строить фронты из сомнительных волн и ничего более, так что нужен какой-то еще метод работы, который позволит нам получать какие-то весомые результаты.

Его усовершенствование - принцип *Гюйгенса-Френеля.*

**Определение 7.2.** *Согласно принципу Гюйгенса-Френеля, вторичные волны Гюйгенса когерентны, а световое поле в результате их интерференции совпадает с полем реальных источников света.*

*)[2])*

*Предполагается, что амплитуда излучения вторичного источника пропорциональна амплитуде пришедшей волны (кст, как понимаю, если в этом пути она ослабнет, то речь именно об ослабшей). Фаза колебания также зависит от фазы пришедшей волны.*

*Самое главное: амплитуда колебания пропорциональна видимой из точки наблюдения площади переизлучающего элемента.*

На этом этапе нам уже удалось объяснить, почему при проникновении в определенные щели волна может в принципе поворачивать, так как без этого принципа, только с уравнением волны, мы ничего сказать не могли.

Здесь также объясняется, почему волна идет только вперед, а сзади она не идет.

(и кстати, как?

понятно, что волна не идет, но нужно посчитать и убедиться, что действительно - все время все гасится.)

(кстати, а как проверить, не противоречит ли это уравнения Максвелла?)

Заметим, что этот подход рассматривает волны вне зависимости от источника,

то есть источник может уже прекратить свое существование, а волна - будет идти и по этому принципу распространяться. Также следует понимать, что точка - это $d\eta d\xi$, где можно считать постоянную фазу и амплитуду.

То есть строго говоря, это никакая не точка, а площадка.

А также на этом этапе мы начинаем думать про интерференцию, так что будем работать только с комплексными амплитудами, как и в интерференции.

Начальные фазы у всех вторичных волн опускаем, так как она ни на что не влияет. Амплитуда вторичных волн также предполагается равной 1 для удобства.

(как переходить к зависимости от A начальной?)

**математический формализм принципа Гюйгенса-Френеля**

Соберем сказанное выше в модель.

Рассмотрим фронт сферической волны.

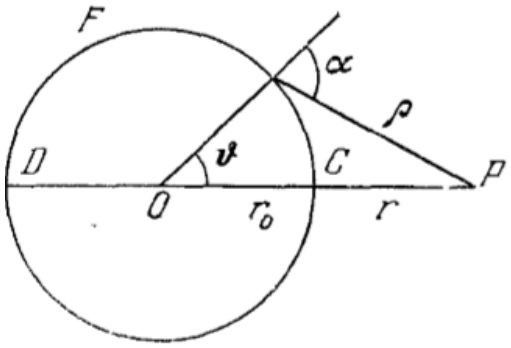



Figure 7.1: геометрия расчетов фронта волны, $\rho = R$

Точка:

$$d\eta d\xi$$

;

Сферическая волна от одной точки:

$$g(x,y,z) = \frac{1}{R}e^{ikR}d\eta d\xi;$$

Пропорциональность ей:

$$g(x,y,z) = C\frac{1}{R}e^{ikR}d\eta d\xi;$$

;

зависимость от угла:

$$g(x,y,z) = C\frac{1}{R}e^{ikR}\cos\alpha d\eta d\xi;$$

;

(заметим, что согласно [2] у Френеля была общая зависимость вида $K)\alpha)$, но он все равно по итоге тупо лесенкой падающих констант потом представлял, так что толку в конкретном $K)\alpha)$ нет,

и более того, часто мы будем искать в зоне, где cos = 1, так что сойдет.)

есть поверхность $M$, каждый ее элемент $dM$ излучает.

в общем, в этой науке мы отвечаем на вопрос качественно, так что нам нужно просто как-то обозначить, что в каком-то месте - свет, так что всегда работаем или с функцией распределения $f_0$ или с коэффициентом интенсивности на площадке $dS$ - $a$.

в данной точке напряжение - интеграл по всей поверхности $M$:

$$E = \int \frac{adF}{\rho} e^{f(\omega t - k\rho)}$$

мы хз, что там с $a$ - с амплитудой вторичных волн.

(говорят, что на этом уже этапе можно решать задачи на дифракцию.

прикольно, что казалось бы, конченная идея + математика = гениальная работающее решение)

дальше идет метод кирхгофа.

посмотрим, что мы имеем:

в общем, нужна $a$ из формулы выше, чему там равна пропорциональность интенсивности в точек $dF$?

в общем, тут идут предположения об интенсивности.

логично, что чем больше площадь - тем от нее больше будет долетать.

а также, что если она (ее нормаль) под углом к нашей точке направлена, то чем больше угол - тем она меньше.

ну и логично, что пропорциональна она интенсивности падающего излучения.

(как видим, проще всего клеить линейные модели, где просто пропорциональность, и тут такое работает.)

таким образом:

$$a = K(\alpha) e^{-ikr_0/r_0}$$

### (??) принцип Гюйгенса-Френеля в описании Кирхгофа

...

[2]

!!!!

хз, не знаю, и не актуально.

(? мб это связано с решением волнового уравнения?)

### строгое обоснование работоспособности

типа почему так можно сделать?

### аналогия с континуальным интегралом

с ним модно вывести и интеграл по траекториям, если понять нормально.

попробую потом записать.

### на самом деле так не происходит, как он утверждает (?)

вроде, это тупо метод поиска дифракции.

а то ни от куда он ведь не следует.

и это бред вообще, каждая точка пространства не может ведь излучать!

## 7.1.4 1.4. Fresnel Diffraction

### 1.4.1. Дифракционные задачи с осевой симметрией. Спираль Френеля

Дифракция Френеля на круглом отверстии. Будем интересоваться полем на оси отверстия в точке $P(0,0,z)$, находящейся на расстоянии $z$ от него (рис. 3.6).

Для начала воспользуемся френелевским приближением: из формулы (3.14) имеем

$$g = A_0 \frac{e^{ikz}}{i\lambda z} \iint_{\rho \leqslant r} e^{i\frac{k}{2z}\rho^2} d\xi d\eta \tag{3.15}$$

где $\rho = \sqrt{\xi^2 + \eta^2}$ - расстояние от площадки $d\xi d\eta$ до центра отверстия, $r$ радиус отверстия. Перейдём, с учётом осевой симметрии, к интегрированию по

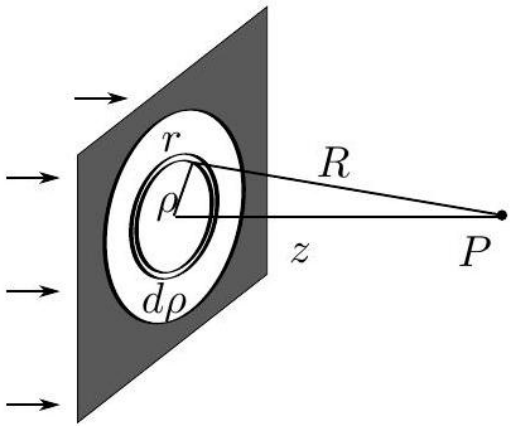



Рис. 3.6. К расчёту поля на оси круглого отверстия кольцам радиуса $\rho$ толщиной $d\rho$ (площадью $2\pi\rho d\rho$). Получаемый интеграл легко берётся, и мы находим комплексную амплитуду поля в точке $P$:

$$g = A_0 \frac{e^{ikz}}{i\lambda z} \int_0^r e^{i\frac{k}{2z}\rho^2} 2\pi\rho d\rho = A_0 e^{ikz} \left(1 - e^{-i\frac{k}{2z}r^2}\right) \tag{3.16}$$

Отсюда интенсивность в точке $P$ равна

$$I = |g|^2 = 2I_0 \left(1 - \cos\left(\frac{k}{2z}r^2\right)\right) \tag{3.17}$$

где $I_0 = A_0^2$ - интенсивность исходной плоской волны. Таким образом, интенсивность $I_P$ на оси отверстия зависит от радиуса отверстия и расстояния до него. Минимумы и максимумы $I_P$ достигаются при значениях радиуса отверстия:

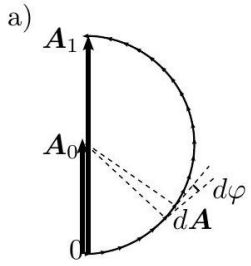
$$r_m = \sqrt{m\lambda z}, \quad m = 1, 2, 3, \ldots \tag{3.18}$$

где для нечётных $m$ имеем максимум $I_{\max} = 4I_0$, для чётных $m$ минимум $I_{\min} = 0$.

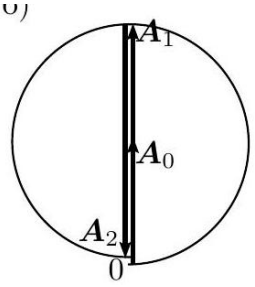
Формулой (3.17) можно пользоваться, если выполняются предположения, сделанные при выводе (3.14), то есть если в точной формулировке принципа Гюйгенса-Френеля (3.9) можно пренебречь отличием $\cos\alpha/R$ от $1/R_0$. Учтём влияние этого множителя на результаты расчётов на качественном уровне (строгое рассмотрение было бы не наглядным, поскольку возникающие при этом интегралы не берутся в элементарных функциях).

Воспользуемся методом векторных диаграмм. Согласно (3.16), нахождение колебания в точке наблюдения сводится к суммированию вкладов от колец радиуса $\rho$ толщиной $d\rho$

(площадью $dS = 2\pi\rho d\rho$). Изобразим колебание, созданное каждым таким кольцом, в виде вектора $d\boldsymbol{A}$, длина которого равна соответствующему вкладу в результирующую амплитуду: $dA = a(\rho)dS$, где функция $a(\rho) = A_0\frac{\cos\alpha}{\lambda R}$ учитывает плавное убывание вклада от колец с большим радиусом $\rho$ согласно формуле (3.9). Угол наклона векторов к горизонтали $\varphi$ соответствует фазе экспоненты в (3.15): $\varphi = \frac{k}{2z}\rho^2$. Каждый последующий вектор при этом повёрнут относительно предыдущего на один и тот же угол $d\varphi = \frac{k}{2z}2\rho d\rho = \frac{1}{\lambda z}dS$. Построенные в последовательную цепочку, эти векторы образуют медленно скручивающуюся (из-за убывания модуля $dA$) к центру спираль (рис. 3.7).




б)




В)

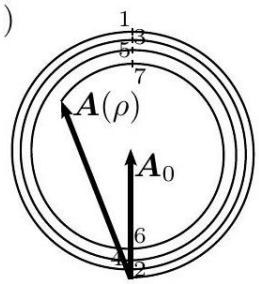



Рис. 3.7. Метод векторных диаграмм для дифракции Френеля на круглом отверстии

Значение $\rho$, соответствующее изменению фазы колебаний на $\pi$, называют радиусом первой зоны Френеля:

$$r_1 = \sqrt{\lambda z}$$

Будем считать, что при интегрировании в пределах первой зоны вклад от каждого кольца не зависит от $\rho$, т. е. $a(\rho) \approx \frac{A_0}{i\lambda z}$. При этом складываемые векторы образуют полуокружность радиуса $A_0$, а результатом сложения цепочки будет колебание, создаваемое всей первой зоной Френеля, - вектор $\boldsymbol{A}_1 = 2\boldsymbol{A}_0$ (см. рис. 3.7а). Этот результат, естественно, соответствует расчёту по формуле (3.17).

Интегрированию по интервалу $r_1 < \rho < r_2$, где $r_2 = \sqrt{2\lambda z}$, на векторной диаграмме соответствует ещё одна полуокружность, на которой фаза векторов меняется от $\pi$ до $2\pi$. Если учесть плавное убывание $a(\rho)$, радиус этой полуокружности немного меньше радиуса предыдущей. Интервал $r_1 < \rho < r_2$ называют второй зоной Френеля, вклад от

неё - вектор $\boldsymbol{A}_2$ - будет по модулю чуть меньше, чем $2A_0$ (см. рис. 3.7б). Суммарная амплитуда колебаний от первых двух ($\boldsymbol{A}_1 + \boldsymbol{A}_2$) зон хоть и мала, но отлична от нуля (в отличие от расчёта по (3.17)).

Дальнейшее интегрирование по всем оставшимся зонам даст на векторной диаграмме спираль, медленно скручивающуюся к центру (т.е. к концу вектора $\boldsymbol{A}_0$, поскольку при отсутствии отверстия амплитуда в точке $P$, очевидно, должна быть равна амплитуде волны, падающей на отверстие), называемую спиралью Френеля (см. рис. 3.7в). Размеры зон Френеля определяются соотношением (3.18). Колебание, создаваемое каждой следующей зоной Френеля, находится в противофазе ($\Delta\varphi = \pi$) с предыдущим.

Если открыты зоны с 1 -й по $n$-ю, результирующая амплитуда представляет собой сумму

$$A = \left| \sum_{i=1}^{n} \boldsymbol{A}_i \right| = A_1 - A_2 + A_3 - \dots$$

где $A_i \approx A_0$ и медленно убывают с ростом $i$. Ясно, что если радиус отверстия соответствует нечётному количеству открытых зон Френеля, то в точке $P$ наблюдается максимум амплитуды, если чётному минимум.

Зонная пластинка Френеля. Поставим на пути плоской волны с амплитудой $A_0$ пластинку, в которой перекрыты все чётные зоны Френеля (для некоторой фиксированной точки наблюдения $P$ на расстоянии $z$) и оставлены прозрачными нечётные зоны. Амплитуда в точке $P$ есть сумма вкладов всех нечётных зон: $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_3 + \boldsymbol{A}_5 + \dots$. Поскольку все
векторы от нечётных зон сонаправлены (т. е. имеют одну фазу), для получения суммарного колебания нужно сложить их модули, из чего получим интенсивность в точке наблюдения $I = |A_1 + A_3 + A_5 + \dots|^2$. Видно, что пластинка обладает фокусирующими свойствами, существенно увеличивая интенсивность света в точке $P$.

Если размер пластинки (её радиус $r$) существенно меньше расстояния $z$ до точки наблюдения, то можно считать, что вклады от всех (нечетных) зон примерно одинаковы и равны $A_1$. Число зон Френеля, укладывающихся на пластинке радиуса $r$, равно $m = r^2/(\lambda z)$ (из них половина зон - нечетные). Положив $m = 1$, можем выразить фокусное расстояние пластинки [4] : $z_0 = r_{\text{П}}^2/\lambda$.

Тонкая линза. Попробуем сконструировать пластинку таким образом, чтобы вклады $d\boldsymbol{A}$ в амплитуду в точке наблюдения $P$ от всех колец $dS$, по которым производится интегрирование, имели одну и ту же фазу. Для этого возьмём прозрачную пластинку с некоторым показателем преломления и подберём её толщину таким образом, чтобы за счёт прохождения пластинки волна набирала фазу $\varphi_{\text{п}} = -\frac{k}{2z}\rho^2$. Это значит, что функция пропускания должна быть равна

$$t(\rho) = e^{-i\frac{k}{2z}\rho^2}$$

Сравнивая это выражение с (3.6), приходим к выводу, что пластинка, делающая синфазными вклады $d\boldsymbol{A}$ от каждого её участка, есть тонкая линза с фокусным расстоянием $f = z$.

Пятно Пуассона. Рассмотрим дифракцию плоской волны на препятствии в виде круглого диска или шарика.

Относительно точки наблюдения $P$ на оси системы препятствие закрывает некоторое количество (не обязательно целое) зон Френеля. Вклад от них изображается вектором $\boldsymbol{A}'$, амплитуда $\boldsymbol{A}$ в точке $P$ равна вкладу от всех оставшихся переферийных зон (рис. 3.8). С учётом того, что в сумме все зоны должны давать $\boldsymbol{A}_0$, запишем

$$\boldsymbol{A} = \boldsymbol{A}_0 - \boldsymbol{A}'$$

Из рис. 3.8 ясно, что, если пренебречь скручиванием спирали Френеля к её центру, $\boldsymbol{A}$ по модулю равен $A \approx A_0$ независимо от того, сколько зон Френеля перекрывает препятствие (лишь бы их число не было слишком большим).

Таким образом, амплитуда в центре геометрической тени малого препятствия (шарика или диска) остаётся почти такой же, как если бы это препятствие отсутствовало.

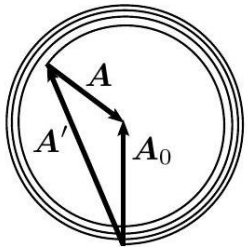



Рис. 3.8. Закрыто $7\frac{1}{3}$ зон Френеля

### 1.4.2. Дифракция Френеля на щели

Рассмотрим дифракцию плоской волны амплитуды $A_0$ на непрозрачном экране, отверстие в котором имеет вид длинной щели, вытянутой вдоль оси $\eta$ (рис. 3.9). Верхний край щели совпадает с прямой $\xi = b_1$, а нижний - с $\xi = b_2$ (ширина щели $b = b_1 - b_2$). Нас интересует световое колебание в точке $P$, расположенной на расстоянии $z$ за щелью на оси $z$. Интегрирование (3.14) по $\eta$ в бесконечных пределах даёт константу, не представляющую интерес. Тогда с точностью до несущественного постоянного множителя получаем

$$g = A_0 \int_{b_1}^{b_2} e^{\frac{ik}{2z}\xi^2} d\xi \tag{3.19}$$

Выражение $e^{\frac{ik}{2z}\xi^2} d\xi$ представляет собой вклад в колебание в точке $P$ узкой полоски толщиной $d\xi$ вторичных источников, параллельной оси $\eta$, а интеграл есть сумма колебаний от всех полосок от нижнего края щели $\xi = b_2$ до верхнего $\xi = b_1$.

Для расчёта светового поля воспользуемся методом векторных диаграмм. Разбив щель на узкие полоски $d\xi$, параллельные краям щели, изобразим колебание, созданное полоской в точке наблюдения $P$, в виде вектора, длина которого равна $d\xi$, а угол наклона (относительно горизонтальной оси) $\varphi = \frac{k}{2z}\xi^2$. В частности, вклад полоски вторичных источников, лежащих на оси $\eta(\xi = 0)$, изображается горизонтальным вектором.

Разность фаз $d\varphi$ колебаний от полосок на расстояниях $\xi$ и $\xi + d\xi$ зависит от $\xi$:

$$d\varphi = \frac{k}{z}\xi d\xi$$

[0]4 Заметим, что в отличие от линзы, у зонной пластинки имеется не один, а множество фокусов. Они находятся в точках $z_n = z_0/(2n+1)$, где $n$ - целое число. Например, при $n = 1$ имеем точку $z_1 = z_0/3$, в которой будет максимум интенсивности, поскольку относительно неё пластинка оставляет открытыми первые три зоны Френеля и перекрывает 4,5 и 6 и т. д.

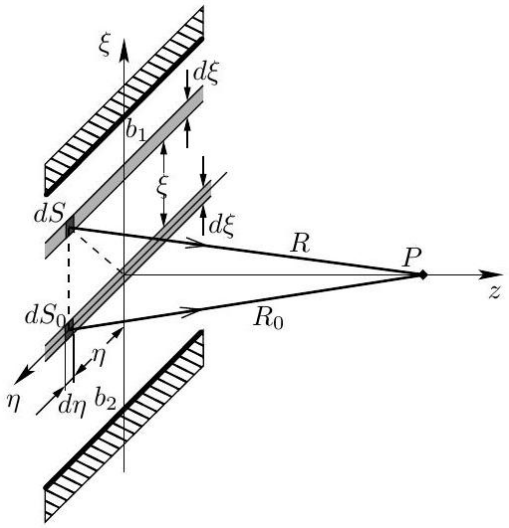



Рис. 3.9. К расчёту дифракции на щели

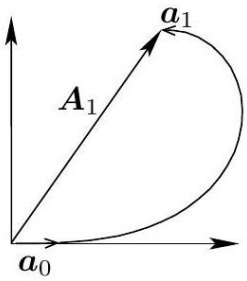



Рис. 3.10. Первая зона Шустера

$\xi_1 = \sqrt{2\pi z/k} = \sqrt{\lambda z}$ вектором $\boldsymbol{A}_1$, проведённым из начала первого вектора $\boldsymbol{a}_0$ в конец последнего вектора $\boldsymbol{a}_1$.

Будем продолжать наше построение, пристраивая на векторной диаграмме вклады последовательно расположенных полосок, все далее отстоящих от оси $\eta$. Мы получим ещё один полувиток, причём вектор $\boldsymbol{a}_2$ изображает вклад полоски, отстоящей от точки наблюдения на расстояние $z+\lambda$ (оно больше, чем расстояние от оси $\eta$ до точки наблюдения, на длину волны $\lambda$), поэтому векторы $\boldsymbol{a}_0$ и $\boldsymbol{a}_2$ коллинеарны - разность фаз колебаний $\Delta\varphi = 2\pi$ (рис. 3.11): $\frac{k}{2z}\xi_2^2 = 2\pi$, откуда $\xi_2 = \sqrt{2\lambda z}$. Вектор $\boldsymbol{A}_2$ изображает суммарный вклад полосы $\xi_1 < \xi < \xi_2$.

В данном случае вместо кольцевых зон Френеля мы имеем зоны в виде полос (их называют зонами Шустера). Вклад от каждой следующей зоны Шустера в колебание в точке наблюдения находится в противофазе (разность фаз $\pi$) с вкладом от предыдущей зоны.

Продолжая дальше построение векторной диаграммы, придём к спирали Корню, изображённой на рис. 3.12. Спираль, быстро скручиваясь (благодаря квадратичной зависимости фазы колебаний $\varphi = \frac{k}{2z}\xi^2$ от координаты $\xi$), при-

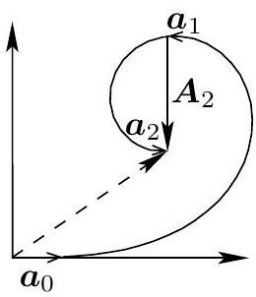


Рис. 3.11. Две зоны Шустера ближается к точке $S_1$, называемой фокусом спирали. Каждый последующий полувиток спирали изображает вклад в суммарное колебание последовательно расположенных зон Шустера, причём $m$-я зона Шустера - это полоса, внешний край которой отстоит от точки наблюдения на расстояние $z + m\lambda/2$ и на расстояние

$$\xi_m = \sqrt{m\lambda z}, \quad m = 1, 2, 3, \ldots \tag{3.20}$$

от оси $\eta$. Нижняя половина спирали Корню (от нуля до точки $S_2$) описывает вклад вторичных источников, расположенных ниже оси $\eta$ ($\xi < 0$). С ростом ширины щели ($b_1 \to \infty, b_2 \to -\infty$ на рис. 3.9) работают все более мелкие витки спирали Корню, которые все ближе прилегают к фокусам $S_1$ и $S_2$. Вектор $\boldsymbol{A}_0$, соединяющий фокусы спирали, определяет амплитуду колебаний в точке наблюдения при отсутствии препятствия.

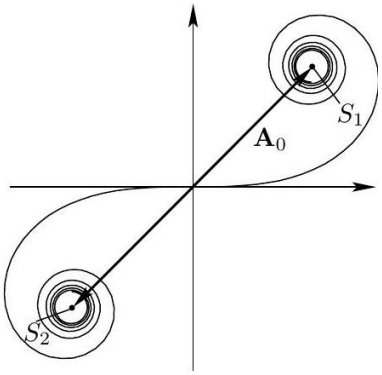



Рис. 3.12. Спираль Корню

Очевидно, амплитуда колебаний в точке $P$ близка к максимальной, если открыты две зоны Шустера (т. е. «работает» один полувиток верх ней половины спирали и один полувиток нижней половины, рис. 3.13). Амплитуда колебаний (длина вектора $\boldsymbol{A}$) при этом больше, чем амплитуда колебаний $A_0$ при отсутствии препятствия. Соответствующая ширина щели равна при этом $b = 2\sqrt{\lambda z}$. Если размер щели увеличить таким образом, что $b/2 = \sqrt{2\lambda z}$ (так, что на верхней и нижней полуспирали начинают «работать» полувитки вторых зон Шустера, рис. 3.14), то амплитуда колебаний в точке наблюдения уменьшается: длина вектора $\boldsymbol{A}$ на рис. 3.14 меньше, чем $A_0$.

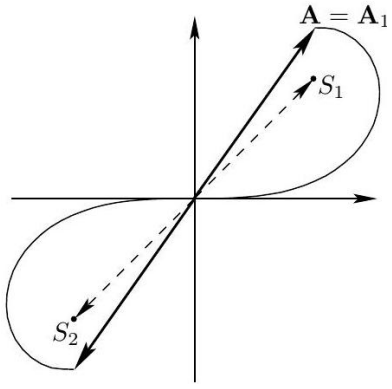



Рис. 3.13. Открыты две первые зоны Шустера

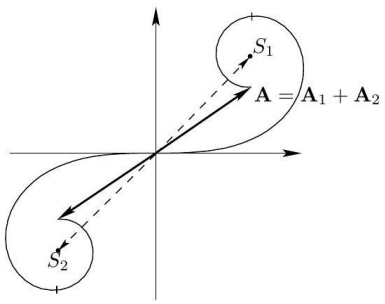



Рис. 3.14. Открыты первые и вторые зоны Шустера

### 1.4.3. Условия наблюдения дифракции Френеля

Выразим количество открытых зон Френеля (или Шустера) в рассмотренных выше задачах дифракции на отверстии и щели. Согласно формулам (3.18) и (3.20)

$$m = \frac{b^2}{\lambda z} \equiv \frac{1}{p^2}$$

где $b$ - характерный размер препятствия (радиус отверстия или полуширина щели соответственно), а $p$ - волновой параметр, определённый согласно формуле (3.1).

Из проведённого рассмотрения следует, что дифракция Френеля имеет место при небольшом количестве зон Френеля (или Шустера), т.е. при $m \sim 1$, что в терминах волнового параметра соответствует условию (см. п. 1.1)

$$p \sim 1$$

Большое количество зон Френеля ($m \gg 1, p \ll 1$) соответствует предельному переходу к геометрической оптике, а в обратном пределе, когда открыта лишь небольшая доля первой зоны ($m \ll 1, p \gg 1$), имеет
место так называемая дифракция Фраунгофера, к рассмотрению которой мы переходим.

**выводы из дифракции????**

так как геометрическая оптика работает на расстояниях от щели $z << b^2/\lambda$,

но что-то у меня получается, что прямо на самом деле дофига она будет работать перед тем, как начнет затухать.

тут место подумать про нее?

и про распространение света вообще?

чего затухает, чего то и се происходит?

**теория дифракции френеля**

(лл2)

Если источник света и точка $P$, в которой мы ищем интенсивность света, находятся на конечном расстоянии от экрана, то длЯ определения интенсивности в точке $P$ играет роль лишь небольшой участок волновой поверхности, по которой происходит интегрирование В (59.2), - участок, лежащий вблизи прямой, соединяющей источник с точкой $P$. Действительно, поскольку отклонения от геометрической оптики слабы, то интенсивность света, приходящего в $P$ из различных точек волновой поверхности, очень быстро падает по мере удалениЯ от указанной прямой.

Дифракционные явления, в которых играют роль лиШь небольшие участки волновой поверхности, носят название Дифракиии френеля.

Рассмотрим дифракцию Френеля от какого-нибудь экрана.

Благодаря указанному свойству при этом играет роль (при заданной точке $P$) только небольшой участок края экрана.

Но на достаточно малых участках край экрана можно всегДа считать прямолинейным.

Ниже под краем экрана будет поэтому подразумеваться именно такой небольшой прямолинейный участок.

Выберем в качестве плоскости $xy$ Плоскость, проходящую через источник света $Q$ (рис. 11) и через линию края экрана.

Перпендикулярную к ней плоскость $xz$ выбираем так, чтобы она прошла через точку $Q$ и точку наблюдения $P$, в которой мы ищем значение интенсивности света.

Наконец, начало координат О выбираем на линии края экрана, после чего положение всех трех осей вполне определено.

Расстояние от источника света $Q$ до начала координат пусть будет $D_q$.$x$ -координату точки наблюдения $P$ обозначим как $D_p$, а ее z-координату, т.е.

расстояние до плоскости $xy$, - как d.

Согласно геометрической оптике свет мог бы попасть только в точки, лежащие над плоскостью $xy$; область же под плоскостью $xy$ есть область, где согласно геометрической оптике была бы тень (область геометрической тени).

Мы определим теперь распределение интенсивности света за экраном вблизи граниЦы геометрической тени, т.е.

при малых (по сравненИЮ с $D_p$ и $D_q$) значениях d.

Отрицательное $d$ означает, что точка $P$ находится в области геометрической тени.

В качестве поверхности интегрирования в (59.2) выберем полуплоскость, проходящую через линию края экрана перпендикулярно К Плоскости ху.

Координаты $x$ и $y$ точек этой поверхности связаны друг с другом соотношением $x = y\,\mathrm{tg}\,\alpha$($\alpha$ - угол между линией края экрана и осьЮ $y$), а координата $z$ положительна.

Поле волны, исходящей из источника $Q$, на расстоянии $R_q$ от него пропорционально множителю $\exp(ikR_q)$. Поэтому поле $u$ на поверхности интегрирования пропорционально

$$u \sim \exp\left[ik\sqrt{y^2+z^2+(D_q+y\,\mathrm{tg}\,\alpha)^2}\right]$$

В интеграле (59.2) длЯ $R$ надо теперь подставить

$$R = \sqrt{y^2+(z-d)^2+(D_p-y\,\mathrm{tg}\,\alpha)^2}$$

В подынтегральном выражении медленно изменяющиеся множители не существенны по сравнению с экспоненЦиальным.

Поэтому мы можем считать $1/R$ постоянным, а вместо $df_n$ пИсать $dydz$. М ы находим тогда, что поле в точке $P$

$$u_P \sim \int\limits_{-\infty}^{+\infty}\int\limits_{0}^{\infty} \exp\left[ik\sqrt{(D_q+y\,\mathrm{tg}\,\alpha)^2+y^2+z^2}+ \right.$$
$$\left. +\sqrt{(D_p-y\,\mathrm{tg}\,\alpha)^2+y^2+(z-d)^2}\right] dydz$$

Как мы уже говорили, в точку $P$ попадает свет главным образом из точек плоскости интегрирования, близких к О.

Поэтому в интеграле (60.1) играют роль малые (по сравнению с $D_q$ и $D_p$) значения $y$ и $z$.

Мы можем написать:::

$$\sqrt{\left(D_q + y\,\mathrm{tg}\,\alpha\right)^2 + y^2 + z^2} \approx D_q + \frac{y^2+z^2}{2D_q} + y\,\mathrm{tg}\,\alpha$$
$$\sqrt{\left(D_p - y\,\mathrm{tg}\,\alpha\right)^2 + y^2 + (z-d)^2} \approx D_p + \frac{y^2+(z-d)^2}{2D_p} - y\,\mathrm{tg}\,\alpha.$$

Подставим это в (60.1).

Поскольку нас интересует поле только как функция от расстояния $d$, то постоянный множитель $\exp\left[ik\left(D_p + D_q\right)\right.$.

опускаем; интеграл по $dy$ тоже дает выражение, не содержащее $d$, которое мы также опустим.

Мы находим тогда:

$$u_P \sim \int_0^\infty \exp\left[ik\left(\frac{1}{2D_q}z^2 + \frac{1}{2D_p}(z-d)^2\right)\right] dz$$

Это выражение можно написать и в таком виде:

$$u_P \sim \exp\left[ik\frac{d^2}{2\left(D_p+D_q\right)}\right]\int_0^\infty \exp\left[ik\frac{\frac{1}{2}\left[\left(\frac{1}{D_p}+\frac{1}{D_q}\right)z - \frac{d}{D_p}\right]^2}{\frac{1}{D_p}+\frac{1}{D_q}}\right] dz$$

Интенсивность света определяется квадратом поля, т.е.

квадратом модуля $|u_P|^2$. Поэтому для нахождения интенсивности стоящий перед интегралом множитель не существен, так как при умножении на сопряженное выражение он дает единицу.

Очевидной подстановкой интеграл приводится к виду

$$u_P \sim \int_{-w}^{\infty} e^{i\eta^2} d\eta$$

где

$$w = d\sqrt{\frac{kD_q}{2D_p\left(D_q+D_p\right)}}$$

Таким образом, интенсивность $I$ в точке $P$ равна

$$I = \frac{I_0}{2}\left|\sqrt{\frac{2}{\pi}}\int_{-w}^{\infty} e^{i\eta^2}d\eta\right|^2 = \frac{I_0}{2}\left[\left(C\left(w^2\right)+\frac{1}{2}\right)^2 + \left(S\left(w^2\right)+\frac{1}{2}\right)^2\right]$$

где

$$C(z) = \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{z}} \cos\eta^2 d\eta, \quad S(z) = \sqrt{\frac{2}{\pi}}\int_0^{\sqrt{z}} \sin\eta^2 d\eta$$

- так называемые интегралы Френеля.

Формула (60.5) решает поставленную задачу, определяя интенсивность света как функцию от $d$

$I_0$ есть интенсивность в освещенной области в точках, достаточно удаленных от края тени, т.е.

при $w \gg 1$ (в пределе $w \to \infty$ имеем

$$C(\infty) = S(\infty) = \frac{1}{2}\Big)$$

Области геометрической тени соответствуют отрицательные $w$.

ЛегКо выяснить асимптотический вид фунКции $I(w)$ при большиХ по абсолютной величине отрицательных значениях $w$. ДлЯ этого поступим следующИм образом.

Интегрируя по частям, имеем:

$$\int\limits_{|w|}^{\infty} e^{i\eta^2} d\eta = -\frac{1}{2i|w|} e^{iw^2} + \frac{1}{2i} \int\limits_{|w|}^{\infty} e^{i\eta^2} \frac{d\eta}{\eta^2}$$

Интегрируя в правой части равенства еще раз по частям и продолжая этот процесс, получим ряд по степеням $1/|w|$:

$$\int\limits_{|w|}^{\infty} e^{i\eta^2} d\eta = e^{iw^2} \left( -\frac{1}{2i|w|} + \frac{1}{4|w|^3} - \dots \right)$$

Хотя бесконечный ряд такого вида и не является сходящимся, но ввиду того, что при больших $|w|$ величина его последовательных членов быстро падает, уже первый его член дает хорошее представление стоящей слева функциИ при достаточно большиХ $|w|$ (ряды такого рода называются асимптотическими).

Таким образом, длЯ интенсивности $I(w)$ (60.5) получим следующую асимптотическую формулу, пригодную для большиХ отрицательных значений $w$:

$$I = \frac{I_0}{4\pi w^2}$$

Мы видим, что в области геометрической тени, вдали от ее края интенсивность стремится к нулю обратно пропорЦионально квадрату расстояния от края тенИ.

Рассмотрим теперь положительные значения $w$, т.е.

область выше Плоскости $xy$. Пишем:

$$\int\limits_{-w}^{\infty} e^{i\eta^2} d\eta = \int\limits_{-\infty}^{+\infty} e^{i\eta^2} d\eta - \int\limits_{-\infty}^{-w} e^{i\eta^2} d\eta = (1+i)\sqrt{\frac{\pi}{2}} - \int\limits_{w}^{\infty} e^{i\eta^2} d\eta$$

При достаточно больших $w$ можно воспользоваться асимПтотическим представлением стоящего в правой части равенства интеграла, и мы будем иметь:

$$\int\limits_{-w}^{\infty} e^{i\eta^2} d\eta \approx (1+i)\sqrt{\frac{\pi}{2}} + \frac{1}{2iw} e^{iw^2}$$

Подставляя это выражение в (60.5), получим:

$$I = I_0 \left[ 1 + \sqrt{\frac{1}{\pi}} \frac{\sin\left(w^2 - \frac{\pi}{4}\right)}{w} \right]$$

Таким образом, в освещенной области, вдали от края тени, интенсивность имеет неограниченный ряд максимумов и минимумов, так что отношение $I/I_0$ колеблется в обе стороны от единицы.

Размах этих колебаний уменьшается с ростом $w$ обратно пропорционально расстоянию от края геометрической тени, а места максимумов и минимумов постепенно сближаются друг с другом.

При небольших $w$ функЦиЯ $I(w)$ имеет качественно тот же характер (рис.
12).

В области геометрической тени интенсивность спадает монОтонно при удалении от гранИЦы тени (на самой этой гранИЦе $I/I_0 = \frac{1}{4}$). При положительных $w$ интенсивность имеет чередуюЩиеся максимумы и минимумы.

В первом, наибольшем из максимумов, $I/I_0 = 1,37$

## 7.1.5 Fresnel Diffraction: Examples

### дифракция на круглом отверстии

(тут речь ведь о плоских волнах?)

#### итоги модели

по сути выше уже задача решена, нужно только придать ясный формуле

$$g(0,0) = A_0 \frac{e^{ikz}}{i\lambda z} \iint e^{i\frac{k}{2z}(\xi^2+\eta^2)} d\xi d\eta$$

(тут нужно порешать на самом деле задачу эту, типа к вопросу, как нахрен мы интегрируем.)

переходим в обще к одному параметру $\rho$, через который считаем окружностями площадь.

в итоге получаем:

$$g = A_0 \frac{e^{ikz}}{i\lambda z} \int_0^r e^{i\frac{k}{2z}\rho^2} 2\pi\rho d\rho = A_0 e^{ikz} \left(1 - e^{-i\frac{k}{2z}r^2}\right)$$

Отсюда интенсивность в точке P равна

$$I = |g|^2 = 2I_0 \left(1 - \cos\left(\frac{k}{2z}r^2\right)\right)$$

ну и дальше идет речь про векторные диаграммы и про всякие нюансы, позже доделаю.

#### модификации

крч, тут разные задачи, вот их и прорешаю.

### дифракция на диске

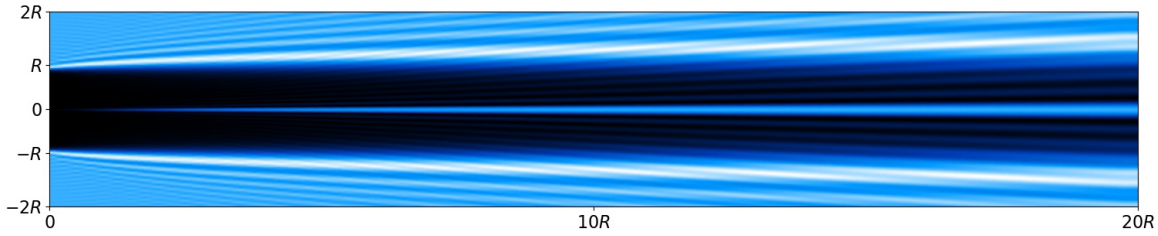
условия его наблюдения????????

### пятно Пуассона-Араго

При некоторых условиях, если поставить диск между источником света и поверхностью, на которой можно смотреть тень этого диска, то можно увидеть в центре тени светлое пятно - пятно Пуассона.

Опишем это явление.

Таким образом, пятно Пуассона видно при $\frac{d^2}{l\lambda} \geq 1$
и определяется




### Дифракция на щели

(начало задачи должно содержать слова, что мы именно Френеля решаем, так как я удалил из главы это слово, чтобы было удобнее)

тоже интересно

$$g = A_0 \int_{b_1}^{b_2} e^{\frac{ik}{2z}} \xi^2 d\xi$$

$$d\varphi = \frac{k}{z} \xi d\xi$$

!!!!!!! и дальше прямо очень важная тема про иллюстрации и спираль Корню, мб ее даже в вольфраме замоделирую и с матаном свяжу.

### На периодических решетках!!!

(а бывают не периодические?)

есть в билетах

ампл и фазовая!!!

(начало задачи должно содержать слова, что мы именно Френеля решаем, так как я удалил из главы это слово, чтобы было удобнее)

здесь как раз и находится эффект саморепродукции.

я хз, что это такое. .

### на краю экрана

(здесь теория про зоны Шустера и красивые спирали, но в билетах этого нет, так что это все потом)

## 7.1.6 1.5. Fraunhofer Diffraction

### 1.5.1. Дифракция Фраунгофера и преобразование Фурье

Рассмотрим дифракцию на отверстии в плоском экране (см. рис. 3.5), находящемся в плоскости $z = 0$. Пусть точка наблюдения имеет координаты $P(x, y, z)$. Расстояние $R$ от площадки $ds$ в точке экрана $(\xi, \eta)$ до точки $P$ равно

$$R = \sqrt{z^2 + (x-\xi)^2 + (y-\eta)^2} = \sqrt{R_0 - (2x\xi + 2y\eta) + (\xi^2 + \eta^2)}$$

где $R_0 = \sqrt{z^2 + x^2 + y^2}$ - расстояние начала координат $O$ до точки наблюдения. Полагая поправку к $R_0$ под знаком радикала малой, получим приближение

$$R \approx R_0 - \frac{x\xi + y\eta}{R_0} + \frac{\xi^2 + \eta^2}{2R_0} \tag{3.21}$$

Используем это выражение для оценки величины $kR$ (показателя экспоненты в (3.10)).

Интересующая нас область изменения переменных $(\xi, \eta)$ ограничена размерами отверстия в непрозрачном экране (в этой области отлично от нуля граничное поле $f_0(\xi, \eta)$). Пусть максимальный размер отверстия равен $b$, так что $\xi^2 + \eta^2 \leqslant b^2$. Если точка наблюдения удалена от препятствия настолько, что выполняется условие

$$\frac{b^2}{R_0} \ll \lambda \tag{3.22}$$

то последним слагаемым в (3.21) можно пренебречь, и тогда

$$R \approx R_0 - \frac{x\xi}{R_0} - \frac{y\eta}{R_0}$$

В этом приближении принцип Гюйгенса-Френеля (3.10) запишется в виде

$$g(x, y) = \frac{e^{ikR_0}}{i\lambda R_0} \iint f_0(\xi, \eta) e^{-i\left(\frac{kx}{R_0}\xi + \frac{ky}{R_0}\eta\right)} d\xi d\eta$$

Условие (3.22), очевидно, выполняется, если $b^2/(\lambda z) \ll 1$. Эту область мы назвали ранее областью дифракции Фраунгофера (или дальней волновой зоной). Введя переменные

$$u = \frac{kx}{R_0}, \quad v = \frac{ky}{R_0}$$

получим

$$g(u, v) = \frac{e^{ikR_0}}{i\lambda R_0} \iint f_0(\xi, \eta) e^{-i(u\xi + v\eta)} d\xi d\eta \tag{3.23}$$

Эта формула показывает, что поле в плоскости наблюдения $g(u, v)$ есть (с точностью до постоянного множителя) двумерное преобразование Ф ръе граничного поля $f_0(\xi, \eta)$.

Одномерный случай. Формула (3.23) становится особенно наглядной, если граничное поле $f_0(\xi, \eta)$ описывается функцией одной переменной $\xi$ : $f_0(\xi, \eta) \equiv f_0(\xi)$. Тогда с точностью до несущественного постоянного множителя

$$g(u) \propto \int_{-\infty}^{+\infty} f_0(\xi) e^{-iu\xi} d\xi \tag{3.24}$$

Таким образом, в данном случае картина дифракции Фраунгофера $g(u)$ представляет собой преобразование Фурье граничного поля $f_0(\xi)$ (т. е. поля в плоскости, примыкающей к препятствию).

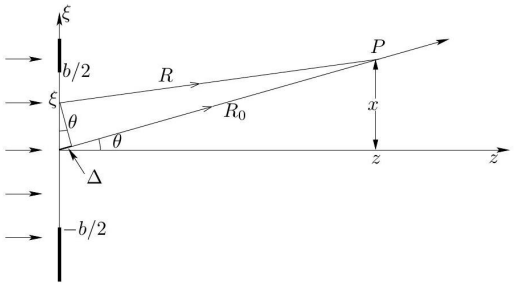


Рис. 3.15. Одномерный случай дифракции Фраунгофера

Определим разность хода волн $\Delta$, приходящих к удалённой точке наблюдения от двух вторичных источников, один из которых находится в точке с координатой $\xi$, а второй - в точке $\xi = 0$ (рис. 3.15). Удалённость точки $P$ позволяет считать направления волн, идущих из этих точек практически параллельными (рис. 3.16), следовательно, разность их хода равна

$$\Delta = \xi \sin\theta$$

Соответственно разность фаз колебаний равна $\varphi = -k\xi \sin\theta$, где $\theta$ - направление на удалённую точку наблюдения, имеющую координату $x$:

$$k \sin\theta = \frac{kx}{R_0} = u$$

Таким образом формула (3.24) представляет собой сумму колебаний, приходящих к точке наблюдения, положение которой на экране определяется направлением $\sin\theta = \frac{x}{R_0}$ :

$$g(u) \propto \int_{-\infty}^{+\infty} f_0(\xi) e^{-ik\sin\theta\cdot\xi} d\xi$$

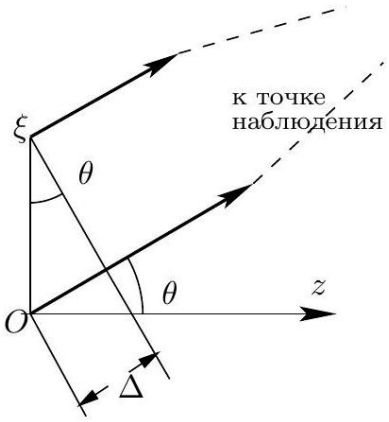



Рис. 3.16. Разность хода

Итак, формулы (3.24) или (3.25) означают, что волны, приходящие к точке наблюдения от всех вторичных источников, заполняющих отверстие, можно считать параллельными, если эта точка удалена в зону Фраунгофера.

Формула (3.25) - полный аналог соотношения

$$C(\omega) = \int f_0(t) e^{-i\omega t} dt$$

где $C(\omega)$ - спектр (преобразование Фурье) процесса $f_0(t)$. Эта аналогия позволяет назвать величину $u = k\sin\theta$ пространственной частотой аналог частоты $\omega$ в спектре колебательного процесса $f_0(t)$.

Ещё раз подчеркнём, что положение точки наблюдения $P$ при условии $b \ll \sqrt{\lambda z}$ определяется углом $\theta$. Если точка наблюдения смещается вдоль фиксированного направления, приближаясь или удаляясь, так что угол $\theta$ остаётся неизменным: $\sin\theta = x/R_0 = \text{const}$ (рис. 3.15) (и при этом остаётся справедливо неравенство $b \ll \sqrt{\lambda z}$), то значение интеграла остаётся неизменным. В этом случае картину дифракции можно характеризовать распределением интенсивности света (потока энергии) по углам $I(\theta)$:

$$I(\theta) \propto \left| \int f_0(\xi) e^{-ik \sin\theta \cdot \xi} d\xi \right|^2 \tag{3.26}$$

Распределение (3.26), изображённое в полярных координатах, называют диаграммой направленности дифракционной картины.

Отсюда следует важное замечание. Для наблюдения дифракционной картины Фраунгофера нет необходимости удаляться от препятствия на большое расстояние (для выполнения условия (3.22)). Достаточно установить за препятствием линзу (причём неважно на каком расстоянии - можно и непосредственно вблизи него) и наблюдать картину в фокальной плоскости линзы. Действительно, волны, идущие от всех вторичных источников в направлении $\theta$, фокусируются линзой с фокусным расстоянием $f$ в точке фокальной плоскости $x = f \operatorname{tg}\theta \approx f \sin\theta$, т. е. каждому направлению соответствует своя точка фокальной плоскости и соответствующее значение $I(\theta)$. Таким образом, картина в фокальной плоскости подобна картине на большом расстоянии $z \gg b^2/\lambda$, отличаясь от неё лишь масштабом (уменьшена в отношении $z/f$ раз).

### 1.5.2. Дифракция Фраунгофера на щели

Пусть щель шириной $b$ освещается слева плоской нормально падающей волной. Граничное поле $f_0(x)$, возникающее в плоскости $z = 0$, примыкающей к непрозрачному экрану со щелью справа, имеет вид, изображённый на рис. 3.17. Согласно (3.25), для нахождения картины Фраунгоферовой дифракции, необходимо найти преобразование Фурье этой функции (эта задача полностью аналогична нахождению спектра прямоугольного импульса - роль длительности импульса здесь играет ширина щели $b$, а роль частоты - пространственная частота $u = k \sin\theta$). Мы получаем

$$g(\theta) \propto \int_{-b/2}^{+b/2} e^{ikx \sin\theta} dx \propto \frac{\sin\left(\frac{kb}{2}\sin\theta\right)}{\frac{kb}{2}\sin\theta} \tag{3.27}$$

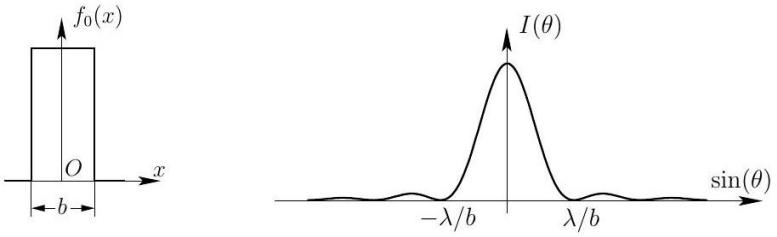



Рис. 3.17.

Поле на

Рис. 3.18. Угловое распределение щели интенсивности при дифракции на щели

Распределение интенсивности $I(\theta) = |g(\theta)|^2$ показано на рис. 3.18. Ближайшие к направлению $\theta = 0$ направления, в которых $I(\theta)$ обраща ется в нуль, определяются условием $\frac{kb}{2}\sin\theta = \pm\pi$, откуда находим

$$\sin\theta = \pm\frac{\lambda}{b} \tag{3.28}$$

Как показывает анализ функции $|g(\theta)|^2$, в угловом конусе

$$|\sin\theta| \leqslant \frac{\lambda}{b}$$

сосредоточена подавляющая величина потока энергии. Этот угловой конус (интервал углов от $-\frac{\lambda}{b}$ до $\frac{\lambda}{b}$) называют главным максимумом дифракционной картины. С помощью (3.27) легко найти, что максимальная интенсивность в ближайшем боковом максимуме, в направлении $\sin\theta = \frac{3}{2}\frac{\lambda}{b}$, составляет примерно 0,05 от интенсивности в направлении $\sin\theta = 0$). Мы видим, что сужение щели (уменьшение $b$) приводит к уширению угла $\theta$: свет расходится от щели во всё более широком интервале углов. Можно сказать и по-другому: с уменьшением ширины щели уширяется пространственный спектр - спектр плоских волн, бегущих от щели.

### 1.5.3. Дифракция Фраунгофера на двух щелях

Рядом со щелью, дифракцию на которой мы рассмотрели выше (эта картина дифракции описывается функцией $g(\theta)$ в формуле (3.27)), расположим параллельно ещё одну щель на расстоянии $d$ от первой (рис. 3.19).

Расстояние от второй щели до точки наблюдения на величину $\Delta = d\sin\theta$ меньше расстояния между первой щелью и точкой

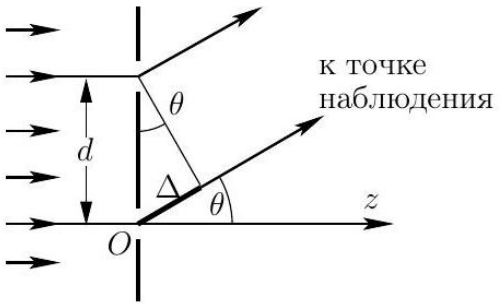



Рис. 3.19. Разность хода наблюдения. Соответствующая фаза колебания отличается на величину

$$\alpha = -k\Delta = -kd\sin\theta$$

Поэтому колебательный процесс, созданный второй щелью в точке наблюдения, описывается функцией $g(\theta)e^{i\alpha}$. Волны, посылаемые в точку наблюдения двумя щелями, интерферируют. Амплитуда суммарного колебательного процесса в точке наблюдения есть $g(\theta) + g(\theta)e^{i\alpha}$.

Картина интенсивности [5]

$$I(\theta) = |g(\theta)|^2 \cdot [1 + \cos(kd\sin\theta)]^2 \tag{3.29}$$

показана на рис. 3.20. Первый сомножитель $|g(\theta)|^2$ (домноженный на 4) показан на рисунке пунктирной огибающей - он описывает дифракцию на отдельной щели. Второй сомножитель связан с интерференцией волн, приходящих от двух щелей к точке наблюдения. Расстояние между максимумами (угловую ширину полосы $\Delta(\sin\theta)$) находим из условия $kd\sin\theta = 2\pi$, откуда $\Delta(\sin\theta) = \frac{\lambda}{d}$.

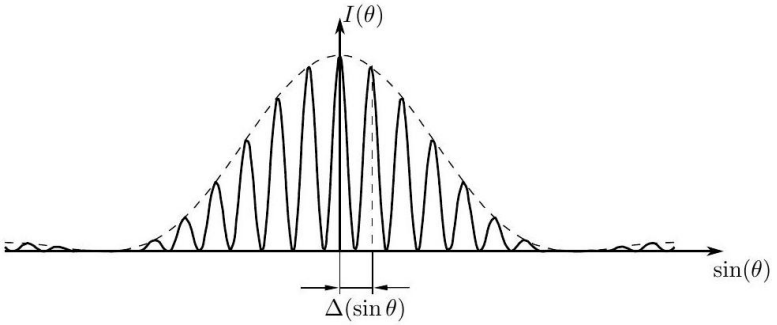


Рис. 3.20. Дифракция Фраунгофера на двух щелях

### 1.5.4. Дифракция Фраунгофера на решётке

Рассмотрим периодическую структуру одинаковых щелей с периодом $d$. По аналогии с дифракцией на двух щелях (см. выше) запишем суммарный колебательный процесс в точке наблюдений как сумму колебаний от каждой щели с учётом сдвига фазы. Фаза колебаний от щели номер $m$ (находящейся на расстоянии $md$ от начала отсчёта) на $\alpha_m = m\alpha = -mkd\sin\theta$ отличается от фазы колебаний начальной щели. Если всего решётка имеет $N$ щелей, суммарное колебание в точке наблюдения есть

$$g_N(\theta) = g(\theta)\sum_{m=0}^{N-1} e^{im\alpha} \tag{3.30}$$

Найдём сумму (3.30), построив векторную диаграмму. Каждое

5 Формулу (3.29) можно получить и как преобразование Фурье (3.25) функции пропускания, которая в данном случае представляет собой два «столбика» шириной $b$, разнесённые на расстояние $d$.

слагаемое $e^{im\alpha}$ изобразим вектором единичной длины, угол поворота которого относительно горизонтальной оси равен $m\alpha$. Получим цепочку вектором, показанную на рис. 3.21. Суммарное колебание изображается вектором, соединяющим начало и конец цепочки. Радиус окружности, в которую вписана цепочка векторов, равен $R = \frac{1}{2|\sin\alpha/2|}$, длина результирующего вектора равна $2R|\sin(N\alpha/2)|$. Отсюда получаем амплитуду колебания в точке наблюдения [6]:

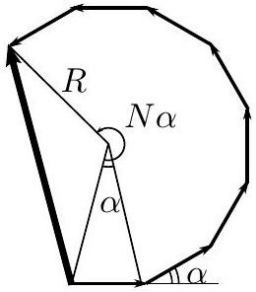



Рис. 3.21

$$|g_N(\theta)| = |g(\theta)|\cdot\left|\frac{\sin\frac{Nkd\sin\theta}{2}}{\sin\frac{kd\sin\theta}{2}}\right| \tag{3.31}$$

Распределение интенсивности $I(\theta) = |g_N(\theta)|^2$ по углам описывается формулой

$$I(\theta) = |g(\theta)|^2\left|\frac{\sin\frac{Nkd\sin\theta}{2}}{\sin\frac{kd\sin\theta}{2}}\right|^2 \tag{3.32}$$

Здесь первый сомножитель $|g(\theta)|^2$ описывает картину дифракции на отдельной щели (3.27), а второй связан с интерференцией волн от разных щелей. График интенсивности $I(\theta)$ представлен на рис. 3.22.

Характерной особенностью решётки является наличие узких максимумов, в которые идёт подавляющая доля общего потока энергии. Их положения определяются условием

$$d\sin\theta_m = m\lambda, \quad m = 0, \pm1, \pm2, \ldots \tag{3.33}$$

означающим, что идущие в направлении $\theta_m$ волны от всех щелей складываются в одной фазе, так как при этом $-\alpha_m = kd\sin\theta_m = 2\pi m$. Поскольку фазы вкладов от всех щелей одинаковы, амплитуда колебаний

6 Сумму (3.30) можно также вычислить, пользуясь известной формулой для суммы геометрической прогрессии $N$ членов с множителем $e^{i\alpha}$:

$$g_N(\theta) = g(\theta)\frac{e^{iN\alpha}-1}{e^{i\alpha}-1}$$

Вынося из числителя $e^{iN\alpha/2}$, а из знаменателя $e^{i\alpha/2}$, упрощаем полученное выражение:

$$g_N(\theta) = g(\theta)e^{i(N-1)\alpha/2}\frac{\sin\frac{Nkd\sin\theta}{2}}{\sin\frac{kd\sin\theta}{2}}$$

из чего, взяв модуль обеих частей, получим (3.31).

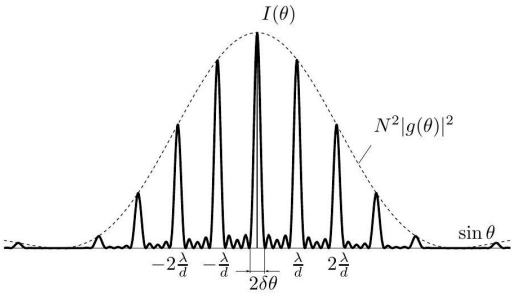



Рис. 3.22. Дифракция Фраунгофера на решётке. Здесь $N = 5, d/b = 5$.

в максимумах в $N$ раз больше амплитуды от одной щели, а интенсивность - в $N^2$ раз больше: $I(\theta) = N^2|g(\theta)|^2$. Этот же результат можно получить непосредственно из (3.32) с учётом того, что

$$\lim_{\alpha\to 2\pi m}\frac{\sin(N\alpha/2)}{\sin(\alpha/2)} = N$$

Угловую полуширину максимумов $\delta\theta$ оценим, найдя ближайшую к какому-либо $\theta_m$ точку $\theta = \theta_m + \delta\theta$, в которой функция (3.32) обращается в нуль. Такая точка соответствует приращению аргумента синуса в числителе (3.32) на $\pi$, то есть

$$\delta(kd\sin\theta) = \frac{2\pi}{N}, \quad \text{или} \quad \delta(\sin\theta) = \frac{\lambda}{Nd}. \tag{3.34}$$

Для небольших углов оценка угловой полуширины главных дифракционных максимумов есть

$$\delta\theta \approx \frac{\lambda}{Nd} \tag{3.35}$$

Целое число $m$ в (3.33) называется порядком главного максимума. Максимальное значение $m$, как ясно из (3.33), ограничено величиной $m_{\max} = \frac{d}{\lambda}$. Реально же заметными являются максимумы, лежащие в пределах углов $|\sin\theta| \leqslant \frac{\lambda}{b}$, поскольку в этих пределах распространяется в основном поток энергии от каждой отдельной щели согласно (3.28). Значит, при $b > \lambda$ имеем

$$m_{\max} \leqslant \frac{d}{b} \tag{3.36}$$

### 1.5.5. Дифракция Фраунгофера на круглом отверстии

При дифракции плоской волны на круглом отверстии в непрозрачном экране (рис. 3.23), на удалённой плоскости наблюдения (т. е. при условии $\frac{\sqrt{\lambda z}}{D} \gg 1, D$ - диаметр отверстия) образуется картина дифракции [7], показанная на рис. 3.23: центральное яркое дифракционное пятно (которое называется пятном Эйри) окружено чередующимися светлыми и тёмными кольцами. Соответствующий график $I(\theta)$ представлен на рис. 3.24. Как показывает расчёт, угловая полуширина пятна Эйри (главного максимума в картине $I(\theta)$) определяется условием

$$\boxed{\sin\theta = 1{,}22\frac{\lambda}{D}.} \qquad (3.37$$

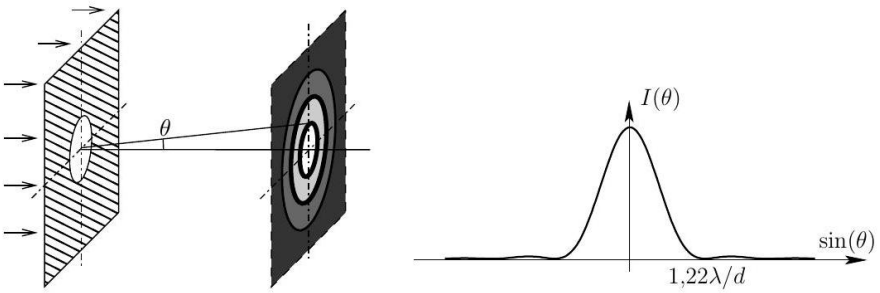



Рис. 3.23. Картина дифракции на круглом отверстии
Рис. 3.24. Зависимость интенсивности от угла при дифракции на отверстии

### Дифракция Фраунгофера

(ходят слухи, что если осветить решетку и поставить за ней линзу - мы получим один большой горб света в ф пл.

то есть дифр фраунгофера.

удивительно!!!! почему это так?

1.5.1 лабник)

в этом случае хз почему, но так:

$$g(x,y) = \frac{e^{ikR_0}}{i\lambda R_0}\iint f_0(\xi,\eta)e^{-i\left(\frac{kx}{R_0}\xi+\frac{ky}{R_0}\eta\right)}d\xi d\eta$$

### Дифракция Фраунгофера по Ландау

**Теория (?)**

Особый интерес для физических применений имеют дифракционные явления, возникающие при падении на экраны плоскопараллельного пучка лучей.

В результате дифракции пучок теряет параллельность и появляется свет, распространяющийся в направлениях, отличных от первоначального.

Поставим задачу об определении распределения по направлениям интенсивности дифрагированного света на больших расстояниях позади экрана (такая постановка вопроса отвечает так называемой дифракции Фраунгофера).

[0]7 Эта картина может быть получена как результат двумерного преобразования Фурье согласно (3.23). Соответствующий интеграл не выражается в элементарных функциях, но может быть найден численно или выражен через спецфункции (т. н. функции Бесселя), см. [5].

При этом мы снова ограничива- емся случаем малых отклонений от геометрической оптики, т.е.

пред- полагаем малыми углы отклонения от первоначального направления лучей (углы дифракции). Поставленную задачу можно было бы решить, исходя из общей формулы (59.2) переходя в ней к пределу бесконечно удаленных

от экрана источника света и точки наблюдения.

Характерной особенностью рассматриваемого случая является при этом то обстоятельство, что в интеграле, определяющем интенсивность дифрагированного света, существенна вся волновая поверхность, по которой производится интегрирование (в противоположность случаю дифракцИи Френеля, В Котором играли роль лишь участки волновой поверхности вблиЗи Проще, однако, рассмотреть поставленный вопрос заново, не прибегая к помощи общей формулы (59.2).

обозначим посредством $u_0$ то поле позади экранов, которое имелось бы при строгом соблюдении геометрической оптики.

оно представляет собой плоскую волну, в поперечном сечении которой, однако, имеются участки (отвечающие «тени» непрозрачных экранов) с равным нулю полем.

обозначим посредством $s$ ту часть плоскости поперечного сече- кость является волновой поверхностью плоской волны, то $u_0 = \text{const}$ вдоль всей площади $s$.

в действительности, однако, волна с ограниченной площадью поперечного сечения не может быть строго плоской (см.

&58).

в ее пространственное разложение фурье входят компоненты с волновыми векторами различных направлений, что и является источником дифракции.

Разложим поле $u_0$ В двухмерный интеграл Фурье по коорДинатам $y, z$ в плоскости поперечного сечения волны.

ДлЯ компонент Фурье НМееМ:

$$u_{\mathbf{q}} = \int\limits_S u_0 e^{-i\mathrm{qr}} dydz$$

где q - постоянный вектор в плоскости $yz$; интегрирование произвоДИТся фактически лишь по той части $S$ плоскости $yz$, на которой $u_0$ то компоненте поля $u_{\mathrm{q}} e^{i\mathrm{qr}}$ отвечает волновой вектор $\mathrm{k}' = \mathrm{k} + \mathrm{q}$.

Таким образом, вектор $\mathbf{q} = \mathbf{k}' - \mathbf{k}$ определяет изменение волнового вектора света при дифракции.

Поскольку абсолютные значения $k = k' = \omega/c$,

то малые углы дифракции $\theta_y, \theta_z$ В ПлоскостЯХ $xy$ И $xz$, связаны с составлЯющими вектора q соотношениЯми

$$q_y = \frac{\omega}{c}\theta_y, \quad q_z = \frac{\omega}{c}\theta_z$$

При малом отклонении от геометрической оптики компоненты разложениЯ поля $u_0$ можно считать совпадающими с компонентами истинного поля дифрагированного света, так что формула (61.1) решает Поставленную задачу.

Распределение интенсивности дифрагированного света определяется квадратом $|u_{\mathbf{q}}|^2$ как функЦией вектора q.

Количественная связь с интенсивностью падающего света устанавливается формулой

$$\iint u_0^2 dydz = \iint |u_{\mathbf{q}}|^2 \frac{dq_y dq_z}{(2\pi)^2}$$

(ср.

(49.8)).

Отсюда видно, что относительная интенсивность дифракЦИИ в элемент телесного угла $do = d\theta_y d\theta_z$ дается величиной

$$\frac{|u_{\mathrm{q}}|^2}{u_0^2}\frac{dq_y dq_z}{(2\pi)^2} = \Big(\frac{\omega}{2\pi c}\Big)^2 \left|\frac{u_{\mathrm{q}}}{u_0}\right|^2 do$$

рассмотрим дифракцию фраунгофера от двух экранов, являющихся «дополнительными» по отношению друг к другу: первый экран имеет отверстия там, где второй непрозрачен, и наоборот.

обозначим посредством $u^{(1)}$ и $u^{(2)}$ поле света, дифрагировавшего на этих экранах (при одинаковом в обоих случаях падающем свете).

поскольку $u_{\mathrm{q}}^{(1)}$

И $u_{\mathrm{q}}^{(2)}$ выражаются интегралами (61.1), взятыми по площадям отверстий в экранах, а отверстия в обоих экранах дополняют друг друга до целой плоскости, то сумма $u_{\mathrm{q}}^{(1)} + u_{\mathrm{q}}^{(2)}$ есть компонента Фурье ПолЯ, получающегося при отсутствии экранов, т.е.

просто падающего света.

Но падающий свет представляет собой строго плоскую волну с определенным направлением распространения, поэтому $u_{\mathrm{q}}^{(1)} + u_{\mathrm{q}}^{(2)} = 0$ при всяком отличном от нуля q.

Таким образом, имеем $u_{\mathrm{q}}^{(1)} = -u_{\mathrm{q}}^{(2)}$ ИлИ, ДлЯ соответствующих интенсивностей,

$$\left|u_{\mathbf{q}}^{(1)}\right|^2 = \left|u_{\mathbf{q}}^{(2)}\right|^2 \quad \text{при } \mathbf{q} \neq 0$$

Это значит, что дополнительные экраны дают одинаковые распределениЯ Интенсивности дифрагированного света (так называемый принцип Бабине).

Упомянем здесь об одном интересном следствии принципа Бабине.

Рассмотрим какое-нибудь черное тело, т.е.

тело, полностью поглОЩаюЩее весь падающий на него свет.

Согласно геометрической оптике при освещении такого тела за ним образовалась бы область геометрической тени, плоШадь сечения которой была бы равна площади сечениЯ тела в направлении, перпендикулярном к направлению падения света.

Наличие дифракции приведет, однако, к частичному отклонению света

от первоначального направлениЯ.

В результате на большом расстоянии позади тела тени не будет, а наряду со светом, распространяющимся в первоначальном направлении, будет также и некоторое количество света, распространяющегося под небольшими углами к своему первоначальному направлению.

Легко определить интенсивность этого, как говорят, рассеянного света.

Для этого замечаем, что согласно принЦиПу Бабине количество света, отклонившегося вследствие дифракции на рассматриваемом теле, равно количеству света, отклоняющегося при дифракции от прорезанного в непрозрачном экране отверстия, форма и площадь которого совПадают с формой и площадью поперечного сечения тела.

Но при дифракции Фраунгофера от отверстия происходит отКлонение всего проходящего через отверстие света.

Отсюда следует, что полное количество света, рассеянного на черном теле, равно количеству света, падающего на его поверхность и поглощаемого им.

**ЛЛ2-?**

Определить дифракцию Фраунгофера при нормальном падении плоской волны на бесконечную щель (ширины $2a$) с параллельными краями, прорезанную в непрозрачном

экране.

Ре ш е н и е.

Выберем плоскость щели в качестве плоскости $yz$ с осью $z$ вдоль длины щели (рис. 13 представляет разрез экрана).

При нормальном падении света плоскость щели является одной из волновых поверхностей, которую мы возьмем в качестве поверхности интегрирования в (61.1).

Ввиду бесконечности длины щели свет отклоняется только в плоскости $xy$ (интеграл (61.1) обращается в нуль при $q_z \neq 0$).

Поэтому разложение поля $u_0$ должно производиться лишь по координате $y$:

$$u_q = u_0 \int\limits_{-a}^{a} e^{-iqy} dy = \frac{2u_0}{q} \sin qa$$

Интенсивность дифрагированного света в интервале углов $d\theta$ есть

$$dI = \frac{I_0}{2a} \left| \frac{u_q}{u_0} \right|^2 \frac{dq}{2\pi} = \frac{I_0}{\pi ak} \frac{\sin^2 ka\theta}{\theta^2} d\theta$$

где $k = \omega/c$, а $I_0$ - полная интенсивность света, падающего на щель.

$dI/d\theta$ как функция угла дифракция имеет вид, изображенный не рис.

14. При увеличении $\theta$ в ту или другую сторону от $\theta = 0$ интенсивность пробегает ряд максимумов с быстро убывающей высотой.

Максимумы разделены в точках $\theta = n\pi/ka$($n$ - целые числа) минимумами, в которых интенсивность обращается в нуль.

**ЛЛ2-?**

То же при дифракции от решетки - плоского экрана с прорезанным В нем рядом одинаковых параллельных щелей (ширина щели $2a$, ширина непрозрачного экрана между соседними щелями $2b$, число щелей $N$).

Р е ш е н и е.

Выберем плоскость решетки в качестве плоскости $yz$ с осью $z$, параллельной щелям.

Дифракция снова происходит лишь в плоскости $xy$, и интегрирование в (61.1) дает:

$$u_q = u'_q \sum_{n=0}^{N-1} e^{-2inqd} = u'_q \frac{1 - e^{-2iNqd}}{1 - e^{-2iqd}}$$

где $d = a + b$, а $u'_q$ есть результат интегрирования по одной щели.

Воспользовавшись результатами задачи 1, получим:

$$dI = \frac{I_0 a}{N\pi} \left( \frac{\sin Nqd}{\sin qd} \right)^2 \left( \frac{\sin qa}{qa} \right)^2 dq = \frac{I_0}{N\pi ak} \left( \frac{\sin Nk\theta d}{\sin k\theta d} \right)^2 \frac{\sin^2 ka\theta}{\theta^2} d\theta$$

($I_0-$ полная интенсивность света, проходящего через все щели).

При большом числе щелей ($N \to \infty$) эту формулу можно написать в ином виде.

При значениях $q = \pi n/d$($n$ - целое число )$dI/dq$ имеет максимумы; вблизи такого максимума (т.е.

при $qd = n\pi + \varepsilon$, $\epsilon$ мало)

$$dI = I_0 a \left( \frac{\sin qa}{qa} \right)^2 \frac{\sin^2 N\varepsilon}{\pi N\varepsilon^2} dq$$

Но при $N \to \infty$ имеет место формула [1])

$$\lim_{N\to\infty} \frac{\sin^2 Nx}{\pi N x^2} = \delta(x)$$

Поэтому вблизи каждого максимума имеем:

$$dI = I_0 \frac{a}{d} \left( \frac{\sin qa}{qa} \right)^2 \delta(\varepsilon) d\varepsilon$$

т.е. максимумы обладают, в пределе, бесконечно малой шириной, а полная интенсивность света в $n$ -м максимуме есть

$$I^{(n)} = I_0 \frac{d}{\pi^2 a} \frac{\sin^2(n\pi a/d)}{n^2}$$

3.

**ЛЛ2-?**

Определить распределение интенсивности по направлениям при дифракции света, падающего в нормальном направлении на круглое отверстие радиуса $a$.

Ре ш е н и е.

Введем цилиндрические координаты $z, r, \varphi$ с осью $z$, проходящей через центр отверстия перпендикулярно к его плоскости.

Очевидно, что дифракция симметрична относительно оси $z$, так что вектор q имеет лишь радиальную компоненту $q_r = q = k\theta$. Отсчитывая угол $\varphi$ от направления q и интегрируя в (61.1) по плоскости отверстия, находим:

$$u_q = u_0 \int_0^a \int_0^{2\pi} e^{-iqr\cos\varphi} r d\varphi dr = 2\pi u_0 \int_0^a J_0(qr) r dr$$

где $J_0$ - функция Бесселя нулевого порядка.

С помощью известной формулы

$$\int_0^a J_0(qr) r dr = \frac{a}{q} J_1(aq)$$

имеем отсюда:

$$u_q = \frac{2\pi u_0 a}{q} J_1(aq)$$

и согласно (61.4) находим окончательно интенсивность света, дифрагировавшего в элемент телесного угла $do$ :

$$dI = I_0 \frac{J_1^2(ak\theta)}{\pi\theta^2} do$$

где $I_0$ - полная интенсивность света, падающего на отверстие.

**связь с преобразованием фурье!!!**

## 7.1.7 Fraunhofer Diffraction: Examples

### 1D дифракция фраунгофера

потом поучу.

$$g(u) \propto \int_{-\infty}^{+\infty} f_0(\xi) e^{-iu\xi} d\xi$$

**Дифракция на щели**

(начало задачи должно содержать слова, что мы именно Фраунгофера решаем, так как я удалил из главы это слово, чтобы было удобнее)

$$g(\theta) \propto \int_{-b/2}^{+b/2} e^{ikx\sin\theta} dx \propto \frac{\sin\left(\frac{kb}{2}\sin\theta\right)}{\frac{kb}{2}\sin\theta}$$

**Дифракция на двух щелях**

(начало задачи должно содержать слова, что мы именно Фраунгофера решаем, так как я удалил из главы это слово, чтобы было удобнее)

$$I(\theta) = |g(\theta)|^2 \cdot [1 + \cos(kd\sin\theta)]^2$$

**Дифракция на решётке**

(вставка теории из последней главы.

уже все характеристики прибора должны быть усвоены.)

(начало задачи должно содержать слова, что мы именно Фраунгофера решаем, так как я удалил из главы это слово, чтобы было удобнее)

прямо очень крутая тема и на самом деле я тут все знаю, нужно только дойти до нее

тр зн:

- положение и интенс гл макс - их ширина и макс порядок

зададимся вопросом о положении главных максимумов.

....(сам я пока вывести нифига не могу)

**Дифракция на круглом отверстии**

есть в билетах

(начало задачи должно содержать слова, что мы именно Фраунгофера решаем, так как я удалил из главы это слово, чтобы было удобнее)

**поле в фокальной плоскости линзы**

(я хз, что тут, но это есть в билетах)

**На синусоидальной решетке (??)**

Рассмотрим применения метода Рэлея на простейших задачах дифракции плоской волны на периодических структурах. Ограничимся сейчас случаем, когда волновые векторы всех волн лежат в плоскости $(x, z)$. 1. Амплитудной синусоидальной решёткой называется объект с функцией пропускания

$$D(x) = D_0(1 + m\cos\Omega x).$$

Период этой решётки равен

$$d = 2\pi/\Omega.$$

Имеется естественное ограничение на значения параметров $D_0$ и $m$:

$$D_0(1+m) \le 1,$$

определяемое тем, что прохождение тонкого экрана не приводит к локальному увеличению амплитуды поля.

Пусть на решётку (8.2.1) слева падает плоская волна под углом $Q_0$ (рис. 8.2.1):

$$A = A_0 e^{i\mathbf{k}_0\mathbf{r}}, u_0 = k\sin\theta_0, w_0 = k\cos\theta_0. \quad (8.2.4)$$

Для нахождения поля в области $z > 0$ воспользуемся граничным условием Кирхгофа:

$$A_S = D(x)A|_{z=-0} = A_0 D(x) e^{iu_0 x}. \quad (8.2.5)$$

Имея в виду выражение (8.2.1) для функции пропускания, получаем пространственное разложение Фурье функции $A_S$ :

$$A_S = A_0 D_0 e^{iu_0 x} + \frac{m}{2} A_0 D_0 e^{i(u_0+\Omega)x} + \frac{m}{2} A_0 D_0 e^{i(u_0-\Omega)x}.$$

В соответствии со сказанным в разделе 8.1 каждое из слагаемых в этой сумме порождает плоскую волну при $z > 0$. Следовательно, получаем

$$\begin{aligned} A(x,z) &= A_0(x,z) + A_{+1}(x,z) + A_{-1}(x,z) \\ A_0(x,z) &= A_0 D_0 \exp\left[iu^{(0)}x + iw^{(0)}z\right] \\ A_{+1}(x,z) &= \frac{m}{2} A_0 D_0 \exp\left[iu^{(+1)}x + iw^{(+1)}z\right] \\ A_{-1}(x,z) &= \frac{m}{2} A_0 D_0 \exp\left[iu^{(-1)}x + iw^{(-1)}z\right] \end{aligned}$$

Здесь введены обозначения

$$\begin{aligned} &u^{(0)} = u_0, w^{(0)} = \sqrt{k^2 - u_0^2}, \\ &u^{(+1)} = u_0 + \Omega, w^{(+1)} = \sqrt{k^2 - (u_0+\Omega)^2}, \\ &u^{(-1)} = u_0 - \Omega, w^{(-1)} = \sqrt{k^2 - (u_0-\Omega)^2}. \end{aligned}$$

Пространственные частоты $u^{(0)}, u^{(+1)}, u^{(-1)}$ определяют направления распространения волн справа от решётки ($z > 0$). Полагая для этих волн $u = k\sin\theta$, записываем уравнения для отределения их направлений: $k\sin\theta^{(0)} = k\sin\theta_0, k\sin\theta^{(+1)} = k\sin\theta_0 + \Omega, k\sin\theta^{(-1)} = k\sin\theta_0 - \Omega$.

Вспоминая равенства $k = 2\pi/\lambda, \Omega = 2\pi/d$, получаем

$$d\,(\sin\theta - \sin\theta_0) = n\lambda, n = 0, \pm 1.$$

Мы пришли к известной формуле, указывающей направления на главные максимумы при дифракции на решётке. Заметим, что в рассматриваемом случае имеется всего три максимума. Это связано прежде всего с тем, что период решётки $d$ и эффективная ширина щели связаны соотношением $b_{\text{eff}} = d/2$. 2. Фазовой синусоидальной решёткой называется объект с функцией пропускания

$$D = e^{im\cos\Omega x}, \Omega = 2\pi/d.$$

Этот объект, не меняя амплитуды падающей на него волны, изменяет её фазу. Пусть на решётку (8.2.10) падает слева плоская волна (рис. 8.2.1):

$$A = A_0 e^{i\mathbf{k}_0\mathbf{r}}.$$

Найдём поле справа от решётки. Используем граничное условие

$$A|_{z=+0} = D(x)A|_{z=-0}\,,$$

или

$$A|_{z=+0} = A_0 e^{iu_0 x + im\cos\Omega x}, u_0 = k\sin\theta_0.$$

Обычно рассматривают фазовые решётки с малыми модуляциями фазы: $m \ll 1$. В этом случае получаем

$$A|_{z=+0} = A_0 e^{iu_0 x}(1 + im\cos\Omega x) =$$
$$= A_0 e^{iu_0 x} + \frac{im}{2} A_0 e^{i(u_0+\Omega)x} + \frac{im}{2} A_0 e^{i(u_0-\Omega)x}$$

Следовательно, от плоскости $z = +0$ исходят три волны, которым отвечают пространственные частоты

$$u = u_0 + n\Omega, n = 0, \pm 1.$$

Полагая

$$u = k\sin\theta, u_0 = k\sin\theta_0, k = 2\pi/\lambda,$$

находим уравнение, определяющее направления волн:

$$d\,(\sin\theta - \sin\theta_0) = n\lambda, n = 0, \pm 1.$$

Таким образом, в пределе $m \ll 1$ остаются, как и в случае амплитудной синусоидальной решётки, три волны. Разница состоит в том, что в случае фазовой решётки фазы волн, отвечающих главным максимумам $n = \pm 1$, имеют дополнительный сдвиг на $\pi/2$ (благодаря наличию множителя $i$ во втором и третьем слагаемых в (8.2.13)).

Заметим, что если провести разложение (8.2.12) до более высоких степеней параметра $m$, то получим волны, распространяющиеся под углами, отвечающими значениям $n = \pm 2, \pm 3, \ldots$ Ограничение на величину $n$ при этом даётся, как и в случае обычной дифракционной решётки, условием $n \leq d/\lambda$.

### Эффект саморепродукции

Рассмотрим экран с периодически расположенными отверстиями (например, дифракционную решетку). Будем для простоты говорить только о структуре на экране, периодичной вдоль оси $x$ и однородной вдоль оси $y$. Обобщение на случай периодичности и вдоль оси $y$ не представляет труда.

Пусть слева на экран падает плоская волна вдоль оси $z$ (рис. 8.2.2):

$$A(x, z) = A_0 e^{ikz}, z < 0$$

Вследствие периодичности структуры экрана справа от него при $z = +0$ поле может быть представлено в виде ряда Фурье:

$$A_S(x) \equiv A(x, z)|_{z=+0} = D(x)A(x, z)|_{z=-0} = \sum_{n=0}^{\infty} a_n e^{iu_n x}$$

где

$$u_n = n\Omega, \Omega = 2\pi/d,$$

$d$ - период структуры (вдоль оси $x$). Каждое слагаемое в сумме (8.2.2) порождает в области $z > 0$ волну

$$a_n e^{iu_n x + iw_n z}, w_n = \sqrt{k^2 - u_n^2}$$

так что

$$A(x,z)|_{z>0} = \sum_{n=0}^{\infty} a_n e^{iu_n x + iw_n z}.$$

Если период структуры $d \gg \lambda$, то есть $\Omega \ll k$, то для не слишком болыших номеров гармоник $n$ можно считать $u_n \ll k$ и записать:

$$w_n \approx k - u_n^2/2k.$$

Соответственно выражение для $n$-й волны принимает вид

$$a_n e^{ikz} \exp\left(iu_n x - i\frac{u_n^2}{2k}z\right)$$

Если на некотором расстоянии $z$ от экрана (периодической структуры) будет выполнено условие

$$\frac{u_n^2}{2k}z = 2\pi p$$

($p$ - целое число), или

$$n^2\frac{\Omega^2}{2k}z = 2\pi p,$$

то в соответствии с (8.2.22), (8.2.23) окажстся

$$A(x,z) = e^{ikz}\sum_{n=0}^{\infty} a_n \exp\left(iu_n x - i\frac{u_n^2}{2k}z\right) =$$

$$= e^{ikz}\sum_{n=0}^{\infty} a_n \exp\left(iu_n x\right) = e^{ikz}A_S(x).$$

Это значит, что с точностью до фазового множителя $e^{ikz}$ на таком расстоянии воспроизводится структура исходного волнового поля $A_S(x)$, поскольку относительный набег фаз всех слагаемых в сумме (8.2.20) оказывается кратным $2\pi$. Условие (8.2.24) выполняется в точках

$$z = mz_T, z_T = 2\pi\frac{2k}{\Omega^2} = \frac{2d^2}{\lambda}, m = 0, 1, 2, \ldots$$

Действительно, в этих точках

$$\frac{u_n^2}{2k}z = n^2\frac{\Omega^2}{2k}mz_T = 2\pi mn^2 = 2\pi p,$$

где $p = mn^2$ - целое число. Заметим, что хотя каждое из слагаемых в сумме (8.2.20) приобретает свой набег фазы, но эти набеги для всех слагаемых кратны $2\pi$.

Таким образом, периодическая структура $A_S(x)$ периодически самовоспроизводится (повторяется) при условии

$$z = mz_T = m\frac{2d^2}{\lambda}.$$

Оденим максимальное число «копий» структуры, которое можно наблюдать при использовании решётки длиной $D$. Эффект саморепродукции возникает благодаря интерференции волн, распространяющихся от решетки под разными углами. Наименьший угол с осью $z$ образуют 0 -й и)1)-й главные дифракционные максимумы. Распространяясь от решетки

конечных размеров $D$, эти три волны перестают перекрываться на расстоянии (рис. 8.2.3):

$$L = \frac{D}{2} \operatorname{ctg} \theta$$

где $\theta$ - угол дифракции, определяемый условием

$$d \sin \theta = \lambda.$$

Считая $\lambda \ll d$, заключаем, что $\theta \approx \lambda/d \ll 1$. Поэтому

$$L \approx \frac{D}{2\theta} \approx \frac{Dd}{2\lambda}.$$

На этом расстоянии число плоскостей саморепродукции составляет

$$N \sim \frac{L}{z_T} \approx \frac{Dd/2\lambda}{2d^2/\lambda} = \frac{D}{4d}$$

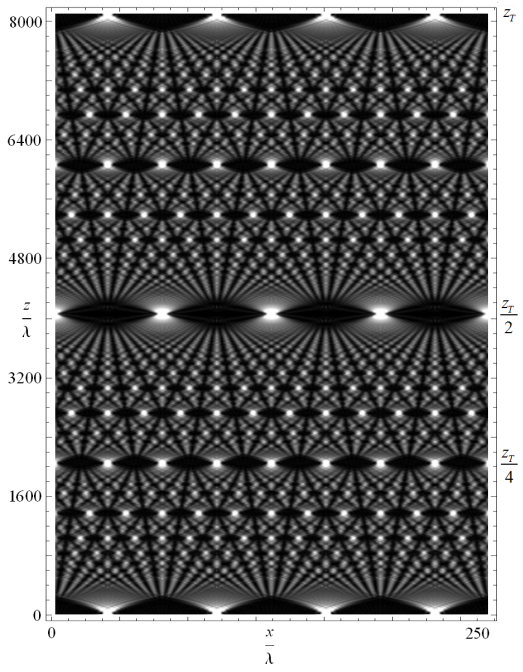
В действительности для установления периодичности объекта по его «копии» достаточно иметь пересечение только 0-го дифракционного максимума и хотя бы одного из боковых лепестков: $(+1)$-го или $(-1)$-го. Соответственно расстояние $L$ можно взять вдвое большее: $L_1 = D \operatorname{ctg} \theta$. Это даёт вдвое большее число наблюдаемых «копий» решётки:

$$N \sim \frac{L_1}{z_1} \approx \frac{D}{2d}.$$

Явление саморепродукции наблюдалось в 1836 г. английским физиком и химиком Генри Фоксом Талботом и получило название эффекта Талбота.

Рассмотренный эффект имеет место на расстояниях $z \geq z_T = 2d^2/\lambda$. По отношению к решётке в целом эта область попадает в зону дифракции Френеля. Действительно, все наблюдаемые «копии» находятся на расстоянии, не превышающем $L = D \operatorname{ctg} \theta \approx Dd/\lambda$. Условие дифракции Френеля имеет вид $D^2/\lambda > L$. Согласно предыдущему равенству это сводится к очевидному требованию $D > d$: размер решётки должен превышать её период.

В ближней зоне $(z < z_T)$ дифракционная картина оказывается весьма сложной. На рис. 8.2.4 приведено распределение интенсивности в области $0 \leq z \leq z_T$. Эту картину называют «ковёр Талбота».




**Функция пропускания линзы**

Рассмотрим на основе метода Рэлея действие линзы. Считаем линзу тонкой и ограничиваемся приближением параксиальных пучков (когда лучи образуют малые углы с главной оптической осью).

Как известно, оптические длины лучей, идущих от линзы к фокусу, одинаковы. Это значит, что плоская волна преобразуется в сферическую, сходящуюся к фокусу. Сформировавшийся сферический волновой фронт показан на рис. 8.3.1 штриховой линией. Отсюда следует, что лучи, наталкивающиеся на линзу, приобретают фазовую задержку до момента достижения сходящегося волнового Фронта:

$$\Delta\varphi = k\left(f - \sqrt{f^2+\rho^2}\right) \approx -\frac{k\rho^2}{2f}$$

Здесь $\rho$ - расстояние от точки входа луча в линзу до оптической оси. Таким образом, функцию пропускания линзы можно записать в виде

$$D = e^{i\Delta\varphi} = \exp\left(-\frac{ik\rho^2}{2f}\right).$$

Если линза рассеивающая, то замена $f \to -f$ даёт

$$D = \exp\left(\frac{ik\rho^2}{2f}\right)$$

В качестве применения формулы (8.3.2) рассмотрим действие собирающей линзы на плоскую волну, распространяющуюся параллельно оптической оси:

$$A = A_0 e^{ikz}.$$

Запишем волну, прошедшую через линзу:

$$A_S(x,y) = D(x,y)A_0 = A_0 \exp\left(-\frac{k\left(x^2+y^2\right)}{2f}\right).$$

Чтобы найти поле в области $z > 0$, нужно выполнить пространственное разложение Фурье функции (8.3.5):

$$a(u,v) = \int\limits_{x=-\infty}^{\infty} \int\limits_{y=-\infty}^{\infty} e^{-i(2x+vy)} A_S(x,y)dxdy.$$

Для вычисления воспользуемся следующим равенством:

$$\int\limits_{k_x=-\infty}^{\infty} e^{-i\left(2x+kx^2/2f\right)}dx = (1-i)\sqrt{\frac{\pi f}{k}}\exp\left(i\frac{f}{2k}u^2\right).$$

Аналогичное равенство имеет место и при интегрировании по $dy$. Соответственно получаем

$$a(u,v) = \frac{2\pi f}{ik}A_0 \exp\left[i\frac{f}{2k}\left(u^2+v^2\right)\right]$$

Чтобы восстановить поле за линзой, нужно продлить в область $z > 0$ каждую из пространственных гармоник:

$$A(x,y,z) = \int\limits_{u=-\infty}^{\infty} \int\limits_{v=-\infty}^{\infty} e^{iux+ivy+iz\sqrt{k^2-u^2-v^2}} a(u,v)\frac{dudv}{(2\pi)^2}.$$

Имея в виду параксиальное приближение: $|u| \ll k, |v| \ll k$, положим

$$\sqrt{k^2-u^2-v^2} \approx k - \frac{u^2+v^2}{2k}.$$

Таким образом, задача сводится к вычислению интеграла

$$A(x,y,z) = A_0 e^{ikz}\frac{2\pi f}{ik} \int\limits_{k_x=-\infty}^{\infty} \int\limits_{y}^{\infty} \exp\left[iux+ivy-i\frac{u^2+v^2}{2k}(z-f)\right]\frac{dudv}{(2\pi)^2}.$$

Расчёт с использованием равенства типа (8.3.7) даёт

$$A(x,y,z) = A_0 e^{ikz}\frac{f}{f-z}\exp\left(-i\frac{k\rho^2}{2(f-z)}\right), z < f$$

Заметим, что допустимость использования параксиального приближения (8.3.10) в рассматриваемой ситуации связана с быстрой сходимостью интеграла (8.3.11).

Таким образом, мы получили очевидный результат: интенсивность вдоль оптической оси меняется по закону

$$I = \frac{I_0}{(1-z/f)^2},$$

достигая бесконечного значения непосредственно в фокусе линзы $z = f, \rho = 0$. В рассмотренной постановке задачи линза предполагалась неограниченной. Поэтому фокусное пятно стягивается практически в точку - эффекты дифракции оказываются пренебрежимо малыми.

## 7.1.8 Numerical Methods for Diffraction

(потом напишу, как если нужно, я считал бы и визуализировал задачи про дифракции. мб П-Л методом?)

## 7.2 2. Principles of Fourier Optics

### 7.2.1 2.1. A Spectral Method for Solving the Diffraction Problem

Спектральные методы являются основой при изучении колебаний различной физической природы (электрический колебательный контур, механический осциллятор, электрон в атоме и т.д.). Суть спектрального метода состоит в представлении внешнего воздействия, возбуждающего колебания в линейной системе, в виде суммы некоторых элементарных воздействий. Свойство линейности системы (принцип суперпозиции) позволяет найти решение задачи в виде соответствующей суммы откликов - вынужденных колебаний, возбуждаемых в системе отдельным элементарным воздействием. Важнейшей проблемой при использовании спектрального метода является проблема поиска «базиса» - элементарных слагаемых. При изучении стационарных колебательных линейных систем широко используется представление внешнего воздействия $f(t)$ в виде суммы гармонических колебаний различных частот $\omega_n$ : $f(t) = \sum c_n e^{i\omega_n t}$. Особая роль гармонических слагаемых $e^{i\omega t}$ обусловлена тем, что гармоническое внешнее воздействие $e^{i\omega t}$ возбуждает в линейной стационарной системе процесс вынужденных колебаний, которые также являются гармоническими и частота которых совпадает с частотой внешнего воздействия.

Применим спектральный метод к исследованию законов распространения волн в задаче дифракции (метод Рэлея), основанный на представлении пространственной структуры дифрагированной волны в виде *су* перпозиции плоских волн разных направлений.

Распространение плоской волны. Рассмотрим плоскую волну, волновой вектор которой $\boldsymbol{k}$ составляет угол $\alpha$ с осью $z$ (рис. 3.25). Комплексная амплитуда такой волны имеет вид

$$f(x,z) = ae^{i(k_x x + k_z z + \varphi)}, \tag{3.38}$$

где $a$ - амплитуда волны, $\varphi$ - начальная фаза, $k_x = k\sin\alpha$ и $k_z = k\cos\alpha$ - проекции вектора $\boldsymbol{k}(k = |\boldsymbol{k}|)$ на оси $x$ и $z$ соответственно (мы рассматриваем для простоты случай $k_y = 0$).

Введём комплексный коэффициент $c = ae^{i\varphi}$ (определяющий и амплитуду, и начальную фазу) и обозначение $k_x = u$. Тогда, имея в виду, что $k_x^2 + k_z^2 = k^2$, т. е. $k_z = \sqrt{k^2 - u^2}$, перепишем комплексную амплитуду волны (3.38) в виде

$$f(x,z) = ce^{i\left(ux + \sqrt{k^2-u^2}\cdot z\right)} \tag{3.39}$$

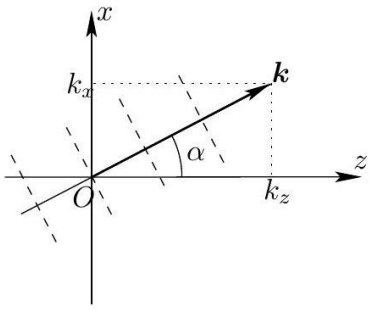



Рис. 3.25. Наклонная плоская волна. Пунктирами показаны волновые поверхности

Волновое поле плоской волны в любой плоскости $z > 0$ можно найти, если известно поле плоской волны в плоскости $z = 0$:

$$f(x,0) = ce^{iux} \tag{3.40}$$

откуда

$$f(x,z) = f(x,0) \cdot e^{i\sqrt{k^2-u^2}z} \tag{3.41}$$

т. е. комплексные амплитуды $f(x,0)$ и $f(x,z)$ отличаются множителем $H(u) = e^{i\sqrt{k^2-u^2}\cdot z}$, определяющим набег фазы плоской волны при распространении между двумя плоскостями, разделенными промежутком $z$.

Обратим внимание на аналогию (3.40) с выражением $f(t) = ce^{i\omega t}$, которое, как мы знаем, есть не что иное, как комплексная форма записи гармонического колебания частоты $\omega$, причём комплексный множитель с определяет амплитуду колебания и его начальную фазу. На основании этой аналогии величина $u = k\sin\alpha$ может быть названа пространственной частотой. Можно сказать, что волны разных направлений $\alpha$ - это волны разных пространственных частот.

Спектр плоских волн. Итак, представим граничное поле $f_0(x)$ в виде суммы плоских волн:

$$f_0(x) = \sum c_n e^{iu_n x} \tag{3.42}$$

Набор чисел $c_n$ представляет собой пространственный спектр волнового поля $f_0(x)$ в (3.42). Каждое слагаемое в сумме (3.42) есть поле плоской волны (в плоскости $z = 0$), направление $\alpha_n$ которой определяется пространственной частотой $u_n = k\sin\alpha_n$. Подчеркнём, что частота волн $\omega$ предполагается заданной и, следовательно, волновое число
$k = |\boldsymbol{k}| = \omega/v$ всех плоских волн в пространственном разложении (3.42) одинаково.

Напомним, что в общем случае произвольное граничное поле $f_0(x)$ представляется в виде интеграла Фурье:

$$f_0(x) = \frac{1}{2\pi}\int C_0(u)e^{iux}du \tag{3.43}$$

т.е. в виде непрерывной суммы плоских волн различных пространственных частот. Функция $C_0(u)$ в (3.43) есть преобразование Фурье функции $f_0(x)$. Пространственный спектр $C_0(u)$ (его можно назвать спектром плоских волн) определяется соотношением

$$C_0(u) = \int\limits_{-\infty}^{\infty} f_0(x)e^{-iux}dx \tag{3.44}$$

Напомним также, что если функция $f_0(x)$ периодична по координате $x$, то она представляется рядом Фурье:

$$f_0(x) = \sum_n c_n e^{inu_0 x} \tag{3.45}$$

причём коэффициенты $c_n$ можно найти с помощью формулы

$$c_n = \frac{1}{d}\int\limits_{-d/2}^{d/2} f_0(x)e^{-inu_0 x}dx \tag{3.46}$$

где $d = \frac{2\pi}{u_0}$ - пространственный период функции $f_0(x)$.

Распространение волн. Важно подчеркнуть, что речь идёт о линейной задаче: распространение волны от плоскости $z = 0_+$до плоскости наблюдения $z = \text{const} > 0$ описывается линейным волновым уравнением (2.1) или, поскольку мы пользуемся комплексным представлением, линейным уравнением Гельмгольца (2.16). Основное свойство линейного уравнения: сумма решений является решением.

Вспомним теперь, что плоская волна

$$f_n(x,z) = c_n e^{i\left(u_n x + \sqrt{k^2-u_n^2}z\right)}$$

есть решение уравнения Гельмгольца, удовлетворяющее на плоскости $z = 0_+$граничному условию

$$f_n(x,0) = c_n e^{iu_n x}$$

поэтому сумма плоских волн

$$f(x,z) = \sum_n c_n e^{i\left(u_n x + \sqrt{k^2-u_n^2}z\right)} \tag{3.47}$$

есть решение, удовлетворяющее граничному условию (3.42).

В общем случае, если граничное поле $f_0(x)$ представляется непрерывной суперпозицией плоских волн (3.43), искомое решение имеет вид

$$f(x,z) = \int C_0(u) e^{i\left(ux + \sqrt{k^2-u^2}z\right)} du \tag{3.48}$$

где $C_0(u)$ - преобразование Фурье граничного поля $f_0(x)$ (т. е. спектр плоских волн граничного поля).

В выражениях (3.47) или (3.48) существенным является следующее обстоятельство: каждая слагаемая плоская волна при распространении до плоскости наблюдения $z = \text{const} > 0$ приобретает свой фазовый набег $\varphi_n = \sqrt{k^2-u_n^2}\cdot z$, зависящий от её пространственной частоты $u_n$. Поэтому фазовые соотношения между слагаемыми плоскими волнами на границе $z = 0$ и в плоскости наблюдения, отстоящей на расстоянии $z$, различны. Изменение фазовых соотношений между слагаемыми плоскими волнами приводит к тому, что изменяется результат интерференции этих волн. Поэтому результирующее поле $f(x,z)$ в плоскости наблюдения может кардинально отличаться от граничного поля $f_0(x)$ (хотя и то, и другое составлено из суперпозиции тех же бегущих плоских волн).

Передаточная функция. Переписав выражение (3.48) в виде

$$f(x,z) = \int C_0(u) e^{i\sqrt{k^2-u^2}z} e^{iux} du = \int C(u,z) e^{iux} du$$

мы видим, что функция

$$C(u,z) = C_0(u) e^{i\sqrt{k^2-u^2}z} \tag{3.49}$$

представляет собой преобразование Фурье (пространственный спектр) монохроматического волнового поля $f(x,z)$ в плоскости наблюдения (это поле рассматривается как функция координаты $x$ в плоскости наблюдения при фиксированном z).

Множитель

$$H(u) = e^{i\sqrt{k^2-u^2}z} \tag{3.50}$$

связывающий пространственные спектры $C_0(u)$ и $C(u,z)$ волнового поля в двух плоскостях, разделённых промежутком свободного пространства $z$, называют передаточной функцией (или частотной характеристикой) свободного пространства (точно так же частотная характеристика
$H(\omega)$ связывает между собой спектры входного и выходного сигнала линейного

временного фильтра). Поэтому участок свободного пространства можно рассматривать как простейший линейный пространственный фильтр, входным сигналом которого является поле $f_0(x)$ во входной плоскости $z = 0_+$, а выходным сигналом - поле $f(x, z)$ в выходной плоскости $z = \text{const} > 0$.

Во многих задачах речь идёт о волновых полях, имеющих достаточно узкий спектр плоских волн $|u| \ll k$. При этом используют приближённое выражение для частотной характеристики (3.50):

$$H(u) \approx e^{ikz} e^{-i\frac{z}{2k}u^2}$$

которое получается разложением радикала $\sqrt{k^2 - u^2}$ в ряд по степеням малого параметра $\left(\frac{u}{k}\right)^2$, причём удерживаются лишь два члена разложения:

$$\sqrt{k^2 - u^2}z \approx kz - \frac{z}{2k}u^2 \tag{3.51}$$

При этом точное равенство (3.49) заменяется приближённым:

$$C(u, z) \approx C_0(u) e^{-i\frac{z}{2k}u^2} \tag{3.52}$$

(постоянный фазовый множитель $e^{ikz}$ отброшен).

Итак, метод Рэлея предлагает следующую последовательность решения задач дифракции.

1. Зная поле сторонних источников $f_s(x)$ и функцию пропускания транспаранта $t(x)$, находим с помощью граничных условий (3.2) граничное поле $f_0(x)$.

2. Используя соотношение (3.46) либо (3.44), определяем пространственный спектр граничного поля: набор коэффициентов $c_n$ ряда Фурье либо фурье-образ $C_0(u)$ функции $f_0(x)$.

3. Наконец, с помощью формул (3.47) либо (3.48) определяем искомое поле $f(x, z)$ на расстоянии $z$ от препятствия - тонкого экрана.

## 7.2.2 2.2. Examples of Calculations of Diffraction by Spectral Method

### 2.2.1. Дифракция Френеля на амплитудной синусоидальной решётке

Функция пропускания решётки с периодом $d = 2\pi/\Omega$ (см. (3.4)):

$$t(x) = 1 + m\cos\Omega x \quad (m < 1)$$

Пусть решётка освещается плоской нормально падающей волной амплитуды $a$, т. е. $f_s(x) = ae^{ikz}$. Решётка установлена в плоскости $z = 0$,
поэтому $f_s = a$. Тогда, согласно граничным условиям (3.2),

$$f_0(x) = a(1 + m\cos\Omega x)$$

Здесь для нахождения пространственного спектра поля $f_0(x)$ нет необходимости использовать общее соотношение (3.46). С помощью тождества

$$\cos\Omega x = \frac{1}{2}e^{i\Omega x} + \frac{1}{2}e^{-i\Omega x}$$

находим

$$f_0(x) = a + \frac{am}{2}e^{i\Omega x} + \frac{am}{2}e^{-i\Omega x} \tag{3.53}$$

Комплексная амплитуда волны в плоскости $z$, то есть решение уравнения Гельмгольца, удовлетворяющего граничному условию (3.53), имеет (как ясно из соотношений (3.42) и (3.47)) вид

$$f(x,z) = ae^{ikz} + \frac{am}{2}e^{i\left(\Omega x + \sqrt{k^2-\Omega^2}z\right)} + \frac{am}{2}e^{i\left(-\Omega x + \sqrt{k^2-\Omega^2}z\right)} \tag{3.54}$$

т. е. каждое слагаемое в граничном поле $f_0(x)$ «ответственно» за «свою» волну в области $z > 0$. Первое слагаемое - волна с амплитудой $a$, бегущая вдоль оси $z$. Два других слагаемых - волны с амплитудами $\frac{am}{2}$ и пространственными частотами $\pm\Omega$. Эти волны бегут в направлениях $\sin\alpha = \pm\Omega/k = \pm\lambda/d$. Отметим, что в начале координат $x = 0$ в плоскости $z = 0_+$все три волны создают синфазные колебания.

Полагаем, что период решётки $d = \frac{2\pi}{\Omega}$ существенно больше длины волны $\lambda$ и, следовательно, частота решётки много меньше $k : \Omega \ll k$. Тогда, используя приближение (3.51), находим

$$f(x,z) = ae^{ikz} + \frac{am}{2}e^{ikz}\cdot e^{i\left(\Omega x - \frac{z}{2k}\Omega^2\right)} + \frac{am}{2}e^{ikz}\cdot e^{i\left(-\Omega x - \frac{z}{2k}\Omega^2\right)}$$

После простых преобразований получаем

$$f(x,z) = ae^{ikz}\left[1 + me^{-i\frac{z}{2k}\Omega^2}\cdot\cos\Omega x\right]$$

Последнюю формулу можно использовать для анализа картины дифракции на различных расстояниях $z$ от решётки.

1. На расстояниях $z_n$, для которых $\frac{z_n}{2k}\Omega^2 = 2\pi n$, т. е. $z_n = \frac{2d^2}{\lambda}n$, имеем $e^{-i\frac{z_n}{2k}\Omega^2} = 1$, поэтому $f(x,z) = ae^{ikz_n}(1 + m\cos\Omega x) = e^{ikz_n}f_0(x)$, т.е. с точностью до фазового множителя воспроизводится граничное поле $f_0(x)$. Воспроизводится, разумеется, наблюдаемая картина интенсивности $I(x,z_n) = I_0(x)$. Найдём её видность $\mathcal{V} = \frac{I_{\max}-I_{\min}}{I_{\max}+I_{\min}}$, полагая
глубину модуляции $m$ малой и пренебрегая членами порядка $m^2$. Тогда $I(x,z_n) = |f(x,z_n)|^2 \approx a^2(1 + 2m\cos\Omega x)$ и $\mathcal{V} \approx 2m$.

2. На расстояниях $z_n$, для которых $e^{-i\frac{z_n}{2k}\Omega^2} = i$, т. е. $\frac{z_n}{2k}\Omega^2 = \frac{\pi}{2} + 2\pi n$, имеем $f(x,z) = a(1 + im\cos\Omega x)$. Картина интенсивности $I(x,z_n) \approx a^2 = \text{const}$. Пренебрегая членами порядка $m^2$, получаем равномерную, не зависящую от $x$ засветку плоскости наблюдения $\mathcal{V} \approx 0$.

Таким образом, периодически по z изменяется видность наблюдаемой дифракционной картины. Причина этих изменений - в различии фазовых набегов трёх плоских волн, бегущих в области $z > 0$ от решётки: осевой волны, бегущей вдоль оси $z$, и двух боковых волн, бегущих в направлениях $\sin\alpha = \pm\Omega/k = \pm\lambda/d$.

### 2.2.2. Diffraction on a Phase Sinusoidal Grating

Функция пропускания (см. (3.5)):

$$t(x) = e^{im\cos\Omega x}$$

Будем полагать, что глубина модуляции фазы мала, т.е $m \ll 1$. Тогда

$$t(x) \approx 1 + im\cos\Omega x = 1 + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x}$$

При освещении этой решётки плоской нормально падающей волной амплитуды $a$ имеем за решёткой (при $z > 0$)

$$f(x,z) = ae^{ikz} + \frac{iam}{2}e^{i\left(\Omega x + \sqrt{k^2-\Omega^2}z\right)} + \frac{iam}{2}e^{i\left(-\Omega x + \sqrt{k^2-\Omega^2}z\right)}$$

Сравнивая с (3.54), видим, что три волны имеют те же амплитуды и направление распространения, что и в случае амплитудной решётки. Отличие лишь в начальных фазах: множитель $i = e^{i\pi/2}$ во втором и третьем слагаемых показывает, что колебания, созданные наклонными волнами (в начале координат $x = 0$ граничной плоскости $z = 0_+$), отстают по фазе на $\pi/2$ от колебания осевой волны (бегущей в направлении оси $z$).

## 7.2.3 2.3. Fresnel Diffraction on Periodic Structures. The Talbot Effect (Self-Production)

Примером периодической структуры является экран с периодически расположенными одинаковыми элементами, например, параллельными щелями одинаковой ширины $b$, расположенными на одинаковом расстоянии $d$ друг от друга. Пусть такой экран (решётка) освещается слева плоской нормально падающей волной (длина волны $\lambda$).

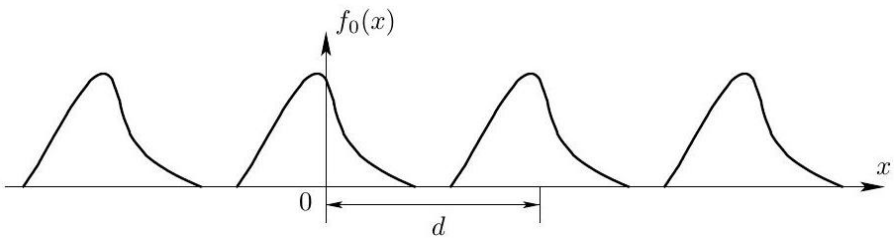



Рис. 3.26. Периодическая структура

Согласно граничным условиям Кирхгофа (3.2) на выходе из экрана (в плоскости, примыкающей к нему справа) получаем световое поле, комплексная амплитуда которого $f_0(x)$ является периодической функцией с периодом $d$ (рис. 3.26), которую можно представить в виде ряда Фурье:

$$f_0(x) = \sum c_n e^{in\frac{2\pi}{d}x} \tag{3.55}$$

Каждое слагаемое ряда (3.55) представляет собой поле плоской волны, пространственная частота которой $u_n = n\frac{2\pi}{d}$ определяет направление волнового вектора $\boldsymbol{k}_n$ ($u_n = k\sin\alpha_n$) этой волны, т. е. $\sin\alpha_n = n\frac{\lambda}{d}$. Согласно (3.47), комплексная амплитуда волны в плоскости наблюдения, отстоящей на расстоянии $z$ от решётки, имеет вид

$$f(x,z) = \sum c_n e^{i\left(u_n x + \sqrt{k^2-u_n^2}z\right)}$$

Фаза $n$-й плоской волны в плоскости $z = \text{const}$ равна $\varphi_n = \sqrt{k^2-u_n^2}z \approx kz - \frac{zu_n^2}{2k}$ (здесь использовано приближение (3.51)). Сравним набег фазы $n$-й плоской волны с набегом фазы $\varphi_0$ плоской волны, бегущей вдоль оси $z$ : $\varphi_0 = kz$. Мы получаем

$$\Delta\varphi_n = \varphi_0 - \varphi_n = \frac{z}{2k}\left(\frac{2\pi}{d}\right)^2 n^2 = \pi\frac{\lambda z}{d^2}n^2 \tag{3.56}$$

Рассмотрим плоскость наблюдения, отстоящую от плоскости решётки на расстояние

$$z_1 = \frac{2d^2}{\lambda} \tag{3.57}$$

(Отметим, что параметр $p = \frac{\sqrt{\lambda z_1}}{d}$ оказывается при этом порядка единицы, т. е. мы имеем дело с областью френелевской дифракции.) В этой плоскости имеем $\Delta\varphi_n = 2\pi n^2$, т.е. относительный набег фазы всех плоских волн кратен величине $2\pi$. Очевидно, что и разность фазовых набегов любых двух волн (с пространственными частотами $n_1 2\pi/d$ и $n_2 2\pi/d$), равная $2\pi\left(n_1^2 - n_2^2\right)$, также кратна $2\pi$. Но изменение разности фаз колебаний на величину, кратную $2\pi$, ничего не меняет в суммарном колебании. Мы пришли к замечательному результату: фазовые соотношения между слагаемыми плоскими волнами оказались одинаковыми как в плоскости, примыкающей к решётке (где сумма плоских волн (3.55) образовала граничное периодическое поле $f_0(x)$), так и в плоскости (3.57). Одинаковость (с точностью до величины, кратной $2\pi$) фазовых соотношений слагаемых плоских волн приводит к тому, что одинаков и результат интерференции этих плоских волн, т. е. световое поле в плоскости $z_1$ отличается от граничного поля $f_0(x)$ лишь постоянным фазовым множителем $e^{ikz}$:

$$f\left(x, z_1\right) = e^{ikz_1} f_0(x)$$

Мы наблюдаем в плоскости $z_1$ периодическую структуру, тождественно повторяющую граничное поле $f_0(x)$. Очевидно также, что такое восстановление изображения периодической структуры повторяется на расстояниях, кратных $z_1$:

$$z_m = m\frac{2d^2}{\lambda} \quad (m = 1, 2, \ldots) \tag{3.58}$$

Описанный эффект называют эффектом самовоспроизведения, или (по имени первооткрывателя) эффектом Талбота.

## 7.2.4 2.4. The Spectrum of Plane Waves in Slit Diffraction. The Uncertainty Relation

Пусть щель в непрозрачном экране (ширина щели $b$) освещается нормально падающей плоской волной единичной амплитуды (рис. 3.27). Тогда согласно (3.3) в плоскости $z = 0_+$, примыкающей к щели справа от неё, имеем

$$f_0(x) = \begin{cases} 1, & |x| \leqslant b \\ 0, & |x| > b \end{cases}$$

Пространственный спектр граничного поля находим с помощью (3.44)

$$C_0(u) = b\frac{\sin\frac{b}{2}u}{\frac{b}{2}u} \tag{3.59}$$

Последняя формула совпадает с (3.27) при $u = k\sin\theta$ - картиной дифракции Фраунгофера на щели.

Спектр $C_0(u)$ показан на рис. 3.28. Область значений $\Delta u$, в которой функция заметно отлична от нуля, называют обычно шириной спектра.

Как видно из графика, её можно оценить по ширине главного максимума функции $C_0(u)$

$$|\Delta u| \lesssim \frac{2\pi}{b} \tag{3.60}$$

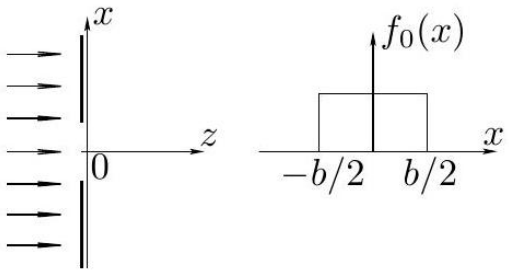



Рис. 3.27. Функция пропускания щели

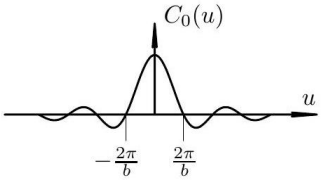



Рис. 3.28. Спектр плоских волн

В общем случае справедливо соотношение неопределённостей, связывающее протяжённость $\Delta x$ граничного поля с шириной $\Delta u$ его пространственного спектра [8]:

$$\Delta x \cdot \Delta u \sim 2\pi \tag{3.61}$$

Пространственная протяжённость граничного поля определяется характерным размером препятствия, в нашем примере - размером $b$ отверстия в непрозрачном экране. Разброс пространственных частот определяет разброс направлений слагаемых плоских волн за отверстием:

$$\Delta u = k\Delta(\sin\alpha)$$

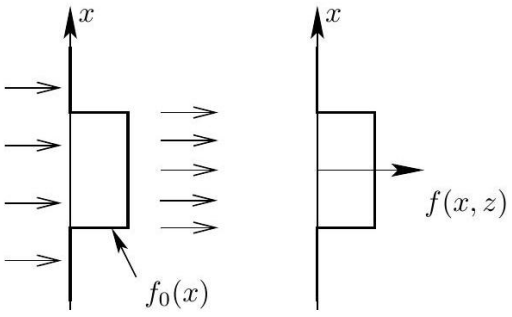
откуда $\Delta(\sin\alpha) \approx \lambda/b$ или, для малых углов,

$$\Delta\alpha \approx \frac{\lambda}{b}$$

Это и есть дифракиионная расходимость пучка света за отверстием размера $b$.

## 7.2.5 2.5. The Limit of Geometric Optics

### 2.5.1. The Condition of Applicability of Geometric Optics




[0]8 Математически задача не отличается от нахождения спектра прямоугольного импульса. Здесь роль длительности импульса $\tau$ играет ширина щели $b$, а роль частоты $\omega$ - пространственная частота $u$. Напомним о соотношении, связывающем длительность сигнала $\Delta t$ с шириной его спектра $\Delta\omega$ (соотношении неопределённостей): $\Delta t\Delta\omega \sim 2\pi$.

Рис. 3.29. Преобразование поля волны в условиях геометрической оптики

Хорошо известно, что в некоторых случаях (когда плоскость наблюдения отстоит недалеко от препятствия или когда размеры препятствия, например отверстия в непрозрачном экране, достаточно велики) никакой дифракции не наблюдается: картина в плоскости, недалеко отстоящей от препятствия, оказывается подобной картине в плоскости, примыкающей к препятствию, то есть поле $f(x,z)$ получается как бы геометрическим проецированием (с помощью лучей) граничного поля $f_0(x)$ - ситуация, показанная на рис. 3.29. Итак, выясним условия, когда выводы, основанные на представлениях геометрической (лучевой) оптики, справедливы.

Пусть щель в непрозрачном экране (ширина щели b) освещается плоской нормально падающей волной с амплитудой $a$. Тогда граничное поле $f_0(x)$ имеет вид, показанный на рис. 3.27, а его пространственный спектр $C(u)$ (фурье-преобразование функции $f_0(x)$) показан на рис. 3.28

Как мы уже отмечали, интервал $\Delta u$ пространственных частот плоских волн, имеющих заметную амплитуду, ограничен величиной $|\Delta u| < \frac{2\pi}{b}$. Эта оценка справедлива по порядку величины и в общем случае, как следует из соотношения неопределённостей.

Для плоской волны, бегущей вдоль оси $z(u = 0)$, набег фазы при распространении до плоскости наблюдения равен $kz$. Для волны с пространственной частотой $u$ он равен $\varphi(u) = \sqrt{k^2 - u^2} \cdot z$, или приближённо (см. (3.51)) $\varphi(u) \approx kz - \frac{z}{2k}u^2$. Разность фазовых набегов не превышает величины

$$\Delta\varphi < \frac{z}{2k}\left(\frac{2\pi}{b}\right)^2 = \pi\frac{\lambda z}{b^2} \tag{3.62}$$

Если эта разность много меньше величины $\pi$, то можно считать, что все плоские волны из интересующего нас интервала $|u| < \frac{2\pi}{b}$ приобретают один и тот же фазовый набег, равный $kz$. Следовательно, фазовые соотношения между плоскими волнами, образующими гранич ное поле $f_0(x)$, и теми же плоскими волнами в плоскости наблюдения $z = \text{const} > 0$ практически одинаковы. Значит, интерференция этих плоских волн даёт в плоскости наблюдения суммарное поле, которое (с точностью до фазового множителя $e^{ikz}$) тождественно повторяет граничное поле $f_0(x)$:

$$f(x,z) = e^{ikz}f_0(x)$$

При этом картина интенсивности $I(x,z)$ тождественно повторяет картину на границе $I(x,z) = I_0(x) = |f_0(x)|^2$, т. е. получается как бы геометрическим проецированием изображения щели на плоскость наблюдения.

Как следует из (3.62), условие $\Delta\varphi \ll \pi$ выполнено, если

$$p = \frac{\sqrt{\lambda z}}{b} \ll 1 \tag{3.63}$$

Это и есть условие применимости геометрической оптики. При любых фиксированных $z$ и $b$ это условие выполняется, если $\lambda \to 0$, а при любом фиксированном $\lambda$ законы геометрической оптики работают, как следует из неравенства (3.63), при достаточно малых $z$ и (или) при достаточно больших $b$.

### 2.5.2. Тонкий транспарант

Найдём условия, при которых транспарант можно считать тонким, то есть поля на входе и выходе транспаранта можно считать связанными граничным условием (3.3). Пусть толщина транспаранта (расстояние между плоскостью, примыкающей к транспаранту слева $z = 0$, и плоскостью $z = 0_+$, примыкающей к нему справа) равна $L$.

Рассмотрим тонкий луч света диаметром $b$, пронизывающий транспарант по нормали (рис. 3.30).

Для того чтобы прохождение луча могло быть описано законами геометрической оптики (как того требует граничное условие (3.3)), необходимо выполнение двух условий.

Во-первых, должна быть мала ди-

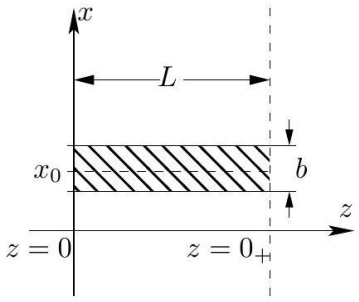



Рис. 3.30. Луч, проходящий через тонкий транспарант фракционная расходимость луча при прохождении транспаранта условие геометрической оптики (3.63):

$$b \ll \sqrt{\lambda L} \tag{3.64}$$

И, во-вторых, луч должен быть «тонким»: амплитуда и фаза волны в сечении луча должны слабо меняться при прохождении транспаранта. Выразим это условие количественно. Пусть $t(x) = a(x)e^{i\varphi(x)}$ - комплексная пропускаемость транспаранта, где $a(x)$ - амплитудная прозрачность, $\varphi(x) = k\,\mathrm{Ln}(x)$ - набег фазы. Разность амплитуд и фаз в сечении луча после прохождения транспаранта (между точками $x = x_0 - b/2$ и $x = x_0 + b/2$ в плоскости $z = 0_+$на рис. 3.30) составит $\Delta a \approx \frac{da}{dx}b$ и $\Delta\varphi \approx \frac{d\varphi}{dx}b = kLb\frac{dn}{dx}$ соответственно. Условие их малости можно записать в виде

$$\Delta a \approx \frac{da}{dx}b \ll a, \quad \Delta\varphi \approx kLb\frac{dn}{dx} \ll \pi \tag{3.65}$$

Условие (3.64) совместно со вторым неравенством (3.65) даёт

$$\sqrt{\lambda L} \ll b \ll \frac{\lambda}{2L\frac{dn}{dx}}$$

Следовательно, достаточно требовать, чтобы толщина транспаранта $L$ и неоднородность показателя преломления $\frac{dn}{dx}$ удовлетворяли условию

$$\frac{dn}{dx} \ll \frac{1}{L}\sqrt{\frac{\lambda}{L}} \tag{3.66}$$

### 2.5.3. Что такое параллельный пучок света

Расположим на пути наклонной плоской волны $f_s(x) = ae^{ik\sin\alpha\cdot x}$ с пространственной частотой $u_0 = k\sin\alpha$ диафрагму-щель ширины $b$ (рис. 3.31). В плоскости $z = 0_+$, примыкающей к щели справа, получаем

$$f_0(x) = \begin{cases} ae^{ik\sin\alpha\cdot x} & \text{при } |x| \leqslant \frac{b}{2} \\ 0 & \text{при } |x| > \frac{b}{2} \end{cases}$$

Таким образом, за щелью в области $z > 0$ вместо идеализированной плоской волны, амплитуда колебаний которой во всех точках волновой поверхности (а следовательно, при всех $-\infty < x < \infty$ отлична от нуля), мы получаем параллельный пучок света,

направление которого (угол $\alpha$) совпадает с направлением плоской волны, падающей на щель-диафрагму, поперечное сечение пучка есть $|x| \leqslant \frac{b}{2}$. Ясно, что пучок света ограниченной протяжённости - это уже не плоская волна с пространственной частотой $u_0$, это - пакет плоских волн, который образован плоскими волнами с пространственными частотами

$$C_0(u) = \frac{\sin\left[\frac{b}{2}(u-u_0)\right]}{\frac{b}{2}(u-u_0)} \tag{3.67}$$

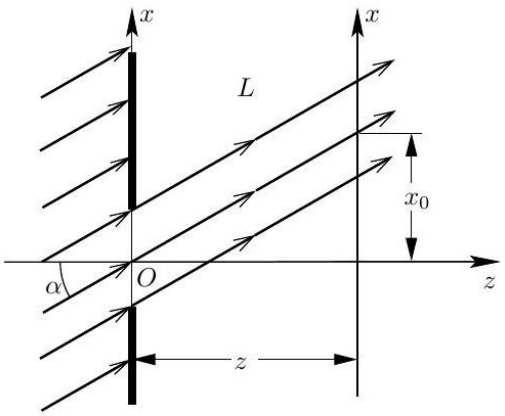



Рис. 3.31. Наклонный параллельный пучок

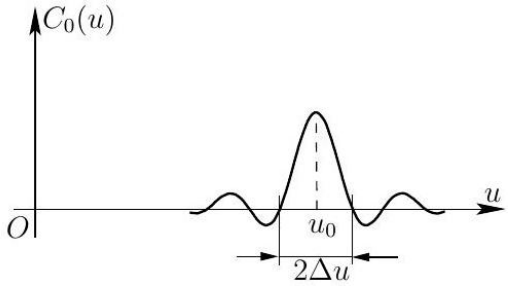



Рис. 3.32. Пространственный спектр параллельного пучка

($C_0(u)$ - фурье-образ поля $f_0(x)$, рис. 3.32). Итак, пропустив плоскую волну через диафрагму шириной $b$, мы получаем пакет плоских волн с несущей частотой $u_0$ (центральной частотой пакета) и разбросом частот $|\Delta u| \leqslant \frac{2\pi}{b}$. Заметим, что при смещении щелевой диафрагмы в плоскости $z = 0$ вверх на расстояние $x_0$ мы получим спектр

$$C_1(u) = C_0(u)e^{-i(u-u_0)x_0} \tag{3.68}$$

(Смещение функции $f_0(x)$ по координате $x$ на величину $x_0$ приводит к умножению её спектра на линейный фазовый множитель $e^{-i(u-u_0)x_0}$.)

Что происходит с волновым пакетом, имеющим пространственный спектр $C_0(u)$ при распространении в области $z > 0$? На расстоянии $z$ набег фазы для несущей частоты пакета $u_0$ равен $\varphi_0 \approx kz - \frac{z}{2k}u_0^2$. Для волны с пространственной частотой $u_0 + \Delta u$ (напоминаем, нас интересует область частот $|\Delta u| < \frac{2\pi}{b}$) набег фазы равен

$$\varphi = kz - \frac{z}{2k}(u_0 + \Delta u)^2 = kz - \frac{z}{2k}u_0^2 - \frac{z}{k}u_0 \cdot \Delta u - \frac{z}{2k}(\Delta u)^2.$$

Разность фазовых набегов

$$\Delta\varphi = \varphi_0 - \varphi = \frac{z}{k}u_0 \cdot \Delta u + \frac{z}{2k}(\Delta u)^2$$

не превышает величины

$$\Delta\varphi = z\sin\alpha \cdot \Delta u + \pi\frac{\lambda z}{b^2}$$

Если условие (3.63) геометрической оптики выполнено, то, отбросив второе слагаемое, получим

$$\Delta\varphi = z \cdot \sin\alpha\,(u - u_0)$$

Таким образом, спектр пакета на расстоянии $z$ от диафрагмы есть

$$C(u) \sim C_0(u)e^{-i\frac{z}{2k}u_0^2}e^{-iz\sin\alpha(u-u_0)}$$

Последнее равенство означает, что форма волнового пакета остаётся неизменной, т. е. параллельный пучок света, распространяясь до плоскости $z$, остаётся параллельным пучком с неизменным поперечным сечением и направлением $\alpha$, лишь приобретает набег фазы (такой же, как у плоской волны с пространственной частотой $u_0$) и смещаясь в поперечном направлении (вдоль оси $x$) на величину $x_0 = z\sin\alpha$ (сравните с (3.68).

Если условие (3.63) не выполнено, то имеет место пространственная дисперсия: фазовые набеги разных плоских волн (разных пространственных частот), образующих волновой пакет, оказываются различными - волновой пакет «расплывается». В частности, на расстояниях $z \gg \frac{b^2}{\lambda}$ (область дифракции Фраунгофера) область значений $\Delta x$, где амплитуда колебаний отлична от нуля, определяется шириной главного максимума дифракционной картины $|\Delta x| = \frac{\lambda}{b}z \gg b$. Таким образом, термин «параллельный пучок света» становится бессмысленным, если речь идёт о распространении волны конечного поперечного сечения $|x| < b$ на расстояния, не удовлетворяющие условию $z \ll \frac{b^2}{\lambda}$.

## 7.2.6 2.6. The Field in the Focal Plane of the Lens. Spatial Transformation. Fourier Analysis

Как известно, линза фокусирует параллельный пучок света: плоская волна, бегущая в направлении $\alpha$ (рис. 3.33), т. е. имеющая пространственную частоту $u = k\sin\alpha$, фокусируется линзой в точку фокальной плоскости с координатой

$$x = f\,\mathrm{tg}\,\alpha \approx f\sin\alpha = \frac{fu}{k} \tag{3.69}$$

Имеется, как мы видим, взаимно однозначное соответствие между точками фокальной плоскости и пространственными частотами плоских волн, которые в эти точки фокусируются. Оказывается, взаимно однозначное соответствие имеет место также и между амплитудами и фазами колебаний в точках фокальной плоскости и соответствующих им плоских волн.

Пусть на линзу падает произвольная волна. Во входной плоскости $\Pi_1$, отстоящей от линзы на расстоянии $z$, волна имеет комплексную амплитуду $f_0(x)$. Представим эту волну в виде суперпозиции плоских волн разных направлений $\alpha_n$, т. е. комплексную амплитуду этой волны
представим в виде (3.42). Плоская волна $c_n e^{iu_n x}$, соответствующая одному из слагаемых, показана на рис. 3.33: её волновой вектор составляет угол $\alpha = \alpha_n\left(\sin\alpha_n = \frac{u_n}{k}\right)$ с оптической осью. Коэффициент $c_n = a_n e^{i\varphi_n}$ определяет амплитуду плоской волны $a_n$ и

начальную фазу $\varphi_n$ (фазу колебаний в точке $\xi = 0$ плоскости $\Pi_1$; такая же и фаза колебаний в точке $P_1$, лежащей на той же волновой поверхности $AA'$).

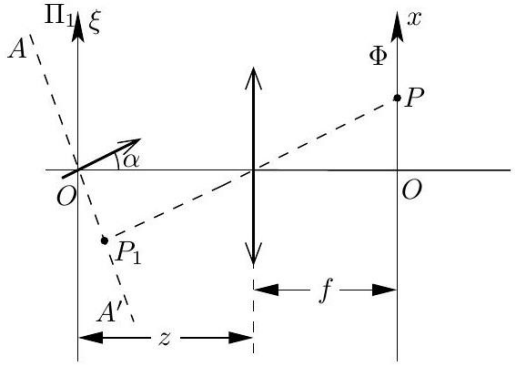



Рис. 3.33. Плоская волна фокусируется в точку фокальной плоскости

Очевидно, что амплитуда колебаний в точке $P$ фокальной плоскости пропорциональна амплитуде $a_n$ плоской волны, которая в эту точку сфокусировалась. Найдём, каковы фазовые соотношения между колебаниями в разных точках фокальной плоскости. Ясно, что колебание в точке $P$ отстаёт по фазе от колебания в точке $P_1$ (а значит, и в точке $\xi = 0$ входной плоскости), причём разность фаз определяется длиной оптического пути $P_1P$: прямая $P_1P$ перпендикулярна в каждой точке волновым поверхностям - плоским поверхностям в волне, падающей на линзу, и сферическим в волне, прошедшей через линзу. (Оптический путь $P_1P$ включает в себя и часть пути, проходящего непосредственно в линзе.) Пусть соответствующая пути $P_1P$ задержка по фазе равна $\psi_n$. Тогда фаза колебания в точке $P$ фокальной плоскости равна $\varphi_n + \psi_n$, а комплексную амплитуду колебаний в этой точке можно записать в виде $f(x_n) \sim a_n e^{i(\varphi_n+\psi_n)} = c_n e^{i\psi_n}$.

Слагаемому $c_0 e^{iu_0 x}$ (при $u_0 = 0$) соответствует плоская волна, бегущая вдоль оптической оси ($\alpha = 0$). Она фокусируется в начало координат $x = 0$ фокальной плоскости Ф. Коэффициент $c_0 = a_0 e^{i\varphi_0}$ определяет амплитуду $a_0$ этой волны и её начальную фазу $\varphi_0$, т. е. фазу колебаний в точке $O$ входной плоскости $\Pi_1$. Задержка по фазе $\psi_0$ в точке $x = 0$ фокальной плоскости Ф, куда эта волна сфокусировалась, определяется длиной оптического пути $OO$. Комплексная амплитуда колебаний в этой точке есть $f(0) \sim a_0 e^{i(\varphi_0+\psi_0)} = c_0 e^{i\psi_0}$.

Пусть $z = f$, т. е. плоскость $\Pi_1$ - это передняя фокальная плоскость линзы. Легко видеть, что в приближении малых углов ($\cos\alpha \approx 1 - \alpha^2/2$) оптические пути $P_1P = z\cos\alpha + \frac{f}{\cos\alpha} + \Delta_0$ и $OO = z + f + \Delta_0$ (где $\Delta_0$ оптический путь, проходящий непосредственно через линзу на главной оптической оси) равны при $z = f$, т.е. $\psi_n = \psi_0$ (для любой $n$-й плоской волны в сумме (3.42)).

Одинаковость для всех плоских волн фазовых задержек $\psi_n$ означает, что фазовые соотношения между колебаниями в разных точках задней фокалъной плоскости, куда эти волны сфокусировались, таковы же, как и фазовые соотношения между колебаниями, которые создают эти волны в начале координат передней фокалъной плоскости. Таким образом, волновое поле в задней фокальной плоскости линзы правильно воспроизводит не только амплитудные соотношения между плоскими волнами разных пространственных частот, но и фазовые соотношения без искажений, т. е. картина поля в фокальной плоскости воспроизводит пространственный спектр (пространственное преобразование Фуръе) падающей на линзу волны.

Опуская несущественный постоянный фазовый множитель $e^{i\psi_0}$ (одинаковый для всех слагаемых плоских волн разных пространственных частот $u_n$), получаем $f(x_n) = c_n = C\left(\frac{kx_n}{f}\right)$. Последнее равенство справедливо для любой точки $x_n$ фокальной плоскости.

Если спектр плоских волн непрерывен, т.е. волна, падающая на линзу, состоит из плоских волн любых направлений, то, опуская индекс $n$, находим

$$f(x) = C\left(\frac{kx}{f}\right) = \int f_0(\xi) e^{-i\frac{kx}{f}\xi} d\xi \tag{3.70}$$

где $C(u)$ - преобразование Фурье поля $f_0(x)$. Итак, световое поле в задней фокальной плоскости линзы $f(x)$ связано с полем волны, падающей на линзу $f_0(\xi)$, преобразованием Фурье.

Если комплексная амплитуда волны, падающей на линзу, задаётся в произвольной плоскости на расстоянии $z \neq f$ от линзы, то легко получить (упражнение для читателя)

$$f(x) = e^{i\frac{k}{2f}\left(1-\frac{z}{f}\right)x^2} \cdot C\left(\frac{kx}{f}\right)$$

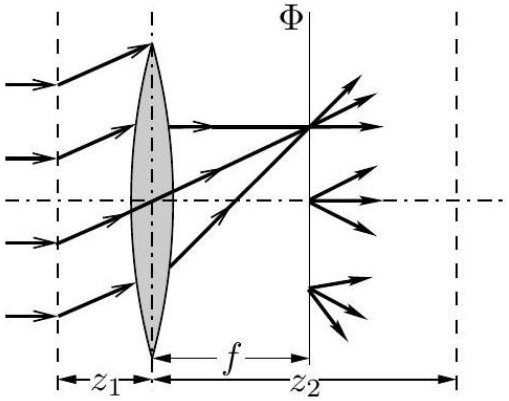
т.е. возникают фазовые искажения, обусловленные дополнительным набегом фазы $\varphi(u) = \sqrt{k^2-u^2}\Delta z$, который приобретает плоская волна, пробегая промежуток свободного пространства, равный $\Delta z = z - f$. Важно обратить внимание, что наблюдаемая картина интенсивности

$$I(x) = \left|C\left(\frac{kx}{f}\right)\right|^2$$

не зависит от $z$ (т. е. не зависит от того, на каком расстоянии от линзы задана комплексная амплитуда падающей волны).

## 7.2.7 2.7. The Principle of Double Diffraction and the Formation of Optical Images (Abbe's Theory)

Формирование изображения с помощью линзы можно рассматривать, основываясь на идее пространственного спектрального разложения. Монохроматическую волну, идущую от предмета, представим в виде суперпозиции плоских волн разных направлений $\alpha$, т. е. разных пространственных частот $u = k\sin\alpha$. Каждая гармоника - плоская волна определённого направления - фокусируется линзой в «свою» точку фокальной плоскости, в которой возникает, таким образом, картина пространственного спектра: амплитуда и фаза колебаний в точке $\xi$ фокальной плоскости однозначно определяются амплитудой и фазой колебаний той




плоской волны, которая в эту точку

Рис. 3.34. Дифракционная картина в фурье-плоскости фокусируется.

По этой причине фокальную плоскость линзы называют фурье-плоскостью. По терминологии Аббе, впервые предложившего такой подход, поле в фокальной плоскости называют первичным изображением. На рис. 3.34 показана ситуация, когда предметом

является решётка, освещаемая плоской нормально падающей волной. При этом в фурьеплоскости, как мы знаем, возникает картина фраунгоферовой дифракции (см. п. 1.5.1): набор ярких точек - дифракционных максимумов. Итак, в процессе распространения света от предмета до фурьеплоскости осуществляется преобразование Фурье светового поля (по терминологии Аббе - первая дифракция).

Далее каждая точка фурье-плоскости (каждый дифракционный максимум) рассматривается как источник сферической волны. Все сферические волны, исходящие из разных точек фурье-плоскости, интерферируя, образуют в плоскости, находящейся на расстоянии $z_2$ за линзой, собственно изображение объекта. Это изображение Аббе назвал вторичным, а процесс распространения света от фурье-плоскости до плоскости изображения - второй дифракцией.

Особенно наглядно принцип двойной дифракции проявляется в оптической схеме, показанной на рис. 3.35. Схема состоит из двух линз

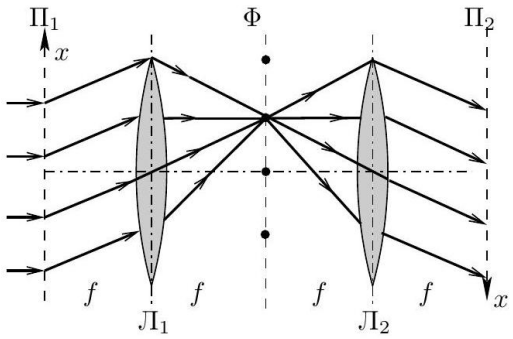



Рис. 3.35. Двойная дифракция Аббе

с общей фокальной плоскостью Ф. Задняя фокальная плоскость линзы $\Omega_1$ совпадает с передней фокальной плоскостью линзы $\Omega_2$. В этом случае первая дифракция - это распространение света от передней фокальной плоскости линзы $_1$, где расположен предмет, к плоскости Ф, где возникает картина пространственного спектра - первичное изображение. Далее, сферическая волна, идущая из любой точки фурьеплоскости, преобразуется линзой $\Omega_2$ в плоскую волну. Таким образом, каждая плоская волна, идущая от предмета, преобразуется системой двух линз в плоскую волну, приходящую к плоскости изображения (задней фокальной плоскости линзы $_2$). Причём, как видно из рис. 3.35, если фокусные расстояния линз одинаковы, то волна с пространственной частотой $u = k \sin \alpha$ преобразуется в волну с пространственной частотой «-u» ($x$-компонента вектора $\boldsymbol{k}$ изменяет знак). Это приводит к инверсии - изображение оказывается перевёрнутым. Можно сказать, что в процессе образования изображения происходит два последовательных преобразования Фурье: от входной плоскости $\Pi_1$ к фурье-плоскости первая дифракция, и затем от фурье-плоскости с помощью линзы $_2$ к выходной плоскости $\Pi_2$ - вторая дифракция.

## 7.2.8 2.8. Spatial Filtering

Особая роль фурье-плоскости обусловлена тем, что именно в этой плоскости возможно избирательное воздействие на разные пространственные гармоники: установив в любой точке $x$ фурье-плоскости маленькую пластинку, вносящую определённое поглощение и (или) фазовую задержку, мы изменим амплитуду и (или) фазу плоской волны с пространственной частотой $u = \frac{kx}{f}$, не изменяя амплитуд и фаз других плоских волн. Устанавливая в фурье-плоскости различные амплитудно-фазовые маски, можно направленно изменять пространственный спектр изображения, влияя таким образом на его характеристики. Этим путём можно решать самые разнообразные задачи:

улучшение качества изображений, разрешающей способности оптических систем, визуализация фазовых объектов, выполнение самых разнообразных преобразований пространственной структуры световых полей и т.д., т.е. решать широкий круг задач оптической обработки информации.

### 2.8.1. Animation (Reproduction) of an Image

Расположим в фурье-плоскости Ф оптической системы (рис. 3.35) фильтрующий транспарант-решётку с узкими щелями и периодом $d$. Во входной плоскости $\Pi_1$ находится объект-транспарант с функцией пропускания $f_0(x)$, освещённый слева плоской нормально падающей волной. В плоскости, примыкающей к объекту справа, возникает световое поле, комплексная амплитуда которого $f_0(x)$ представляется в общем случае непрерывной суммой плоских волн - интегралом Фурье:

$$f_0(x) = \frac{1}{2\pi}\int C_0(u)e^{iux}dx$$

Спектр объекта $C_0(u)$ находится с помощью (3.44). На щелях фильтрующей решётки, расположенных в точках фурье-плоскости $x_n = nd$, фокусируются плоские волны с пространственными частотами

$$u_n = k\sin\alpha_n = k\frac{x_n}{f} = \frac{knd}{f} = 2\pi n\frac{d}{\lambda f}$$

(см. (3.69)), т. е. волны с кратными частотами $u_n = nu_0$, где $u_0 = 2\pi\frac{d}{\lambda f}$, и спектральными амплитудами $C_0(nu_0)$.

Только эти волны, отфильтрованные (т. е. пропущенные) решёткой, формируют изображение $f(x)$, возникающее в выходной плоскости $\Pi_2$ оптической системы:

$$f(x) \sim \sum C_0(nu_0)\, e^{inu_0x}$$

Таким образом, из непрерывного спектра $C_0(u)$ объекта, показанного на рис. 3.36 пунктиром, фильтрующая решётка пропускает дискретный спектр компонент, показанный стрелками. Спектральные амплитуды $C_0(nu_0)$ компонент, формирующих изображение, очевидно, пропорцио нальны (см. (3.44)) величинам

$$C_0(nu_0) = \int\limits_{-\frac{a}{2}}^{\frac{a}{2}} f_0(x)e^{-inu_0x}dx \tag{3.71}$$

($a$ размер объекта: $f_0(x) \equiv 0$ при $|x| > \frac{a}{2}$).

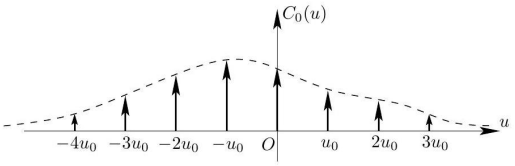



Рис. 3.36. Дискретный спектр, полученный в результате фильтрации

Из сравнения (3.71) с формулой (3.46), определяющей коэффициенты $c_n$ ряда Фурье периодической функции, можно сделать вывод: изображение, возникающее в плоскости $\Pi_2$, представляет собой периодически повторяющееся с периодом $d_0 = \frac{\lambda f}{d}$ изображение объекта $f_0(x)$. Соседние элементы периодической структуры $f(x) = \sum f_0(x - nd_0)$ не налагаются друг на друга при условии $d_0 > a$. Число элементов $N$ размноженного

изображения определяется шириной главного максимума картины дифракции Фраунгофера на отдельной щели решётки $|x| < \frac{\lambda f}{b}$ ($b-$ ширина щели) : $N \approx \frac{2b}{d_0}$.

### 2.8.2. Visualization of Phase Objects. The Phase Contrast Method And The Dark Field Method

Рассмотрим проблему визуализации фазовых объектов, которую можно решить, используя метод фазового контраста, предложенный Цернике. Пусть фазовый объект - тонкая прозрачная пластинка, имеющая разный в разных точках показатель преломления (или толщину), но не изменяющая амплитуду прошедшей волны, находится во входной плоскости $\Pi_1$ оптической системы, показанной на рис. 3.35. Функция пропускания такой пластинки $t(x) = e^{i\varphi(x)}$, где $\varphi(x) = kn(x)d(d$ - толщина, $n(x)$ - распределение показателя преломления). При освещении пластинки (фазового объекта) плоской нормально падающей волной комплексная амплитуда волны в плоскости, примыкающей к пластинке справа, согласно (3.3), есть $f_0(x) = e^{i\varphi(x)}$. Если оптическая система
идеальна, то комплексная амплитуда в выходной плоскости $\Pi_2$ тождественно повторяет (с точностью до инверсии) входное поле $f_0(x)$, а наблюдаемая картина интенсивности $I_0(x) = |f_0(x)|^2 \equiv 1$, т. е. в плоскости $\Pi_2$ мы наблюдаем равномерную засветку: информация о фазовой структуре предмета потеряна, фазовый объект невидим.

Для визуализации фазового объекта Цернике предложил установить в фурье-плоскости, на оптической оси, маленькую фильтрующую пластинку, которая, не изменяя амплитуды прошедшей волны, вносит фазовую задержку, равную $\pi/2$. Проанализируем структуру светового поля в выходной плоскости $\Pi_2$, рассмотрев в качестве примера объект - фазовую синусоидальную решётку с малой глубиной модуляции ($m \ll 1$). В этом случае

$$f_0(x) = e^{im\cos\Omega x} \approx 1 + im\cos\Omega x = 1 + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x} \tag{3.72}$$

Итак, входное поле представляется в виде суммы трёх слагаемых, в соответствии с этим от входной плоскости $\Pi_1$ вправо распространяются три плоские волны. Первое слагаемое ответственно за появление плоской волны единичной амплитуды $f_1(x,z) = e^{ikz}$, бегущей вдоль оси оптической системы (оси $z$). Второе и третье слагаемые - плоские волны с амплитудой $m/2$, направления распространения которых составляют углы $\pm\alpha$ с оптической осью, где $\sin\alpha = \pm\Omega/k$ (т. е. $\Omega$ и $-\Omega$ - пространственные частоты этих волн). Обратим внимание, что в точке $x = 0$ входной плоскости колебание первой волны отличается по фазе на $\pi/2$ от колебаний двух наклонных волн (т. к. $i = e^{i\pi/2}$).

Три слагаемых в (3.72) можно изобразить в виде векторов. На рис. 3.37а показано положение этих векторов в начале координат ($x = 0$). Горизонтальный вектор единичной длины изображает колебание, созданное осевой волной, два вектора длины $m/2$, повёрнутые на угол $\pi/2$, - колебания боковых волн. При смещении из точки $x = 0$ перпендикулярно оси $z$ (вдоль оси $x$) фаза колебания осевой волны не меняется, поэтому соответствующий вектор остаётся горизонтальным. Фазы боковых (наклонных) волн изменяются: $\varphi = \pm\Omega x$, поэтому изображающие векторы поворачиваются (по и против часовой стрелки). На рис. 3.37б показано положение векторов в точке, где $\Omega x = \pi/2$: векторы повернулись (из начального положения) на угол $\pm\pi/2$, а на рис. 3.37 в в точке $x$, где $\Omega x = \pi$, векторы повернулись на угол $\pm\pi$. При смещении на расстояние, равное периоду решётки ($\Omega x = 2\pi$), восстанавливается исходное расположение векторов (рис. 3.37г). Легко видеть, что суммарный вектор, не меняя своей длины (с точностью до величины порядка $m^2$), изменяет угол наклона $\varphi$ от $-m$ до $+m$, что и соответствует фазовой структуре волнового поля.

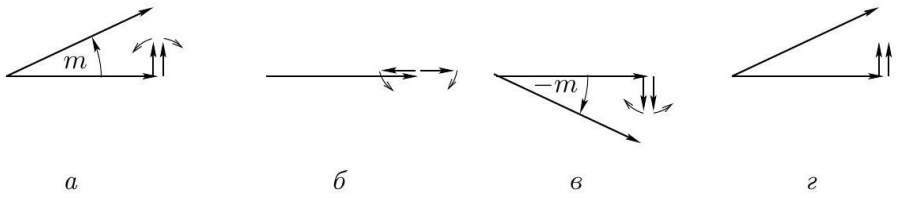



Рис. 3.37. Векторные диаграммы для фазовой решётки

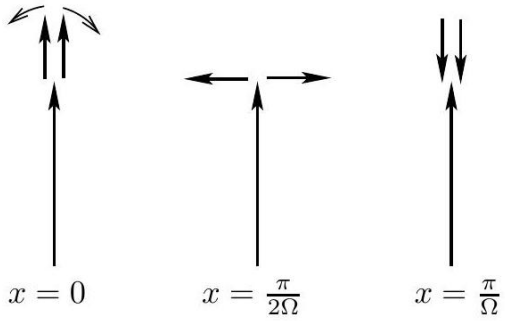



Рис. 3.38. Векторные диаграммы метода фазового контраста

Осевая плоская волна, фокусируясь линзой в начало координат фурье-плоскости ($\xi = 0$), проходит через фазовую фильтрующую пластинку, а две наклонные волны, фокусируясь в точки $\xi_{1,2} = \pm f\alpha = \pm f\frac{\Omega}{k}$, не «задевают» пластинку. Далее линза $_2$ преобразует сферические волны, исходящие из точек $\xi = 0$ и $\xi_{1,2} = f\frac{\Omega}{k}$, в плоские волны, которые, интерферируя, образуют изображение.

Наличие маленькой фазовой пластинки в фурье-плоскости на оптической оси приводит к относительной фазовой задержке в $\pi/2$ осевой волны (относительно боковых наклонных волн), поэтому поле в выходной плоскости можно записать в виде

$$f(x) = e^{i\frac{\pi}{2}} + \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x}$$

или

$$f(x) = i\left(1 + \frac{m}{2}e^{i\Omega x} + \frac{m}{2}e^{-i\Omega x}\right) = i(1 + m\cos\Omega x)$$

Изменение фазовых соотношений между осевой и наклонными волнами иллюстрируют векторные диаграммы, показанные на рис. 3.38.

Поворот вектора единичной длины (изображающего колебания осевой волны) на $\pi/2$ приводит к тому, что в точке $x = 0$ плоскости изображения все три изображающих вектора оказываются коллинеарными. При смещении из начала координат изменяется фаза боковых волн.

Как видно из векторных диаграмм, направление суммарного вектора остаётся неизменным, а его длина при смещении по координате $x$ изменяется, что и соответствует чисто амплитудной структуре, т. е. полю с плоским волновым фронтом и меняющейся от точки к точке амплитудой. Таким образом, метод фазового контраста позволяет преобразовать исходную фазовую решётку в амплитудную решётку в плоскости изображения.

Наблюдаемая картина интенсивности имеет вид

$$I(x) = |f(x)|^2 \approx (1 + m\cos\Omega x)^2 \approx 1 + 2m\cos\Omega x$$

(так как $m \ll 1$, членами порядка $m^2$ и выше мы пренебрегаем). Итак, фазовые изменения входного поля $f_0(x)$ оказались визуализированы: мы наблюдаем изменения

интенсивности, повторяющие изменения фазы входного поля. Обратите внимание на аналогию описанного здесь метода фазового контраста с методом преобразования колебания, модулированного по фазе, в амплитудно-модулированное колебание (т. н. приём с изменением фазы несущей).

Читатель может самостоятельно проанализировать «метод тёмного поля». В методе тёмного поля вместо фазовой пластинки в $\pi/2$ в фурье-плоскости на оптической оси устанавливается непрозрачный маленький экран. Осевая плоская волна, фокусируясь линзой в начало координат фурье-плоскости, поглощается непрозрачным экраном и не участвует в формировании изображения. Боковые же волны остаются без изменения. Поле в выходной плоскости в этом случае имеет вид

$$f(x) = \frac{im}{2}e^{i\Omega x} + \frac{im}{2}e^{-i\Omega x} = im\cos\Omega x$$

а картина интенсивности -

$$I(x) = m^2\cos^2\Omega x$$

Метод тёмного поля аналогичен методу, который в радиотехнике используется для преобразования фазовой модуляции в амплитудную и называется «приёмом без несущей».

## 7.2.9 2.9. The Resolution of Optical Systems

Во всякой реальной оптической системе линзы имеют конечные размеры. Учёт конечных размеров используемых объективов чрезвычайно важен для оценки предельных возможностей оптических систем. Если диаметр линзы $D$, то её функция пропускания определяется формулой (3.8). То есть всё происходит так, как если бы идеальная бесконечная линза (функция пропускания которой описывается формулой (3.7)) была задиафрагмирована непрозрачным экраном с отверстием диаметра $D$, оставляющим открытой центральную часть линзы.

### 2.9.1. Дифракционное изображение удалённого источника

Итак, пусть на линзу падает плоская волна, распространяющаяся вдоль оптической оси. Волна излучается удалённым источником, настолько удалённым, что в пределах площади зрачка, т. е. в пределах диафрагмы диаметра $D$ (рис. 3.39), волновой фронт неотличим от плоского.

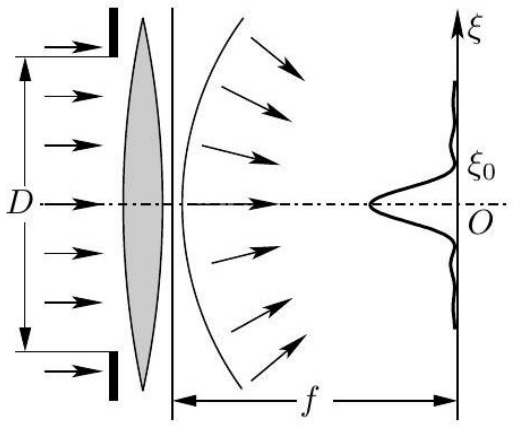



Рис. 3.39. Дифракционное изображение удалённого источника

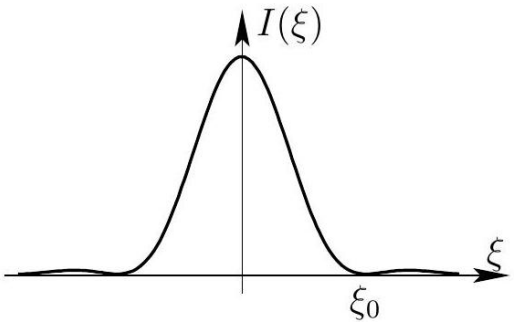



Рис. 3.40. Распределение интенсивности изображения удалённого источника

Пройдя через диафрагму, волна падает на идеальную (бесконечную) линзу. Изображённая на рис. 3.39 оптическая схема есть не что иное, как схема наблюдения дифракции Фраунгофера плоской волны на круглом отверстии диаметра $D$. Волна, пройдя через отверстиедиафрагму, уже не является плоской волной: за отверстием мы имеем «пакет плоских волн» (т.е. набор плоских волн разных направлений). Каждая плоская волна из этого пакета фокусируется идеальной линзой в свою точку фокальной плоскости, а результирующее распределение интенсивности, показанное на рис. 3.40, - дифракционное пятно Эйри (в котором концентрируется подавляющая доля светового потока), окружённое чередующимися светлыми (очень слабыми) и тёмными кольцами. Полуширина пятна Эйри (см. (3.37)):

$$\xi_0 = 1,22\frac{\lambda}{D}f \tag{3.73}$$

Это и есть изображение бесконечно удалённого точечного источника.

### 2.9.2. Дифракционное изображение точечного источника

Если точечный источник света находится на конечном расстоянии $z_1$ от линзы, то дифракционное пятно Эйри возникает не в фокальной плоскости, а на расстоянии $z_2$ за линзой, где $z_1$ и $z_2$ связаны формулой линзы: $\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}$. При этом полуширина пятна в дифракционном изображении точечного источника определяется формулой

$$\xi_0 = 1,22\frac{\lambda}{D}z_2 \tag{3.74}$$

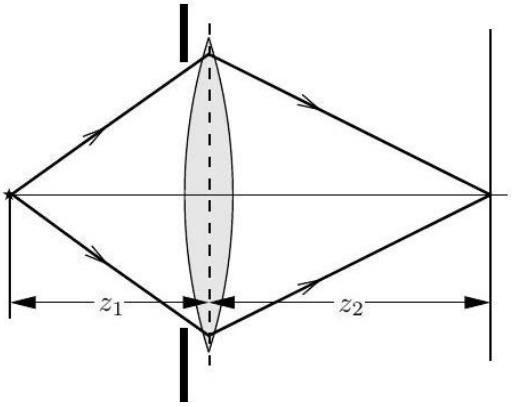



Рис. 3.41. Изображение точечного источника

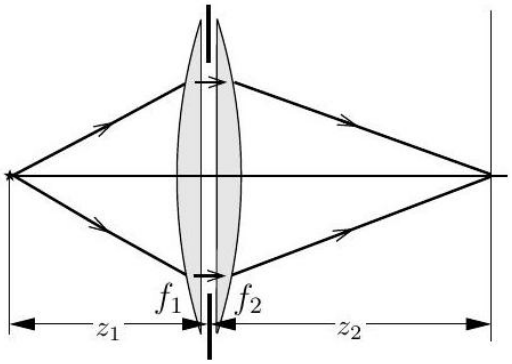



Рис. 3.42. Вспомогательная модель

Это утверждение можно пояснить, если в качестве модели реальной оптической схемы на рис. 3.41 рассмотреть оптическую систему, показанную на рис. 3.42: линзу с фокусным расстоянием $f$ заменим системой двух линз $_1$ и $_2$ с фокусными расстояниями $f_1 = z_1$ и $f_2 = z_2$. Поскольку оптические силы двух вплотную расположенных линз складываются (см. раздел I), то должно выполняться равенство $\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{z_1} + \frac{1}{z_2}$, а это есть не что иное, как формула линзы, формирующей изображение на рис. 3.41, то есть модель рис. 3.42 действительно эквивалентна системе рис. 3.41. Первая линза $_1$ преобразует сферическую волну от источника, расположенного в её фокусе, в плоскую волну. Эта волна затем испытывает дифракцию на диафрагме диаметром $D$, после чего падает на вторую линзу $_2$. Тогда, как следует из приведённого ранее анализа, в фокусе второй линзы (на расстоянии $z_2$) образуется изображение - дифракционное пятно, размер которого определяется формулой (3.74).

### 2.9.3. Разрешающая способность. Критерий Рэлея

Если источник сместить на малое расстояние $\ell$ перпендикулярно оптической оси, то изображение пятна Эйри также сместится. Центр пятна окажется в точке, расположенной на

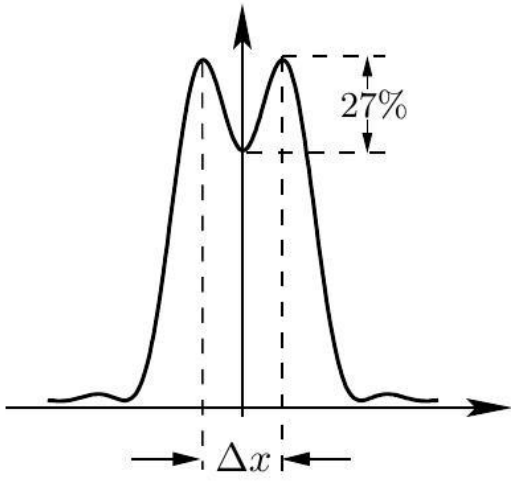



Рис. 3.43. Разрешение двух пятен Эйри расстоянии $\Delta x = -\frac{z_2}{z_1}\ell$ от оптической оси.

Пусть во входной плоскости (на расстоянии $z_1$ от линзы) находятся два близко расположенных некогерентных источника света $S_1$ и $S_2$. В плоскости изображения (на расстоянии $z_2$) возникают два пятна Эйри. В силу некогерентности источников картина интенсивности в плоскости изображения представляет собой сумму интенсивностей, созданных каждым из источников. Если расстояние между центрами пятен Эйри мало (меньше их размера), то пятна накладываются друг на друга.

Согласно качественному критерию, предложенному Рэлеем, два источника света различимы, если дифракционный максимум одного источника приходится на минимум другого. Иными словами, расстояние между центрами пятен $\Delta x$ равно полуширине пятна Эйри (см. рис. 3.43):

$$\Delta x = \xi_0 = 1,22\frac{\lambda}{D}z_2$$

При этом, как показывает расчёт, в суммарной картине интенсивности между двумя максимумами возникает небольшой $\sim 27\%$ ($\sim 20\%$ при дифракции на щели) провал, по наличию которого можно визуально разрешить два источника 9.

Смещению пятен $\Delta x$ в плоскости изображения соответствует расстояние между точечными источниками во входной плоскости (на расстоянии $z_1$ от линзы), равное $l = \frac{z_1}{z_2}\Delta x$. Мы получаем оценку минимального разрешимого расстояния между точечными источниками:

$$\ell_{\min} \approx 1,22\frac{\lambda}{D}z_1 \tag{3.75}$$

Если рассматриваются удалённые объекты, то обычно говорят об угловом разрешении $\alpha = \frac{\ell}{z_1}$. Изображение оказывается при этом в фокальной плоскости линзы (при $z_1 \to \infty, z_2 \to f$). Согласно (3.74), имеем

$$\alpha_{\min} \approx 1,22\frac{\lambda}{D} \tag{3.76}$$

Разрешающая способность глаза. Если рассматривать с расстояния наилучшего зрения $z_1 \approx 25$ см две светящиеся точки, то, принимая диаметр зрачка равным $D \approx 3$ мм и $\lambda = 500$ нм (наиболее чувствительная для глаза область спектра), получим оценку минимально разрешимого расстояния для человека с нормальным зрением: $\ell_{\min} \approx 0,5 \cdot 10^{-2}$ mm.

## 7.2.10 2.10. Image Formation. Transfer Function. The Resolution of the Microscope

Исследуя в 2.7 формирование изображения, основанное на теории Аббе (принцип двойной дифракции), мы рассматривали идеальную оптическую систему (рис. 3.35). При этом первичное изображение - поле в фурье-плоскости - представляет собой неискажённое преобразование Фурье предметного поля $f_0(x)$. Здесь мы обсудим дифракционные искажения, обусловленные конечными размерами обдектива $\Omega_1$. Согласно теории Рэлея, волну, дифрагированную на предмете, представляем суперпозицией плоских волн разных направлений $\alpha$ (разных пространственных частот $u = k\sin\alpha$):

$$f_0(x) = \frac{1}{2\pi}\int C_0(u)e^{iux}du \tag{3.77}$$

где $C_0(u)$ - преобразование Фурье функции $f_0(x)$ (формула (3.44)).

[0]9 При оценке возможностей любой оптической системы следует принимать во внимание характеристики непосредственно регистрирующей среды: фотоэмульсии, матрицы цифровых устройств или чувствительных к свету элементов сетчатки человеческого глаза (палочки и колбочки). Любая регистрирующая среда обладает определённой дискретностью, и свет, фокусируясь оптической системой в малое пятнышко, меньшее размера «зерна» системы, засвечивает зерно целиком. Так что минимальный размер площадки в фоточувствительной среде, реагирующий на свет, не может быть меньше размера зерна регистрирующей среды.

Одна из таких плоских волн, с пространственной частотой $u = k\sin\alpha$, бегущая от предметной плоскости $\Pi_0$, показана на рис. 3.44 параллельными стрелками. Волна, бегущая под углом $\alpha$, сфокусировалась бы в точку $x_0 = f\,\mathrm{tg}\,\alpha \approx f\sin\alpha = f\frac{u}{k}$ фокальной плоскости Ф. Рисунок подчёркивает важнейшее обстоятельство: каждая слагаемая плоская волна, возникающая при дифракции на предмете конечного размера $l$ ($f_0(x) \equiv 0$ при $x \geqslant \frac{l}{2}$), имеет волновые поверхности неограниченной протяжённости по координате $x$: колебания поля этой идеализированной волны имеют одинаковую амплитуду во всех точках $x$ предметной плоскости $\Pi_0$ от $-\infty$ до $+\infty$. Однако в результате интерференции всех плоских волн образуется световое поле $f_0(x)$, так что интеграл (3.77) обращается в нуль при $|x| \geqslant \frac{l}{2}$ (подобно тому,

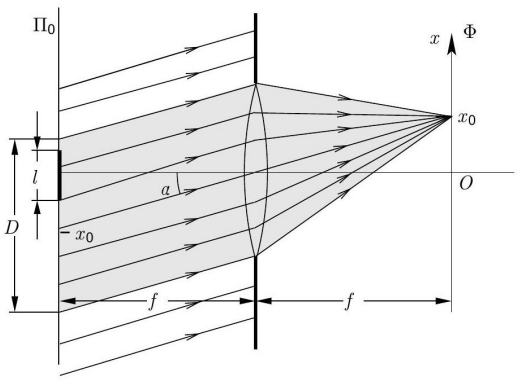



Рис. 3.44. Плоская волна и формирование изображения

как бесконечно длящиеся гармонические колебания образуют в сумме импульс конечной длительности).

Заметим, однако, что поле на объективе диаметра $D$ обусловлено лишь участком волновой поверхности в предметной плоскости

$$-x_0 - \frac{D}{2} \leqslant x \leqslant -x_0 + \frac{D}{2}, \quad \text{т. е.} \quad |x + x_0| \leqslant \frac{D}{2}$$

Действительно, если диафрагму диаметра $D$, с помощью которой мы моделируем конечные размеры объектива, перенести из плоскости, примыкающей к объективу, в предметную плоскость $\Pi_0$, расположив центр диафрагмы в точке $x = -x_0$ (см. рис. 3.44), то поле на объективе, т. е. в области $|x| \leqslant \frac{D}{2}$, при этом не изменяется: плоская волна неограниченной по $x$ протяжённости, пройдя через диафрагму, превращается в «пакет плоских волн» с несущей частотой $u_0 = k\sin\alpha \approx k\,\mathrm{tg}\,\alpha = \frac{kx_0}{f}$ и полушириной пространственного спектра $\Delta u = \frac{2\pi}{D}$ (см. рис. 3.32). Этот пакет плоских волн образует пучок света диаметра $D$, бегущий в направлении $\alpha$ (заштрихован на рисунке), и, что существенно, распространяясь от предметной плоскости $\Pi_1$ до плоскости, где расположен объектив, этот пучок остаётся параллелъным пучком неизменного поперечного сечения с той же плоской волновой поверхностью и той же амплитудой, поскольку условие геометрической оптики $\frac{\lambda f}{D^2} \ll 1$ выпол няется для любого реального объектива с огромным запасом (например, для объектива микроскопа при $D = 1$ см и $f = 2$ см $\frac{\lambda f}{D^2} \simeq 10^{-4}$).

Таким образом, для расчёта поля в фурье-плоскости Ф можно объектив считать идеальным (неограниченного размера), а каждую слагаемую плоскую волну $C_0(u)e^{iux}$, входящую в состав предметного поля, заменить пучком ограниченного сечения $C_0(u)P_D\,(x + x_0)\,e^{iux}$, где $x_0 = \frac{fu}{k}$, $P_D(\xi)$ - единичная нулевая функция:

$$P_D(\xi) = \begin{cases} 1 & \text{при } |\xi| \leqslant \frac{D}{2} \\ 0 & \text{при } |\xi| \geqslant \frac{D}{2} \end{cases}$$

Следовательно, предметное поле $f_0(x)$ (формула 3.77) следует заменить полем

$$f(x) = \frac{1}{2\pi} \int C_0(u) P_D \left( x + \frac{fu}{k} \right) e^{iux} du \tag{3.78}$$

В этом выражении интерес представляют лишь значения $|x| \leqslant \frac{l}{2}$, поскольку интерференция плоских волн в формуле (3.77) так же, как и интерференция пучков в формуле (3.78), даёт в сумме поле, тождественно равное нулю при $|x| > \frac{l}{2}$.

Пространственный спектр этого поля, наблюдаемый в фурьеплоскости идеальной линзы, есть $C(u) = C_0(u) P_{\mathrm{D}} \left( x + \frac{fu}{k} \right)$. Если речь идёт об объектах малого размера $l \ll D$ (в микроскоп обычно рассматривают объекты, размеры которых существенно меньше диаметра объектива), то единично нулевую функцию, рассматриваемую как функцию частоты $u$, можно с хорошей точностью заменить функцией $P_D \left( \frac{fu}{k} \right)$. При этом ошибка в полосе частот $\Delta u$ (в которой функция $P_D \left( \frac{fu}{k} \right) = 1$) окажется незначительной. Следовательно,

$$C(u) = C_0(u) P_D \left( \frac{fu}{k} \right) \tag{3.79}$$

и

$$f(x) = \frac{1}{2\pi} \int C_0(u) P_D \left( \frac{fu}{k} \right) e^{iux} du \tag{3.80}$$

где

$$P_D \left( \frac{fu}{k} \right) = \begin{cases} 1 & \text{при } |u| \leqslant \frac{kD}{2f} \\ 0 & \text{при } |u| > \frac{kD}{2f} \end{cases} \tag{3.81}$$

Итак, спектр изображения, получаемого в оптической системе с объективом конечного размера $D$, ограничен полосой частот $|u| \leqslant \frac{kD}{2f}$. Только плоские волны с пространственными частотами $u = k \sin \alpha$, удовлетворяющими последнему условию, формируют изображение (в частности, изображение, наблюдаемое в микроскоп).

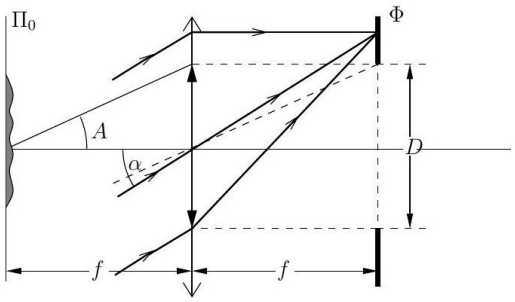



Рис. 3.45. Перенос диафрагмы в фурье-плоскость

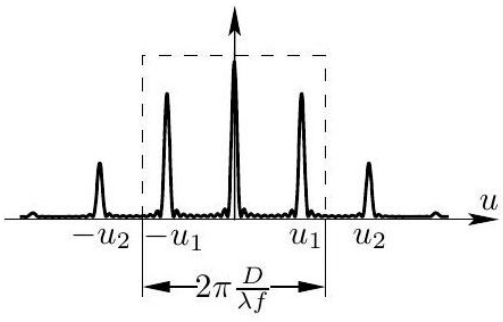


Рис. 3.46. Спектр решётки и функция пропускания диафрагмы

Равенство (3.81) означает, что при расчёте дифракционных искажений, обусловленных конечными размерами объектива, можно диафрагму размером $D$ перенести в фурье-плоскость, а объектив считать идеальным. При этом волны, направления которых $|\sin\alpha| > \frac{D}{2f}$, фокусируются идеальной линзой на непрозрачную часть диафрагмы и не участвуют в формировании изображения (рис. 3.45).

Пусть объект, рассматриваемый в микроскоп, - решётка периода $d$. Спектр такой решётки - картина дифракции Фраунгофера, наблюдаемая в фурье-плоскости идеальной линзы, имеет вид, показанный на рис. 3.46. Положение дифракционных максимумов определяется условием $\sin\alpha_m = m\frac{\lambda}{d}$. Изображение, наблюдаемое в микроскоп, формируется лишь теми из них, которые сфокусировались в точки фурье-плоскости, оказавшиеся внутри диафрагмы: $m\frac{\lambda}{d}f \leqslant \frac{D}{2}$ (или $|u_m| \leqslant \pi\frac{D}{\lambda f}$). Пунктиром показана передаточная функция $P_D\left(\frac{fu}{k}\right)$, определяющая «полосу пропускания» пространственных частот $|u| < \pi\frac{D}{\lambda f}$.

В частности, для получения информации о периоде решётки необходимо, чтобы в полосу пропускания попали $\pm1$ максимумы ($m = \pm1$):

$$d_{\min} \geqslant \frac{2\lambda f}{D} \approx \frac{\lambda}{\sin A} \tag{3.82}$$

где $\sin A$ - апертура микроскопа (рис. 3.45). В этом случае три дифракционных максимума $m = 0$ и $m = \pm1$ формируют, как известно,
изображение синусоидальной решётки периода $d$. Формула (3.82) определяет минимальную апертуру микроскопа $A$, необходимую для получения информации о периоде решётки.

# 8 Typical Spectral Devices

## 8.0.1 1. The Principle of Operation of Spectral Devices

Оптические приборы, в которых осуществляется физическое разложение электромагнитного излучения на монохроматические составляющие, называются спектральными. Спектральный состав электромагнитного излучения может много рассказать о своём источнике и о среде, в которой излучение распространялось. Существуют сотни приборов, анализирующих спектры - от радиодиапазона до высокоэнергетических $\gamma$-квантов. По характеру распределения интенсивности в спектральном разложении спектры могут быть разделены на линейчатые (например, спектры излучения атомов), непрерывные или сплошные (например, спектр излучения абсолютно чёрного тела), а также смешанные, представляющие собой наложение линейчатых и непрерывных спектров (например, спектр излучения Солнца). В данном лабораторном практикуме наблюдение спектров проводится на установках, основными элементами которых являются дифракционная решётка, интерферометр Фабри-Перо и призма. Эти установки позволяют проводить визуальное наблюдение линейчатых спектров, состоящих из набора узких спектральных линий в оптическом диапазоне длин волн, и на них можно довольно точно определить значение длины волны, но интенсивность спектральной линии определяется грубо, «на глаз».

Принципиальная схема установки для изучения спектров приведена на рис. 4.1. Свет от источника $S$ попадает на экран, в котором имеется отверстие в виде щели. Экран располагают в фокальной плоскости линзы или системы линз. Коллиматор формирует пучок света, близкий к параллельному. После коллиматора пучок лучей попадает на диспергирующий элемент (ДЭ): амплитудную или фазовую дифракционную решётку,

интерферометр Фабри-Перо или призму. Наблюдаются изображения с помощью зрительной трубы, установленной на бесконечность.

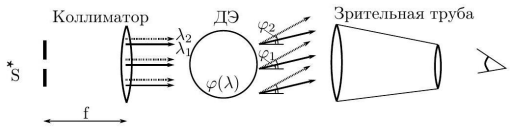



Рис. 4.1. Схема прибора: источник-коллиматор диспергирующий элемент - зрительная труба

Если удалить из схемы диспергирующий элемент, а коллиматор и зрительную трубу расположить на одной оси, то можно увидеть чёткое изображение входной щели коллиматора.

Диспергирующий элемент перераспределяет интенсивность падающего на него излучения по углам в зависимости от длины волны: каждой монохроматической компоненте излучения с длиной волны $\lambda$ соответствует один или несколько углов $\varphi(\lambda)$ на выходе прибора, в направлении которых интенсивность прошедшей волны максимальна. Иными словами, диспергирующий элемент пространственно разделяет монохроматические составляющие падающего на него излучения, осуществляя тем самым его физическое разложение по спектру. При известной зависимости $\varphi(\lambda)$ по измеряемому углу поворота $\varphi$ зрительной трубы можно определить длину волны спектральной линии.

Каждый спектральный прибор предназначен для решения конкретной задачи спектроскопии. Выбор прибора для исследования спектра какого-либо источника должен заключаться в сравнении его характеристик с требуемыми. Наиболее важными характеристиками являются угловая дисперсия, разрешающая способность и дисперсионная область.

Разрешающая способность $R = \frac{\lambda}{\delta\lambda}$ характеризует возможность прибора различать две близкие спектральные линии с длинами волн $\lambda$ и $\lambda + \delta\lambda$.

Угловая дисперсия $D = \frac{d\varphi}{d\lambda}$ - производная зависимости угла отклонения $\varphi(\lambda)$ волны диспергирующим элементом по $\lambda$. По величине угловой дисперсии можно определить угловое расстояние между двумя близкими спектральными линиями: $\delta\varphi \approx D\delta\lambda$.

Дисперсионная область (или область дисперсии) - предельная ширина спектрального интервала $\Delta\lambda$ прибора, для которой дифракционные максимумы соседних порядков не перекрываются. Она определяет
диапазон длин волн, при которых прибор может быть использован для анализа спектра.

## 8.0.2 2. Description of Spectral Devices

### 2.1. Амплитудная дифракционная решётка

Амплитудную решётку (см. также раздел III, п. 1.5.4) можно представить в виде непрозрачного экрана, в котором прорезано большое число $N$ параллельных щелей - штрихов (рис. 4.2). Постоянство расстояний между штрихами $d$ (период решётки, или шаг решётки) и шириной штриха $b$ должно выдерживаться с большой точностью.

Наблюдение изображения спектра проводится с помощью зрительной трубы, настроенной на бесконечность (дифракции Фраунгофера на штрихах решётки). В этом случае амплитуда и интенсивность поля световой волны определяются углом $\varphi$ между нормалью к решётке и направлением дифрагировавших лучей. Будем считать, что амплитуды всех интерферирующих волн одинаковы, т. е. фиксирована амплитуда падающей волны и постоянна площадь всех штрихов. Интенсивность дифрагированного

света максимальна для углов $\varphi_m$, при которых волны, приходящие в точку наблюдения от всех щелей, оказываются в фазе:

$$d \sin \varphi_m = m\lambda$$

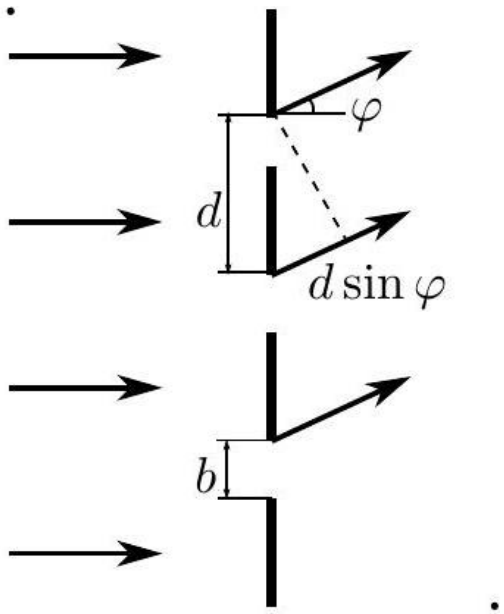



Рис. 4.2. Дифракция световой волны на амплитудной решётке

Величина $m = 0, \pm 1, \pm 2, \pm 3, \ldots$ называется порядком спектра.

Рассмотрим качественный пример. Пусть падающее на решётку излучение содержит две спектральные линии одинаковой интенсивности. Одну линию условно назовем «красной», а другую «фиолетовой». Длина волны «красной» линии больше, чем длина волны «фиолетовой». Изображение спектра приведено на рис. 4.3. Для угла дифракции $\varphi_0 = 0 (m = 0)$, когда ось зрительной трубы параллельна оси коллиматора, наблюдается наложение изображений входной щели коллиматора в «красном» и «фиолетовом» цвете друг на друга. При повороте зрительной трубы вокруг решётки в поле зрения возникает «фиолетовая» щель коллиматора, затем «красная» и т. д. Для малых углов дифракции $\varphi_m$ угловое расстояние между порядками $\varphi_{m+1} - \varphi_m \approx \lambda/d$ пропорционально длине волны, поэтому «фиолетовые» линии следуют чуть чаще, чем «красные», и возможна ситуация, когда они вновь налагаются друг
на друга (на рис. 4.3 это происходит при $m = 5$ для «красной» линии и $m = 6$ для «фиолетовой»).

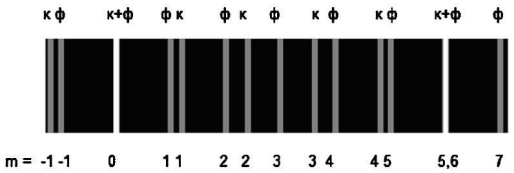



Рис. 4.3. Изображение спектра двух линий

Уточним форму распределения интенсивности света наблюдаемой спектральной линии. Для очень узкой входной щели коллиматора эта форма зависит от числа штрихов решётки $N$, шага решётки $d$ и ширины её штриха $b$ (рис. 4.2). Пусть интенсивность квазимонохроматического света (длина волны $\lambda$) равна $I_0$ [ $\mathrm{B_т}$ /$\mathrm{cm}^2$]. Учёт дифракции Фраунгофера на отдельной щели решётки и когерентного сложения полей от

всех штрихов решётки даёт следующее выражение для мощности света на единицу высоты изображения спектральной линии в малом интервале углов дифракции $d\varphi$:

$$dW = I_0 \frac{b^2}{\lambda} \left[ \frac{\sin\left(\frac{\pi}{\lambda} b \sin\varphi\right)}{\frac{\pi}{\lambda} b \sin\varphi} \right]^2 \left[ \frac{\sin\left(\frac{\pi}{\lambda} N d \sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\varphi\right)} \right]^2 d\varphi \tag{4.2}$$

Здесь величина $W$ имеет размерность [$\text{В}_\text{т}$ /cm].

В выражении (4.2) множитель

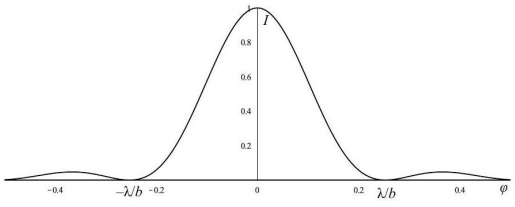
$$\left[ \frac{\sin\left(\frac{\pi}{\lambda} b \sin\varphi\right)}{\frac{\pi}{\lambda} b \sin\varphi} \right]^2$$

описывает дифракцию света на отдельной щели решётки (рис. 4.4а). Для ширины щели $b \gg \lambda$ область углов центрального, или главного максимума ограничена отрезками $|\varphi| \leqslant \lambda/b$ (координаты нулей распределения). В этой области сосредоточено около 90% мощности света. Множитель

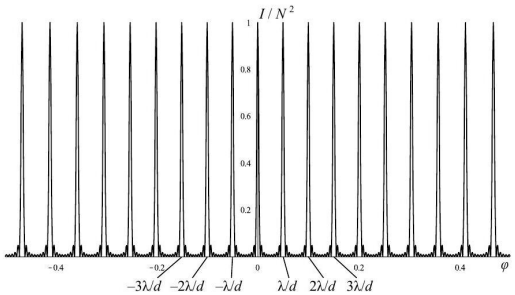
$$\left[ \frac{\sin\left(\frac{\pi}{\lambda} N d \sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\varphi\right)} \right]^2$$

описывает когерентное сложение полей от $N$ штрихов (рис. 4.4б). Максимумы этого распределения возникают при обращении знаменателя в ноль, что имеет место при выполнении условия (4.1). Область локализации мощности отдельного максимума $|\delta\varphi| \leqslant \lambda/Nd$ много меньше

а)




б)




В)

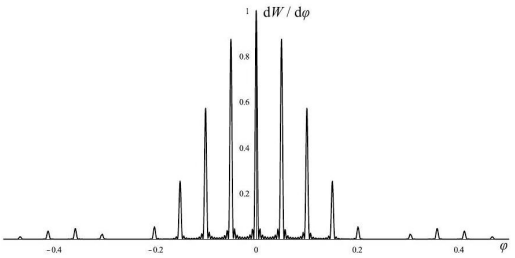


Рис. 4.4. Дифракция света на решётке ($\lambda/b \approx 0,25, \lambda/d \approx 0,05$). Зависимость интенсивности света от угла: а) дифракция на отдельной щели;

б) дифракция на решётке в пределе бесконечно узких щелей;

в) распределение мощности света на единицу высоты изображения, отнесённое к малому диапазону углов

ширины главного дифракционного максимума $2\lambda/b$. На рис. 4.4в приведена форма спектра квазимонохроматического света.

Мощность света на единицу высоты изображения спектральной линии в порядке $m$ :[1]

$$W_m = I_0 \frac{b^2}{\lambda} \left[ \frac{\sin\left(\frac{\pi}{\lambda} b \sin\varphi_m\right)}{\left(\frac{\pi}{\lambda} b \sin\varphi_m\right)} \right]_{\varphi_m - \frac{\lambda}{2d}}^{2\varphi_m + \frac{\lambda}{2d}} \left[ \frac{\sin\left(\frac{\pi}{\lambda} N d \sin\varphi\right)}{\sin\left(\frac{\pi}{\lambda} d \sin\varphi\right)} \right]^2 d\varphi \approx I_0 \frac{b^2}{\lambda} N \left[ \frac{\sin \pi \frac{b}{d} m}{\pi \frac{b}{d} m} \right]^2$$

Суммарная мощность всех порядков

$$I_0 \frac{b^2}{d} N \sum_{-d/\lambda}^{d/\lambda} \left[ \frac{\sin \pi \frac{b}{d} m}{\pi \frac{b}{d} m} \right]^2 \approx I_0 \frac{b^2}{d} N \int_{-d/\lambda}^{d/\lambda} \left( \frac{\sin \pi \frac{b}{d} x}{\pi \frac{b}{d} x} \right)^2 dx \approx I_0 N b$$

равна мощности света за решёткой. Разумное ограничение на рабочий порядок спектра $m_p : 0 < |m_p| < d/b$, т. е. порядок не нулевой и лежит в области главного дифракционного максимума отдельной щели. Если на единицу высоты решётки падает мощность света $I_0 N d$, то в рабочий порядок попадает только её часть:

$$\frac{I_0 b^2 N/d}{I_0 N d} \left[ \frac{\sin \pi \frac{b}{d} m_p}{\pi \frac{b}{d} m_p} \right]^2 = \left( \frac{b}{d} \right)^2 \left[ \frac{\sin \pi \frac{b}{d} m_p}{\pi \frac{b}{d} m_p} \right]^2$$

Пусть, для примера, $d/b = 4$ и рабочий порядок $m_p = 2$, в этом случае только 3% падающей на решётку мощности света попадает в нужный порядок. Ограниченное использование в спектроскопии амплитудных дифракционных решёток связано с большими потерями света. Этот недостаток отсутствует у фазовых решёток.

Отметим, что наблюдение в опытах идеальных линий, приведённых на рис. 4.4в, весьма затруднительно, так как угловой размер щели коллиматора $l/f$($l$ - ширина щели, $f$ - фокусное расстояние коллиматора) должен быть заметно меньше ширины линии $\lambda/Nd$. Как уменьшение ширины щели $l$, так и уменьшение размера решётки $Nd$ приводят к большим потерям света. Практически при измерениях наблюдается слегка размытое изображение щели коллиматора.

### 2.2. Фазовая дифракционная решётка

В современных спектральных приборах широко используются отражательные решётки с треугольным профилем штриха, они способны концентрировать до $70 - 80\%$ падающего излучения в рабочий порядок спектра. Отражательная решётка (рис. 4.5), в которой угол $\Omega$ между рабочей гранью и плоскостью решётки (угол скоса) не превышает 20°, называется эшелеттом (от французского echelle - лестница). Для эшелетта, варьируя угол скоса и шаг решётки, получают рабочий порядок $m_p \leqslant 10$. Если угол скоса лежит в пределах 20-60, то решётка называется эшелле, её рабочий порядок $m_p = 5 - 500$.

[0]1 После замены переменных и с учётом малого соотношения ширины главного максимума функций к размеру области интегрирования интегралы сводятся к известному: $\int_{-\infty}^{\infty} \left(\frac{\sin \pi x}{\pi x}\right)^2 dx = 1$.

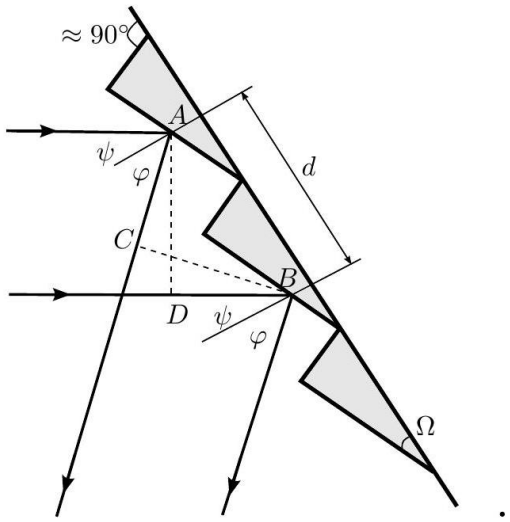



Рис. 4.5. Дифракция световой волны на эшелетте

Эшелетт можно представить как зеркало в форме лестницы. Найдём разность хода между лучами (рис. 4.5). Пусть на эшелетт падает плоская монохроматическая волна с длиной волны $\lambda$, угол падения, отсчитанный от нормали к решётке, равен $\psi$, угол дифракции - $\varphi$. Условие возникновения спектра порядка $m$:

$$AC - BD = d\,(\sin\varphi_m - \sin\psi) = m\lambda$$

Для нулевого порядка $\varphi_0 = \psi$. В отличие от амплитудной решётки, нулевой порядок не будет самым ярким. Угол $\varphi_б$ , под которым наблю дается максимум интенсивности света, соответствует зеркальному отражению падающей волны от одной ступеньки, он называется углом блеска и определяется формулой

$$\varphi_\sigma = \psi + 2\Omega \tag{4.4}$$

Для эшелетта рабочим порядком спектра $m_p$ будет то целое число, которое соответствует минимальной ошибке решения уравнения

$$d\,(\sin\varphi_\sigma - \sin\psi) - \lambda m_p = 0$$

В паспорте эшелетта указаны рабочий порядок спектра $m_p$ и рабочая длина волны $\lambda_p$. Чтобы устранить произвол в выборе угла падения света при расчёте угла блеска (4.4), считаем, что эшелетт работает в автоколлимационном режиме. В этом режиме свет падает перпендикулярно рабочей грани решётки ($\psi = -\Omega$) и отражается в обратном направлении ($\varphi = \varphi_m = \varphi_p = \Omega$), условие (4.3) возникновения рабочего спектра даёт

$$2d\sin\Omega = m_p\lambda_p \tag{4.5}$$

В автоколлимационном режиме дифракция на одной ступенькезеркальце описывается так же, как и дифракция на отдельной щели амплитудной решётки, рис. 4.4а, с максимумом вблизи $\varphi \approx 0$. В отличие от амплитудной решётки, при углах $\varphi \approx 0$ располагается не нулевой порядок спектра (так называемая ахроматическая полоса), а порядок, определяемый соотношением (4.5), т. е. нумерацию порядков, приведённую на рис. 4.4б для амплитудной решётки, следует сместить на величину $m_p$. Доля мощности, попадающей в рабочий порядок, составляет $(b/d)^2$ от мощности светового потока, падающего на эшелетт ($b$ - ширина отдельной ступеньки, чуть меньшая шага решётки $d$).

### 2.3. Интерферометр Фабри-Перо

Как спектральный прибор высокой разрешающей способности интерферометр Фабри-Перо широко используется в физических экспериментах. Он применяется для исследования тонкой структуры спектральных линий, является неотъемлемым элементом лазера, выполняя роль оптического резонатора, и т. д. Интерферометр Фабри-Перо состоит из двух стеклянных или кварцевых пластин с хорошо отполированными поверхностями (с шероховатостью до $10^{-2}\lambda$), которые установлены параллельно друг другу на некотором расстоянии. На одну поверхность каждой пластины нанесены хорошо отражающие свет покрытия. Для получения коэффициента отражения $r \approx 0,9$ используют

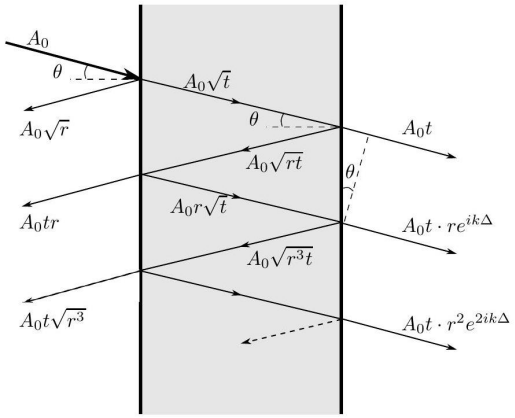



Рис. 4.6. Амплитуды волн в интерферометре Фабри-Перо. Для прошедших волн также указаны набеги фаз

металлические покрытия (Ag, Al), для достижения $r \approx 0,99$ наносятся многослойные диэлектрические интерференционные покрытия.

Интерферометр Фабри-Перо можно рассматривать как плоскопараллельную пластину, в которой происходят многократные отражения и интерференция световых волн. На рис. 4.6 приведена схема интерферирующих волн. Коэффициенты пропускания и отражения по интенсивности отдельного зеркала интерферометра равны $t$ и $r$ соответственно (из закона сохранения энергии следует, что $t + r = 1$). Пусть $A_0$ - амплитуда падающей на интерферометр волны, тогда амплитуда отражённой от первого зеркала волны равна $A_0\sqrt{r}$, амплитуда прошедшей внутрь интерферометра волны - $A_0\sqrt{t}$, амплитуда волны, отражённой от второго зеркала, $-A_0\sqrt{rt}$, амплитуда первой прошедшей волны равна $A_0t$ и т. д. В результате многократных переотражений на выходе интерферометра будем иметь набор волн, амплитуды которых равны $A_0t, A_0tr, A_0\,\mathrm{tr}^2, \ldots$ Фазовая задержка между двумя «соседними» волнами равна $k\Delta$, где $k = \frac{2\pi}{\lambda}$ - волновое число; $\Delta$ - разность хода для угла падения $\theta$. Интерференционная картина, наблюдаемая с помощью зрительной трубы, настроенной на бесконечность, состоит из концентрических колец равного наклона (см. раздел II, п. 4.5). Най дём условие возникновения интерференционной картины для световой длины волны $\lambda$. Выразим разность хода двух интерферирующих волн, падающих на интерферометр под углом $\theta$:

$$\Delta = 2L\left(\frac{1}{\cos\theta} - \mathrm{tg}\,\theta\sin\theta\right) = 2L\cos\theta \tag{4.6}$$

где $L$ - расстояние между зеркалами, или база интерферометра. Интерференционные максимумы будут наблюдаться для волн, падающих под углами $\theta_m$, удовлетворяющими условию

$$2L\cos\theta_m = m\lambda \tag{4.7}$$

Отметим отличия условий возникновения спектров для дифракционных решёток и интерферометра Фабри-Перо. Для решёток с увеличением угла дифракции порядок спектра растёт начиная с $m = 0$. Для интерферометра Фабри-Перо углу падения $\theta = 0$ соответствует максимальный порядок спектра $M = 2L/\lambda$ и с увеличением угла порядок спектра уменьшается. Если зрительную трубу вращать вокруг дифракционной решётки, то в поле зрения трубы возникает сначала «фиолетовая», затем «красная» линия (рис. 4.3). Для интерферометра Фабри-Перо первоначально возникнет кольцо для длинноволновой («красной») линии, затем коротковолновой («фиолетовой»).

Просуммируем комплексные амплитуды световых волн, прошедших интерферометр. Можно видеть (см. рис. 4.6), что амплитуды прошедших волн образуют геометрическую прогрессию. Считая её бесконечной, получим комплексную амплитуду суммарной прошедшей волны:

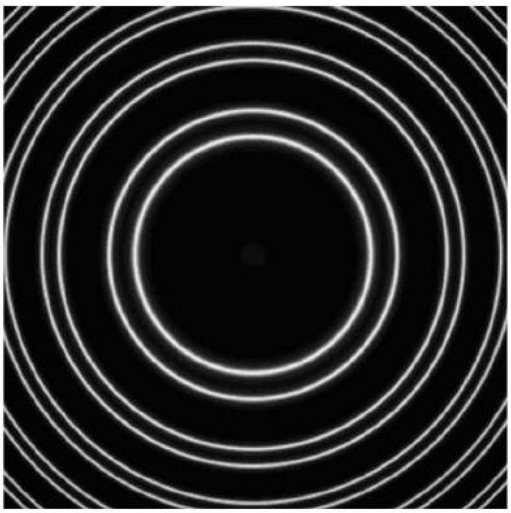
$$A = \frac{A_0 t}{1 - re^{ik\Delta}}$$

и её интенсивность:

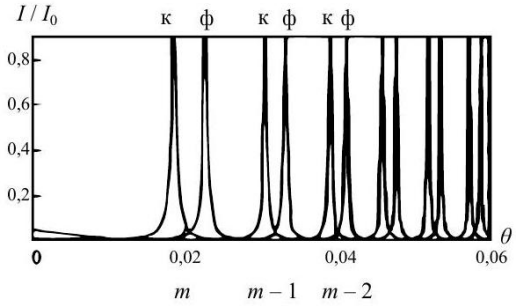
$$I = |A|^2 = \frac{|A_0|^2 t^2}{1 + r^2 - 2r\cos\left[\frac{4\pi}{\lambda}L\cos\theta\right]} = \frac{I_0}{1 + \frac{4r}{(1-r)^2}\sin^2\left(\frac{2\pi}{\lambda}L\cos\theta\right)} \tag{4.8}$$

Интерферометр Фабри-Перо становится почти прозрачным (просветляется) для световой волны, удовлетворяющей условию резонанса (4.7).

В качестве примера вида спектра рассмотрим интерференционные кольца для двух близких спектральных линий одинаковой интенсивности (рис. 4.7). Пусть длина волны первой линии $\lambda = 5890$Å (близкая к жёлтой линии натрия), а другая длина волны отличается на $0,5$Å. База интерферометра $L = 1$ мм (максимальный порядок спектра $2L/\lambda = 3,4 \cdot 10^3$), коэффициент отражения зеркал $r = 0,85$. Видно,
что каждый порядок спектра (здесь их 5) представлен парой колец. Угловое расстояние между парами уменьшается при увеличении угла наблюдения или уменьшении порядка спектра. Кроме того, уменьшается расстояние между кольцами в одном порядке и их ширина. Отметим, что приведённому изображению соответствует диапазон углов $\pm 0,06$ рад, или всего $\pm 3,4°$.



а)




б)

Рис. 4.7. Интерференционные кольца для двух линий: а) изображение спектра, б) зависимость интенсивности света от угла

### 2.4. Призма

Принцип действия этого спектрального прибора основан на зависимости коэффициента преломления стекла от длины световой волны. Комбинируя различные виды стекла, количество элементов, их форму, можно сконструировать множество призм для решения задач спектроскопии. Мы рассмотрим простую призму из тяжёлого флинта (марка стекла ТФ-1) с углом между гранями 60°. Для стекла ТФ-1 показатель преломления на длине волны $\lambda_1 = 4880$Å (голубая линия) $n_1 = 1,661$, на длине волны $\lambda_2 = 6327$Å (красная линия) $n_2 = 1,644$.

По этим данным можно оценить важную характеристику стекла дисперсию показателя преломления:

$$\frac{dn}{d\lambda} \approx \frac{n_1 - n_2}{\lambda_1 - \lambda_2} = -1,17 \cdot 10^3 \text{ см}^{-1}$$

Оценим угол $\delta\varphi$ между волновыми фронтами двух близких линий $\lambda$ и $\lambda + \delta\lambda$. Пусть на призму с основанием $a$ падает параллельный пучок света шириной $h$, и этот пучок целиком заполняет призму. Показанный на рис. 4.8 симметричный ход лучей (внутри призмы лучи распространяются параллельно основанию) соответствует углу наименьшего отклонения падающего пучка $\theta$, который в свою очередь зависит от преломляющего угла призмы $\alpha$ и показателя преломления $n(\lambda)$.

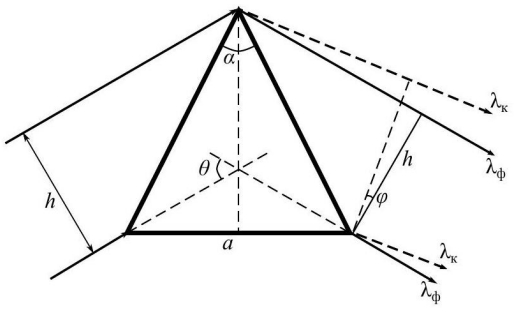



Рис. 4.8. Ход лучей в призме для угла наименьшего отклонения

Максимальная разность хода двух лучей с длиной волны $\lambda$ и $\lambda + \delta\lambda$ возникает вблизи основания призмы:

$$\Delta = a[n(\lambda + \delta\lambda) - n(\lambda)] \approx a\frac{dn}{d\lambda}\delta\lambda = h\delta\varphi \tag{4.9}$$

После прохождения призмы «фиолетовые» лучи отклоняются больше, чем «красные». В этом полезно убедиться практически, рассматривая резкую границу

белого света и тени (освещённый подоконник или крышку стола). Если граница горизонтальна и основание призмы ниже её вершины, то граница тени будет казаться красной, если призму перевернуть на 180°, - фиолетовой.

Типичная зависимость $n(\lambda)$, или закон дисперсии показателя преломления, приведена на рис. 4.9. В области, закрашенной серым, показатель преломления растёт с длиной волны $dn/d\lambda > 0$ - это так называемая область аномальной дисперсии. Аномальная дисперсия имеет место на частотах, близких к резонансным, и соответственно в этой области велико поглощение света. В области, далёкой от резонансов, вещество прозрачно, а показатель преломления убывает с ростом длины волны, $dn/d\lambda < 0$, - это область нормальной дисперсии. Стекло в оптическом диапазоне длин волн имеет нормальную дисперсию (аномальная дисперсия в стекле имеет место в ультрафиолетовой области спектра).

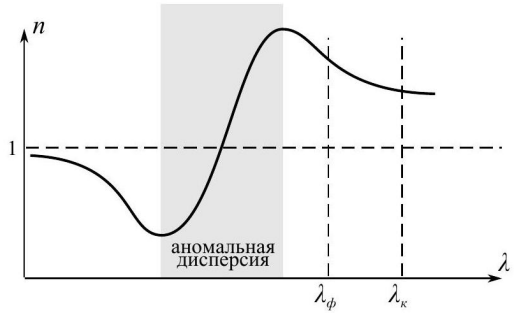



Рис. 4.9. Зависимость показателя преломления стекла от длины волны

### 8.0.3 3. Angular Dispersion of Spectral Devices

Угловая дисперсия $D(\lambda)$ характеризует угловое расстояние между близкими спектральными линиями:

$$D(\lambda) = \frac{d\varphi}{d\lambda} \tag{4.10}$$

В современных приборах спектроскопии регистрация изображения спектров проводится не глазом, а линейкой или матрицей чувствительных к свету элементов. Угловая дисперсия позволяет определить минимальное расстояние между ячейками приёмного устройства: если требуется разрешить две спектральные линии с разностью длин волн $\delta\lambda$, то расстояние между элементами приемного устройства должно быть заметно меньше $D\delta\lambda f$, где $f$ - фокусное расстояние объектива зрительной трубы.

Амплитудная решётка. Выражение для угловой дисперсии дифракционной решётки следует из (4.1):

$$D = \frac{d\varphi}{d\lambda} = \frac{m}{d\cos\varphi} = \frac{m}{\sqrt{d^2 - m^2\lambda^2}} \tag{4.11}$$

Дисперсия возрастает с увеличением порядка спектра. Для малых углов дифракции $\varphi \ll 1$ дисперсия пропорциональна порядку спектра: $D \approx m/d$.

Эшелетт. Выражение угловой дисперсии эшелетта получается так же, как для амплитудной решётки, но с учётом углов падения; см (4.3):

$$D = \frac{d\varphi}{d\lambda} = \frac{m}{d\cos\varphi} = \frac{m}{\sqrt{d^2 - (m\lambda + d\sin\psi)^2}} \tag{4.12}$$

В отличие от амплитудной решётки углы дифракции нельзя считать малыми. В опытах углы падения $\psi \approx 30 - 60°$.

Интерферометр Фабри-Перо. Выражение для угловой дисперсии следует из (4.6). Для малых углов $\theta_m \ll 1$ и больших порядков спектров имеем $m \approx M = 2L/\lambda$,

$$-2L\sin\theta_m d\theta = -2L\theta_m d\theta = md\lambda \approx \frac{2L}{\lambda}d\lambda$$

и угловая дисперсия

$$D = \frac{d\theta}{d\lambda} = -\frac{m}{2L\sin\theta_m} \approx -\frac{1}{\lambda\theta_m} \tag{4.13}$$

Призма. Для разности хода (4.9) получаем угловую дисперсию:

$$D = \frac{d\varphi}{d\lambda} = \frac{a}{h}\frac{dn}{d\lambda} \tag{4.14}$$

Проведём сравнительную оценку значений угловой дисперсии для рассматриваемых спектральных приборов: амплитудной и фазовой дифракционной решёток, интерферометра Фабри-Перо, призмы. Вычисления выполним для длины волны $\lambda = 578$ нм (средняя длина волны для двух близких жёлтых линий ртути).

Амплитудная дифракционная решётка с плотностью штрихов $n = 5001$ мм для порядков спектра $m = 1, 2, 3$ имеет в соответствии с (4.11) угловую дисперсию $5,2 \cdot 10^3; 1,2 \cdot 10^4; 3 \cdot 10^4$ см$^{-1}$. Эшелетт для угла падения $\psi = 30^\circ$ имеет только один порядок спектра $m = 1$, в котором угловая дисперсия (4.12) равна $8,1 \cdot 10^3$ см$^{-1}$. Интерферометр Фабри-Перо с базой $L = 1$ мм для системы колец максимального порядка спектра и меньшего на 1 и 2 имеет угловую дисперсию (4.13), равную $-1 \cdot 10^6, -5,9 \cdot 10^5, -4,6 \cdot 10^5 1/$ см. Призма вблизи выбранной длины волны имеет показатель преломления $n = 1,64$, и его дисперсия $\frac{dn}{d\lambda} \approx -920$ см$^{-1}$. Расчёты по законам геометрической оптики дают отношение стороны равносторонней призмы к ширине пучка $a/h \approx 1,7$. Угловая дисперсия призмы (4.14) при этом равна $-1,6 \cdot 10^3$ см$^{-1}$.

Для решёток угловая дисперсия положительна, с увеличением длины волны спектральных линий угол, под которым они наблюдаются,
растёт. Для интерферометра Фабри-Перо и призмы угол наблюдения увеличивается с уменьшением длины волны линии. Для рассмотренных примеров минимальной угловой дисперсией обладает призма, дифракционные решётки превосходят её по этому параметру в несколько раз, а интерферометр Фабри-Перо - на порядки.

## 8.0.4 4. Resolution Capacity

Рассмотрим изображения спектра для двух узких спектральных линий с длинами волн $\lambda$ и $\lambda+\delta\lambda$. Для минимального значения $\delta\lambda$, которое может быть определено по результатам измерений, вводят важнейшую характеристику спектрального прибора - разрешающую способность:

$$R = \frac{\lambda}{\delta\lambda} \tag{4.15}$$

Если спектральным прибором является интерферометр ФабриПеро, то пары колец в каждом порядке спектра (рис. 4.7) для минимальных значений $\delta\lambda$ должны заметно накладываться друг на друга. Минимальное значение $\delta\lambda$ определяется не только качеством коллиматора и зрительной трубы (рис. 4.1), но главным образом точностью изготовления интерферометра. Точность полировки поверхностей до $\lambda/100$, строгая параллельность плоскостей (малейший клин заметно искажает изображение спектра), малые температурные деформации конструкции - лишь небольшая часть требований к интерферометру. Высоки технические требования и к дифракционным решёткам.

Решётку изготавливают с помощью делительной машины или методами литографии, далее её размножают или распечатывают. Копия эталонной решётки называется репликой. В учебных лабораториях используются именно реплики. На большие решётки делительная машина наносит $10^4 - 10^5$ штрихов, шаг решётки составляет величину порядка микрометра. Если на одном участке решётки шаг чуть больше, а на других чуть меньше среднего, то в изображении спектра могут возникать ложные линии спектра. Они называются «духами». Естественной причиной уменьшения разрешающей способности спектральных приборов является их старение и нарушение правил обращения с ними.

### 4.1. Амплитудная решётка, эшелетт, призма

Приведённые выше соображения относятся к техническим требованиям изготовления спектрального прибора. Рассмотрим физические ограничения разрешающей способности. На рис. 4.3 изображения спектральных линий являются изображениями входной щели коллимато ра. Начнём уменьшать размер щели. В начале этого процесса будет уменьшаться интенсивность линий и их ширина. Начиная с некоторого момента будет уменьшаться только интенсивность, а ширина линий изменяться не будет. Достигнут физический предел ширины линии, и он определяется дифракцией света на апертуре решётки. Определим угловое расстояние между максимумом линии и её первым нулем - полуширину линии $\delta\varphi$. Пусть на решётку, состоящую из $N$ штрихов, падает параллельный пучок света перпендикулярно её поверхности. Если $N = 2$, то две волны погасят друг друга, если между ними возникнет разность хода $\lambda/2$, если $N = 3$, то $\lambda/3$. В общем случае $N$ штрихов для полуширины линии $\delta\varphi$ получаем уравнение, решение которого совместно с уравнением (4.1) при $\delta\varphi \ll 1$ имеет вид

$$d\sin(\varphi_m + \delta\varphi) = m\lambda + \frac{\lambda}{N}$$
$$\delta\varphi = \frac{\lambda}{Nd\cos\varphi_m} \tag{4.16}$$

здесь $Nd\cos\varphi_m$ - видимый под углом $\varphi_m$ размер решётки. Тем же выражением описывается полуширина линии для эшелетта. Угловое расстояние между двумя линиями определяется дисперсией (4.11), (4.12):

$$\Delta\varphi \approx D\delta\lambda = \frac{m}{d\cos\varphi_m}\delta\lambda \tag{4.17}$$

Для сравнения между собой различных спектральных приборов Релей предложил приравнять полуширину $\delta\varphi$ и расстояние между линиями $\Delta\varphi$. Критерий Релея удобен для различных оценок. Согласно ему для дифракционных решёток разрешающая способность определяется порядком спектра и числом штрихов:

$$R = Nm \tag{4.18}$$

Здесь под $N$ следует понимать число одновременно работающих штрихов решётки, которое, вообще говоря, не равно суммарному числу штрихов освещённого участка решётки. Число штрихов $N$ определяется качеством реплики, размером источника света и т. д. Например, если источник является однородной линией шириной $b$ на расстоянии $z$ от решётки, то размер области решётки, засвеченной когерентно, определяется формулой $\rho \sim \lambda/\psi$, где $\psi = b/z$ - угловой размер источника (см. раздел II, п. 4.3).

На рис. 4.10а показана зависимость от угла интенсивности двух линий. Форма линий определяется дифракцией Фраунгофера на апертуре

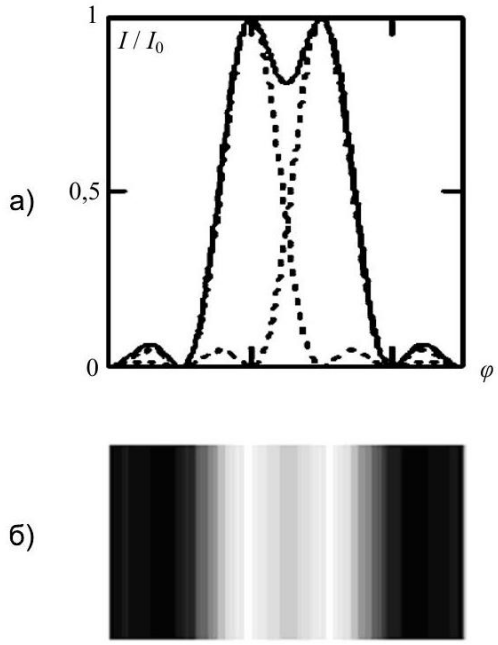



Рис. 4.10. Условие Релея для дифракционной решётки: а) интенсивности двух близких линий и их сумма, б) изображение спектра

решётки, угловая координата максимума одной линии совпадает с первым минимумом для другой (условие Релея). Там же приведена форма суммарной интенсивности. Основная числовая характеристика условия Релея - отношение интенсивности света в провале между линиями к максимальному значению. Численное значение этой характеристики для решётки равно $2\left(\frac{\sin \pi/2}{\pi/2}\right)^2 \approx 0,81$. На рис. 4.86 приведено изображение спектра этих линий. Можно отметить, что глаз легко различает две линии. В экспериментальной спектроскопии разрешающая способность спектрального прибора превосходит разрешение по критерию Релея при использовании вместо глаза малошумящих приемников света. Измеряется отношение ширины суммарной к ширине отдельной линии на уровне половинной мощности: для критерия Релея это отношение равно 2 (рис. 4.10а).

Аналогичная процедура для призмы с учётом (4.14) даёт

$$\delta\varphi = \frac{a}{h}\left|\frac{\partial n}{\partial \lambda}\right| \delta\lambda = \frac{\lambda}{h}$$

и разрешающая способность

$$R = \frac{\lambda}{\delta\lambda} = a\left|\frac{dn}{d\lambda}\right| \tag{4.19}$$

При полном заполнении призмы светом разрешающая способность пропорциональна размеру её основания.

### 4.2. Интерферометр Фабри-Перо

Дифракция Фраунгофера на апертуре интерферометра также влияет на ширину интерференционного кольца и разрешающую способность. Строго говоря, имеет значение не размер освещённой области зеркала, а размер её однородно обработанного участка. Напомним, что погрешность обработки и настройки базы интерферометра $L$ должна быть много меньше $\lambda$.

Мы ограничимся рассмотрением зависимости разрешающей способности интерферометра Фабри-Перо от величины базы $L$ (порядка спектра) и коэффициента отражения зеркал $r$, определим угловое расстояние между двумя близкими линиями, соответствующее условию Релея. Пусть, как и в случае дифракционной решётки, две линии пересекаются на уровне половинной мощности каждой линии. Запишем это условие для угла $\theta_m + \delta\theta/2$ (угол $\theta_m$ соответствует максимуму для длины волны $\lambda$ порядка $m$). Используя равенство (4.8), имеем

$$\frac{4r}{(1-r)^2}\sin^2\left[\frac{\pi}{\lambda}2L\cos\left(\theta_m+\delta\theta/2\right)\right]=1$$

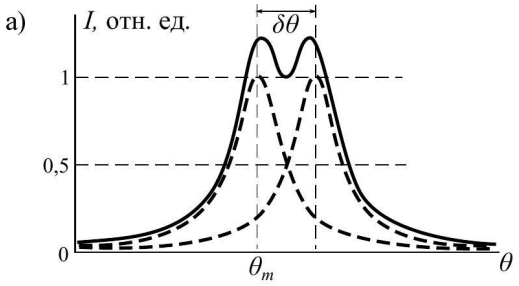
Учитывая условие резонанса (4.6) и малость углов $\theta_m$ и $\delta\theta$, $\left(\left|\frac{\pi}{\lambda}\theta_m\delta\theta\right|\ll 1\right)$, получаем

$$\sin^2\left[\frac{\pi}{\lambda}2L\cos\left(\theta_m+\delta\theta/2\right)\right]=$$
$$=\sin^2\left(\frac{2\pi L}{\lambda}\cos\theta_m-\frac{\pi L}{\lambda}\sin\theta_m\cdot\delta\theta\right)\approx\left[\frac{\pi L}{\lambda}\theta_m\delta\theta\right]^2$$

Пусть угловому радиусу $\theta_m+\delta\theta$ соответствует максимум интерференционного кольца с тем же порядком спектра $m$ и длиной волны $\lambda+\delta\lambda$ (при пересечении двух линий на уровне половинной мощности угловое расстояние между максимумами линий в два раза больше полуширины отдельной линии, см. рис. 4.10а). Величину $\delta\lambda$ можно определить используя выражение для угловой дисперсии (4.13): связь величин $\delta\theta$ и $\delta\lambda$ имеет вид

$$\delta\theta=\frac{1-r}{2\pi\sqrt{r}}\frac{\lambda}{2L}\frac{1}{\theta_m}\approx D\delta\lambda/2=\frac{m}{4L\theta_m}\delta\lambda$$

а)




б)

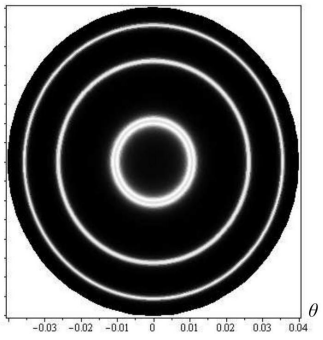



Рис. 4.11. Условие Релея для интерферометра Фабри-Перо:

а) интенсивности близких линий и их сумма (схематично); б) расчётное изображение спектра двух близких линий (здесь $r=0,85, \lambda=5760,0\text{Å}$,

$$\delta\lambda=0,09\text{Å}, L=1\text{m})$$

Разрешающая способность для порядка спектра $m \approx 2L/\lambda$ равна

$$R = \frac{\lambda}{\delta\lambda} = \frac{\pi\sqrt{r}}{1-r}m \tag{4.20}$$

Разрешающая способность пропорциональна порядку спектра, который велик и практически постоянен, и эффективному числу волн $\frac{\pi\sqrt{r}}{1-r}$, аналогу числа штрихов дифракционной решётки $N$.

На рис. 4.11а показана зависимость от угла интенсивности пары колец для линий равной мощности. Напомним, что для решётки максимум интенсивности второй линии попадал на минимум интенсивности первой линии. Отношение величины провала интенсивности между кольцами к максимальному значению интенсивности составляет $\approx 0,83$, что близко к числовой характеристике критерия Релея 0,81. На рис. 4.11 приведено изображение колец. По первой паре колец глаз легко разрешает линии.

Рассчитаем разрешающую способность спектральных приборов по условию Релея. Дифракционная решётка с плотностью штрихов $500\text{m}^{-1}$ и размером однородной области $1$ cm, $N \approx 5 \cdot 10^3$ в первом порядке спектра имеет $R = 5 \cdot 10^3$ и $\delta\lambda = 1,1$Å для $\lambda = 5760$Å, во втором $R = 10^4; \delta\lambda = 0,5$Å. Призма с $\left|\frac{dn}{d\lambda}\right| \approx 10^3$ cm$^{-1}$ и размером основания $a = 3$ см: $R \approx 3 \cdot 10^3, \delta\lambda \approx 2$Å. Интерферометр Фабри-Перо: $L = 1$, $r \approx 0,85; R \approx 6,6 \cdot 10^4, \delta\lambda \approx 0,09$Å, порядок спектра $2L/\lambda \approx 3,5 \cdot 10^3$, эффективное число волн $\frac{\pi\sqrt{r}}{1-r} \approx 19$.

## 8.0.5 5. The Dispersion Region

При большой ширине спектра спектры различных порядков могут накладываться друг на друга. На рис. 4.3 отмечено наложение «красной» границы спектра 5-го порядка на «фиолетовую» границу спектра 6 -го порядка. Предельная ширина спектрального интервала $\Delta\lambda$, при которой спектры соседних порядков перекрываются своими границами, называется дисперсионной областью. При этом $m(\lambda + \Delta\lambda) = (m+1)\lambda$ и дисперсионная область

$$\Delta\lambda = \frac{\lambda}{m} \tag{4.21}$$

Ограничений на оптические спектры не возникает при использовании призмы и дифракционных решёток, работающих в первом порядке спектра, для них $\Delta\lambda \approx \lambda$.

Возможная для исследования ширина спектра источника уменьшается при увеличении рабочего порядка спектра дифракционной решётки. Особые требования к источнику возникают при использовании интерферометра Фабри-Перо, так как его рабочий порядок спектра весьма высок $(2L/\lambda)$ и дисперсионная область

$$\Delta\lambda = \frac{\lambda^2}{2L} \tag{4.22}$$

очень мала. Для порядка спектра $2L/\lambda \approx 5 \cdot 10^3$ дисперсионная область $\Delta\lambda \approx 1$Å. Интерферометр Фабри-Перо обладает очень высокой разрешающей способностью, но при этом может использоваться только для анализа узкополосных спектров. Обычно перед интерферометром располагают светофильтры или другие спектральные приборы, вырезающие вблизи $\lambda$ полосу шириной менее $\Delta\lambda$.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§36, 46, 47, 49.

2. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. §8.9.

## 8.1 Laser Radiation

Лазер, или оптический квантовый генератор, - источник квазимонохроматического и узконаправленного высококогерентного потока излучения, работающий за счёт квантово-механического эффекта вынужденного (индуцированного) излучения (LASER = Light Amplification by Stimulated Emission of Radiation - «усиление света посредством вынужденного излучения»).

Главными элементами лазера являются оптический резонатор и расположенная в нём активная среда, способная усиливать проходящее через неё излучение.

Простейший резонатор представляет собой интерферометр Фабри-Перо (см. раздел IV), состоящий из двух плоских зеркал с высокими коэффициентами отражения, размещённых параллельно друг другу на фиксированном расстоянии. Благодаря наличию активной среды, в резонаторе многократно усиливаются волны, распространяющиеся вдоль оси системы и набирающие за один полный проход резонатора фазу, кратную $2\pi$ (т.е. на оптической длине резонатора укладывается целое число полуволн, в системе при этом образуются стоячие волныи). Таким образом, резонатор обеспечивает создание положительной обратной связи в лазере и превращает его в генератор излучения. Также в резонаторе происходит накопление энергии излучения и отбор узких резонансных линий из спектра излучения, рождающегося в среде. Одно из зеркал резонатора обычно имеет несколько меньший коэффициент отражения, что позволяет выпускать через него часть излучения в виде узконаправленного высокомонохроматического пучка (см. рис. 5.1).

Принцип усиления в активной среде основан на использовании эффекта вынужденного (индуцированного) излучения. Вынужденное излу чение - процесс, обратный поглощению фотона атомом. В нём атом, находящийся в возбуждённом состоянии, под действием налетающего на него фотона переходит в состояние с меньшей энергией, испуская при этом ещё один фотон, частота, поляризация и фаза которого совпадают с таковыми у исходного фотона. Иными словами, падающий и рождённый в результате индуцированного излучения фотоны являются полностью когерентными.




Рис. 5.1. Схема лазера

Вероятность поглощения фотона при взаимодействии с атомом в невозбуждённом состоянии равна вероятности индуцированного испускания при взаимодействии фотона с возбуждённым атомом. Поэтому для того, чтобы среда в целом усиливала излучение, количество возбуждённых атомов должно быть больше, чем количество невозбуждённых. Такое состояние среды не является термодинамически равновесным и называется состоянием с инверсной заселённостью уровней. Для создания инверсной заселённости используют различные методы накачки. Источником первичных фотонов служит процесс спонтанного (т.е. самопроизвольного) излучения возбуждённых атомов.

Рассмотрим подробнее различные аспекты работы лазеров.

## 8.1.1 1.1. Elementary Processes in Laser

Как известно из квантовой механики, энергия электронов в атомах способна принимать лишь некоторый дискретный набор значений энергетических уровней $E_i$. Если энергия электронов атома минимальна (и равна $E_0$), то говорят, что атом находится в основном состоянии. Если атом находится на любом другом энергетическом уровне $E_i > E_0$ (т. е. его электроны имеют энергию $E_i$), то говорят, что атом находится в возбуждённом состоянии. Рассмотрим, для определённости, переходы электрона между основным $E_0$ и первым возбуждённым $E_1$ уровнями энергии (всё сказанное в этом параграфе справедливо и для любого другого перехода между атомными уровнями).

Изменение энергии электрона в атоме может сопровождаться испусканием или поглощением энергии в виде кванта электромагнитного излучения - фотона. Энергия фотона равна $\hbar\omega$, где $\omega$ - частота коле баний электромагнитного поля, $\hbar = h/2\pi$ ($h \approx 6,6 \cdot 10^{-34}$ Дж с с постоянная Планка). Согласно принципам квантовой механики, электрон в атоме, находящийся в состоянии $E_0$ или $E_1$, может провзаимодействовать с фотоном энергии $\hbar\omega$, если выполняется равенство $E_1 - E_0 = \hbar\omega$.

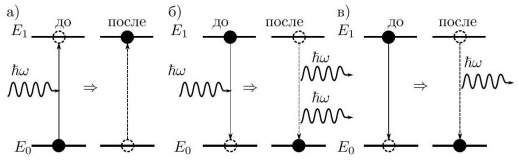



Рис. 5.2. Переходы между уровнями $E_1$ и $E_0$

Можно выделить три главных элементарных процесса, определяющих взаимодействие излучения с активной средой лазера.

1. Поглощение фотона невозбуждённым атомом $\hbar\omega + A \rightarrow A^*$. Атом при взаимодействии с внешним электромагнитным полем поглощает квант энергии $\hbar\omega$ и переходит из состояния $E_0$ в состояние $E_1 = E_0 + \hbar\omega$ (рис. 5.2а).

2. Вынужденное (индуцированное) излучение при взаимодействии атома с падающим на него электромагнитным полем $\hbar\omega + A^* \rightarrow A + 2\hbar\omega$ (рис. 5.2б). При этом взаимодействии атом переходит в основное состояние, испуская ещё один фотон той же частоты $\omega$. Кроме того, кванты электромагнитного поля, излучённые при индуцированных переходах, полностью тождественны квантам поля, вызвавшего переход. Иными словами, вновь родившийся фотон имеет ту же фазу, поляризацию и направление распространения (волновой вектор), что и исходный фотон, то есть является полностью когерентным с ним.

Согласно законам квантовой механики, вероятности того, что атом поглотит (переход $0 \rightarrow 1$) или вынужденно испустит (переход $1 \rightarrow 0$) фотон (квант энергии электромагнитного поля $\hbar\omega$), отличны от нуля только для поля резонансной частоты $\hbar\omega = E_1 - E_0$. Кроме того, вероятности переходов $0 \rightarrow 1$ и $1 \rightarrow 0$ равнъ между собой (в силу принципа детального равновесия) и пропорциональны спектралъной плотности энергии внешнего поля $\rho_\omega$ (которая пропорциональна количеству фотонов, имеющих данную частоту). Это можно записать как

$$W_{0\rightarrow1} = W_{1\rightarrow0} = B\rho_\omega \tag{5.1}$$

где $W_{0\leftrightarrow1}$ - вероятности в единицу времени [1] для соответствующих переходов, $B$ - коэффициент, характеризующий переходы между рассматриваемыми уровнями энергии, не зависящий от величины поля (так называемый коэффициент Эйнштейна для вынужденных переходов).

3. Спонтанное излучение фотона возбуждённым атомом $A^* \to A + \hbar\omega$ (рис. 5.2в). Время жизни возбуждённых состояний - конечная величина, поэтому по прошествии некоторого времени возбуждённый атом испустит фотон частоты $\omega = (E_1 - E_0)/\hbar$. Скорость этого процесса является константой, не зависящей от наличия внешнего поля. Фаза и направление распространения испущенного фотона будут случайны, так что данный процесс препятствует сохранению когерентности излучения лазера. Кроме того, он мешает поддержанию инверсной заселённости (см. п. 1.5) уровней энергии атомов. С другой стороны, спонтанное излучение выполняет важную функцию источника первичных фотонов, являясь «затравкой» для формирования направленного пучка лазерного излучения.

## 8.1.2 1.2. Gain Factor

Изменение интенсивности электромагнитной волны $dI$, прошедшей участок поглощающей (или усиливающей) среды толщиной $dx$, пропорционально интенсивности волны на этом участке $I(x) : dI = -\alpha I dx$, где $\alpha$ - коэффициент поглощения на единицу длины среды. Если $\alpha = \text{const}$, то в результате интегрирования получаем, что интенсивность волны в среде изменяется по закону

$$I(x) = I_0 \exp(-\alpha x)$$

(закон Бугера-Ламберта-Бера). Если $\alpha < 0$, то говорят о коэффициенте усиления среды $\gamma = -\alpha$, интенсивность при этом экспоненциально нарастает:

$$I(x) = I_0 \exp(\gamma x) \tag{5.2}$$

Рассмотрим прохождение через среду плоской электромагнитной волны, частота $\omega$ которой удовлетворяет резонансному условию $\hbar\omega = E_1 - E_0$, где $E_1$ и $E_0$ - некоторые уровни энергии атомов среды.

Пусть $N_1$ и $N_0$ - концентрации атомов с энергиями $E_1$ и $E_0$ соответственно. Согласно формуле (5.1), доля «невозбуждённых» атомов (находящихся на нижнем энергетическом уровне), поглощающих квант излучения в единицу времени, равна

$$\frac{dN_0}{N_0} = W_{0\to1}dt = B\rho_\omega dt$$

Убыль фотонов в единице объёма при этом равна

$$dN_\Phi^- = -dN_0 = -B\rho_\omega N_0 dt$$

Аналогично, количество фотонов $dN_\phi^+$, рождённых за счёт вынужденного излучения [2], равно

$$dN_\Phi^+ = B\rho_\omega N_1 dt$$

При прохождении волны сквозь среду изменение интенсивности волны пропорционально суммарному количеству рождённых и поглощённых квантов (в расчёте на единицу длины в направлении движения волны). Тогда для волны, прошедшей участок длиной $dx$ за время $dt = dx/v$, где $v = c/n$ - скорость волны, относительное изменение интенсивности равно

[0]1 Под «вероятностью в единицу времени» имеется в виду доля атомов $dN/N$, совершающих соответствующий переход за время $dt$, или, иными словами, скорость процесса, нормированная на полное число атомов: $W \equiv \frac{1}{N}\frac{dN}{dt}$.

$$\frac{dI}{I} = \frac{dN_{\Phi}^{+} + dN_{\Phi}^{-}}{\rho_{\omega}/\hbar\omega} = B\frac{\hbar\omega}{v}\left(N_1 - N_0\right)dx$$

Таким образом, коэффициент усиления волны с частотой $\omega$ в активной среде лазера равен

$$\gamma = B\frac{\hbar\omega}{v}\Delta N \tag{5.3}$$

где $\Delta N = N_1 - N_0$. Отсюда видно, что среда является усиливающей ($\gamma > 0$), если концентрация атомов на верхнем уровне больше, чем на нижнем: $N_1 > N_0$.

## 8.1.3 1.3. The Generation Spectrum. Doppler Broadening

Всё сказанное в предыдущем параграфе относилось к монохроматической волне и бесконечно узкой линии поглощения/излучения $\omega = \left(E_1 - E_0\right)/\hbar$. В реальности эта линия имеет конечную ширину $\Delta\omega_{\gamma}$, и коэффициент усиления (5.3) должен быть домножен на функцию, пропорциональную форме контура этой спектральной линии. Иными словами, коэффициент усиления $\gamma(\omega)$ есть функция частоты с острым максимумом вблизи резонансной частоты $\omega = \left(E_1 - E_0\right)/\hbar$, обладающая некоторой конечной шириной $\Delta\omega_{\gamma}$.

Ширина спектра усиления активной среды лазера $\Delta\omega_{\gamma}$ (спектра генерации) определяется естественной шириной резонансной линии лазерного перехода и различными механизмами уширения.

Естественная ширина линии - внутренняя характеристика атома, определяемая строением его энергетических уровней. Если известно время жизни возбуждённого состояния $\tau_{\text{e}}$, естественная ширина $\Delta\omega_{\text{e}}$ может быть оценена из соотношения неопределённостей $\Delta\omega_{\text{e}} \sim 2\pi/\tau_{\text{e}}$. В модели цугов (см. раздел II, п. 3) время $\tau_{\text{e}}$ может быть отождествлено с длительностью цуга, излучаемого отдельным атомом.

Основным механизмом уширения в газовых лазерах, определяющим ширину спектральной линии излучения среды, является эффеетm Доплер*а*, действие которого схематически изображено на рис. 5.3.

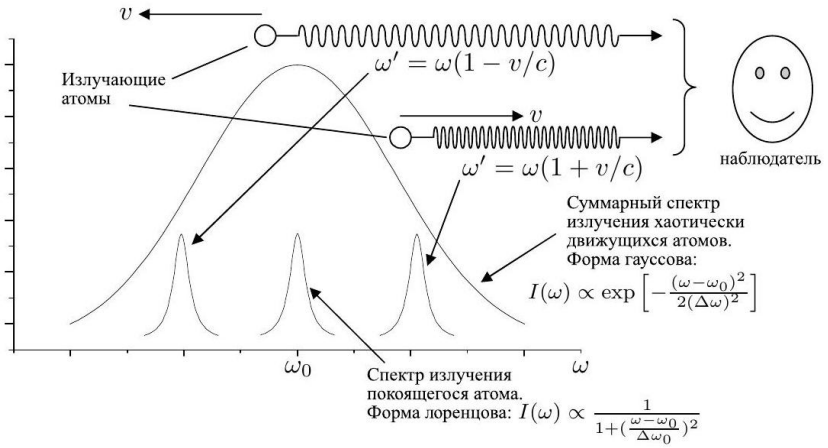



Рис. 5.3. Доплеровское уширение спектра излучения атомов вследствие теплового движения

[0]2 Заметим, что вероятность спонтанного излучения пропорциональна концентрации возбуждённых атомов и не зависит от плотности энергии излучения $\rho_{\omega}$, поэтому при больших $\rho_{\omega}$ ей можно пренебречь по сравнению с вероятностью вынужденного излучения.

В лазерной трубке атомы среды участвуют в хаотическом тепловом движении. Частота излучения движущегося источника сдвинута относительно неподвижного источника; в нерелятивистском случае

$$\frac{\Delta\omega_{\text{д}}}{\omega} \sim \frac{v}{c}\cos\alpha$$

где $\alpha$ - угол между вектором скорости и направлением на наблюдателя. Результирующий спектр излучения есть сумма вкладов от большого количества атомов среды (движущихся с различными скоростями и
в разных направлениях). Поскольку при хаотическом движении $\cos\alpha$ принимает значения от 1 до -1, то полуширина линии приблизительно составляет

$$\Delta\omega_{\text{д}} \sim \omega\frac{v_T}{c} = \frac{\omega}{c}\sqrt{\frac{8k_{\text{B}}T}{\pi m}} \tag{5.4}$$

где $v_T$ - средняя тепловая скорость движения атомов [3].

## 8.1.4 1.4. The Condition for Reaching the Generation Threshold

Предположим, что активная среда с коэффициентом усиления $\gamma$ находится между двумя зеркалами с коэффициентами отражения по энергии $r_1$ и $r_2$ (рис. 5.1). Пусть световой пучок распространяется из некоторой точки внутри резонатора и после отражения от двух зеркал возвращается в эту же точку. При длине промежутка $L$ усиление за один полный проход резонатора с учётом формул (5.2) и (5.3) равно

$$G = e^{2\gamma L}$$

Учтём, что в лазере имеются дополнительные потери энергии, так что за один проход интенсивность из-за них составит долю $T$ от исходной [4]. Для самовозбуждения лазера необходимо, чтобы усиление было достаточным для компенсации всех потерь при полном обходе резонатора: $r_1 r_2 TG \geqslant 1$, откуда

$$e^{\gamma L} \geqslant \frac{1}{\sqrt{Tr_1r_2}} \tag{5.5}$$

или

$$2\gamma L \geqslant -\ln T - \ln r_1 r_2$$

[3] Строгий вывод на основе распределения Максвелла позволяет найти форму спектра излучения, представляющую собой гауссову функцию $I \propto \exp\left[\frac{(\omega-\omega_0)^2}{2\Delta\omega^2}\right]$ со стандартным отклонением $\Delta\omega = \frac{\omega}{c}\sqrt{\frac{kT}{m}}$.

[4] Потери могут быть связаны с рассеянием и поглощением на дефектах и примесях среды, а также на дополнительных элементах лазера, таких как стеклянные окна лазерной трубки. Возможны и другие виды потерь, приводящие к уменьшению пропускания, например, за счёт дифракционного расплывания пучка. Дифракционные потери становятся существенными при малых поперечных размерах зеркал или активного элемента $D$, сравнимых с размером одной зоны Френеля для расстояния, равного длине резонатора, то есть $D \sim \sqrt{\lambda L}$.

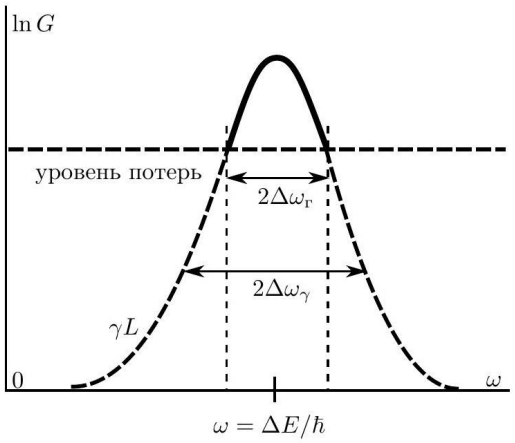



Рис. 5.4. Контур усиления и уровень потерь

Заметим, что коэффициент усиления $\gamma(\omega)$ зависит от частоты и генерация может происходить в диапазоне частот $\omega \pm \Delta\omega_{\text{г}}$ , в котором справедливо условие (5.5) (см. рис. 5.4).

В лазерах с непрерывной генерацией в установившемся режиме потери излучения в точности компенсируются усилением, то есть усиление активного элемента за один проход равно $G = 1/(Tr_1r_2)$. При этом следует иметь в виду, что для того, чтобы лазер имел ненулевую мощность выходного излучения, активный элемент лазера должен иметь запас по усилению, то есть при отключении положительной обратной связи (удалении зеркал резонатора) усиление должно превышать порог, задаваемый правой частью неравенства (5.5). При подключении обратной связи и выходе на стационарный режим генерации усиление автоматически падает до величины $1/(\text{Tr}_1\, r_2)$ за счет сброса возбужденных атомов с верхнего лазерного уровня на нижний лазерным излучением.

## 8.1.5 1.5. Reverse Occupancy. Pumping

Как показано в п. 1.2, для создания среды, которая способна усиливать проходящее через неё излучение, необходимо, чтобы концентрация атомов на верхнем лазерном уровне была больше, чем на нижнем: $N_1 > N_0$. В таком случае говорят, что уровни энергий атомов в активной среде имеют инверсную заселённость.

Напомним, что в равновесии при температуре $T$ отношение заселённостей энергетических уровней $E_1$ и $E_0$, согласно распределению

Больцмана, равно $\frac{N_1}{N_0} = \exp\left(-\frac{E_1-E_0}{k_5T}\right)$. Если $N_1 > N_0$, то формально $T < 0$ и такое состояние не может быть равновесным. Система в состоянии инверсной заселённости, приведённая в контакт с окружающей средой, быстро отдаст ей избыток энергии и вернётся к нормальной заселённости. Однако если имеется постоянная подкачка энергии извне (система незамкнута), то неравновесное состояние может быть стационарным.

Для создания инверсной заселённости уровней используют различные методы накачки: атомы «перекачиваются» из основного состояния в возбуждённое с помощью внешних источников энергии. В качестве таких источников часто используются источники электромагнитного излучения нужной частоты (оптическая накачка).

Ввиду того, что вероятность поглощения равна вероятности вынужденного испускания ($B_{10} = B_{01}$), инверсную заселённость в двухуровневой системе невозможно создать оптической накачкой. Поэтому для создания инверсной населённости чаще всего используют трёх- и четырёхуровневые системы. На рис. 5.5 изображена трехуровневая система, используемая в оптическом диапазоне (например, в рубиновом лазере).

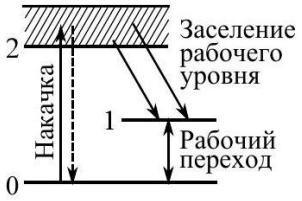



Рис. 5.5. Схема получения инверсной заселённости в трёхуровневой системе Излучение накачки переводит атомы из состояния 0 в состояние 2. Из состояния 2 атомы могут вернуться либо в состояние 0, либо перейти в 1. На уровне 1 атомы будут накапливаться в том случае, если время жизни на нём велико (метастабильный уровень). При этом условии на этом уровне можно накопить достаточное количество атомов, превосходящее числом количество атомов в основном состоянии. Время жизни $\tau$ связано с шириной уровня соотношением неопределённостей $\Delta E\tau \sim \hbar$, поэтому желательно, чтобы уровень 1 был узким. Легко видеть, что уровень 2, наоборот, должен быть широким - в этом случае удаётся полезно использовать заметную часть спектра оптического излучения накачки. Чаще всего роль уровня 2 играет широкая полоса. Переход $2 \to 1$ нередко является безызлучательным; в таких переходах излишняя энергия атома передаётся не кванту излучения, а кристаллической решётке (твёрдые тела) или сталкивающимся атомам (газы).

## 8.2 2. The Mode Composition of the Laser Radiation

Модами (англ. mode - состояние, режим) называют стационарные типы колебаний электромагнитного поля в резонаторе, различающиеся частотой и пространственным распределением амплитуды поля.

Рассмотрим моды в открытом резонаторе Фабри-Перо с плоскими зеркалами, расстояние между которыми равно $L$. Будем считать, для простоты, что активная среда заполняет весь резонатор и имеет показатель преломления [5] $n = 1$.

### 8.2.1 2.1. Longitudinal Modes

Зеркала обеспечивают возможность многократного прохода плоской волны, если её волновой вектор направлен по оси интерферометра $\boldsymbol{k} = (0, 0, k)$ (продольные моды). Если же он имеет наклон к оси, то после нескольких проходов волна покинет область пространства между зеркалами и не будет в достаточной мере усилена.

Большая часть энергии поля в резонаторе накапливается в виде *стоя* - яиих волн, образованных встречными волнами, бегущими вдоль оси системы $x$ : $E \propto \sin(\omega t)\sin(kx)$ (вектор $\boldsymbol{E}$ лежит в плоскости зеркал и перпендикулярен оси $x$). Если зеркала проводящие (металлические), то минимальные потери будут испытывать моды, у которых напряжённость поля на поверхности зеркал ($x = 0, x = L$) равна нулю, т. е. выполняется условие $k_q L = \pi q$, где $q = 1, 2, 3, \ldots$, и $k_q = \omega_q/c$ - волновое число внутри резонатора. Отсюда имеем

$$L = q\frac{\lambda}{2}, \qquad \text{где } q = 1, 2, 3, \ldots \tag{5.6}$$

то есть на длине резонатора укладывается целое число полуволн [6]. Набор резонансных частот (собственных частот резонатора):

$$\omega_q = q\frac{\pi c}{L}, \quad \left(\nu_q = q\frac{c}{2L}\right) \quad q = 1, 2, 3, \ldots \tag{5.7}$$

Таким образом, резонатор выделяет из спектра генераиии активной среды лазера ряд узких спектралъных линий центры которых соответствуют собственным частотам $\omega_q$ резонатора (5.7).

5 Если $n \neq 1$, необходимо учитывать, что длина волны $\lambda$ в резонаторе, заполненном активной средой, отличается от излучаемой $\lambda_0 : \lambda = \lambda_0/n$.

6 В общем случае (например, для диэлектрических зеркал) максимальным усилением в резонаторе обладают волны, для которых набег фазы при полном обходе резонатора кратен $2\pi$. Отсюда нетрудно получить, что формулы (5.6) и (5.7) остаются справедливыми, если под $L$ понимать оптическую длину резонатора, учитывающую показатель преломления среды и возможный набег фазы при отражении от зеркала.

## 8.2.2 2.2. The Spectrum of Laser Radiation

Моды колебаний в резонаторе не являются строго монохроматическими и каждая содержит узкий спектр частот в некотором интервале $\omega_q \pm \Delta\Omega$. Как известно из теории колебаний, полуширина резонансного пика $\Delta\Omega$ связана с добротностью системы $Q : \Delta\Omega \sim \omega_q/Q$.

Рассчитаем сначала добротность резонатора Фабри-Перо $Q_{\text{рез и}}$ ширину его линий пропускания $\Delta\Omega_{\text{рез без учёта усиления в активной}}$ среде и потерь, не связанных с отражением от зеркал. Воспользуемся энергетическим смыслом добротности: $Q_{\text{рез}} = 2\pi\frac{W}{\Delta W_T}$, где $W$ - запасённая энергия в резонаторе, $\Delta W_T$ - её потери за период колебаний. Пользуясь обозначениями, введёнными в п. 1.4, запишем относительную потерю энергии волны за один проход пустого резонатора: $\Delta W/W = 1 - r_1 r_2$, откуда с учётом того, что за время одного прохода совершается $L/\lambda$ колебаний, получим

$$Q_{\text{рез}} \sim \frac{2\pi L}{\lambda}\frac{1}{1-r_1 r_2}, \quad \Delta\Omega_{\text{рез}} \sim \frac{\omega_q}{Q_{\text{рез}}}$$

(ср. с формулой разрешающей способности интерферометра ФабриПеро (4.20)). Для типичных параметров лазерного резонатора $r = 0,99, L = 0,5$ m, $\nu = 5 \cdot 10^{14}$ Гц получим $Q_{\text{рез}} \sim 5 \cdot 10^8$, $\Delta\nu_{\text{рез}} = \Delta\Omega_{\text{рез}}/2\pi \sim 10^6$ Гц. Функция пропускания резонатора Фабри-Перо для продольных мод (см. формулу (4.8) при $\theta = 0$) представлена на рис. 5.6.

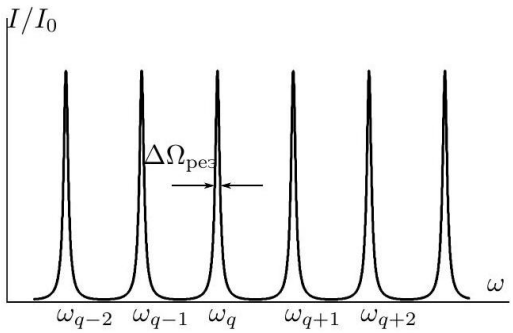



Рис. 5.6. Спектр пропускания резонатора Фабри-Перо

Резюмируя сказанное выше, мы теперь можем описать процесс формирования спектра излучения лазера. Накладывая на спектр генерации активной среды (рис. 5.4) спектр пропускания резонатора Фабри-Перо (рис. 5.6), получим набор мод, которые могут генерироваться лазером. Заметим, что ввиду наличия усиления в активной среде, реальная
ширина $\Delta\Omega$ генерируемых лазером спектральных линий может быть и значительно меньше ширины линии пропускания резонатора $\Delta\Omega_{\text{рез}}$ (см. далее п. 2.3).

Рассмотрим в качестве примера типичный спектр излучения газовых лазеров. Ширина спектра генерации определяется эффектом Доплера $\Delta\omega_\gamma \sim \Delta\omega_д$ (см. (5.4)) и обычно больше межмодового расстояния $\omega_{q+1} - \omega_q = \frac{\pi c}{L}$. В таком случае формируется спектр излучения лазера, представленный на рис. 5.7: из доплеровского контура (полушириной $\omega_д$ «вырезается» ряд узких спектральных линий, причём «выживают» только те линии, усиление которых выше уровня потерь. Такой режим работы лазера называют многомодовым. Количество генерируемых мод зависит от соотношения усиления и потерь. Если усиление лишь немного выше уровня потерь, то возможна ситуация, когда будет возбуждена только центральная линия и режим работы лазера будет одномодовым.

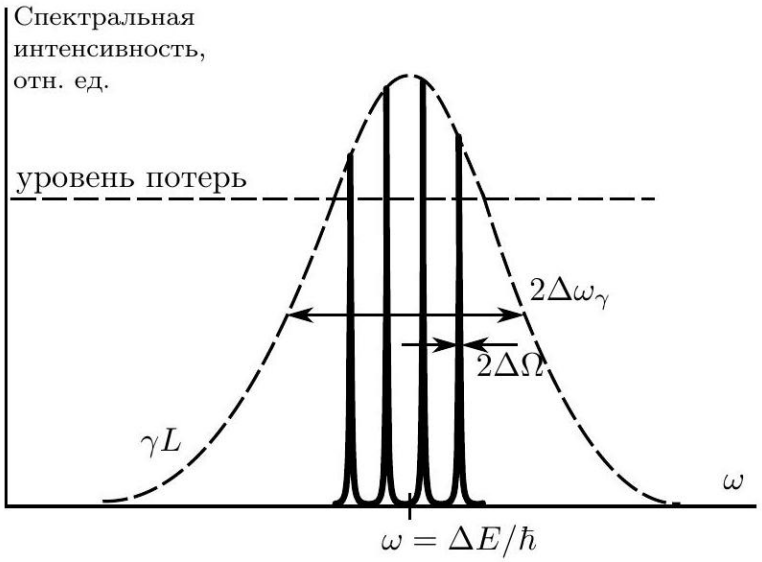



Рис. 5.7. Схематическое изображение генерируемых лазером мод

В реальности амплитуды и фазы генерируемых лазером мод могут испытывать сильные флуктуации ввиду влияния процессов спонтанного излучения и конкуренции мод за использование инверсной заселённости (излучение любой моды сбрасывает часть возбуждённых атомов в основное состояние, уменьшая разность заселённостей $\Delta N$, а значит, и усиление для себя и для других мод согласно формуле (5.3)). Тепловые флуктуации длины $L$ резонатора приводят к смещению положений всех мод (межмодовое расстояние при этом практически не меняется) по спектру и, следовательно, к изменению их количества и средней амплитуды.

## 8.2.3 2.3. Coherence of Laser Radiation

Как было указано в предыдущем параграфе, ширина спектра $\Delta\Omega$ излучения лазера может быть во много раз меньше ширины линии пропускания резонатора $\Delta\Omega_{рез}$. Дело в том, что поскольку при вынужден- . ном излучении активной среде рождаются фотоны, тождественные исходным, эффективное время жизни фотона в лазере многократно возрастает (по сравнению с временем его жизни в пустом резонаторе), из чего, ввиду соотношения неопределённостей, следует сужение результирующего спектра излучения. С другой стороны, с точки зрения теории колебаний наличие усиления в резонаторе повышает результирующую добротность системы, что опять-таки означает сужение резонансной линии.

Опишем, как происходит процесс сужения спектра линий излучения во времени. Прежде всего заметим, что усиление в центре контура усиления $\gamma(\omega)$ больше, чем на любой частоте, отличной от центральной. В начальный момент после возбуждения

активной среды из имеющегося в среде спонтанного излучения выделяются те фотоны, для которых усиление максимально. Они будут размножаться, опустошая верхний лазерный уровень и не оставляя возможности размножаться другим фотонам. С каждым проходом резонатора фотоны, частоте которых соответствует максимальное усиление, размножаются, а фотоны с меньшим усилением «выбывают из игры», что и приводит к постепенному сужению спектра [7].

Фундаментальным фактором, препятствующим уменьшению ширины спектра излучения до нуля (и росту времени когерентности лазера до бесконечности) является спонтанное излучение. С каждым обходом резонатора такие характеристики излучения, как спектральный состав, форма огибающей и фаза, слегка изменяются, поскольку часть излучения постоянно покидает резонатор через выходное зеркало и на его место приходит та часть усиленного спонтанного излучения, которая попадает в полосу усиления лазерной среды и совпадает по частотному и пространственному спектру со спектром мод резонатора. При этом фаза «подмешанного» спонтанного излучения имеет случайное значение. Ясно, что ширина линии и степень когерентности будут зависеть от доли спонтанного излучения в выходном излучении лазера. Вклад спонтанного излучения определяется в основном выходной мощностью лазера (мощность спонтанного излучения определяется концентрацией возбуждённых атомов) и, конечно, шириной линии пропускания используемого резонатора $\Delta\Omega_{\text{рез}}$. Чем выше выходная мощность, тем меньше вклад спонтанного излучения и уже линия. Шавлов и Таунс получили формулу для вычисления ширины линии в зависимости от этих факторов. Приведём её в упрощённом виде:

$$\frac{\Delta\Omega}{\omega} \sim \frac{(\Delta\Omega_{\text{рез}})^2}{P} \cdot \frac{N_2}{\Delta N}$$

где $P$ - выходная мощность лазера. Расчёты, выполненные по формуле Шавлова-Таунса, дают для типичного гелий-неонового лазера с выходной мощностью 1 мВт относительную ширину линии одной моды $\Delta\Omega/\omega \sim 10^{-15}$, и $\Delta\nu = \Delta\Omega/2\pi \sim 1$ Гц. Это чрезвычайно малая величина: оценки показывают, что для того, чтобы уход частоты лазерного излучения вследствие тепловых нестабильностей длины резонатора не превышал этой ширины, необходимо поддерживать длину резонатора постоянной с точностью $10^{-13}$ см. Это требует стабилизации температуры $\Delta T \leqslant 10^{-8}$ K, что обеспечить крайне трудно. Поэтому в обычном лазере ширина линии и связанное с ней время когерентности определяются техническими нестабильностями резонатора. Типичная ширина спектра одной моды гелий-неонового лазера составляет $\sim 10^6$ Гц, соответственно время когерентности $\sim 10^{-6}$ с, а длина когерентности несколько сотен метров. Для сравнения: полупроводниковый лазер, например, лазерная указка, без принятия специальных мер по сужению спектра и стабилизации частоты имеет длину когерентности менее одного сантиметра.

Интерференция от разных источников. При использовании специальных методов стабилизации частоты в гелий-неоновых лазерах удалось достичь ширины линии порядка нескольких Гц и ещё недавно казавшейся фантастической относительной стабильности частоты $10^{-14}$. Наличие источников света с такими параметрами позволило наблюдать эффекты, считавшиеся ранее невозможными, например, интерференцию между пучками от двух разных лазеров. Длительное время среди

[7] Ввиду того, что среди возможных мод одна всегда имеет наибольшее усиление по сравнению с остальными, со временем эта мода должна, казалось бы, остаться единственной, «перехватив» на себя всю инверсную заселённость активной среды. Однако этого не происходит и многие лазеры излучают стационарно в многомодовом режиме. Это объясняется тем, что моды различаются не только частотой, но и имеют разные пространственные распределения электромагнитного поля: узлы и пучности стоячих волн соседних мод не совпадают, что и позволяет одновременное сосуществование разных мод.

физиков бытовало мнение, что интерферировать могут только пучки, полученные от одного источника. Дираку принадлежит ещё более сильное утверждение: «... каждый фотон интерферирует только с самим собой. Интерференция между различными фотонами невозможна». Считалось, что вследствие случайных уходов фазы излучения, происходящих независимо у разных источников света, интерференционная картина от двух источников столь быстро «дёргается», что её невозможно наблюдать. И это было фактически верно в долазерную эпоху, поскольку
не было методов получения излучения достаточной мощности с временем сбоя фазы (фактически временем когерентности) более $10^{-8} \div 10^{-7}$ с и не существовало методов регистрации изображений с временным разрешением лучше $10^{-6}$ с. Поэтому интерференционная картина смазывалась быстрее, чем её можно было зарегистрировать. Развитие лазерной техники позволило создать источники света с временем когерентности порядка секунды, что позволило провести успешные эксперименты по наблюдению интерференции пучков от разных лазеров.

### 8.2.4 2.4. Transverse Modes*

Излучение, распространяющееся под небольшим углом к оси резонатора, может испытать достаточное число проходов, чтобы заметно усилиться и присутствовать в выходном излучении. При этом допустимые углы должны удовлетворять определённым соотношениям, вытекающим из свойств резонаторов. Анализ углового распределения лазерного излучения удобнее проводить для случая квадратных зеркал резонатора, считая, что поперечный размер активной среды превышает ширину $D$ зеркал, т. е. лазерный пучок ограничивается зеркалами.

Простейшая теория лазерных резонаторов, предложенная Шавловым и Таунсом, основана на сходстве открытых (без боковых стенок) резонаторов с квадратными зеркалами с закрытыми. В процессе усиления излучения от спонтанных шумов до стационарного уровня «выживают» только моды, имеющие минимальные потери на дифракцию. Для этого интенсивность излучения на краях зеркал должна быть минимальной, в идеале равной нулю. Если какая-либо мода содержит две волны, распространяющиеся под малыми углами $\pm\theta$ к оси резонатора, то при сложении этих волн на зеркале создаётся интерференционная картина с периодом $\lambda/2\theta$. Для обращения интенсивности в ноль на краях зеркала необходимо, чтобы на ширине зеркала $D$ укладывалось целое число периодов, поэтому $\theta = \pm m\lambda/2D$. Такое же условие должно выполняться по второй координате в плоскости зеркала: $\theta = \pm p\lambda/2D$. Здесь $m$ и $p$ - целые числа. Результирующий угол равен

$$\theta = \pm\frac{\lambda}{2D}\sqrt{m^2+p^2} = \frac{\theta_д}{2}\sqrt{m^2+p^2} \tag{5.8}$$

где $\theta_д \approx \lambda/D$ - дифракционная расходимость лазерного пучка с поперечным размером $D$.

Для классификации лазерных мод применяют обозначение $\mathrm{TEM}_{mpq}$ (TEM = transverse electromagnetic mode), имея в виду, что и электрическое и магнитное поля $\boldsymbol{E}$ и $\boldsymbol{H}$ перпендикулярны направлению распространения волны. Типы колебаний с различающимися значениями
поперечных индексов $m$ или $p$ называются поперечными модами, с различными $q$ - продольными. Фактически значение поперечного индекса даёт число нулей в распределении интенсивности излучения по соответствующей координате на поверхности зеркала, не считая нулей на краях. Значение продольного индекса $q$ равно числу полуволн, укладывающихся на длине резонатора.

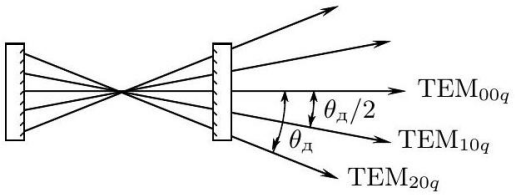



Рис. 5.8. Направления распространения мод с разными поперечными индексами

На рис. 5.8 и 5.9 схематически показаны направления распространения разных поперечных мод и распределение интенсивности излучения на зеркале по одной из поперечных координат.

Найдём сдвиг частоты поперечных мод относительно чисто продольных. Для моды TEM $_{mpq}$, распространяющейся под углом $\theta$ к оси резонатора, составляющая волнового вектора по оси резонатора равна $k_z = q\pi/L$, где $L$ - длина резонатора. Составляющая в плоскости зеркал равна $k_{x,y} = k\sin\theta \approx k_z\theta$. Учитывая, что $m, p \ll q, \theta \ll 1$, $\omega = kc$ и используя выражение (5.8) для $\theta$, нетрудно получить частоту поперечной моды с поперечными индексами $m, p$ и продольным $q$:

$$\omega_{mpq} \simeq \omega_{00q}\left[1 + \frac{(m^2+p^2)\,\lambda^2}{8D^2}\right] \tag{5.9}$$

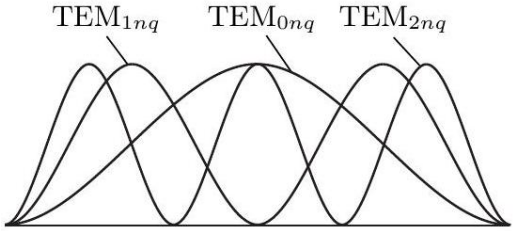



Рис. 5.9. Распределение интенсивности лазерного излучения на зеркале резонатора по одной из поперечных координат для трёх низших поперечных мод

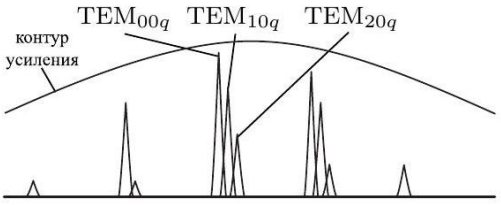
Сравним разницу частот продольной и поперечной мод (с тем же продольным индексом) с межмодовым расстоянием для продольных мод:

$$\frac{\omega_{mpq} - \omega_{00q}}{\omega_{00q+1} - \omega_{00q}} \simeq \frac{\lambda L\,(m^2+p^2)}{4D^2} \tag{5.10}$$

При длине резонатора лазера $L \approx 0,5$ м и длине волны $\lambda \approx 0,6$ мкм для низшей поперечной моды ($m = 1, p = 0$) имеем $\frac{\omega_{10q}-\omega_{00q}}{\omega_{00,q+1}-\omega_{00q}} \approx 10^{-2}$. Это означает, что спектр мод с ненулевыми поперечными индексами лишь незначительно сдвинут относительно спектра чисто продольных мод, этот сдвиг составляет единицы процентов от межмодового расстояния и не разрешается большинством спектральных приборов.

Типичный спектр излучения лазера с плоскими зеркалами схематически изображён на рис. 5.10. Для наглядности пропорции искажены: в действительности поперечные моды ближе к соответствующей чисто продольной моде, а ширина спектра отдельных мод меньше, чем изображено на рисунке. Соотношение между амплитудами различных мод носит случайный характер и меняется со временем.

Если резонатор составлен не из плоских, а из сферических вогнутых зеркал, то спектр мод с ненулевыми поперечными индексами сдвинут относительно чисто продольных мод гораздо сильнее. Например, в случае конфокального резонатора (радиус кривиз-




ны зеркал равен расстоянию меж-

Рис. 5.10. Примерный вид спектра излучения гелий-неонового лазера ду ними) изменение суммы поперечных индексов на 2 приводит точно к такому же спектральному сдвигу, как и изменение продольного индекса на единицу, в результате имеет место вырождение, т. е. совпадение частот разных мод. В реальных гелий-неоновых лазерах, как правило, используется резонатор, составленный из сферических зеркал, но с меньшей кривизной, чем в конфокальном резонаторе, что приводит к меньшему сдвигу частот при изменении поперечного индекса.

## 8.2.5 2.5. Lamb's Failure and Mode Competition*

Как было отмечено выше, полная ширина спектра излучения гелий-неонового лазера определяется доплеровским уширением (рис. 5.3). В диапазоне времён порядка времени обхода резонатора это уширение неоднородно. Неоднородным называют уширение, вызванное различным смещением спектров индивидуальных атомов или молекул относительно среднего положения. При этом спектр коллектива атомов шире спектра индивидуальных атомов. Противоположный случай, когда уширен спектр каждого атома относительно естественной ширины линии, но смещения спектров атомов друг относительно друга не происходит, называется однородным уширением, при этом спектр коллектива совпадает со спектром отдельных атомов. Следует отметить, что отнесение к определённому типу уширения зависит от времени наблюдения. Например, в случае гелий-неонового лазера характерное время столкновения между атомами составляет $10^{-6}$ с, а время обхода резонатора $10^{-8}$ с. За $10^{-8}$ с атомы не успевают изменить направление и скорость движения и каждый атом имеет своё положение спектра излучения, смещённое относительно центра полосы, что приводит к неоднородному уширению, как изображено на рис. 5.3. Если же наблюдать за атомами существенно дольше, чем время столкновения между ними, то за время наблюдения каждый атом успеет неод нократно изменить направление движения и скорость, спектр каждого атома будет совпадать со спектром коллектива и уширение будет однородным.

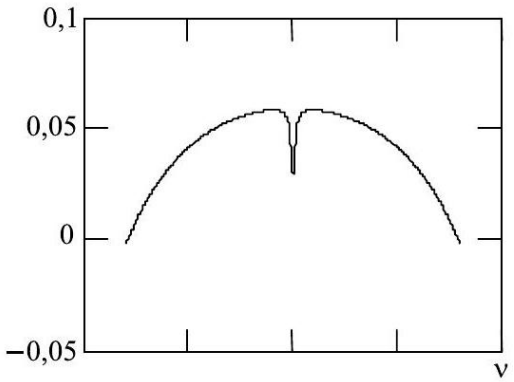



Рис. 5.11. Расчётный спектр коэффициента усиления гелий-неонового лазера при одной продольной моде в центре контура усиления

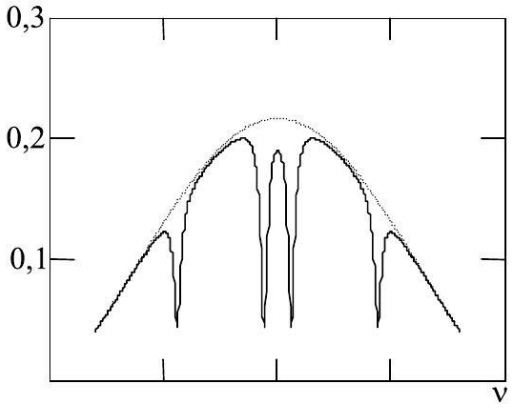



Рис. 5.12. Расчётный спектр коэффициента усиления гелий-неонового лазера при двух продольных модах

Рассмотрим световую волну с частотой, не совпадающей с центром спектра усиления лазерной среды, например, её частота больше центральной частоты. Такая волна при перемещении от одного зеркала к другому взаимодействует только с атомами, летящими ей навстречу с определённой скоростью, сбрасывая их на нижний лазерный уровень и уменьшая усиление на данной частоте. После одного отражения от зеркала эта волна распространяется в противоположном направлении и взаимодействует уже с другими атомами, летящими с той же скоростью также в противоположном направлении, поскольку за время обхода резонатора атомы не успевают изменить направление движения. В результате в спектре усиления «выедается» два провала, расположенные симметрично относительно максимума спектра усиления (провалы Беннета). Ширина этих провалов определяется шириной спектра генерирующей моды. При этом усиление в других участках спектра, где нет провалов, оказывается больше, что создаёт условия для генерации других мод. Поэтому если не приняты специальные меры, например, не выбран очень короткий резонатор с большим межмодовым расстоянием, в лазере обычно генерируется одновременно несколько продольных мод. Если в лазере генерируется одна мода, расположенная в центре спектра усиления, то два провала сливаются в один. В центре контура усиления образуется так называемый лэмбовский провал, что приводит к уменьшению выходной мощности лазера при точной настройке частоты генерирующей моды на максимум спектра усиления. На рис. 5.11 представлен расчётный спектр коэффициента усиления типичного гелий-неонового лазера при наличии только одной генерирующей продольной моды в центре спектра усиления, а на рис. 5.12 - для двух мод.

Лэмбовский провал используется практически для точной стабилизации
частоты излучения лазеров путём минимизации выходной мощности за счёт точной подстройки длины резонатора. Если в лазере генерируются две продольные моды, частоты которых расположены симметрично относительно максимума спектра усиления, то они попеременно взаимодействуют с одними и теми же атомами и «конкурируют за усиление» друг с другом. В результате возникают неустойчивости этих мод, они «переползают» к частотам с большим усилением и их интенсивности начинают сильно флуктуировать.

## 8.2.6 3. Common Types of Lasers*

**Инжекционный полупроводниковый лазер.**

Это самый распространённый тип лазеров, во всем мире их выпускается порядка $10^9$ штук в год. Такие лазеры используются во всех CD, DVD и Blue-Ray приводах, а также в

спектроскопии, лазерных дальномерах, нивелирах, лазерных указках, для накачки других лазеров и т.д.

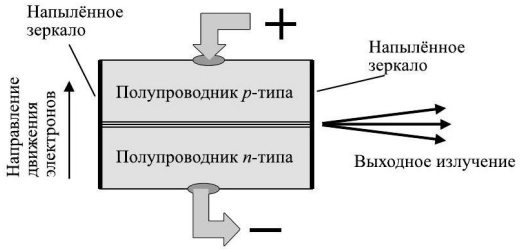



Рис. 5.13. Устройство инжекционного полупроводникового лазера

Схематически устройство инжекционного полупроводникового лазера изображено на рис. 5.13.

В полупроводнике $n$-типа есть некоторое количество электронов в зоне проводимости, которые переносят заряд при протекании электрического тока. Перейти вниз, в валентную зону они не могут, потому что в валентной зоне все состояния заняты электронами. В полупроводнике $p$-типа есть свободные состояния на самом верху валентной зоны, а находящаяся выше зона проводимости практически пуста. В области $p-n$-перехода, т. е. на границе раздела полупроводников, электроны из области $n$-типа впрыскиваются (инжектируются) в область $p$-типа и «сваливаются» из зоны проводимости на свободные уровни валентной зоны с излучением фотонов как за счёт спонтанного, так и за счёт вынужденного излучений. Усиливающая область представляет собой слой микронной толщины. Длина кристалла обычно составляет порядка миллиметра, таким образом, инжекционные полупроводниковые лазеры являются самыми миниатюрными из всех типов лазеров. Устройство современных полупроводниковых лазеров несколько сложнее, в частности,
$p-n$-переход представляет собой многослойную структуру (гетероструктуру), усиливающая область делается предельно тонкой (заметно меньше длины волны), а чтобы излучение не уходило в стороны за счёт дифракции, формируется канал с повышенным показателем преломления, таким образом, образуется волновод, в котором каналируется излучение. Поскольку размер выходного пятна порядка длины волны, дифракционная расходимость излучения очень велика, обычно десятки градусов, поэтому сразу на выходе излучения из лазера всегда ставится короткофокусная линза или объектив для коллимирования пучка.

К настоящему времени изготавливаются полупроводниковые лазеры, излучающие в области от 0,63 мкм до $\sim$ 10 мкм и в области $0,40-0,52$ мкм. Разработаны также лабораторные образцы ультрафиолетовых полупроводниковых лазеров с длиной волны излучения до 0,34 мкм. Красные, фиолетовые (404-404 нм) и синие)450 нм) лазерные указки устроены чрезвычайно просто: они представляют собой полупроводниковый лазер с источником тока на батарейках и короткофокусную линзу, собирающую лазерное излучение в параллельный пучок. На желтую и оранжевую область инжекционные полупроводниковые лазеры пока не созданы, а лазерные указки зеленого цвета с длиной волны 532 нм чаще всего представляют собой неодимовые лазеры с удвоением частоты.

Твердотельные лазеры на ионах неодима $Nd^{3+}$. Неодим в виде соли вносится в какую-либо матрицу: либо в кристалл в процессе его роста, либо в расплавленное стекло в относительной концентрации порядка $10^{-2}$. Схема энергетических уровней ионов неодима изображена на рис. 5.14 (на рис. 5.14 использованы применяемые в спектроскопии обозначения уровней, смысл которых здесь не существенен).

Накачка среды обычно осуществляется импульсными лампами. Под действием накачки ионы $Nd^{3+}$ переходят из основного состояния $^4I_{9/2}$ на несколько сравнительно

узких уровней, образованных возбуждёнными состояниями. Из этих уровней энергия возбуждения быстро передаётся в результате безызлучательных процессов на метастабильный уровень $^4\,\mathrm{F}_{3/2}$, время жизни которого составляет 0,2 мс в кристалле иттрий-алюминиевого граната ($\mathrm{Y_2Al_5O_{12}}$ или ИАГ) и $0,1-1$ мс в стеклах. Этот уровень является верхним уровнем рабочего лазерного перехода. Наибольшей вероятностью обладает лазерный переход $^4\,\mathrm{F}_{3/2} \rightarrow {}^4\mathrm{I}_{11/2} (\lambda = 1,05 - 1,08$ мкм в зависимости от матрицы). Такое излучение находится в ближней инфракрасной области и невидимо человеческим глазом. Время жизни на уровне $^4\mathrm{I}_{11/2}$, то есть на нижнем лазерном уровне данного перехода, составляет всего $10^{-9}$ с, поэтому данный уровень всегда практически пуст. Это облегчает достижение инверсной населённости и позволяет считать, что разность заселённостей равна $\Delta N = N_2$. Таким образом, генерация лазерного излучения в неодимовом лазере описывается четырёхуровневой схемой (основной уровень, уровень, на который производится накачка, верхний рабочий уровень и нижний рабочий уровень). Созданы твердотельные лазеры и на других элементах: Yb, Cr, Er, Ho, Ce, Ti и т. д. Но самыми распространёнными являются лазеры на ионах неодима в кристалле иттрий-алюминиевого граната.

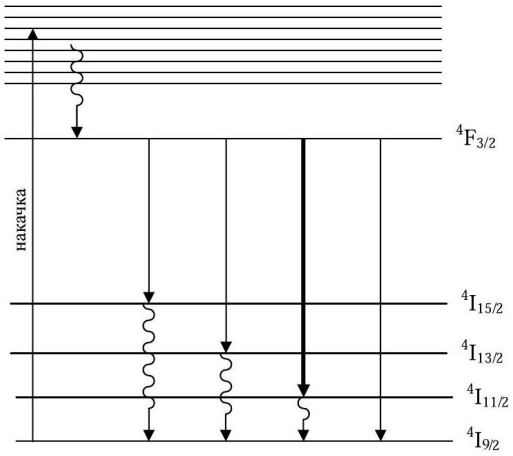



Рис. 5.14. Схема энергетических уровней иона неодима. Волнистыми линиями показаны безызлучательные переходы. Выделен наиболее интенсивный лазерный переход

В последнее время становятся весьма популярными неодимовые и иттербиевые лазеры, накачиваемые не лампами, а полупроводниковыми инжекционными лазерами с длиной волны излучения, попадающей в полосу поглощения активной среды)808 нм для Nd и $\sim$ 980 нм для Yb). Полупроводниковая накачка позволяет более чем на порядок уменьшить габариты лазера по сравнению с ламповой накачкой. Появились даже указки зелёного и голубого света размером с авторучку и менее. Типичное устройство зелёного лазера показано на рис. 5.15.

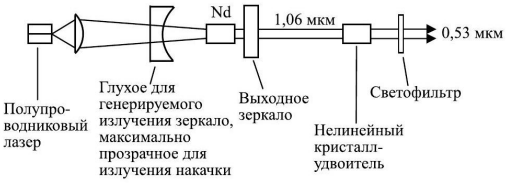



Рис. 5.15. Устройство зелёного лазера на Nd с полупроводниковой накачкой

Излучение такого миниатюрного неодимового лазера удваивается по частоте при прохождении через нелинейный кристалл. В таком кристалле отклик среды, т. е. наведённая полем поляризация $P$, зависит нелинейно от напряжённости поля световой

волны, в результате отклик становится несинусоидальным во времени, а это означает появление дополнительных гармоник в спектре распространяющейся волны. Набор гармоник зависит от симметрии кристаллической решетки, в частности, в кристаллах без центра инверсии появляется вторая гармоника ( воение частоти). Путём подбора направления распространения света в некоторых кристаллах удаётся обеспечить сложение рождающихся в разных местах кристалла волн второй гармоники синфазно и достичь коэффициента преобразования энергии во вторую гармонику на уровне десятков процентов. Оставшаяся непреобразованная часть инфракрасного излучения поглощается цветным светофильтром. Голубые указки с $\lambda = 473$ нм устроены аналогично, только для удвоения используется другой лазерный переход иона неодима с длиной волны $\sim 0,9$ мкм. Часто кристаллудвоитель размещают внутри резонатора неодимового лазера, что позволяет увеличить эффективность удвоения частоты. В последнее время появились зеленые лазеры-указки без неодимового лазера, в них излучение инфракрасного полупроводникового лазера сразу подается на кристалл-удвоитель.

**Литература**

1. Звелто О. Принципы лазеров. - М.: Лань, 2008.

2. Квантовая электроника. Маленькая энциклопедия. - М.: Сов. Энциклопедия, 1969.

# 9 Optics by Born Wolf

## 9.1 Historical Introduction

The physical principles underlying the optical phenomena with which we are concerned in this treatise were substantially formulated before 1900 . Since that year, optics, like the rest of physics, has undergone a thorough revolution by the discovery of the quantum of energy. While this discovery has profoundly affected our views about the nature of light, it has not made the earlier theories and techniques superfluous; rather, it has brought out their limitations and defined their range of validity. The extension of the older principles and methods and their applications to very many diverse situations has continued, and is continuing with undiminished intensity.

In attempting to present in an orderly way the knowledge acquired over a period of several centuries in such a vast field it is impossible to follow the historical development, with its numerous false starts and detours. It is therefore deemed necessary to record separately, in this preliminary section, the main landmarks in the evolution of ideas concerning the nature of light.*

The philosophers of antiquity speculated about the nature of light, being familiar with burning glasses, with the rectilinear propagation of light, and with refraction and reflection. The first systematic writings on optics of which we have any definite knowledge are due to the Greek philosophers and mathematicians [Empedocles (c. 490-430 b.c.), Euclid (c. 300 b.c.)].

Amongst the founders of the new philosophy, Ren Descartes (1596-1650) may be singled out for mention as having formulated views on the nature of light on the basis of his metaphysical ideas. † Descartes considered light to be essentially a pressure transmitted through a perfectly elastic medium (the aether) which fills all space, and he attributed the diversity of colours to rotary motions with different velocities of the particles in this medium. But it was only after Galileo Galilei (1564-1642) had, by his development of mechanics, demonstrated the power of the experimental method that optics was put on a firm foundation. The law of reflection was known to the Greeks; the law of refraction was discovered experimentally in 1621 by Willebrord

Snell ‡ (Snellius, c. 1580-1626). In 1657 Pierre de Fermat (1601-1665) enunciated the celebrated Principle of Least Time§ in the form "Nature

always acts by the shortest course". According to this principle, light always follows that path which brings it to its destination in the shortest time, and from this, in turn, and from the assumption of varying "resistance" in different media, the law of refraction follows. This principle is of great philosophical significance, and because it seems to imply a teleological manner of explanation, foreign to natural science, it has raised a great deal of controversy.

The first phenomenon of interference, the colours exhibited by thin films now known as "Newton's rings", was discovered independently by Robert Boyle* (1627-1691) and Robert Hooke † (1635-1703). Hooke also observed the presence of light in the geometrical shadow, the "diffraction" of light but this phenomenon had been noted previously by Francesco Maria Grimaldi ‡ (1618-1663). Hooke was the first to advocate the view that light consists of rapid vibrations propagated instantaneously, or with a very great speed, over any distance, and believed that in an homogeneous medium every vibration will generate a sphere which will grow steadily.§ By means of these ideas Hooke attempted an explanation of the phenomena of refraction, and an interpretation of colours. But the basic quality of colour was revealed only when Isaac Newton (1642-1727) discovered|| in 1666 that white light could be split up into component colours by means of a prism, and found that each pure colour is characterized by a specific refrangibility. The difficulties which the wave theory encountered in connection with the rectilinear propagation of light and of polarization (discovered by Huygens Φ ) seemed to Newton so decisive that he devoted himself to the development of an emission (or corpuscular) theory, according to which light is propagated from a luminous body in the form of minute particles.

At the time of the publication of Newton's theory of colour it was not known whether light was propagated instantaneously or not. The discovery of the finite speed of light was made in 1675 by Olaf Römer (1644-1710) from the observations of the eclipses of Jupiter's satellites.**

The wave theory of light which, as we saw, had Hooke amongst its first champions was greatly improved and extended by Christian Huygens **Φ** ( **1629 − 1695** ). He enunciated the principle, subsequently named after him, according to which every point of the "aether" upon which the luminous disturbance falls may be regarded as the centre of a new disturbance propagated in the form of spherical waves; these secondary waves combine in such a manner that their envelope determines the wavefront at any later time. With the aid of this principle he succeeded in deriving the laws of reflection and refraction. He was also able to interpret

0

- For a more extensive account of the history of optics, reference may be made to: J. Priestley, History and Present State of Discoveries relating to Vision, Light and Colours (2 Vols., London, 1772); Thomas Young, A Course of Lectures on Natural Philosophy and the Mechanical Arts (London, 1845, Vol. 1, pp. 374-385); E. Wilde, Geschichte der Optik vom Ursprung dieser Wissenschaft bis auf die gegenwrtige Zeit (2 Vols., Berlin, 1838, 1843); Ernst Mach, The Principles of Physical Optics, A historical and philosophical treatment (First German edition 1913. English translation 1926, reprinted by Dover Publications, New York, 1953); E. Hoppe, Geschichte der Optik (Leipzig, Weber, 1926); V. Ronchi, Storia della Luce (Bologna: Zanichelli, 2nd. Ed., 1952). A comprehensive historical account up to recent times is E. T. Whittaker's A History of the Theories of Aether and Electricity, Vol. I (The Classical Theories), revised and enlarged edition 1952; Vol. II (The Modern Theories 1900-1926), 1953, published by T. Nelson and Sons, London and Edinburgh. The first volume was used as the chief source for this introductory section.
  † R. Descartes, Dioptrique, Mtores [published (anonymously) in Leyden in 1637 with prefactory essay "Discours de la mthode"]. Principia Philosophiae (Amsterdam, 1644).
  ‡ Snell died in 1626 without making his discoveries public. The law was first published by Descartes in his Dioptrique without an acknowledgement to Snell, though it is generally believed that Descartes had seen Snell's manuscript on this subject.
  § In a letter to Cureau de la Chambre. It is published in Oeuvres de Fermat (Paris, 1891, 2, 354).

the double refraction of calc-spar (discovered in 1669 by Erasmus Bartholinus (1625-1698)) by assuming that in the crystal there is, in addition to a primary spherical wave, a secondary ellipsoidal wave. It was in the course of this investigation that Huygens made the fundamental discovery of polarization: each of the two rays arising from refraction by calc-spar may be extinguished by passing it through a second crystal of the same material if the latter crystal be rotated about the direction of the ray. It was, however, left to Newton to interpret these phenomena; he assumed that rays have "sides";

and indeed this "transversality" seemed to him an insuperable objection to the acceptance of the wave theory, since at that time scientists were familiar only with longitudinal waves (from the propagation of sound).

The rejection of the wave theory on the authority of Newton led to its abeyance for nearly a century. But it still found an occasional supporter, such as the great mathematician Leonhard Euler (1707-1783).*

It was not until the beginning of the 19 th century that the decisive discoveries were made which led to general acceptance of the wave theory. The first step towards this was the enunciation in 1801 by Thomas Young (1773-1829) of the principle of interference and the explanation of the colours of thin films. † But as Young's views were expressed largely in a qualitative manner they did not gain general recognition.

About this time polarization of light by reflection was discovered by tienne Louis Malus ‡ (1775-1812). Apparently, one evening, in 1808, he observed the reflection of the sun from a window pane through a calc-spar crystal, and found that the two images obtained by double refraction varied in relative intensities as the crystal was rotated about the line of sight. But Malus did not attempt an interpretation of this phenomenon, being of the opinion that current theories were incapable of providing an explanation.

In the meantime the emission theory had been developed further by Pierre Simon de Laplace (1749-1827) and Jean-Baptiste Biot (1774-1862). Its supporters proposed the subject of diffraction for the prize question set by the Paris Academy for 1818, in the expectation that a treatment of this subject would lead to the crowning triumph of the emission theory. But their hopes were disappointed, for, in spite of strong opposition, the prize was awarded to Augustin Jean Fresnel (1788-1827), whose treatment§ was based on the wave theory, and was the first of a succession of investigations which, in the course of a few years, were to discredit the corpuscular theory completely. The substance of his memoir consisted of a synthesis of Huyaens's Envelope Construction with Young's Principle of Interference. This, as Fresnel showed, was sufficient to explain not only the "rectilinear propagation" of light but also the minute deviations from it-diffraction phenomena. Fresnel calculated the diffraction caused by straight edges, small apertures, and screens; particularly impressive was the experimental confirmation by Arago of a prediction, deduced by Poisson from Fresnel's theory, that in the centre of the shadow of a small circular disc there should appear a bright spot.

In the same year (1818) Fresnel also investigated the important problem of the influence of the earth's motion on the propagation of light, the question being whether there was any

0

- The Philosophical Works of Robert Boyle (abridged by P. Shaw), Second Ed. (London, 1738), Vol. II, p. 70.

† R. Hooke, Micrographia (1665), 47.

‡ F. M. Grimaldi, Physico-Mathesis de lumine, coloribus, et iride (Bologna, 1665).

§ The early wave theories of Hooke and Huygens operate with single "pulses" rather than with wave trains of definite wavelengths.

|| I. Newton, Phil. Trans. No. 80 (Feb. 1672), 3075.

If Chr. Huygens, Trait de la lumire (completed in 1678, published in Leyden in 1690).

** Olaf Römer, Mm. de l'Acad. Sci. Paris, 10 (1666-1699), 575; J. de Sav. (1676), 223.

difference between the light from stellar and terrestrial sources. Dominique Franois Arago (1786-1853) found from experiment that (apart from aberration) there was no difference. On the basis of these findings Fresnel developed his theory of the partial convection of the luminiferous aether by matter, a theory confirmed in 1851, by direct experiment carried out by Armand Hypolite Louis Fizenu (1819-1896). Together with Arago, Fresnel investigated the interference of polarized rays of light and found (in 1816) that two rays polarized at right angles to each other never interfere. This fact could not be reconciled with the assumption of

longitudinal waves, which had hitherto been taken for granted. Young, who had heard of this discovery from Arago, found in 1817 the key to the solution when he assumed that the vibrations were transverse.

Fresnel at once grasped the full significance of this hypothesis, which he sought to put on a more secure dynamical basis* and from which he drew numerous conclusions. For, since only longitudinal oscillations in a fluid are possible, the aether must behave like a solid body; but at that time a theory of elastic waves in solids had not yet been formulated. Instead of developing such a theory and deducing the optical consequences from it, Fressel proceeded by inference, and sought to deduce the properties of the luminiferous aether from the observations. The peculiar laws of light propagation in crystals were Fresnel's starting point; the elucidation of these laws and their reduction to a few simple assumptions about the nature of elementary waves represents one of the greatest achievements of natural science. In 1832, William Rowan Hamilton † (1805-1865), who himself made important contributions to the development of optics, drew attention to an important deduction from Fresnel's construction, by predicting the so-called conical refraction, whose existence was confirmed experimentally shortly afterwards by Humphrey Lloyd ‡ (1800-1881).

It was also Fresnel who (in 1821) gave the first indication of the cause of dispersion, by taking into account the molecular structure of matter§, a suggestion elaborated later by Cauchy. Dynamical models of the mechanism of aether vibrations led Fresnel to deduce the laws which now bear his name, governing the intensity and polarization of light rays produced by reflection and refraction.||

Fresnel's work had put the wave theory on such a secure foundation that it seemed almost superfluous when in 1850 Foucault and Fizeau and Breguet** undertook a crucial experiment first suggested by Arago. The corpuscular theory explains refraction in terms of the attraction of the light-corpuscles at the boundary towards the optically denser medium, and this implies a greater velocity in the denser medium; on the other hand the wave theory demands, according to Huygens's construction, that a smaller velocity obtains in the optically denser medium. The direct measurement of the velocity of light in air and water decided unambiguously in favour of the wave theory.

The decades that followed witnessed the development of the elastic aether theory. The first step was the formulation of a theory of the elasticity of solid bodies. Claude Louis Marie Henri Navier (1785-1836) developed such a theory ††, discerning that matter consists of countless particles (mass points, atoms) exerting on each other forces along the lines joining them. The now customary derivation of the equations of elasticity by means of the continuum concept is due to Augustine Louis Cauchy ‡‡ (1789-1857). Of other scientists who participated in the development of optical theory mention must be made of Simon Denis Poisson§§ (1781-1840),

0

- L. Euleri Opuscula varii argumenti, Berlin (1746), 169.
  † Th. Young, Phil. Trans. Roy. Soc., London xcii (1802) 12, 387. Young's Works, Vol. 1, p. 202.
  ‡ . L. Malus, Nouveau Bull. d. Sci., par la Soc. Philomatique, Vol. 1 (1809), 266. Mm. de la Soc. d'Arcueil, Vol. 2 (1809).
  § A. Fresnel, Ann. Chim. et Phys., (2), 1 (1816) 239; Oeuvres, Vol. 1, 89, 129.

Georqe

Green* (1793-1841), James MacCullagh † (1809-1847) and Franz Neumann ‡ (1798-1895). To-day it is no longer relevant to enter into the details of these theories or into the difficulties which they encountered; for the difficulties were all caused by the requirement that optical processes should be explicable in mechanical terms, a condition which has long since been abandoned. The following indication will suffice. Consider two contiguous elastic media, and assume that in the first a transverse wave is propagated towards their common boundary. In the second medium the wave will be resolved, in accordance with the laws of mechanics, into longitudinal and transverse waves. But, according to Arago's and Fresnel's experiments, elastic longitudinal waves must be ruled out and must therefore be eliminated somehow. This, however, is not possible without violating the laws of mechanics expressed by the boundary conditions for strains and stresses. The various theories put forward by the authors mentioned above differ in regard to the assumed boundary conditions, which always conflicted in some way with the laws of mechanics.

An obvious objection to regarding the aether as an elastic solid is expressed in the following query: How is one to imagine planets travelling through such a medium at enormous speeds without any appreciable resistance? George Gabriel Stokes (18191903) thought that this objection could be met on the grounds that the planetary speeds are very small compared to the speeds of the aetherial particles in the vibrations constituting light; for it is known that bodies like pitch or sealing wax are capable of rapid vibrations but yield completely to stresses applied over a long period. Such controversies seem superfluous today since we no longer consider it necessary to have mechanical pictures of all natural phenomena.

A first step away from the concept of an elastic aether was taken by MacCullagh, § who postulated a medium with properties not possessed by ordinary bodies. The latter store up energy when the volume elements change shape, but not during rotation. In MacCullagh's aether the inverse conditions prevail. The laws of propagation of waves in such a medium show a close similarity to Maxwell's equations of electromagnetic waves which are the basis of modern optics.

In spite of the many difficulties, the theory of an elastic aether persisted for a long time and most of the great physicists of the 19th century contributed to it. In addition to those already named, mention must be made of William Thomson || (Lord Kelvin, 1824-1908), Carl Neumann (1832-1925), John William Strutt** (Lord Rayleigh, 1842-1919), and Gustay Kirchhoff †† (1832-1887). During this period many optical problems were solved, but the foundations of optics remained in an unsatisfactory state.

In the meantime researches in electricity and magnetism had developed almost independently of optics, culminating in the discoveries of Michael Faraday ‡‡ (1791-1867). James Clerk Maxwell§§ (1831-1879) succeeded in summing up all previous experiences in this field in a system of equations, the most important

0

- A. Fresnel, Oeuvres, 2, 261, 479.
  † W. R. Hamilton, Trans. Roy. Irish Acad., 17 (1833), 1. Also Hamilton's Mathematical Papers, edited by J. L. Synge and W. Conway (Cambridge University Press, 1931), Vol. 1, p. 285.
  ‡ H. Lloyd, Trans. Roy. Irish Acad., 17 (1833), 145.
  § A. Fresnel, Oeuvres, 2, 438.
  || A. Fresnel, Mm. de l'Acad., 11 (1832), 393; Oeuvres, 1, 767.
  IT L. Foucault, Compt. Rend. Acad. Sci. Paris, 30 (1850), 551.
  ** H. Fizeau and L. Breguet, Compt. Rend. Acad. Sci. Paris, 30 (1850), 562, 771.
  †† C. L. M. H. Navier, Mm. de l'Acad., 7, 375 (submitted in 1821, published in 1827).
  ‡‡ A. L. Cauchy, Exercise de Mathmatiques, 3 (1828), 160.
  § § S. D. Poisson, Mm. de l'Acad., Vol. 8 (1828), 623.

0

consequence of which was to establish the possibility of electromagnetic waves, propagated with a velocity which could be calculated from the results of purely electrical measurements. When Rudolph Kohlrausch (1809-1858) and Wilhelm Weber* (1804-1891) carried out these measurements, the velocity turned out to be that of light. This led Maxwell to conjecture that light waves are electromagnetic waves; a conjecture verified by direct experiment in 1888 by Heinrich Hertz † (1857-1894). In spite of this, Maxwell's electromagnetic theory had a long struggle to gain general acceptance. It seems to be a characteristic of the human mind that familiar concepts are abandoned only with the greatest reluctance, especially when a concrete picture of phenomena has to be sacrificed. Maxwell himself, and his followers, tried for a long time to describe the electromagnetic field with the aid of mechanical models. It was only gradually, as Maxwell's concepts became more familiar, that the search for an "explanation" of his equations in terms of mechanical models was abandoned; today there is no conceptual difficulty in regarding Maxwell's field as something which cannot be reduced to anything simpler.

But even the electromagnetic theory of light has attained the limits of its serviceability. It is capable of explaining, in their main features, all phenomena connected with the propagation of light. However, it fails to elucidate the processes of emission and absorption, in which the finer features of the interaction between matter and the optical field are manifested.

The laws underlying these processes are the proper object of modern optics, indeed of modern physics. Their story begins with the discovery of certain regularities in spectra. The first step was Josef Fraunhofer's (1787-1826) discovery ‡ (1814-1817) of the dark lines in the solar spectrum, since named after him; and their interpretation as absorption lines given in 1861 on the basis of experiments by Robert Wilhelm Bunsen (1811-1899) and Gustay Kirchhoff§ (1824-1887). The light of the continuous spectrum of the body of the sun, passing the cooler gases of the sun's atmosphere, loses by absorption just those wavelengths which are emitted by the gases. This discovery was the beginning of spectrum analysis, which is based on the recognition that every gaseous chemical element possesses a characteristic line spectrum. The investigation of these spectra has been a major object of physical research up to and including the present; and since light is its subject and optical methods are employed, it is often considered as a part of optics. The problem of how light is produced or destroyed in atoms is, however, not exclusively of an optical nature, as it involves equally the mechanics of the atom itself; and the laws of spectral lines reveal not so much the nature of light as the structure of the emitting particles. Thus, from being a part of optics, spectroscopy has gradually evolved into a separate discipline which provides the empirical foundations for atomic and molecular physics. This field is, however, beyond the scope of this book.

Concerning methods, it has become apparent that classical mechanics is inadequate for a proper description of events occurring within the atom and must be replaced by

---

• G. Green, Trans. Camb. Phil. Soc. (1838); Math. Papers, 245.
† J. MacCullagh, Phil. Mag. (3), 10 (1837), 42, 382; Proc. Roy. Irish Acad., 18 (1837).
‡ F. Neumann, Abh. Berl. Akad., Math. Kl. (1835), 1.
§ J. MacCullagh, Trans. Roy. Irish Acad., 21, Coll. Works, Dublin (1880), 145.
|| W. Thomson, Phil. Mag., (5), 26 (1888), 414. Baltimore Lectures (London, 1904).
II C. Neumann, Math. Ann., 1 (1869), 325, 2 (1870), 182.
** J. W. Strutt, (Lord Rayleigh), Phil. Mag., (4) 41 (1871), 519; 42 (1871), 81.
†† G. Kirchnoff, Berl. Abh. Physik., Abteilg. 2 (1876), 57; Ges. Abh. 352; Berl. Ber. (1882), 641; Pogg. Ann. Physik. u. Chem. (2), 18 (1883), 663; Ges. Abh., Nachtrag. 22.
‡‡ M. Faraday, Experimental Researches in Electricity (London, 1839).
§§ J. C. Maxwell, A Treatise on Electricity and Magnetism, 2 Vols. (Oxford, 1873).

0

• R. Kohlrausch and W. Weber, Pogg. Ann. Physik u. Chem. (2), 99 (1856), 10.
† H. Hertz, Sitzb. Berl. Akad. Wiss., Feb. 2, 1888; Wiedem. Ann. 34 (1888), 551; English translation in his Electric Waves (London, Macmillan, 1893), 107.

the quantum theory, originated in 1900 by Max Planck* (1858-1947). Its application to the atomic structure, led in 1913, to the explanation by Niels Bohr † (born in 1885) of the simple laws of line spectra of gases. From these beginnings and from the ever increasing experimental material, modern quantum mechanics developed (Heisenberg, Born, Jordan, de Broglie, Schrödinger, Dirac). ‡ By its means considerable insight has been obtained into the structure of atoms and molecules.

However, our concept of the nature of light has also been greatly influenced by quantum theory. Even in its first form due to Planck there appears a proposition which is directly opposed to classical ideas, namely that an oscillating electric system does not impart its energy to the electromagnetic field in a continuous manner but in finite amounts, or "quanta" $\varepsilon = hv$, proportional to the frequency $v$ of the light, where $h = 6.55 \times 10^{-27}$erg/sec is Planck's constant. We may say that the occurrence of the constant $h$ is the feature which distinguishes modern physics from the old.

It was only gradually that the paradoxical, almost irrational, character of Planck's equation $\varepsilon = h\nu$ was fully realized by physicists. This was brought about mainly by the work of Einstein and Bohr. On the basis of Planck's theory, Einstein (18791955) revived in 1905 the corpuscular theory of light in a new form § by assuming that Planck's energy quanta exist as real light-particles, called "light quanta" or "photons". He thereby succeeded in explaining some phenomena which had been discovered more recently in connection with the transformation of light into corpuscular energy, phenomena which were inexplicable by the wave theory. Chief among these are the so-called photo-electric effect and the fundamentals of photochemistry. In phenomena of this type light does not impart to a detached particle an energy proportional to its intensity, as demanded by the wave theory, but behaves rather like a hail of small shots. The energy imparted to the secondary particles is independent of the intensity, and depends only on the frequency of the light (according to the law $\varepsilon = hv$ ). The number of observations confirming this property of light increased year by year and the situation arose that the simultaneous validity of both wave and corpuscular theories had to be recognized, the former being exemplified by the phenomena of interference, the latter by the photo-electric effect. It is only in more recent years that the development of quantum mechanics has led to a partial elucidation of this paradoxical state of affairs, but this has entailed giving up a fundamental principle of the older physics, namely the principle of deterministic causality.

The detailed theory of the interaction between field and matter required the extension of the methods of quantum mechanics (field quantization). For the electromagnetic radiation field this was first carried out by Dirac [2] and these investigations form the basis of quantum optics. A discussion of these topics is, however, outside the scope of this book.

Another branch of optics, not touched upon in this work, is the optics of moving bodies. Like the quantum theory it has grown into a vast independent field of study.

---

‡ J. Fraunhofer, Gilberts Ann., 56 (1817), 264. W. H. Wollaston (1766-1828) observed these lines in 1802 (Phil. Trans. Roy. Soc., London, 1802, 365) but had not appreciated his discovery and interpreted them incorrectly.

§ R. Bunsen and G. KirchHoff, Untersuchungen ber das Sonnenspektrum und die Spektren der Chemischen Elemente., Abh. kgl. Akad. Wiss., Berlin, 1861, 1863.

0

- M. Planck, Verh. d. deutsch phys. Ges., 2, 1900, 202, 237. Ann. d. Physik, (4), 4 (1901), 553.

† N. Bohr, Phil. Mag., (6), 26 (1913), 1, 476, 857.

‡ W. Heisenberg, Z.f. Phys., 33 (1925), 879; M. Born and P. Jordan, ibid., 34 (1925), 858; M. Born, W. Heisenberg, and P. Jordan, ibid., 35 (1926), 557; L. de Broglie, Thse, Paris, 1924; Ann. de Physique (10), 3 (1925), 22; E. Schrödinger, Ann. d. Physik, (4), 79 (1926), 361, 489 and 734; 80 (1926), 437; 81 (1926), 109. English translation: "Collected Papers on Wave Mechanics" by E. Schrödinger (London and Glasgow, Blackie, 1928); P. A. M. Dirac, Proc. Roy. Soc. A, 109 (1925), 642 ; ibid., 110 (1926), 561.

§ A. Einstein, Ann. d. Physik, (4), 17 (1905), 132; 20 (1906), 199.

The first observed phenomenon in this field, recorded in 1728 by James Bradley* (1692-1762), was the aberration of "fixed stars", i.e. the observation of slightly different angular positions of the stars according to the motion of the earth relative to the direction of the light ray. Bradley correctly interpreted this phenomenon as being due to the finite velocity of light and thus was able to determine the velocity. We have already mentioned other phenomena belonging to optics of moving media: Fresnel was the first to enquire into the convection of light by moving bodies and to show that it behaved as if the luminiferous aether participated in the movement only with a fraction of the speed of the moving bodies; Fizead then demonstrated this partial convection experimentally with the aid of flowing water. The effect of the motion of the light source or of the observer was investigated by Christian Doppler † (1803-1853) who formulated the well-known principle named after him. So long as the elastic theory of light held the field and the precision of measurements was sufficiently limited, Fresnel's ideas on partial convection sufficed for a satisfactory explanation of all the phenomena. But the electromagnetic theory of light encountered difficulties of a fundamental nature. Hertz was the first to attempt to generalize Maxwell's laws to moving bodies. His formulae were, however, in conflict with some electromagnetic and optical experiments. Of great importance was the theory of Hendrik Antoon Lorentz (1853-1928) who assumed an "aether in a state of absolute rest" to be the carrier of the electromagnetic field and deduced the properties of material bodies from the interaction of elementary electric particles-the electrons. He was able to show that Fresnel's coefficient of convection could be obtained correctly from his theory and that in general all phenomena known at the time (1895) lent themselves to explanation by this hypothesis. ‡ But the enormous increase of precision in the determination of optical paths by means of the interferometer of Albert Abraham Michelson (1852-1931) led to a new anomaly: it proved impossible to demonstrate the existence of an "aether drift" required by the theory of the "stationary aether".§ The anomaly was resolved by Albert Einstein|| in 1905 in his special theory of relativity. The theory is founded on a critique of the concepts of time and space and leads to the abandonment of euclidian geometry and the intuitive conception of simultaneity. Its further development into the so-called general theory of relativity [4] led to a completely new conception of gravitational phenomena by a "geometrization" of the space-time manifold. The application of this theory involves the use of special mathematical and physical methods which, although relevant to optics in many cases, may easily be considered separately from it. The number of optical phenomena in which the motion of bodies (e.g. light sources) plays a significant part is rather small.

| T P. A. M. Dirac, Proc. Roy. Soc., A114 (1927), 243, 710.

0

- J. Bradley, Phil. Trans., 35 (1728), 637.
  † Chr. Doppler, Abh. Köngl. böhm. Geselsch, 2 (1842), 466; Pogg. Ann. 68 (1847), 1.
  ‡ H. A. Lorentz, Versuch einer Theorie der electrischen und optischen Erscheinungen in bewegten Körpern (Leiden, E. J. Brill 1895; reprinted by Teubner, Leipzig, 1906).
  § A. A. Michelson, Amer. Jour. Sci., (3), 22 (1881), 20; A. A. Michelson and E. W. Morley, Amer. Jour. Sci., (3), 34 (1887), 333; Phil. Mag., 24 (1887), 449.
  || A. Einstein, Ann. d. Physik (4), 17 (1905), 891.
  || A. Einstein, Berl. Sitz. (1915), 778, 799, 831, 844. Ann. d. Physik, (4) 49 (1916), 769

## 9.2 Basic Properties of the Electromagnetic Field

### 9.2.1 The Electromagnetic Field

#### 1.1.1 Maxwell's equations

The state of excitation which is established in space by the presence of electric charges is said to constitute an electromagnetic field. It is represented by two vectors, $\boldsymbol{E}$ and $\boldsymbol{B}$, called the electric vector and the magnetic induction respectively.*

To describe the effect of the field on material objects, it is necessary to introduce a second set of vectors, viz. the electric current density $\boldsymbol{j}$, the electric displacement $\boldsymbol{D}$, and the magnetic vector $\boldsymbol{H}$.

The space and time derivatives of the five vectors are related by Maxwell's equations, which hold at every point in whose neighbourhood the physical properties of the medium are continuous: †

$$\operatorname{curl} H - \frac{1}{c}\dot{D} = \frac{4\pi}{c} j \tag{1}$$

$$\operatorname{curl} E + \frac{1}{c}\dot{B} = 0 \tag{2}$$

the dot denoting differentiation with respect to time.
They are supplemented by two scalar relations:

$$\operatorname{div} D = 4\pi\rho, \tag{3}$$

$$\operatorname{div} B = 0. \tag{4}$$

Eq. (3) may be regarded as a defining equation for the electric charge density $\rho$ and (4) may be said to imply that no free magnetic poles exist.

From (1) it follows (since div curl $\equiv 0$ ) that

$$\operatorname{div} j = -\frac{1}{4\pi} \operatorname{div} \dot{D}$$

or, using (3),

$$\frac{\partial \rho}{\partial t} + \operatorname{div} j = 0. \tag{5}$$

0

- In elementary considerations $E$ and $H$ are, for historical reasons, usually regarded as the basic field vectors, and $\boldsymbol{D}$ and $\boldsymbol{B}$ as describing the influence of matter. In general theory, however, the present interpretation is compulsory for reasons connected with the electrodynamics of moving media.

The four Maxwell equations (1)-(4) can be divided into two sets of equations, one consisting of two homogeneous equations (right-hand side zero), containing $\boldsymbol{E}$ and $\boldsymbol{B}$, the other of two nonhomogeneous equations (right-hand side different from zero), containing $D$ and $H$. If a coordinate transformation of space and time (relativistic Lorentz transformation) is carried out, the equations of each group transform together, the equations remaining unaltered in form if $j/c$ and $\rho$ are transformed as a four-vector, and each of the pairs $E, B$ and $D, H$ as a six-vector (antisymmetric tensor of the second order). Since the nonhomogeneous set contains charges and currents (which represent influence of matter), one has to attribute the corresponding pair ( $\boldsymbol{D}, \boldsymbol{H}$ ) to the influence of matter. It is, however, customary to refer to $\boldsymbol{H}$ and not to $\boldsymbol{B}$ as the magnetic field vector; we shall conform to this terminology when there is no risk of confusion.

† The so-called Gaussian system of units is used here, i.e. the electrical quantities ( $E, D, j$ and $p$ ) are measured in electrostatic units, and the magnetic quantities ( $\boldsymbol{H}$ and $\boldsymbol{B}$ ) in electromagnetic units. The constant $c$ in (1) and (2) relates the unit of charge in the two systems; it is the velocity of light in the vacuum and is approximately equal to $3 \times 10^{10}$ cm/sec. (A more accurate value is given in § 1.2.)

By analogy with a similar relation encountered in hydrodynamics, (5) is called equation of continuity. It expresses the fact that the charge is conserved in the neighbourhood of any point. For if one integrates (5) over any region of space, one obtains, with the help of Gauss' theorem,

$$\frac{d}{dt}\int \rho dV + \int j \cdot n dS = 0 \tag{6}$$

the second integral being taken over the surface bounding the region and the first throughout the volume, $n$ denoting the unit outward normal. This equation implies that the total charge

$$e = \int \rho dV \tag{7}$$

contained within the domain can only increase on account of the flow of electric current

$$\mathscr{J} = \int j \cdot n dS \tag{8}$$

If all the field quantities are independent of time, and if, moreover, there are no currents ($j = 0$), the field is said to be static. If all the field quantities are time independent, but currents are present ( $\boldsymbol{j} \neq 0$ ), one speaks of a stationary field. In optical fields the field vectors are very rapidly varying functions of time; but the sources of the field are usually such that, when averages over any macroscopic time interval are considered rather than the instantaneous values, the properties of the field are found to be independent of the instant of time at which the average is taken. The word stationary is often used in a wider sense to describe a field of this type. An example is a field constituted by the steady flux of radiation (say from a distant star) through an optical system.

### 1.1.2 Material equations

The Maxwell equations (1)-(4) connect the five basic quantities $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{B}, \boldsymbol{D}$ and $\boldsymbol{j}$. To allow a unique determination of the field vectors from a given distribution of currents and charges, these equations must be supplemented by relations which describe the behaviour of substances under the influence of the field. These relations are known as material equations* (or constitutive relations). In general they are rather complicated, but if the bodies are at rest (or in very slow motion) relative to each other and if the material is isotropic (i.e. when its physical properties at each point are independent of direction) they take usually the relatively simple form †

$$\boldsymbol{j} = \boldsymbol{\sigma E} \tag{9}$$

$$\boldsymbol{D} = \varepsilon \boldsymbol{E} \tag{10}$$

$$\boldsymbol{B} = \mu \boldsymbol{H} \tag{11}$$

Here $\sigma$ is called the specific conductivity, $\varepsilon$ is known as the dielectric constant (or permittivity), and $\mu$ is called the magnetic permeability.

0

- There is an alternative way of describing the behaviour of matter. Instead of the quantities $\varepsilon = D/E, \mu = B/H$ one considers the differences $D - E$ and $B - H$; these have a simpler physical significance and will be discussed in Chapter II.
  † The more general relations, applicable also to moving bodies, are studied in the theory of relativity. We shall only need the following result from the more general theory: that in the case of moving charges there is also, in addition to the conduction current $\sigma E$, a convection current $\rho V$, where $V$ is the velocity of the moving charges and $\rho$ the charge density (cf. p. 8).

Eq. (9) is the differential form of Ofm's law. Substances for which $\sigma \neq 0$ (or more precisely is not negligibly small; the precise meaning of this cannot however be discussed here) are called conductors. Metals are very good conductors, but there are other classes of good conducting materials such as ionic solutions in liquids and also in solids. In metals the conductivity decreases with increasing temperature. However, in other classes of materials, known as semiconductors (e.g. germanium), conductivity increases with temperature over a wide range.

Substances for which $\sigma$ is negligibly small are called insulators or dielectrics. Their electric and magnetic properties are then completely determined by $\varepsilon$ and $\mu$. For most substances the magnetic permeability $\mu$ is practically unity. If this is not the case, i.e. if $\mu$ differs appreciably from unity, we say that the substance is magnetic. In particular, if $\mu > 1$, the substance is said to be paramagnetic (e.g. platinum, oxygen, nitrogen), while if $\mu < 1$ it is said to be diamagnetic (e.g. bismuth, copper, hydrogen, water).

If the fields are exceptionally strong, such as are obtained, for example, by focusing light that is generated by an optical maser, the right-hand sides of the material equations may have to be supplemented by terms involving components of the field vectors in powers higher than the first.*

In many cases the quantities $\sigma, \varepsilon$ and $\mu$ will be independent of the field strengths; in other cases, however, the behaviour of the material cannot be described in such a simple way. Thus, for example, in a gas of free ions the current, which is determined by the mean speed of the ions, depends, at any moment, not on the instantaneous value of $E$, but on all its previous values. Again, in so-called ferromagnetic substances (substances which are very highly magnetic, e.g. iron, cobalt and nickel) the value of the magnetic induction $B$ is determined by the past history of the field $H$ rather than by its instantaneous value. The substance is then said to exhibit hysteresis. A similar history-dependence will be found for the electric displacement in certain dielectric materials. Fortunately hysteretic effects are rarely significant for the highfrequency field encountered in optics.

In the main part of this book we shall study the propagation in substances which light can penetrate without appreciable weakening (e.g. air, glass). Such substances are said to be transparent and must be electrical nonconductors ( $\sigma = 0$ ), since conduction implies the evolution of Joule heat (cf. §1.1.4) and therefore loss of electromagnetic energy. Optical properties of conducting media will be discussed in Chapter XIII.

### 1.1.3 Boundary conditions at a surface of discontinuity

Maxwell's equations were only stated for regions of space throughout which the physical properties of the medium (characterized by $\varepsilon$ and $\mu$ ) are continuous. In optics one often deals with situations in which the properties change abruptly across one or more surfaces. The vectors $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{B}$ and $\boldsymbol{D}$ may then be expected also to become discontinuous, while $\rho$ and $j$ will degenerate into corresponding surface quantities. We shall derive relations describing the transition across such a discontinuity surface.

---

0

- Non-linear relationship between the displacement vector $\boldsymbol{D}$ and the electric field $\boldsymbol{E}$ was first demonstrated in this way by P. A. Franken, A. E. Hill, C. W. Peters and G. Weinrich, Phys. Rev. Let., 7 (1961), 118. The theory of propagation, of refraction and reflection of electromagnetic waves under conditions when such non-linear effects have to be taken into account has been developed by J. A. Armstrong, N. Bloembergen, J. Ducuing and P. S. Pershan, Phys. Rev., 127 (1962), 1918 and N. Bloembergen and P. S. Pershan, ibid, 128 (1962), 606.

For systematic treatments of non-linear effects see N. Bloembergen, Nonlinear Optics (New York, W. A. Benjamin, Inc., 1965) and P. S. Pershan, Progress in Optics, Vol. 5, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 83. See also P. A. Franken and J. F. Ward, Rev. Mod. Phys., 35 (1963), 23.

Let us replace the sharp discontinuity sur-

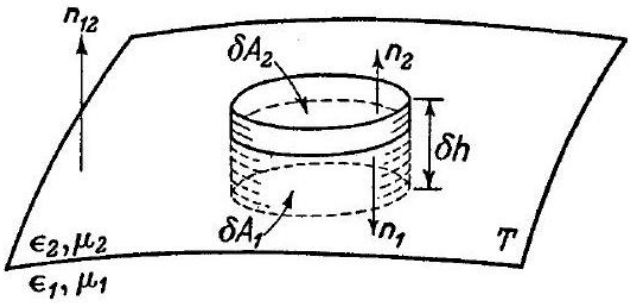



Fig. 1.1. Derivation of boundary conditions for the normal components of $\boldsymbol{B}$ and $\boldsymbol{D}$.

face $T$ by a thin transition layer within which $\varepsilon$ and $\mu$ vary rapidly but continuously from their values near $T$ on one side to their value near $T$ on the other. Within this layer we construct a small near-cylinder, bounded by a stockade of normals to $T$, roofed and floored by small areas $\delta A_1$ and $\delta A_2$ on each side of $T$, at constant distance from it, measured along their common normal (Fig. 1.1). Since $\boldsymbol{B}$ and its derivatives may be assumed to be continuous throughout this cylinder, we may apply Gauss' theorem to the integral of $\operatorname{div} B$ taken throughout the volume of the cylinder and obtain, from (4),

$$\int \operatorname{div} B dV = \int B \cdot n dS = 0 \tag{12}$$

the second integral is taken over the surface of the cylinder, and $n$ is the unit outward normal.

Since the areas $\delta A_1$ and $\delta A_2$ are assumed to be small, $\boldsymbol{B}$ may be considered to have constants values $\boldsymbol{B}^{(1)}$ and $\boldsymbol{B}^{(2)}$ on $\delta A_1$ and $\delta A_2$, and (12) may then be replaced by

$$B^{(1)} \cdot n_1 \delta A_1 + B^{(2)} \cdot n_2 \delta A_2 + \text{ contribution from walls } = 0. \tag{13}$$

If the height $\delta h$ of the cylinder decreases towards zero, the transition layer shrinks into the surface and the contribution from the walls of the cylinder tends to zero, provided that there is no surface flux of magnetic induction. Such flux never occurs, and consequently in the limit,

$$\left(B^{(1)} \cdot n_1 + B^{(2)} \cdot n_2\right) \delta A = 0 \tag{14}$$

$\delta A$ being the area in which the cylinder intersects $T$. If $n_{12}$ is the unit normal pointing from the first into the second medium, then $n_1 = -n_{12}, n_2 = n_{12}$ and (14) gives

$$n_{12} \cdot \left(B^{(2)} - B^{(1)}\right) = 0, \tag{15}$$

i.e. the normal component of the magnetic induction is continuous across the surface of discontinuity.

The electric displacement $D$ may be treated in a similar way, but there will be an additional term if charges are present. In place of (12) we now have from (3)

$$\int \operatorname{div} D dV = \int D \cdot n dS = 4\pi \int \rho dV \tag{16}$$

As the areas $\delta A_1$ and $\delta A_2$ shrink together, the total charge remains finite, so that the volume density becomes infinite. Instead of the volume charge density $\rho$ the concept of surface charge density $\hat{\rho}$ must then be used. It is defined by*

$$\lim_{\delta h \to 0} \int \rho dV = \int \hat{\rho} dA \tag{17}$$

We shall also need later the concept of surface current density $\hat{\boldsymbol{j}}$, defined in a similar way:

$$\lim_{\delta h \to 0} \int j dV = \int \hat{j} dA \tag{18}$$

If the area $\delta A$ and the height $\delta h$ are taken sufficiently small, (16) gives

$$D^{(1)} \cdot n_1 \delta A_1 + D^{(2)} \cdot n_2 \delta A_2 + \text{ contribution from walls } = 4\pi\hat{\rho}\delta A.$$

The contribution from the walls tends to zero with $\delta h$, and we therefore obtain in the limit as $\delta h \to 0$,

$$n_{12} \cdot \left(D^{(2)} - D^{(1)}\right) = 4\pi\hat{\rho}, \tag{19}$$

i.e. in the presence of a layer of surface charge density $\hat{\rho}$ on the surface, the normal component of the electric displacement changes abruptly across the surface, by an amount equal to $4\pi\hat{\rho}$.

Next, we examine the behaviour of the tangential components. Let us replace the sharp discontinuity surface by a continuous transition layer. We also replace the

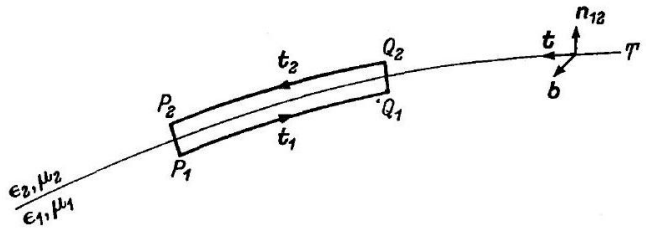



Fig. 1.2. Derivation of the boundary conditions for the tangentia components of $\boldsymbol{E}$ and $\boldsymbol{H}$.

cylinder of Fig. 1.1 by a "rectangular" area with sides parallel and perpendicular to $T$ (Fig. 1.2).

Let $\boldsymbol{b}$ be the unit vector perpendicular to the plane of the rectangle. Then it follows from (2) and from Stokes' theorem that

$$\int \operatorname{curl} \boldsymbol{E} \cdot \boldsymbol{b} dS = \int \boldsymbol{E} \cdot d\boldsymbol{r} = -\frac{1}{c} \int \dot{\boldsymbol{B}} \cdot \boldsymbol{b} dS \tag{20}$$

0

- For later purposes we note a representation of the surface charge density and the surface current density in terms of the Dirac delta function (see Appendix IV). If the equation of the surface of discontinuity is $F(x, y, z) = 0$, then

$$\rho = \hat{\rho} |\operatorname{grad} F| \delta(F), \tag{17a}$$

$$\boldsymbol{j} = \hat{\boldsymbol{j}} |\operatorname{grad} F| \delta(F). \tag{18a}$$

These relations can immediately be verified by substituting from (17a) and (18a) into (17) and (18) and using the relation $dF = |\operatorname{grad} F| dh$ and the sifting property of the $\delta$ function.

2

the first and third integrals being taken throughout the area of the rectangle, and the second along its boundary. If the lengths $P_1Q_1\,(=\delta s_1)$, and $P_2Q_2\,(=\delta s_2)$ are small, $E$ may be replaced by constant values $E^{(1)}$ and $E^{(2)}$ along each of these segments. Similarly $\dot{B}$ may be replaced by a constant value. (20) then gives

$$E^{(1)}\cdot t_1\delta s_1 + E^{(2)}\cdot t_2\delta s_2 + \text{ contribution from ends } = -\frac{1}{c}\dot{B}\cdot b\delta s\delta h \tag{21}$$

where $\delta s$ is the line element in which the rectangle intersects the surface. If now the height of the rectangle is gradually decreased, the contribution from the ends $P_1P_2$ and $Q_1Q_2$ will tend to zero, provided that $E$ does not in the limit acquire sufficiently sharp singularities; this possibility will be excluded. Assuming also $\dot{B}$ to remain finite, we obtain in the limit as $\delta h \to 0$,

$$\left(E^{(1)}\cdot t_1 + E^{(2)}\cdot t_2\right)\delta s = 0 \tag{22}$$

If $t$ is the unit tangent along the surface, then (see Fig. 1.2) $t_1 = -t = -b\wedge n_{12}$, $t_2 = t = b\wedge n_{12}$, and (22) gives

$$b\cdot\left[n_{12}\wedge\left(E^{(2)}-E^{(1)}\right)\right] = 0.$$

Since the orientation of the rectangle and consequently that of the unit vector $\boldsymbol{b}$ is arbitrary, it follows that

$$n_{12}\wedge\left(E^{(2)}-E^{(1)}\right) = 0 \tag{23}$$

i.e. the tangential component of the electric vector is continuous across the surface.
Finally consider the behaviour of the tangential component of the magnetic vector. The analysis is similar, but there is an additional term if currents are present. In place of (21) we now have

$$H^{(1)}\cdot t_1\delta s_1 + H^{(2)}\cdot t_2\delta s_2 + \text{ contribution from ends } = \frac{1}{c}\dot{\boldsymbol{D}}\cdot\boldsymbol{b}\delta s\delta h + \frac{4\pi}{c}\hat{\boldsymbol{j}}\cdot\boldsymbol{b}\delta s \tag{24}$$

On proceeding to the limit $\delta h \to 0$ as before, we obtain

$$n_{12}\wedge\left(H^{(2)}-H^{(1)}\right) = \frac{4\pi}{c}\hat{j} \tag{25}$$

From (25) it follows that in the presence of a surface current of density $\hat{\boldsymbol{j}}$, the tangential component (considered as a vector quantity) of the magnetic vector changes abruptly, its discontinuity being $\frac{4\pi}{c}\hat{j}\wedge n_{12}$.

Apart from discontinuities due to the abrupt changes in the physical properties of the medium, the field vectors may also be discontinuous because of the presence of a source which begins to radiate at a particular instant of time $t = t_0$. The disturbance then spreads into the surrounding space, and at any later instant $t_1 > t_0$ will have filled a well-defined region. Across the (moving) boundary of this region, the field vectors will change abruptly from finite values on the boundary to the value zero outside it.

The various cases of discontinuity may be covered by rewriting Maxwell's equations in an integral form.* The general discontinuity conditions may also be

written in the form of simple difference equations; a derivation of these equations is given in Appendix VI.

0

- See for example A. Sommerfeld, Electrodynamics (New York, Academic Press, 1952), p. 11; or J. A. Stratton, Electromagnetic Theory (New York, McGraw-Hill, 1941), p. 6.

### 1.1.4 The energy law of the electromagnetic field

Electromagnetic theory interprets the light intensity as the energy flux of the field. It is therefore necessary to recall the energy law of Maxwell's theory.

From (1) and (2) it follows that

$$E.\operatorname{curl} H - H.\operatorname{curl} E = \frac{4\pi}{c} j.E + \frac{1}{c} E.\dot{D} + \frac{1}{c} H.\dot{B}. \tag{26}$$

Also, by a well-known vector identity, the term on the left may be expressed as the divergence of the vector product of $\boldsymbol{H}$ and $\boldsymbol{E}$ :

$$E.\operatorname{curl} H - H.\operatorname{curl} E = -\operatorname{div}(E \wedge H). \tag{27}$$

From (26) and (27) we have that

$$\frac{1}{c}(E \cdot \dot{D} + H \cdot \dot{B}) + \frac{4\pi}{c} j \cdot E + \operatorname{div}(E \wedge H) = 0. \tag{28}$$

When we multiply this equation by $c/4\pi$, integrate throughout an arbitrary volume and apply Gauss's theorem, this gives

$$\frac{1}{4\pi} \int (E.\dot{D} + H.\dot{B}) dV + \int j.E dV + \frac{c}{4\pi} \int (E \wedge H).n dS = 0 \tag{29}$$

where the last integral is taken over the boundary of the volume, $n$ being the unit outward normal.

The relation (29) is a direct consequence of Maxwell's equations and is therefore valid whether or not the material equations (9)-(11) hold. It represents, as will be seen, the energy law of an electromagnetic field. We shall discuss it here only for the case where the material equations (9)-(11) are satisfied. Generalizations to anisotropic media, where the material equations are of a more complicated form, will be considered later (Chapter XIV).

We have, on using the material equations,

$$\left.\begin{array}{l} \frac{1}{4\pi}(\boldsymbol{E} \cdot \dot{D}) = \frac{1}{4\pi} \boldsymbol{E} \cdot \frac{\partial}{\partial t}(\varepsilon \boldsymbol{E}) = \frac{1}{8\pi} \frac{\partial}{\partial t} \left(\varepsilon \boldsymbol{E}^2\right) = \frac{1}{8\pi} \frac{\partial}{\partial t}(\boldsymbol{E} \cdot \boldsymbol{D}), \\ \frac{1}{4\pi}(\boldsymbol{H} \cdot \dot{\boldsymbol{B}}) = \frac{1}{4\pi} \boldsymbol{H} \cdot \frac{\partial}{\partial t}(\mu \boldsymbol{H}) = \frac{1}{8\pi} \frac{\partial}{\partial t} \left(\mu \boldsymbol{H}^2\right) = \frac{1}{8\pi} \frac{\partial}{\partial t}(\boldsymbol{H} \cdot \boldsymbol{B}). \end{array}\right\} \tag{30}$$

Setting

$$w_e = \frac{1}{8\pi} E.D, \quad w_m = \frac{1}{8\pi} H.B, \tag{31}$$

and

$$W = \int (w_e + w_m)\, dV \tag{32}$$

(29) becomes,

$$\frac{dW}{dt} + \int j \cdot E dV + \frac{c}{4\pi} \int (E \wedge H) \cdot n dS = 0 \tag{33}$$

We shall show that $W$ represents the total energy contained within the volume, so that $w_e$ may be identified with the electric energy density and $w_m$ with the magnetic energy density of the field.*

To justify the interpretation of $W$ as the total energy we have to show that, for a closed system (i.e. one in which the field on the boundary surface may be neglected), the change in $W$ as defined above is due to the work done by the field on the material charged bodies which are embedded in it. It suffices to do this for slow motion of the material bodies, which themselves

may be assumed to be so small that they can be regarded as point charges $e_k (k = 1, 2, \ldots)$. Let the velocity of the charge $e_k$ be $\mathbf{v}_k$ ($\boldsymbol{v}_k \ll c$).

The force exerted by a field $(\boldsymbol{E}, \boldsymbol{B})$ on a charge $e$ moving with velocity $\mathbf{v}$ is given by the so-called Lorentz law,

$$F = e\left(E + \frac{1}{c} \vee \wedge B\right) \tag{34}$$

which is based on experience. It follows that if all the charges $e_k$ are displaced by $\delta x_k (k = 1, 2, \ldots)$ in time $\delta t$, the total work done is

$$\begin{aligned}\delta A &= \sum_k \boldsymbol{F}_k \cdot \delta \boldsymbol{x}_k = \sum_k e_k \left(\boldsymbol{E}_k + \frac{1}{c}\mathbf{v}_k \wedge \boldsymbol{B}\right) \cdot \delta \boldsymbol{x}_k \\ &= \sum_k e_k \boldsymbol{E}_k \cdot \delta \boldsymbol{x}_k = \sum_k e_k \boldsymbol{E}_k \cdot \mathbf{v}_k \delta t\end{aligned}$$

since $\delta x_k = \mathbf{v}_k \delta t$. If the number of the charged particles is large, we can consider the distribution to be continuous. We introduce the charge density $\rho$ (i.e. total charge per unit volume) and the last equation becomes

$$\delta A = \delta t \int \rho \mathbf{v} \cdot \boldsymbol{E} dV \tag{35}$$

the integration being carried throughout an arbitrary volume. Now the velocity $v$ does not appear explicitly in Maxwell's equations, but it may be introduced by using an experimental result found by Röntgen †, according to which a convection current (i.e. a set of moving charges) has the same electromagnetic effect as a conduction current in a wire. Hence the current density $\boldsymbol{j}$ appearing in Maxwell's equations can be split into two parts

$$\boldsymbol{j} = \boldsymbol{j}_c + \boldsymbol{j}_v \tag{36}$$

where

$$\boldsymbol{j}_c = \boldsymbol{\sigma} \boldsymbol{E}$$

is the conduction current density, and

$$\boldsymbol{j}_v = \rho \boldsymbol{v}$$

- In the general case the densities are defined by the expressions

$$w_e = \frac{1}{4\pi} \int E \cdot dD, \quad w_m = \frac{1}{4\pi} \int H \cdot dB$$

When the relationship between $\boldsymbol{E}$ and $\boldsymbol{D}$ and between $\boldsymbol{H}$ and $\boldsymbol{B}$ is linear, as here assumed, these expressions reduce to (31).

† W. C. Röntgen, Ann. d. Physik., 35 (1888), 264; 40 (1890), 93.

represents the convection current density. (35) may therefore be written as

$$\delta A = \delta t \int \boldsymbol{j}_v \cdot \boldsymbol{E} dV \tag{37}$$

Let us now define a vector $S$ and a scalar $Q$ by the relations

$$S = \frac{c}{4\pi}(E \wedge H) \tag{38}$$

$$Q = \int j_c \cdot E dV = \int \sigma E^2 dV \tag{39}$$

Then by (35) and (36)

$$\begin{aligned} \int \boldsymbol{j} \cdot \boldsymbol{E} dV &= Q + \int j_v \cdot \boldsymbol{E} dV \\ &= Q + \frac{\delta A}{\delta t} \end{aligned} \tag{40}$$

where the second function is not, of course, a total derivative of a space time function. (33) now takes the form

$$\frac{dW}{dt} = -\frac{\delta A}{\delta t} - Q - \int S.ndS. \tag{41}$$

For a nonconductor ($\sigma = 0$) we have that $Q = 0$. Assume also that the boundary surface is so far away that we can neglect the field on it, due to the electromagnetic processes inside; then $\int$ S.ndS $= 0$, and integration of (41) gives

$$W + A = \text{ constant } . \tag{42}$$

Hence, for an isolated system, the increase of $W$ per unit time is due to the work done on the system during this time. This result justifies our definition of electromagnetic energy by means of (32).

The term $Q$ represents the resistive dissipation of energy (called Joule's heat) in a conductor ( $\sigma \neq 0$ ). According to (41) there is a further decrease in energy if the field extends to the boundary surface. The surface integral must therefore represent the flow of energy across this boundary surface. The vector $S$ is known as the Poynting vector and represents the amount of energy which crosses per second a unit area normal to the directions of $\boldsymbol{E}$ and $\boldsymbol{H}$.

It should be noted that the interpretation of $S$ as energy flow (more precisely as the density of the flow) is an abstraction which introduces a certain degree of arbitrariness. For the quantity which is physically significant is, according to (41), not $S$ itself, but the integral of $\boldsymbol{S}$. $\boldsymbol{n}$ taken over a closed surface. Clearly, from the value of the integral, no unambiguous conclusion can be drawn about the detailed distribution of $S$, and alternative definitions of the energy flux density are therefore possible. One can always add to $\boldsymbol{S}$ the curl of an arbitrary vector, since such a term will not contribute to the surface integral as can be seen from Gauss' theorem and the identity div curl $\equiv 0$.* However, when the definition has been applied cautiously, in particular

for averages over small but finite regions of space or time, no contradictions with experiments have been found. We shall therefore accept the above definition in terms of the Poynting vector of the density of the energy flow.

---

0

- According to modern theories of fields the arbitrariness is even greater, allowing for alternative expressions for both the energy density and the energy flux, but consistent with the change of the Lagrangian density of the field by the addition of a four-divergence.

Finally we note that in a nonconducting medium ( $\sigma = 0$ ) where no mechanical work is done ($A = 0$), the energy law may be written in the form of a hydrodynamical continuity equation for noncompressible fluids:

$$\frac{\partial w}{\partial t} + \operatorname{div} \boldsymbol{S} = 0, \quad (w = w_e + w_m)\,. \tag{43}$$

A description of propagation of light in terms of a hydrodynamical model is often helpful, particularly in the domain of geometrical optics and in connection with scalar diffraction fields, as it gives a picture of the energy transport in a simple and graphic manner. In optics, the (averaged) Poynting vector is the chief quantity of interest. The magnitude of the Poynting vector is a measure of the light intensity, and its direction represents the direction of propagation of the light.

## 9.2.2 1.2 The Wave Equation and the Velocity of Light

Maxwell's equations relate the field vectors by means of simultaneous differential equations. On elimination we obtain differential equations which each of the vectors must separately satisfy. We shall confine our attention to that part of the field which contains no charges or currents, i.e. where $j = 0$ and $\rho = 0$.

We substitute for $\boldsymbol{B}$ from the material equation § 1.1 (11) into the second Maxwell equation § 1.1 (2), divide both sides by $\mu$ and apply the operator curl. This gives

$$\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl} E\right) + \frac{1}{c}\operatorname{curl}\dot{H} = 0. \tag{1}$$

Next we differentiate the first Maxwell equation § 1.1 (1) with respect to time, use the material equation §1.1 (10) for $D$, and eliminate curl $\dot{H}$ between the resulting equation and (1); this gives

$$\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl} E\right) + \frac{\varepsilon}{c^2}\ddot{E} = 0. \tag{2}$$

If we use the identities curl $uv = u$ curl $v + (\operatorname{grad} u) \wedge v$ and curl curl $= \operatorname{grad}\operatorname{div} - \nabla^2$, (2) becomes

$$\nabla^2 E - \frac{\varepsilon\mu}{c^2}\ddot{E} + (\operatorname{grad}\log\mu) \wedge \operatorname{curl} E - \operatorname{grad}\operatorname{div} E = 0. \tag{3}$$

Also from § 1.1 (3), using again the material equation for $D$ and applying the identity $\operatorname{div} u\mathbf{v} = u \operatorname{div}\mathbf{v} + \mathbf{v}\cdot\operatorname{grad} u$ we find

$$\varepsilon \operatorname{div} E + E \cdot \operatorname{grad}\varepsilon = 0. \tag{4}$$

Hence (3) may be written in the form

$$\nabla^2 E - \frac{\varepsilon\mu}{c^2}\ddot{E} + (\operatorname{grad}\log\mu) \wedge \operatorname{curl} E + \operatorname{grad}(E.\operatorname{grad}\log\varepsilon) = 0. \tag{5}$$

In a similar way we obtain an equation for $H$ alone:

$$\nabla^2 \boldsymbol{H} - \frac{\varepsilon\mu}{c^2}\ddot{\boldsymbol{H}} + (\operatorname{grad}\log\varepsilon) \wedge \operatorname{curl}\boldsymbol{H} + \operatorname{grad}(\boldsymbol{H}.\operatorname{grad}\log\mu) = 0. \tag{6}$$

In particular, if the medium is homogeneous, $\operatorname{grad}\log\varepsilon = \operatorname{grad}\log\mu = 0$, and (5) and (6) reduce to

$$\nabla^2 \boldsymbol{E} - \frac{\varepsilon\mu}{c^2}\ddot{\boldsymbol{E}} = 0, \quad \nabla^2 \boldsymbol{H} - \frac{\varepsilon\mu}{c^2}\ddot{\boldsymbol{H}} = 0. \tag{7}$$

These are standard equations of wave motion and suggest the existence of electromagnetic waves propagated with a velocity*

$$v = c/\sqrt{\varepsilon\mu} \tag{8}$$

The constant $c$ was first determined by R. Kohlrausch and W. Weber in 1856 from ratio of the values of the capacity of a condenser measured in electrostatic and electromagnetic units, and it was found to be identical with the velocity of light in free space. Using this result, Maxwell developed his electromagnetic theory of light, predicting the existence of electromagnetic waves; the correctness of this prediction was confirmed by the celebrated experiments of H. Hertz (cf. Historical Introduction).

As in all wave theories of light, the elementary process which produces the optical impression is regarded as being a harmonic wave in space-time (studied in its simplest form in § 1.3 and § 1.4). If its frequency is in the range from $4 \times 10^{14}\text{sec}^{-1}$ to $7.5 \times 10^{14}\text{sec}^{-1}$ (approximately) it gives rise to the psychological impression of a definite colour. (The opposite however is not true: coloured light of a certain subjective quality may be a composition of harmonic waves of very different frequency distributions.) The actual connection between colour and frequency is very involved and will not be studied in this book. †

The first determination of the velocity of light ‡ was made by Römer in 1675 from observations of the eclipses of the first satellite of Jupiter and later in a different way (from aberration of fixed stars) by Bradley (1728).

The first measurements of the velocity of light from terrestrial sources were carried out by Fizeau in 1849. It is necessary to employ a modulator, which marks off a portion of the beam§ and for this purpose Fizeau used a rotating wheel. Later methods employed rotating mirrors or electronic shutters. The rotating mirror method was suggested by Wheatstone in 1834 and was used by Foucault in 1860. It was later systematically developed over a period of many years by Michelson. The average value based on about 200 measurements by Michelson gave $c$ as 299,796 km/sec. An optical shutter method employing a Kerr cell was developed by Karolus and Mittelstaedt (1928), Anderson (1937), and Httel (1940). The values of $c$ obtained from these measurements are in excellent agreement with those based on

indirect methods, such as determinations from the ratio of an electric charge measured in electrostatic and electromagnetic units; for example Rosa and Dorsey (1907) found in this way $c$ as $299,784$ km/sec. Measurements of the velocity of electromagnetic waves on wires carried out by Mercier (1923) gave the value of $c$ equal to 299,782 km/sec. The value adopted by R. T. Birge (1941, loc. cit.) from a careful analysis of all the available data is

$$c = 299,776 \pm 4 \text{ km/sec.} \tag{9}$$

0

- The concept of a velocity of an electromagnetic wave has actually an unambiguous meaning only in connection with waves of very simple kind, e.g. plane waves. That $v$ does not represent the velocity of propagation of an arbitrary solution of (7) is obvious if we bear in mind that these equations also admit standing waves as solutions.

In this introductory section it is assumed that the reader is familiar with the concept of a plane wave, and we regard $v$ as the velocity with which such a wave advances. The mathematical representation of a plane wave will be discussed in § 1.3 and § 1.4.

† The sensitivity of the human eye to different colours is, however, briefly discussed in §4.8.1.

‡ For description of the methods used for determination of the velocity of light, see for example, E. Bergstrand, Encyclopedia of Physics, edit. S. Flgge, (Berlin, Springer), 24 (1956), 1.

Detailed analysis of the results obtained by different methods is also given by R. T. Birge in Rep. Progr. Phys. (London, The Physical Society), 8 (1941), p. 90.

§ Such determinations give essentially the group velocity (see § 1.3.4). The difference between the group velocity and the phase velocity in air at standard temperature and pressure is about 1 part in 50,000 .

The close agreement between the values of $c$ obtained from measurements of very different kinds (and in some cases using radiation whose frequencies differ by a factor

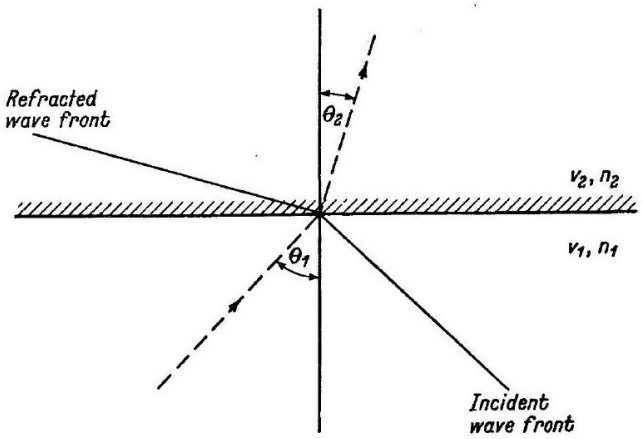



Fig. 1.3. Illustrating the refraction of a plane wave.

of hundreds of thousands from those used in the optical measurements) gives a striking confirmation of Maxwell's theory.

The dielectric constant $\varepsilon$ is usually greater than unity, and $\mu$ is practically equal to unity for transparent substances, so that the velocity $v$ is then according to (8) smaller than the vacuum velocity $c$. This conclusion was first demonstrated experimentally for propagation of light in water in 1850 by Foucault and Fizeau.

The value of $v$ is not as a rule determined directly, but only relative to $c$, with the help of the law of refraction. According to this law, if a plane electromagnetic wave falls on to a plane boundary between two homogeneous media, the sine of the angle $\theta_1$ between the normal to the incident wave and the normal to the surface bears a constant ratio to the sine of the angle $\theta_2$ between the normal of the refracted wave and the surface normal (Fig. 1.3), this constant ratio being equal to the ratio of the velocities $v_1$ and $v_2$ of propagation in the two media:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_1}{v_2}. \tag{10}$$

This result will be derived in § 1.5. Here we only note that it is equivalent to the assumption that the wave-front, though it has a kink at the boundary, is continuous, so that the line of intersection between the incident wave and the boundary travels at the same speed ( $v'$, say) as the line of intersection between the refracted wave and the boundary. We then have

$$v_1 = v' \sin\theta_1, \quad v_2 = v' \sin\theta_2, \tag{11}$$

from which, on elimination of $v', (10)$ follows. This argument, in a slightly more elaborate form, is often given as an illustration of Huygens' construction (cf. § 3.3).

The value of the constant ratio in (10) is usually denoted by $n_{12}$ and is called the refractive index, for refraction from the first into the second medium. We also define an "absolute refractive index" $n$ of a medium; it is the refractive index for refraction from vacuum into that medium,

$$n = \frac{c}{v}. \tag{12}$$

If $n_1$ and $n_2$ are the absolute refractive indices of two media, the (relative) refractive index $n_{12}$ for refraction from the first into the second medium then is

$$n_{12} = \frac{n_2}{n_1} = \frac{v_1}{v_2}. \tag{13}$$

Comparison of (12) and (8) gives Maxwell's formula:

$$n = \sqrt{\varepsilon\mu} \tag{14}$$

Since for all substances with which we shall be concerned, $\mu$ is effectively unity (nonmagnetic substances), the refractive index should then be equal to the square root of the dielectric constant, which has been assumed to be a constant of the material. On the other hand, well-known experiments on prismatic colours, first carried out by Newton, show that the index of refraction depends on the colour, i.e. on the frequency of the light. If we are to retain Maxwell's formula, it must be supposed that $\varepsilon$ is not a constant characteristic of the material, but is a function of the frequency of the field. The dependence of $\varepsilon$ on frequency can only be treated by taking into account the atomic structure of matter, and will be briefly discussed in § 2.3.

Maxwell's formula (with $\varepsilon$ equal to the static dielectric constant) gives a good approximation for such substances as gases with a simple chemical structure which do not disperse light substantially, i.e. for those whose optical properties do not strongly depend on the colour of the light. Results of some early measurements for such gases, carried out by L. Boltzmann* are given in Table I. (14) also gives a good approximation for liquid hydrocarbons; for example benzene $C_6H_6$ has $n = 1.482$ for yellow light whilst $\sqrt{\varepsilon} = 1.489$.

Table I

| | $n$ (yellow light) | $\sqrt{\varepsilon}$ |
|---|---|---|
| Air | | |
| Hydrogen $H_2$ · | $1 \cdot 000294$ | $1 \cdot 000295$ |
| Carbon dioxide $CO_2$ · | $1 \cdot 000138$ | $1 \cdot 000132$ |
| Carbon monoxide CO · | $1 \cdot 000449$ | $1 \cdot 000473$ |
| | | $1 \cdot 000345$ |

On the other hand, there is a strong deviation from the formula for many solid bodies (e.g. glasses), and for some liquids, as illustrated in Table II.

Table II

| | $n$ (yellow light) | $\sqrt{\varepsilon}$ |
|---|---|---|
| Methyl alcohol $CH_3OH$. | 1.34 | 5.7 |
| Ethyl alcohol $C_2H_5OH$. | 1.36 | 5.0 |
| Water $H_2O$. | 1.33 | 9.0 |

## 9.2.3 1.3 Scalar Waves

In an homogeneous medium in regions free of currents and charges, each rectangular component $V(r,t)$ of the field vectors satisfies, according to § 1.2 (7), the homogeneous wave equation

0

- L. Boltzmann, Wien. Ber., 69 (1874), 795; Pogg. Ann., 155 (1875), 403; Wiss. Abh. Physiktechn. Reichsanst., 1, Nr. 26, 537.

$$\nabla^2 V - \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2} = 0. \tag{1}$$

We shall now briefly examine the simplest solution of this equation.

### 1.3.1 Plane waves

Let $r(x, y, z)$ be a position vector of a point $P$ in space and $s\,(s_x, s_y, s_z)$ a unit vector in a fixed direction. Any solution of (1) of the form

$$V = V(r.s, t) \tag{2}$$

is said to represent a plane wave, since at each instant of time $V$ is constant over each of the planes

$$r.s = \text{ constant}$$

which are perpendicular to the unit vector $s$.

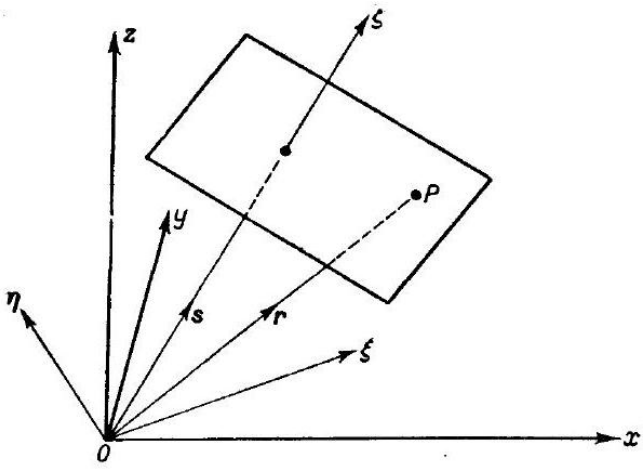



Fig. 1.4. Propagation of a plane wave.

It will be convenient to choose a new set of Cartesian axes $O\xi, O\eta, O\zeta$ with $O\zeta$ in the direction of $s$. Then (see Fig. 1.4)

$$\boldsymbol{r} \cdot \boldsymbol{s} = \zeta, \tag{3}$$

and one has

$$\frac{\partial}{\partial x} = s_x \frac{\partial}{\partial \zeta}, \quad \frac{\partial}{\partial y} = s_y \frac{\partial}{\partial \zeta}, \quad \frac{\partial}{\partial z} = s_z \frac{\partial}{\partial \zeta}.$$

From these relations one easily finds that

$$\nabla^2 V = \frac{\partial^2 V}{\partial \zeta^2} \tag{4}$$

so that (1) becomes

$$\frac{\partial^2 V}{\partial \zeta^2} - \frac{1}{v^2}\frac{\partial^2 V}{\partial t^2} = 0 \tag{5}$$

If we set

$$\zeta - vt = p, \quad \zeta + vt = q \tag{6}$$

(5) takes the form

$$\frac{\partial^2 V}{\partial p \partial q} = 0 \tag{7}$$

The general solution of this equation is

$$\begin{aligned} V &= V_1(p) + V_2(q) \\ &= V_1(r.s - vt) + V_2(r.s + vt), \end{aligned} \tag{8}$$

where $V_1$ and $V_2$ are arbitrary functions.
We see that the argument of $V_1$ is unchanged when ( $\zeta, t$ ) is replaced by ( $\zeta + v\tau$, $t + \tau$), where $\tau$ is arbitrary. Hence $V_1$ represents a disturbance which is propagated with velocity $v$ in the positive $\zeta$ direction. Similarly $V_2(\zeta + vt)$ represents a disturbance which is propagated with velocity $v$ in the negative $\zeta$ direction.

### 1.3.2 Spherical waves

Next we consider solutions representing spherical waves, i.e. solutions of the form

$$V = V(r, t) \tag{9}$$

where $r = |r| = \sqrt{x^2 + y^2 + z^2}$.
Using the relations $\frac{\partial}{\partial x} = \frac{\partial r}{\partial x}\frac{\partial}{\partial r} = \frac{x}{r}\frac{\partial}{\partial r}$, etc., one finds after a straightforward calculation that

$$\nabla^2 V = \frac{1}{r}\frac{\partial^2}{\partial r^2}(rV) \tag{10}$$

so that the wave equation (1) now becomes

$$\frac{\partial^2}{\partial r^2}(rV) - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}(rV) = 0 \tag{11}$$

Now this equation is identical with (5), if $\zeta$ is replaced in the latter by $r$ and $V$ by $rV$. Hence the solution of (11) can immediately be written down from (8):

$$V = \frac{V_1(r - vt)}{r} + \frac{V_2(r + vt)}{r} \tag{12}$$

$V_1$ and $V_2$ being again arbitrary functions. The first term on the right hand side of (12) represents a spherical wave diverging from the origin, the second a spherical wave converging towards the origin, the velocity of propagation being $v$ in both cases.

### 1.3.3 Harmonic waves. The phase velocity

At a point $r_0$ in space the wave disturbance is a function of time only:

$$V\,(r_0, t) = F(t) \tag{13}$$

As will be evident from our earlier remarks about colour, the case when $F$ is periodic is of particular interest. Accordingly we consider the case when $F$ has the form

$$F(t) = a\cos[\omega t + \delta] \tag{14}$$

Here $a(>0)$ is called the amplitude, and the argument $\omega t + \delta$ of the cosine term is called the phase. The quantity

$$v = \frac{\omega}{2\pi} = \frac{1}{T} \tag{15}$$

is called the frequency and represents the number of vibrations per second. $\omega$ is called the angular frequency and gives the number of vibrations in $2\pi$ seconds. Since $F$ remains unchanged when $t$ is replaced by $t + T, T$ is the period of the vibrations. Wave functions (i.e. solutions of the wave equation) of the form (14) are said to be harmonic with respect to time.

Let us first consider a wave function which represents a harmonic plane wave propagated in the direction specified by a unit vector $s$. According to § 1.3.1 it is obtained on replacing $t$ by $t - r.s/v$ in (14):

$$V(r,t) = a\cos\left[\omega\left(t - \frac{r\cdot s}{v}\right) + \delta\right]. \tag{16}$$

Eq. (16) remains unchanged when $r.s$ is replaced by $r.s + \lambda$, where

$$\lambda = v\frac{2\pi}{\omega} = vT \tag{17}$$

The length $\lambda$ is called the wavelength. It is also useful to define a reduced wavelength $\lambda_0$ as

$$\lambda_0 = cT = n\lambda \tag{18}$$

this is the wavelength which corresponds to a harmonic wave of the same frequency propagated in vacuo. In spectroscopy one uses also the concept of a wave number* $\kappa$, which is defined as the number of wavelengths in vacuo, per unit of length (cm):

$$\kappa = \frac{1}{\lambda_0} = \frac{v}{c} \tag{19}$$

It is also convenient to define vectors $\boldsymbol{k_0}$ and $\boldsymbol{k}$ in the direction $\boldsymbol{s}$ of propagation, whose lengths are respectively

$$k_0 = 2\pi\kappa = \frac{2\pi}{\lambda_0} = \frac{\omega}{c} \tag{20}$$

and

$$k = nk_0 = \frac{2\pi}{\lambda} = \frac{n\omega}{c} = \frac{\omega}{v} \tag{21}$$

The vector $k = ks$ is called the wave vector or the propagation vector in the medium, $k_0 = k_0 s$ being the corresponding vector in the vacuum.

Instead of the constant $\delta$ one also uses the concept of path length $l$, which is the distance through which a wave-front recedes when the phase increases by $\delta$ :

$$l = \frac{v}{\omega}\delta = \frac{\lambda}{2\pi}\delta = \frac{\lambda_0}{2\pi n}\delta \tag{22}$$

Let us now consider time harmonic waves of more complicated form. A general time harmonic real scalar wave of frequency $\omega$ may be defined as a real solution of the wave equation, of the form

0

- We shall refer to $\kappa$ as the "spectroscopic wave number" and reserve the term "wave number" for $k_0$ or $k$, defined by (20) and (21), as customary in optics.

$$V(\boldsymbol{r},t) = a(\boldsymbol{r})\cos[\omega t - g(\boldsymbol{r})], \tag{23}$$

$a(>0)$ and $g$ being real scalar functions of positions. The surfaces

$$g(r) = \text{ constant} \tag{24}$$

are called co-phasal surfaces or wave surfaces. In contrast with the previous case, the surfaces of constant amplitude of the wave (23) do not, in general, coincide with the surfaces of constant phase. Such a wave is said to be inhomogeneous.

Calculations with harmonic waves are simplified by the use of exponential instead of trigonometric functions. Eq. (23) may be written as

$$V(\boldsymbol{r},t) = \mathscr{R}\left\{U(\boldsymbol{r})e^{-i\omega t}\right\} \tag{25}$$

where

$$U(r) = a(r)e^{ig(r)} \tag{26}$$

and $\mathscr{R}$ denotes the real part. On substitution from (26) into the wave equation (1), one finds that $U$ must satisfy the equation

$$\nabla^2 U + n^2 {k_0}^2 U = 0. \tag{27}$$

$U$ is called the complex amplitude* of the wave. In particular, if the wave is plane, one has

$$g(r) = \omega\left(\frac{r\cdot s}{v}\right) - \delta = k(r\cdot s) - \delta = k\cdot r - \delta \tag{28}$$

If the operations on $V$ are linear, one may drop the symbol $\mathscr{R}$ in (25) and operate directly with the complex function, the real part of the final expression being then understood to represent the physical quantity in question. However, when dealing with expressions which involve non-linear operations such as squaring, etc. (e.g. in calculations of the electric or magnetic energy densities), one must in general take the real parts first and operate with these alone. †

Unlike a plane harmonic wave, the more general wave (25) is not periodic with respect to space. The phase $\omega t - g(r)$ is, however, seen to be the same for $(r,t)$ and $(\boldsymbol{r}+d\boldsymbol{r}, t+dt)$, provided that

$$\omega dt - (\operatorname{grad} g)\cdot dr = 0. \tag{29}$$

If we denote by $\boldsymbol{q}$ the unit vector in the direction of $\boldsymbol{dr}$, and write $\boldsymbol{dr} = \boldsymbol{q}ds$, then (29) gives

$$\frac{ds}{dt} = \frac{\omega}{q\cdot\operatorname{grad} g} \tag{30}$$

This expression will be numerically smallest when $\boldsymbol{q}$ is the normal to the co-phasal surface, i.e. when $\boldsymbol{q} = \operatorname{grad} g/|\operatorname{grad} g|$, the value then being

$$v^{(p)}(r) = \frac{\omega}{|\operatorname{grad} g|} \tag{31}$$

0

- In the case of plane waves, one often separates the constant factor $e^{-i\delta}$ and implies by complex amplitude only the variable part aeik.r.
  † This is not necessary when only a time average of a quadratic expression is required [cf. § 1.4, eq. (54)-(56)].

$v^{(p)}(r)$ is called the phase velocity and is the speed with which each of the co-phasal surfaces advances. For a plane electromagnetic wave one has from (28) that $\operatorname{grad} g = k$, and therefore

$$v^{(p)} = \frac{\omega}{k} = \frac{c}{\sqrt{\varepsilon\mu}}$$

because of (21). For waves of more complicated form, the phase velocity $v^{(p)}$ will in general differ from $c/\sqrt{\varepsilon\mu}$ and will vary from point to point even in a homogeneous medium. However, it will be seen later (§ 3.1.2) that, when the frequency is sufficiently large, the phase velocity is approximately equal to $c/\sqrt{\varepsilon\mu}$, even for waves whose co-phasal surfaces are not plane.

It must be noted that the expression for $ds/dt$ given by (30) is not the resolute of the phase velocity in the $\boldsymbol{q}$-direction, i.e. the phase velocity does not behave as a vector. On the other hand its reciprocal, i.e. the quantity

$$\frac{dt}{ds} = \frac{q \cdot \operatorname{grad} g}{\omega} \tag{32}$$

is seen to be the component of the vector $(\operatorname{grad} g)/\omega$ in the $\boldsymbol{q}$ direction. The vector $(\operatorname{grad} g)/\omega$ is sometimes called phase slowness.

The phase velocity may in certain cases be greater than $c$. For plane waves this will be so when $n = \sqrt{\varepsilon\mu}$ is smaller than unity, as in the case of dispersing media in regions of the so-called anomalous dispersion* (cf. §2.3.4). Now according to the theory of relativity, signals can never travel faster than $c$. This implies that the phase velocity cannot correspond to a velocity with which a signal is propagated. It is, in fact, easy to see that the phase velocity cannot be determined experimentally and must therefore be considered to be void of any direct physical significance. For in order to measure this velocity, it would be necessary to affix a mark to the infinitely extended smooth wave and to measure the velocity of the mark. This would, however, mean the replacement of the infinite harmonic wave train by another function of space and time.

### 1.3.4 Wave packets. The group velocity

The monochromatic waves considered in the preceding section are idealizations never strictly realized in practice. It follows from Fourier's theorem that any wave $V(r,t)$

(provided it satisfies certain very general conditions), may be regarded as a superposition of monochromatic waves of different frequencies:

$$V(r,t) = \int_0^\infty a_\omega(r)\cos\left[\omega t - g_\omega(r)\right] d\omega \tag{33}$$

It will again be convenient to use a complex representation, in which $V$ is regarded as the real part of an associated complex wave:*

$$V(r,t) = \mathscr{R}\int_0^\infty a_\omega(r) e^{-i[\omega t - g_\omega(r)]} d\omega \tag{33a}$$

0

- The problem of propagation of electromagnetic signals in dispersive media has been very fully investigated by A. Sommerfeld, Ann. d. Physik, 44 (1914), 177 and by L. Brillouin, ibid., 44 (1914), 203. English translations of these papers are included in L. Brillouin, Wave Propagation and Group Velocity (New York, Academic Press, 1960), 17, 43. See also H. G. Baerwald, ibid. (5), 6 (1930), 295; 7 (1930), 731; 8 (1931), 565.

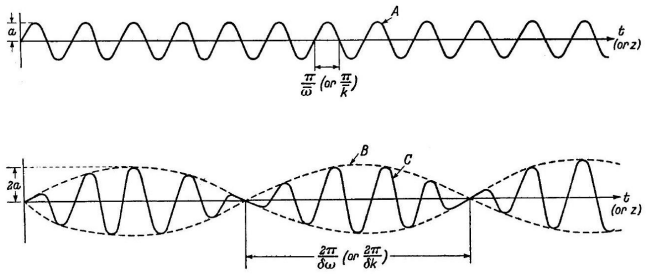



Fig. 1.5. A simple wave group.
(A) The wave $a \cos (\bar{\omega} t-\bar{k} z)$.
(B) The wave $2a \cos \left[\frac{1}{2}(t\delta\omega - z\delta k)\right]$.
(C) The wave group $2a \cos \left[\frac{1}{2}(t\delta\omega - z\delta k)\right] \cos(\bar{\omega}t - \bar{k}z)$.
The ordinate represents one of the two independent variables ( $t$ or $z$ ) whilst the other is kept constant.

A wave may be said to be "almost monochromatic," if the Fourier amplitudes $a_\omega$ differ appreciably from zero only within a narrow range

$$\bar{\omega} - \frac{1}{2}\Delta\omega \leqslant \omega \leqslant \bar{\omega} + \frac{1}{2}\Delta\omega \quad (\Delta\omega/\bar{\omega} \ll 1)$$

around a mean frequency $\bar{\omega}$. In such a case one usually speaks of a wave group or a wave packet. †

To illustrate some of the main properties of a wave group, consider first a wave formed by the superposition of two plane monochromatic waves of the same amplitudes and slightly different frequencies and wave numbers, propagated in the direction of the $z$-axis:

$$V(z,t) = ae^{-i(\omega t - kz)} + ae^{-i[(\omega+\delta\omega)t-(k+\delta k)z]}. \tag{34}$$

The symbol $\mathscr{R}$ is omitted here in accordance with the convention explained earlier. Eq. (34) may be written in the form

$$\begin{aligned} V(z,t) &= a\left\{e^{\frac{1}{i}i(t\delta\omega - z\delta k)} + e^{-\frac{1}{i}i(\delta\omega - z\delta k)}\right\} e^{-i(\bar{\omega}t - \bar{k}z)} \\ &= 2a\cos\left[\frac{1}{2}(t\delta\omega - z\delta k)\right] e^{-i(\bar{\omega}t - \bar{k}z)} \end{aligned} \tag{35}$$

where

$$\bar{\omega} = \omega + \frac{1}{2}\delta\omega, \quad \bar{k} = k + \frac{1}{2}\delta k \tag{36}$$

are the mean frequency and the mean wave number respectively. (35) may be interpreted as representing a plane wave of frequency $\bar{\omega}$ and wavelength $2\pi/\bar{k}$ propagated in the $z$ direction. The amplitude of this wave is, however, not constant, but varies with time and position, between the values $2a$ and 0 (Fig. 1.5), giving rise to the wellknown phenomenon of beats. The successive maxima of the amplitude function are at intervals

0

- For a fuller discussion of the complex representation of real polychromatic waves see §10.2.
† Strictly speaking, in order that $V$ should exhibit properties commonly attributed to a wave group, one should also assume that over the effective frequency range the phase function $g_\omega$ can be approximated to by a linear function of $\omega$.

$$\delta t = \frac{4\pi}{\delta\omega}(\text{ with } z \text{ fixed }) \quad \text{or} \quad \delta z = \frac{4\pi}{\delta k}(\text{ with } t \text{ fixed }) \tag{37}$$

from each other, whilst the maxima of the phase function are at intervals

$$\delta t = \frac{2\pi}{\bar{\omega}}(\text{ with } z \text{ fixed }) \quad \text{or} \quad \delta z = \frac{2\pi}{\bar{k}}(\text{ with } t \text{ fixed }) \tag{38}$$

Hence, since $\delta\omega/\bar{\omega}$ and $\delta k/\bar{k}$ are assumed to be small compared with unity, the amplitude will vary slowly in comparison with the other term.

From (35) it follows that the planes of constant amplitude and, in particular, the maxima of the amplitude are propagated with the velocity

$$v^{(g)} = \frac{\delta\omega}{\delta l} \tag{39}$$

whilst the planes of constant phase are propagated with the velocity

$$v^{(p)} = \frac{\bar{\omega}}{\bar{k}} \tag{40}$$

$v^{(g)}$ is called the group velocity of the wave. Since $V$ obeys the wave equation, the frequency $\omega$ and the wave number $k$ are related; in a medium of refractive index $n$, one has (see (21))

$$k = n(\omega)\frac{\omega}{c} \tag{41}$$

where $n$ is the refractive index function. Equation (41) expresses the dispersion of the wave. In a non-dispersive medium, $n$ is independent of $\omega$; the phase velocity $v^{(p)}$ and the group velocity $v^{(g)}$ are then both equal to $c/n$. In a dispersive medium, however, the two velocities will, in general, be different.

Since $\delta\omega$ is assumed to be small, $\delta\omega/\delta k$ may be replaced by the differential coefficient $d\omega/dk$, so that the expression for the group velocity may be written as

$$v^{(g)} = \frac{d\omega}{dk} \tag{42}$$

This relation is, in fact, valid under more general conditions, as we shall now show. Consider a one-dimensional wave group

$$V(z,t) = \int_{(\Delta\omega)} a_\omega e^{-i(\omega t - kz)} d\omega \tag{43}$$

where $\Delta\omega$ denotes the small interval around a mean frequency $\bar{\omega}(\Delta\omega/\bar{\omega} \ll 1)$ for which $a_\omega$ differs appreciably from zero. Let $\bar{k} = n(\bar{\omega})\bar{\omega}/c$ be the corresponding wave number. Then (43) may be expressed in the form

$$V(z,t) = A(z,t)e^{-i(\bar{\omega}t - \bar{k}z)} \tag{44}$$

where

$$A(z,t) = \int_{(\Delta\omega)} a_\omega e^{-i(\omega-\bar{\omega})t-(k-\bar{k})z]} d\omega \sim \int_{(\Delta\omega)} a_\omega e^{-i\left[(\omega-\bar{\omega})\left\{t-\left(\frac{dk}{d\omega}\right)_{\bar{\omega}}t\right\}\right]} d\omega \tag{45}$$

if $\Delta\omega$ is sufficiently small. $V$ may again be interpreted as a plane wave with variable amplitude, of frequency $\bar{\omega}$ and wave number $\bar{k}$, propagated in the $z$ direction. The amplitude $A(z,t)$ is represented as a superposition of harmonic waves of frequencies $\omega - \bar{\omega}$. Since $\Delta\omega/\bar{\omega}$ is assumed to be small compared with unity, $A$ will vary slowly in comparison with the other

term. In general $A$ is complex, so that there is a contribution ( $\arg A$ ) to the phase $\bar{\omega}t - kz$. The surfaces

$$t = \left(\frac{dk}{d\omega}\right)_{\bar{\omega}} z \tag{46}$$

are seen to play a special role: on each such surface $A(z,t)$ is constant. Hence the velocity of advance of a definite value of $A$ and also of the maximum of $|A|$ is given by the group velocity

$$v^{(g)} = \left(\frac{d\omega}{dk}\right)_{\bar{k}} \tag{47}$$

as before.
The following relations are seen to hold between the group velocity and the phase velocity:

$$v^{(g)} = \frac{d}{dk}\left(v^{(p)}k\right) = v^{(p)} + k\frac{dv^{(p)}}{dk} = v^{(p)} - \lambda\frac{dv^{(p)}}{d\lambda}, \tag{48}$$

all the quantities here referring to mean frequency $\bar{\omega}$.
Finally, let us consider the general three-dimensional group

$$V(\boldsymbol{r},\boldsymbol{t}) = \mathscr{R}\int_{(\Delta\omega)} a_\omega(\boldsymbol{r})e^{-i[\omega t - g_\omega(\boldsymbol{r})]}d\omega \tag{49}$$

By analogy with (43), we separate a term corresponding to the mean frequency $\bar{\omega}$, and write

$$V(\boldsymbol{r},t) = A(\boldsymbol{r},t)e^{-i[\bar{\omega}t - g_{\bar{\omega}}(\boldsymbol{r})]}, \tag{50}$$

where

$$\begin{aligned} A(r,t) &= \int_{(\Delta\omega)} a_\omega(r)e^{-i((\omega-\bar{\omega})t - [g_\omega(r) - g_{\bar{\omega}}(r)]\}}d\omega \\ &\sim \int_{(\Delta\omega)} a_\omega(r)e^{-i\left\{(\omega-\bar{\omega})\left[t - \left(\frac{\partial g(r)}{\partial\omega}\right)_{\bar{\omega}}\right]\right\}}d\omega \end{aligned} \tag{51}$$

if $\Delta\omega$ is sufficiently small. Eq. (50) represents a wave of frequency $\bar{\omega}$ whose (generally complex) amplitude $A(r,t)$ varies both in space and time, this variation again being slow in comparison with the other term. By analogy with (46) the surface

$$t = \left[\frac{\partial g(\boldsymbol{r})}{\partial\omega}\right]_{\bar{\omega}} \tag{52}$$

may be expected to play a special part. However, the amplitude function $A$ is now not necessarily constant over each such surface, since the Fourier amplitudes $a_\omega$ are now functions not only of the frequency but also of the position. The significance of (52) is seen by considering the absolute amplitude $M = |A|$. We have

$$M^2(\boldsymbol{r},t) = A(\boldsymbol{r},t)A^\star(\boldsymbol{r},t) = \int_{(\Delta\omega)}\int_{(\Delta\omega)} a_\omega(\boldsymbol{r})a_{\omega'}(\boldsymbol{r})e^{-i\left\{(\omega-\omega')\left[t - \left(\frac{\partial g(\boldsymbol{r})}{\partial\omega}\right)_{\bar{\omega}}\right]\right\}}d\omega d\omega' \tag{53}$$

Obviously the imaginary part of the double integral vanishes since $M^2$ is real. (Formally this may be verified by interchanging the independent variables $\omega$ and $\omega'$ and noting that the imaginary part of the integrand then changes sign.) Hence

$$M^2(\boldsymbol{r},t) = \int_{(\Delta\omega)}\int_{(\Delta\omega)} a_\omega(\boldsymbol{r})a_{\omega'}(\boldsymbol{r})\cos\left\{(\omega-\omega')\left[t - \left(\frac{\partial g(\boldsymbol{r})}{\partial\omega}\right)_{\bar{\omega}}\right]\right\}d\omega d\omega' \tag{54}$$

Fixing our attention on any particular point $r = r_0$ and remembering that $a_\omega$ is either positive or zero, we see that $M^2(r_0, t)$ attains its maximum when the argument of the cosine term is zero, i.e. when $t = \left(\frac{\partial g(r_0)}{\partial \omega}\right)_{\bar{\omega}}$. Thus (52) represents the surfaces on which the absolute amplitude attains its maximum at time $t$ in the sense just explained. It is therefore appropriate to define the group velocity of the general three-dimensional wave group as the velocity with which these surfaces advance. Considering a small displacement $\delta \boldsymbol{r} = \boldsymbol{q}\delta s$, where $\boldsymbol{q}$ is a unit vector in the direction normal to the surface, we have from (52) that the corresponding change $\delta t$ is given by

$$\delta t = \delta s \left| \operatorname{grad} \left( \frac{\partial g(\boldsymbol{r})}{\partial \omega} \right)_{\bar{\omega}} \right| \tag{55}$$

so that the group velocity of the general three-dimensional group is given by

$$v^{(g)} = \frac{1}{\left| \operatorname{grad} \left( \frac{\partial g}{\partial \omega} \right)_{\bar{\omega}} \right|} \tag{56}$$

This expression should be compared with the expression (31)

$$v^{(p)} = \frac{1}{\left| \operatorname{grad} \frac{g}{\omega} \right|} \tag{57}$$

for the phase velocity of a general harmonic wave. In the special case of a group of plane waves propagated in the $z$ direction, $g_\omega = kz$, and (56) reduces to (47).

It is evident from the preceding discussion that the effective frequency range $\Delta\omega$ is an important parameter relating to a wave group; it is this quantity which substantially determines the rate of fluctuation of the amplitude and the phase. If the medium is not strongly dispersive, a wave group will travel a considerable distance
without appreciable "diffusion". In such circumstances, the group velocity, which may be considered as the velocity of the propagation of the group as a whole, will also represent the velocity at which the energy is propagated.* This, however, is not true in general. In particular, in regions of anomalous dispersion (cf. § 2.3.4) the group velocity may exceed the velocity of light or become negative, and in such cases it has no longer any appreciable physical significance.

## 9.2.4 1.4 Vector Waves

### 1.4.1 The general electromagnetic plane wave

The simplest electromagnetic field is that of a plane wave; then each Cartesian component of the field vectors and consequently $\boldsymbol{E}$ and $\boldsymbol{H}$ are, according to § 1.3.1, functions of the variable $u = r.s - vt$ only:

$$\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{r}.\boldsymbol{s} - \boldsymbol{v}t), \quad \boldsymbol{H} = \boldsymbol{H}(\boldsymbol{r}.\boldsymbol{s} - \boldsymbol{v}t), \tag{I}$$

$s$ denoting as before a unit vector in the direction of propagation.
Denoting by a dot differentiation with respect to $t$, and by a prime differentiation with respect to the variable $u$, we have

$$\left.\begin{aligned} \dot{\boldsymbol{E}} &= -v\boldsymbol{E}' \\ (\operatorname{curl} \boldsymbol{E})_x &= \tfrac{\partial E_z}{\partial y} - \tfrac{\partial E_y}{\partial z} = E'_z s_y - E'_y s_z = (s \wedge \boldsymbol{E}')_x . \end{aligned}\right\} \tag{2}$$

Substituting these expressions into Maxwell's equations § 1.1 (1), (2) with $\boldsymbol{j} = 0$, and using the material equations § 1.1 (10), (11) we obtain

$$\left.\begin{array}{l} s \wedge \boldsymbol{H}' + \frac{\varepsilon v}{c}\boldsymbol{E}' = 0 \\ s \wedge \boldsymbol{E}' - \frac{\mu v}{c}\boldsymbol{H}' = 0 \end{array}\right\} \tag{3}$$

If we set the additive constants of integration equal to zero (i.e. neglect a field constant in space) and set, as before, $v/c = 1/\sqrt{\varepsilon\mu}$, (3) gives, on integration,

$$\left.\begin{array}{l} E = -\sqrt{\frac{\mu}{\varepsilon}} s \wedge H, \\ H = \sqrt{\frac{\varepsilon}{\mu}} s \wedge E. \end{array}\right\} \tag{4}$$

Scalar multiplication with $s$ gives

$$\boldsymbol{E}.\boldsymbol{s} = \boldsymbol{H}.\boldsymbol{s} = 0. \tag{5}$$

This relation expresses the "transversality" of the field, i.e. it shows that the electric and magnetic field vectors lie in planes normal to the direction of propagation.

From (4) and (5) it is seen that $\boldsymbol{E}, \boldsymbol{H}$ and $\boldsymbol{s}$ form a right-handed orthogonal triad of vectors. We also have from (4) that

$$\sqrt{\mu} H = \sqrt{\varepsilon} E \tag{6}$$

where $E = |\boldsymbol{E}|, H = |\boldsymbol{H}|$.

Let us now consider the amount of energy which crosses, in unit time, an element of area perpendicular to the direction of propagation. Imagine a cylinder whose axis is parallel to $s$ and whose cross-sectional area is unity. The amount of energy which crosses the base of the cylinder in unit time is then equal to the energy which was contained in the portion of the cylinder of length $v$. Hence the energy flux is equal to $vw$ where $w$ is the energy density. According to (6) and § 1.1 (31), the energy density is given by

$$w = \frac{\varepsilon}{4\pi} E^2 = \frac{\mu}{4\pi} H^2 \tag{7}$$

The Poynting vector, on the other hand, is according to § 1.1 (38), given by

$$S = \frac{c}{4\pi} EHs = \frac{c}{4\pi}\sqrt{\frac{\varepsilon}{\mu}} E^2 s = \frac{c}{4\pi}\sqrt{\frac{\mu}{\varepsilon}} H^2 s. \tag{8}$$

Comparison of (7) and (8) shows that

$$S = \frac{c}{\sqrt{\varepsilon\mu}} ws = vws \tag{9}$$

We see that, in agreement with § 1.1.4, the Poynting vector represents the flow of energy both with regard to its magnitude and direction of propagation.

### 1.4.2 The harmonic electromagnetic plane wave

Of particular interest is the case when the plane wave is time harmonic, i.e. when each Cartesian component of $\boldsymbol{E}$ and $\boldsymbol{H}$ is of the form

$$a\cos(\tau + \delta) = \mathscr{R}\left\{ae^{-i(\tau+\delta)}\right\}. \quad (a > 0) \tag{10}$$

0

- See, for example, F. Borgnis, Z. f. Phys., 117 (1941), 642; L. J. F. Broer, Appl. Sci. Res., A2 (1951), 329.

Here $\tau$ denotes the variable part of the phase factor, i.e.

$$\tau = \omega\left(t - \frac{r \cdot s}{v}\right) = \omega t - k \cdot r \tag{11}$$

We choose the $z$-axis in the $s$ direction. Then, since according to (5) the field is transversal, only the $x$ and $y$ components of $E$ and $H$ are different from zero. We shall now consider the nature of the curve which the end point of the electric vector describes at a typical point in space; this curve is the locus of the points whose coordinates ( $E_x, E_y$ ) are

$$\left.\begin{aligned} E_x &= a_1 \cos(\tau + \delta_1), \\ E_y &= a_2 \cos(\tau + \delta_2), \\ E_z &= 0. \end{aligned}\right\} \tag{12}$$

(a) Elliptic polarization

In order to eliminate $\tau$ between the first two equations of (12), we re-write them in the form

$$\left.\begin{aligned} \tfrac{E_x}{a_1} &= \cos\tau\cos\delta_1 - \sin\tau\sin\delta_1, \\ \tfrac{E_y}{a_2} &= \cos\tau\cos\delta_2 - \sin\tau\sin\delta_2. \end{aligned}\right\} \tag{13}$$

Hence

$$\left.\begin{aligned} \tfrac{E_x}{a_1}\sin\delta_2 - \tfrac{E_y}{a_2}\sin\delta_1 &= \cos\tau\sin(\delta_2 - \delta_1), \\ \tfrac{E_x}{a_1}\cos\delta_2 - \tfrac{E_y}{a_2}\cos\delta_1 &= \sin\tau\sin(\delta_2 - \delta_1). \end{aligned}\right\} \tag{14}$$

Squaring and adding gives

$$\left(\frac{E_x}{a_1}\right)^2 + \left(\frac{E_y}{a_2}\right)^2 - 2\frac{E_x}{a_1}\frac{E_y}{a_2}\cos\delta = \sin^2\delta, \tag{15}$$

where

$$\delta = \delta_2 - \delta_1. \tag{16}$$

(15) is the equation of a conic. It is an ellipse, since the associated determinant is not negative:

$$\begin{vmatrix} \frac{1}{a_1^2} & -\frac{\cos\delta}{a_1 a_2} \\ -\frac{\cos\delta}{a_1 a_2} & \frac{1}{a_2^2} \end{vmatrix} = \frac{1}{a_1^2 a_2^2}\left[1 - \cos^2\delta\right] = \frac{\sin^2\delta}{a_1^2 a_2^2} \geqslant 0.$$

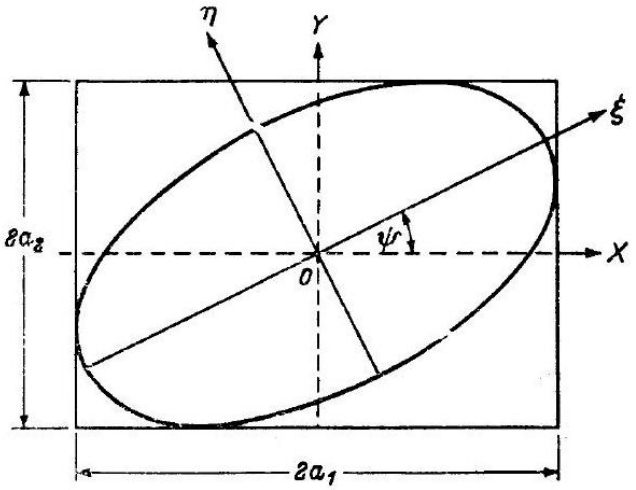



Fig. 1.6. Elliptically polarized wave. The vibrational ellipse for the electric vector.

The ellipse is inscribed into a rectangle whose sides are parallel to the co-ordinate axes and whose lengths are $2a_1$ and $2a_2$ (Fig. 1.6). The ellipse touches the sides at the points ( $\pm a_1, \pm a_2 \cos\delta$ ) and ( $\pm a_1 \cos\delta, \pm a_2$ ).

The wave (12) is then said to be elliptically polarized. It is easily seen that the wave associated with the magnetic vector is also elliptically polarized. For by (5) and (12),

$$\left.\begin{aligned} H_x &= -\sqrt{\tfrac{\varepsilon}{\mu}}E_y = -\sqrt{\tfrac{\varepsilon}{\mu}}a_2\cos(\tau+\delta_2)\,, \\ H_y &= \sqrt{\tfrac{\varepsilon}{\mu}}E_x = \sqrt{\tfrac{\varepsilon}{\mu}}a_1\cos(\tau+\delta_1)\,, \\ H_z &= 0. \end{aligned}\right\} \tag{17}$$

The end point of the magnetic vector therefore describes an ellipse which is inscribed into a rectangle whose sides are parallel to the $x$ and $y$ directions and whose lengths are $2\sqrt{\tfrac{\varepsilon}{\mu}}a_2, 2\sqrt{\tfrac{\varepsilon}{\mu}}a_1$.

In general the axes of the ellipse are not in the $Ox$ and $Oy$ directions. Let $O\xi, O\eta$ be a new set of axes along the axes of the ellipse and let $\psi(0 \leqslant \psi < \pi)$ be the angle between $Ox$ and the direction $O\xi$ of the major axis (Fig. 1.6). Then the components $E_\xi$ and $E_\eta$ are related to $E_x$ and $E_y$ by

$$\left.\begin{aligned} E_\xi &= E_x\cos\psi + E_y\sin\psi, \\ E_\eta &= -E_x\sin\psi + E_y\cos\psi. \end{aligned}\right\} \tag{18}$$

If $2a$ and $2b(a \geqslant b)$ are the lengths of the axes of the ellipse, the equation of the ellipse referred to $O\xi, O\eta$ is:

$$\left.\begin{aligned} E_\xi &= a\cos(\tau+\delta_0)\,, \\ E_\eta &= \pm b\sin(\tau+\delta_0)\,. \end{aligned}\right\} \tag{19}$$

The two signs distinguish the two possible senses in which the end point of the electric vector may describe the ellipse.

To determine $a$ and $b$ we compare (18) and (19) and use (13):

$$\begin{aligned} &a(\cos\tau\cos\delta_0 - \sin\tau\sin\delta_0) = a_1[\cos\tau\cos\delta_1 - \sin\tau\sin\delta_1]\cos\psi \\ &+ a_2[\cos\tau\cos\delta_2 - \sin\tau\sin\delta_2]\sin\psi. \\ &\pm b(\sin\tau\cos\delta_0 + \cos\tau\sin\delta_0) = -a_1[\cos\tau\cos\delta_1 - \sin\tau\sin\delta_1]\sin\psi \\ &+ a_2[\cos\tau\cos\delta_2 - \sin\tau\sin\delta_2]\cos\psi. \end{aligned}$$

Next we equate the coefficients of $\cos\tau$ and $\sin\tau$ :

$$a\cos\delta_0 = a_1\cos\delta_1\cos\psi + a_2\cos\delta_2\sin\psi, \tag{20a}$$

$$a\sin\delta_0 = a_1\sin\delta_1\cos\psi + a_2\sin\delta_2\sin\psi, \tag{20b}$$

$$\pm b\cos\delta_0 = a_1\sin\delta_1\sin\psi - a_2\sin\delta_2\cos\psi, \tag{21a}$$

$$\pm b\sin\delta_0 = -a_1\cos\delta_1\sin\psi + a_2\cos\delta_2\cos\psi. \tag{21b}$$

On squaring and adding (20a) and (20b) and using (16) we obtain

$$\left.\begin{aligned} a^2 &= {a_1}^2\cos^2\psi + {a_2}^2\sin^2\psi + 2a_1a_2\cos\psi\sin\psi\cos\delta, \\ &\qquad \text{and similarly from (21a) and (21b)} \\ b^2 &= {a_1}^2\sin^2\psi + {a_2}^2\cos^2\psi - 2a_1a_2\cos\psi\sin\psi\cos\delta. \end{aligned}\right\} \tag{22}$$

Hence

$$a^2 + b^2 = a_1^2 + a_2^2 \tag{23}$$

Next we multiply (20a) by (21a), (20b) by (21b) and add. This gives

$$\pm ab = a_1 a_2 \sin\delta. \tag{24}$$

Further on dividing (21a) by (20a) and (21b) by (20b) we obtain

$$\pm\frac{b}{a} = \frac{a_1 \sin\delta_1 \sin\psi - a_2 \sin\delta_2 \cos\psi}{a_1 \cos\delta_1 \cos\psi + a_2 \cos\delta_2 \sin\psi} = \frac{-a_1 \cos\delta_1 \sin\psi + a_2 \cos\delta_2 \cos\psi}{a_1 \sin\delta_1 \cos\psi + a_2 \sin\delta_2 \sin\psi},$$

and these relations give the following equation for $\psi$ :

$$\left(a_1^2 - a_2^2\right) \sin 2\psi = 2a_1 a_2 \cos\delta \cos 2\psi.$$

It will be convenient to introduce an auxiliary angle $\alpha(0 \leqslant \alpha \leqslant \pi/2)$, such that

$$\frac{a_2}{a_1} = \tan\alpha \tag{25}$$

The preceding equation then becomes

$$\tan 2\psi = \frac{2a_1 a_2}{a_1^2 - a_2^2} \cos\delta = \frac{2\tan\alpha}{1 - \tan^2\alpha} \cos\delta$$

i.e.

$$\tan 2\psi = (\tan 2\alpha) \cos\delta \tag{26}$$

Now from (23) and (24) we also have

$$\pm\frac{2ab}{a^2 + b^2} = \frac{2a_1 a_2}{a_1^2 + a_2^2} \sin\delta = (\sin 2\alpha) \sin\delta \tag{27}$$

Let $\chi(-\pi/4 \leqslant \chi \leqslant \pi/4)$ be another auxiliary angle, such that

$$\pm\frac{b}{a} = \tan\chi \tag{28}$$

The numerical value of $\tan\chi$ represents the ratio of the axes of the ellipse and the sign of $\chi$ distinguishes the two senses in which the ellipse may be described. Eq. (27) may be written in the form

$$\sin 2\chi = (\sin 2\alpha) \sin\delta \tag{29}$$

It will be useful to summarize the results. If $a_1, a_2$ and the phase difference $\delta$ are given, referred to an arbitrary set of axes, and if $\alpha(0 \leqslant \alpha \leqslant \pi/2)$ denotes an angle such that

$$\tan\alpha = \frac{a_2}{a_1}, \tag{30}$$

then the principal semi-axes $a$ and $b$ of the ellipse and the angle $\psi(0 \leqslant \psi < \pi)$ which the major axis makes with $Ox$ are specified by the formulae

$$a^2 + b^2 = a_1^2 + a_2^2 \tag{31a}$$
$$\tan 2\psi = (\tan 2\alpha) \cos\delta \tag{31b}$$
$$\sin 2\chi = (\sin 2\alpha) \sin\delta \tag{31c}$$

where $\chi(-\pi/4 < \chi \leqslant \pi/4)$ is an auxiliary angle which specifies the shape and orientation of the vibrational ellipse:

$$\tan\chi = \pm b/a. \tag{32}$$

Conversely, if the lengths $a$ and $b$ of the axes and the orientation of the ellipse is known (i.e. $a, b$ and $\psi$ given) these formulae enable the determination of the amplitudes $a_1, a_2$ and the phase difference $\delta$. In Chapter XIV, instruments will be described by means of which these quantities may be directly determined.

Before discussing some important special cases, we must say a few words about the terminology. We distinguish two cases of polarization, according to the sense in which the end point of the electric vector describes the ellipse. It seems natural to call the polarization right-handed or left-handed according as to whether the rotation of $\boldsymbol{E}$ and the direction of propagation form a right-handed or left-handed screw. But the traditional terminology is just the opposite-being based on the apparent behaviour of $\boldsymbol{E}$ when "viewed" face on by the observer. We shall conform throughout this book to this customary usage. Thus we say that the polarization is right-handed when to an observer looking in the direction from which the light is coming, the endpoint of the electric vector would appear to describe the ellipse in the clockwise sense. If we consider the values of (12) for two time instants separated by a quarter of a period, we see that in this case $\sin\delta > 0$, or by (29), $0 < \chi \leqslant \pi/4$. For left-handed polarization the opposite is the case, i.e. to an observer looking in the direction from which the light is propagated, the electric vector would appear to describe the ellipse anticlockwise; in this case $\sin\delta < 0$, so that $-\pi/4 \leqslant \chi < 0$.

For reasons connected with the historical development of optics, the direction of the magnetic vector is often called the direction of polarization and the plane containing the magnetic vector and the direction of propagation is known as the plane of polarization. This terminology is, however, not used by all writers; some define these quantities with respect to the electric rather than the magnetic vector. This lack of uniformity arises partly from the fact that there is no single physical entity which could be described without ambiguity as "the light vector". When particular attention is paid to the physical effect of the field vectors, there would actually be some grounds for regarding $\boldsymbol{E}$ as the light vector. For every action is a consequence of the motion of elementary charged particles (electrons, nuclei) set into motion by the electromagnetic field. The mechanical force $F$ of the field on the particle is then given by Lorentz' law, § 1.1 (34).

$$F = e\left(E + \frac{\mu}{c} v \wedge H\right),$$

$e$ being the charge and $v$ the velocity of the particle. Hence the electric vector is seen to act even when the particle is at rest. On the other hand, the magnetic vector plays a part only when the particle is in motion; however, since $v/c$ is usually very small compared to unity this effect may often be neglected. Nevertheless the "direction of polarization" and the "plane of polarization" are usually associated with the magnetic vector. The reason for this nomenclature will become apparent in the next section when polarization on reflection is discussed.

To avoid confusion we shall, in accordance with more recent practice, not use the terms "direction of polarization" and "plane of polarization"; instead we shall speak of direction of vibration and plane of vibration to denote the direction of a field vector and the plane containing the field vector and the direction of propagation, the vector in question being specified in each case.

(b) Linear and circular polarization

Two special cases are of particular importance, namely when the polarization ellipse degenerates into a straight line or a circle.

According to (12) the ellipse will reduce to a straight line, when

$$\delta = \delta_2 - \delta_1 = m\pi \quad (m = 0, \pm 1, \pm 2, \ldots) \tag{33}$$

Then

$$\frac{E_y}{E_x} = (-1)^m \frac{a_2}{a_1} \tag{34}$$

and we say that $\boldsymbol{E}$ is linearly polarized.* One of the coordinate axes, $\boldsymbol{x}$ say, may be chosen along this line. Then only one component ($E_x$) remains. Moreover, since the electric and magnetic vectors are orthogonal and lie in the plane perpendicular to the $z$ direction, the component $H_x$ then vanishes, so that $H$ is linearly polarized in the $y$ direction.

The other special case of importance is that of a circularly polarized wave, the ellipse then degenerating into a circle. Clearly a necessary condition for this is that the circumscribed rectangle shall become a square:

$$a_1 = a_2 = a \tag{35}$$

Also, one of the $\boldsymbol{E}$ components must be zero when the other has an extreme value; this demands that

$$\delta = \delta_2 - \delta_1 = m\pi/2 \quad (m = \pm 1, \pm 3, \pm 5 \ldots) \tag{36}$$

and (15) then reduces to the equation of the circle

$$E_x^2 + E_y^2 = a^2 \tag{37}$$

When the polarization is right-handed $\sin\delta > 0$, so that

$$\left.\begin{array}{rl} \delta & = \frac{\pi}{2} + 2m\pi \quad (m = 0, \pm 1, \pm 2, \ldots) \\ E_x & = a\cos(\tau + \delta_1) \\ E_y & = a\cos(\tau + \delta_1 + \pi/2) = -a\sin(\tau + \delta_1)\,. \end{array}\right\} \tag{39}$$

For left-handed polarization $\sin\delta < 0$, so that

$$\left.\begin{array}{l} \delta = -\frac{\pi}{2} + 2m\pi \quad (m = 0, \pm 1, \pm 2, \ldots) \\ E_x = a\cos(\tau + \delta_1) \\ E_y = a\cos(\tau + \delta_1 - \pi/2) = a\sin(\tau + \delta_1)\,. \end{array}\right\} \tag{41}$$

If, instead of the real representation, the complex one is used (i.e. if the exponential instead of the cosine function is written in (12)), then

$$\frac{E_y}{E_x} = \frac{a_2}{a_1} e^{i(\delta_2 - \delta_1)} = \frac{a_2}{a_1} e^{i\delta} \tag{42}$$

and one can immediately determine from the value of this ratio the nature of the polarization. One has for
(a) Linearly polarized electric wave ( $\delta = m\pi, m = 0, \pm 1, \pm 2, \ldots$ ):

$$\frac{E_y}{E_x} = (-1)^m \frac{a_2}{a_1}$$

0

- The less appropriate term plane polarized is also used.

(b) Right-handed circularly polarized electric wave* ( $a_1 = a_2, \delta = \pi/2$ ):

$$\frac{E_y}{E_x} = e^{i\pi/2} = i$$

(c) Left-handed circularly polarized electric wave ( $a_1 = a_2, \delta = -\pi/2$ ):

$$\frac{E_y}{E_x} = e^{-i\pi/2} = -i.$$

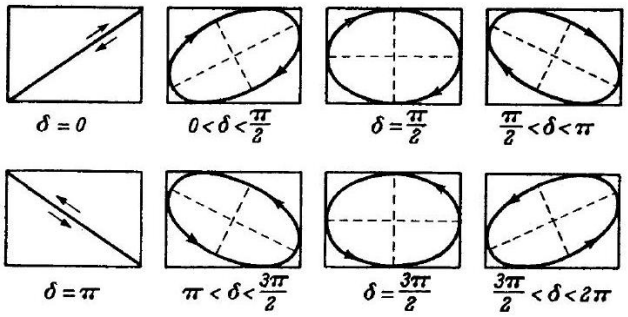



Fig. 1.7. Elliptical polarization with various values of the phase difference $\delta$.

More generally it may be shown that for right-handed elliptical polarization, the ratio $E_y/E_x$ has a positive imaginary part, whereas for left-handed elliptical polarization the imaginary part is negative.

Fig. 1.7 illustrates how the polarization ellipse changes with varying $\delta$.

(c) Characterization of the state of polarization by Stokes parameters

To characterize the polarization ellipse three independent quantities are necessary, e.g. the amplitudes $a_1$ and $a_2$ and the phase difference $\delta$, or the major and minor axes $a, b$ and the angle $\chi$ which specifies the orientation of the ellipse. For practical purposes it is convenient to characterize the state of polarization by certain parameters which are all of the same physical dimensions, and which were introduced by G. G. Stokes in 1852, in his investigations relating to partially polarized light. We shall define them later (§ 10.8.3) in their full generality. We will also show there, that for any given wave these parameters may be determined from simple experiments.

The Stokes parameters of a plane monochromatic wave are the four quantities

$$\left.\begin{aligned} s_0 &= a_1^2 + a_2^2, \\ s_1 &= a_1^2 - a_2^2, \\ s_2 &= 2a_1 a_2 \cos\delta, \\ s_3 &= 2a_1 a_2 \sin\delta. \end{aligned}\right\} \tag{43}$$

Only three of them are independent since they are related by the identity

$$s_0^2 = s_1^2 + s_2^2 + s_3^2. \tag{44}$$

The parameter $s_0$ is evidently proportional to the intensity of the wave. The parameters $s_1, s_2$ and $s_3$ are related in a simple way to the angle $\psi (0 \leqslant \psi < \pi)$ which specifies the orientation of the ellipse and the angle $\chi (-\pi/4 \leqslant \chi \leqslant \pi/4)$

0

• N.B. This is right-handed according to the traditional, not the natural nomenclature.

which characterizes the ellipticity and the sense in which the ellipse is being described. In fact the following relations hold:

$$s_1 = s_0 \cos 2\chi \cos 2\psi, \tag{45a}$$
$$s_2 = s_0 \cos 2\chi \sin 2\psi, \tag{45b}$$
$$s_3 = s_0 \sin 2\chi. \tag{45c}$$

The relation (45c) follows from (25) and (29). To derive the other two relations we note that according to the equation preceding (26),

$$s_2 = s_1 \tan 2\psi. \tag{46}$$

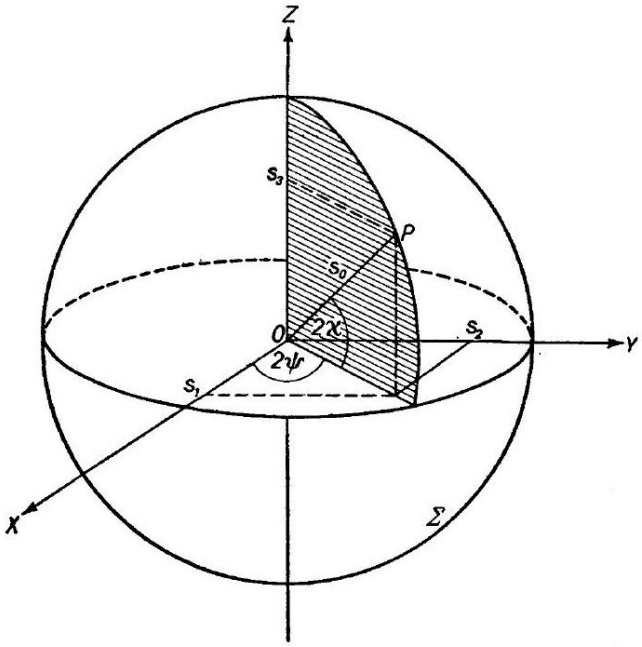



Fig. 1.8. Poncar's representation of the state of polarization of a monochromatic wave. (The Poincar sphere.)

The relation (45a) follows on substitution from (46) and from (45c) into (44). Finally, (45b) is obtained on substitution from (45a) into (46).

The relations (45) indicate a simple geometrical representation of all the different states of polarization: $s_1, s_2$ and $s_3$ may be regarded as the Cartesian coordinates of a point $P$ on a sphere $\Sigma$ of radius $s_0$, such that $2\chi$ and $2\psi$ are the spherical angular coordinates of this point (see Fig. 1.8). Thus to every possible state of polarization of a plane monochromatic wave of a given intensity ( $s_0$ = constant), there corresponds one point on $\Sigma$ and vice versa. Since $\chi$ is positive or negative according as the polarization is right-handed or left-handed, it follows from (45c) that right-handed polarization is represented by points on $\Sigma$ which lie above the equatorial plane ( $xy$-plane), and left-handed polarization by points on $\Sigma$ which lie below this plane. Further, for linearly polarized light the phase difference $\delta$ is zero or an integral multiple of $\pi$; according to (43) the Stokes parameter $s_3$ is then zero, so that linear polarization is represented by points in the equatorial plane. For circular polarization $a_1 = a_2$ and $\delta = \pi/2$ or $-\pi/2$ according as the polarization is right- or left-handed; hence right-handed circular polarization is represented by the north pole ( $s_1 = s_2 = 0$, $s_3 = s_0$ ) and left-handed circular polarization by the south pole ( $s_1 = s_2 = 0$,

$s_3 = -s_0$ ). This geometrical representation of different states of polarization by points on a sphere is due to Poincar,* and is very useful in crystal optics for determining the effect of crystalline media on the state of polarization of light traversing it. † The sphere $\Sigma$ is called the Poincar sphere.

### 1.4.3 Harmonic vector waves of arbitrary form

The main results of the preceding sections can easily be extended to time-harmonic waves of more complicated form.

A general time-harmonic real vector wave $V(r,t)$ is a solution of the vector wave equation, whose Cartesian components $V_x, V_y, V_z$ are represented by expressions of the form § 1.3 (23):

$$\left.\begin{aligned} V_x(\boldsymbol{r},t) &= a_1(\boldsymbol{r})\cos\left[\omega t - g_1(\boldsymbol{r})\right], \\ V_y(\boldsymbol{r},t) &= a_2(\boldsymbol{r})\cos\left[\omega t - g_2(\boldsymbol{r})\right], \\ V_z(\boldsymbol{r},t) &= a_3(\boldsymbol{r})\cos\left[\omega t - g_3(\boldsymbol{r})\right]. \end{aligned}\right\} \tag{47}$$

where $a_s$ and $g_s (s = 1, 2, 3)$ are real functions of position. For the plane harmonic wave considered in the preceding section the $a$ 's were constant, and $g_s(r) = k.r - \delta_s$.

It will be convenient to write (47) in the form

$$V_x(\boldsymbol{r},t) = p_x(\boldsymbol{r})\cos\omega t + q_x(\boldsymbol{r})\sin\omega t \quad \text{etc.} \tag{48}$$

where

$$\left.\begin{aligned} p_x(r) &= a_1(r)\cos g_1(r), \\ q_x(r) &= a_1(r)\sin g_1(r). \end{aligned}\right\} \tag{49}$$

If a new set of axes is chosen, then each component of $V$ with respect to the new set will again be of the form (47), since each new component is a linear combination of the old ones and can therefore involve only the sum of terms in $\cos\omega t$ and $\sin\omega t$.

We may regard ( $p_x, p_y, p_z$ ) and ( $q_x, q_y, q_z$ ) as components of two real vectors $\boldsymbol{p}$ and $\boldsymbol{q}$. Then

$$V(r,t) = p(r)\cos\omega t + q(r)\sin\omega t \tag{50}$$

By Fourier analysis, an arbitrary vector wave may be expressed as superposition of waves of this type.

As in the case of scalar waves, it is often convenient to use a complex representation. We write (50) in the form

$$V(r,t) = \mathscr{R}\left\{U(r)e^{-i\omega t}\right\} \tag{51}$$

where $\boldsymbol{U}$ is the complex vector

$$U(r) = p(r) + iq(r) \tag{52}$$

$\mathscr{R}$ denoting the real part. When the operations on $V$ are linear one can, as in the corresponding scalar case, operate directly on the complex quantity, omitting the symbol $\mathscr{R}$ altogether. The real part of the final expression is then understood to represent the physical quantity in question.

Operation with complex vectors follow the usual rules of vector algebra and of algebra of complex numbers. For example, the conjugate of $U$ is the vector

$$U^\star = p - iq$$

0

Similarly

$$U^2 = U \cdot U = p^2 - q^2 + 2ip \cdot q$$
$$U \cdot U^\star = (p + iq) \cdot (p - iq) = p^2 + q^2,$$

etc.

To illustrate calculations with complex vectors we shall derive formulae needed later for the energy densities and the Poynting vector in a time-harmonic electromagnetic field. The electric and magnetic vectors then are of the form

$$\left.\begin{aligned} \boldsymbol{E}(\boldsymbol{r},t) &= \mathscr{R}\left\{\boldsymbol{E}_0(\boldsymbol{r})e^{-i\omega t}\right\} = \tfrac{1}{2}\left[\boldsymbol{E}_0(\boldsymbol{r})e^{-i\omega t} + \boldsymbol{E}_0^\star(\boldsymbol{r})e^{i\omega t}\right], \\ \boldsymbol{H}(\boldsymbol{r},t) &= \mathscr{R}\left\{\boldsymbol{H}_0(\boldsymbol{r})e^{-i\omega t}\right\} = \tfrac{1}{2}\left[\boldsymbol{H}_0(\boldsymbol{r})e^{-i\omega t} + \boldsymbol{H}_0^\star(\boldsymbol{r})e^{i\omega t}\right], \end{aligned}\right\} \tag{53}$$

$E_0$ and $H_0$ being complex vector functions of position. Since the optical frequencies are very large ( $\omega$ is of order $10^{15}\text{sec}^{-1}$ ), one cannot observe the instantaneous values of any of the rapidly oscillating quantities, but only their time average taken over a time interval (say $-T' \leqslant t \leqslant T'$ ) which is large compared to the fundamental period $T = 2\pi/\omega$. In particular, the time-averaged electric energy density is

$$\langle w_e \rangle = \frac{1}{2T'}\int_{-T}^{T} \frac{\varepsilon}{8\pi} E^2 dt = \frac{\varepsilon}{8\pi}\frac{1}{2T'}\int_{-T'}^{T'} \frac{1}{4}\left[{E_0}^2 e^{-2i\omega t} + 2E_0.E_0^\star + E_0^{\star 2} e^{2i\omega t}\right] dt$$

Now

$$\frac{1}{2T'}\int_{-T}^{T} e^{-2i\omega t} dt = -\frac{1}{4i\omega T'}\left[e^{-2i\omega t}\right]_{-T'}^{T'} = \frac{1}{4\pi}\frac{T}{T'}\sin 2\omega T'$$

Since $T'$ is assumed to be large compared to $T, T/T'$ will be small compared to unity, so that the integral involving $e^{-2i\omega t}$ may be neglected. Similarly the integral involving $e^{2i\omega t}$ may also be neglected, and we finally obtain

$$\langle w_e \rangle = \frac{\varepsilon}{16\pi} E_0.E_0^\star. \tag{54}$$

In a similar way the time average of the magnetic energy density is seen to be

$$\langle w_m \rangle = \frac{\mu}{16\pi} H_0 \cdot H_0^\star. \tag{55}$$

The average of the Poynting vector is given by

$$\begin{aligned} \langle S \rangle &= \frac{1}{2T'}\int_{-T'}^{T'} \frac{c}{4\pi}(E \wedge H) dt \\ &= \frac{c}{4\pi}\frac{1}{2T'}\int_{-T'}^{T'} \left[E_0 \wedge H_0 e^{-2i\omega t} + E_0 \wedge H_0^\star + E_0^\star \wedge H_0 + E_0^\star \wedge H_0^\star e^{2i\omega t}\right] dt \\ &\simeq \frac{c}{16\pi}\left(E_0 \wedge H_0^\star + E_0^\star \wedge H_0\right) \\ &= \frac{c}{8\pi}\mathscr{R}\left(E_0 \wedge H_0^\star\right) \end{aligned} \tag{56}$$

---

• H. Poincar, Thorie Mathmatique de la Lumire (Paris, Georges Carr), Vol. 2, (1892) Chap. 12.
† Examples illustrating the method and references to the relevant literature are given in H. G. Jerrard, J. Opt. Soc. Amer., 44 (1954), 634. See also the paper by M. J. Walker, referred to in § 10.8.3 and S. Pancharatnam, Proc. Ind. Acad. Sci., A, 44 (1956), 247.

The law of conservation of energy also takes a simple form. For a nonconducting medium ( $\sigma = 0$ ) where no mechanical work is done, we find, on taking the time average of eq. (43) of § 1.1 that

$$\operatorname{div}\langle S\rangle = 0 \tag{57}$$

If we integrate (57) throughout an arbitrary volume which contains no radiator or absorber of energy, and apply Gauss' theorem, it follows that

$$\int \langle \boldsymbol{S}\rangle \cdot n d\boldsymbol{S} = 0 \tag{58}$$

$n$ being the outward normal to the surface; the integration is taken over the boundary. Thus the averaged total flux of energy through any closed surface is zero.

We now return to the general time-harmonic vector wave (50), and consider the behaviour of $V$ at a point $r = r_0$ in space. In general, as time varies, the end point of $V$ describes an ellipse, so that, like the plane wave, the wave (50) is in general also elliptically polarized. To see this we note first that, with varying time, the end point describes a curve in the plane specified by $\boldsymbol{p}(\boldsymbol{r}_0)$ and $\boldsymbol{q}(\boldsymbol{r}_0)$. Since $\boldsymbol{V}$ is periodic, the curve must be closed. Now we may set

$$(p + iq) = (a + ib)e^{i\varepsilon} \tag{59}$$

where $\varepsilon$ is any scalar. In terms of $\boldsymbol{p}, \boldsymbol{q}$ and $\varepsilon$,

$$\left.\begin{aligned} a &= p\cos\varepsilon + q\sin\varepsilon \\ b &= -p\sin\varepsilon + q\cos\varepsilon \end{aligned}\right\} \tag{60}$$

Let us choose $\varepsilon$ so that the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are perpendicular to each other and let $|\boldsymbol{a}| \geqslant |\boldsymbol{b}|$. For $\boldsymbol{a}$ and $\boldsymbol{b}$ to be orthogonal, $\varepsilon$ must satisfy the equation

$$(\boldsymbol{p}\cos\varepsilon + \boldsymbol{q}\sin\varepsilon)\cdot(-\boldsymbol{p}\sin\varepsilon + \boldsymbol{q}\cos\varepsilon) = 0 \tag{61}$$

i.e.

$$\tan 2\varepsilon = \frac{2p\cdot q}{p^2 - q^2} \tag{62}$$

We now consider as parameters specifying the wave the 5 independent components of the orthogonal vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ and the corresponding phase factor $\varepsilon$, instead of the 6 rectangular components of $\boldsymbol{p}$ and $\boldsymbol{q}$. Then from (51), (52) and (59),

$$\begin{aligned} V &= \mathscr{R}\left\{[a + ib]e^{-i(\omega t - s)}\right\} \\ &= a\cos(\omega t - \varepsilon) + b\sin(\omega t - \varepsilon) \end{aligned} \tag{63}$$

If we take Cartesian axes with origin at $r_0$ and with the $x$ and $y$ directions along $\boldsymbol{a}$ and $\boldsymbol{b}$, then

$$V_x = a\cos(\omega t - \varepsilon), \quad V_y = b\sin(\omega t - \varepsilon), \quad V_z = 0 \tag{64}$$

This represents an ellipse (the polarization ellipse)

$$\frac{V_x^2}{a^2} + \frac{V_y^2}{b^2} = 1 \tag{65}$$

with semi-axes of lengths $a$ and $b$ along the $x$ and $y$ co-ordinate axes. By elementary geometry it may be shown that $p$ and $q$ are a pair of conjugate semi-diameters of the ellipse.

As in the case of plane waves, the ellipse may be described in two possible senses, corresponding to left- and right-handed polarization; these are distinguished by the sign of the scalar triple product $[\boldsymbol{a}, \boldsymbol{b}, \nabla\varepsilon] = [\boldsymbol{p}, \boldsymbol{q}, \nabla\varepsilon]$.

The lengths of the semi-axes of the polarization ellipse are easily obtained from (60) and (62). From (60)

$$\begin{aligned} a^2 &= p^2 \cos^2\varepsilon + q^2 \sin^2\varepsilon + 2p \cdot q \sin\varepsilon \cos\varepsilon \\ &= \frac{1}{2}\left(p^2 + q^2\right) + \frac{1}{2}\left(p^2 - q^2\right)\cos 2\varepsilon + p \cdot q \sin 2\varepsilon \end{aligned}$$

From (62)

$$\sin 2\varepsilon = \frac{2\boldsymbol{p} \cdot \boldsymbol{q}}{\sqrt{\left(p^2 - q^2\right)^2 + 4(\boldsymbol{p} \cdot \boldsymbol{q})^2}}, \quad \cos 2\varepsilon = \frac{p^2 - q^2}{\sqrt{\left(p^2 - q^2\right)^2 + 4(\boldsymbol{p} \cdot \boldsymbol{q})^2}}$$

Hence
Similarly one finds

$$\left.\begin{aligned} a^2 &= \tfrac{1}{2}\left[p^2 + q^2 + \sqrt{\left(p^2 - q^2\right)^2 + 4(\boldsymbol{p} \cdot \boldsymbol{q})^2}\right] \\ b^2 &= \tfrac{1}{2}\left[p^2 + q^2 - \sqrt{\left(p^2 - q^2\right)^2 + 4(\boldsymbol{p} \cdot \boldsymbol{q})^2}\right] \end{aligned}\right\} \tag{66}$$

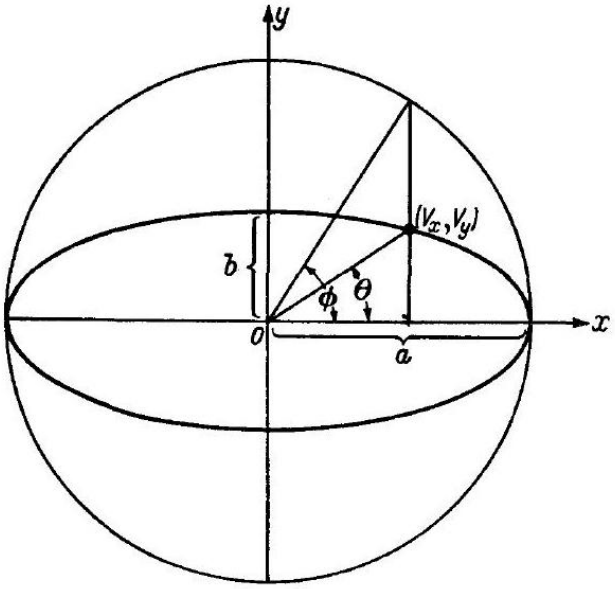



Fig. 1.9. Illustrating the notation relating to eqs. (67) and (68).

To find an expression for the angle between $\boldsymbol{a}$ and $\boldsymbol{p}$ we express the equation of the ellipse in parametric form:

$$V_x = a\cos\phi, \quad V_y = b\sin\phi \tag{67}$$

where $\phi$ is the eccentric angle (see Fig. 1.9). From elementary geometry we learn that this angle is related to the polar angle $\theta$ of the point $(V_x, V_y)$ by

$$\tan\theta = \frac{b}{a}\tan\phi \tag{68}$$

Comparison of (64) and (67) shows that in the present case $\phi = \omega t - \varepsilon$. Now according to (50), $\boldsymbol{V} = \boldsymbol{p}$ when $t = 0$, so that the eccentric angle of $\boldsymbol{p}$ is $-\varepsilon$. Hence the angle $\psi$ between $p$ and $a$ is given by

$$\tan\psi = \frac{b}{a}\tan\varepsilon \tag{69}$$

If $\gamma$ denotes the angle between $\boldsymbol{p}$ and $\boldsymbol{q}$, and we introduce an auxiliary angle $\beta$ such that

$$\frac{q}{p} = \tan\beta \tag{70}$$

then (62) becomes

$$\begin{aligned}\tan 2\varepsilon &= \frac{2pq}{p^2 - q^2}\cos\gamma \\ &= \tan 2\beta\cos\gamma\end{aligned} \tag{71}$$

Let us again summarize the results: If $\boldsymbol{p}$ and $\boldsymbol{q}$ are given, $\gamma$ denotes the angle between these vectors and $\beta$ denotes the auxiliary angle defined by (70), then the principal semi-axes of the ellipse and the angle $\psi$ which the major axis makes with $p$ are given by

$$\left.\begin{aligned} a^2 &= \tfrac{1}{2}\left[p^2 + q^2 + \sqrt{(p^2 - q^2)^2 + 4p^2q^2\cos^2\gamma}\right] \\ b^2 &= \tfrac{1}{2}\left[p^2 + q^2 - \sqrt{(p^2 - q^2)^2 + 4p^2q^2\cos^2\gamma}\right] \\ \tan\psi &= \tfrac{b}{a}\tan\varepsilon \end{aligned}\right\} \tag{72}$$

where

$$\tan 2\varepsilon = \tan 2\beta\cos\gamma \tag{73}$$

As for plane waves, there are two cases of special interest, namely when the ellipse degenerates into a circle or a straight line. For a circularly polarized wave, $\boldsymbol{a}$ and $\boldsymbol{b}$ and consequently $\varepsilon$ are not determined. According to (62), for this to be the case,

$$\boldsymbol{p}\cdot\boldsymbol{q} = \boldsymbol{p}^2 - \boldsymbol{q}^2 = 0 \tag{74}$$

For a linearly polarized wave, the minor axis vanishes ( $b^2 = 0$ ) and ( 66 ) then gives

$$p^2q^2 = (p\cdot q)^2 \tag{75}$$

Finally we stress that the term polarization refers to the behaviour at a particular point in the field, and that the state of polarization will therefore in general be different at different points of the field. Thus a wave may be linearly or circularly polarized at some points and elliptically at others.* Only in special cases, as for example for the homogeneous plane wave, will the state of polarization be the same at every point in the field.

## 9.2.5 1.5 Reflection and Refraction of a Plane Wave

In § 1.1.3 relations were derived which the field vectors must satisfy across surfaces at which the physical properties of the medium are discontinuous. These formulae will now be applied to the study of the propagation of a plane wave incident on a plane boundary between two homogeneous isotropic media.

### 1.5.1 The laws of reflection and refraction

When a plane wave falls on to a boundary between two homogeneous media of different optical properties, it is split into two waves: a transmitted wave proceeding into the second medium and a reflected wave propagated back into the first medium. The existence of these two waves can be demonstrated from the boundary conditions, since it is easily seen that these conditions cannot be satisfied without postulating both the transmitted and the reflected wave. We shall tentatively assume that these waves are

also plane, and derive expressions for their directions of propagation and their amplitudes.

A plane wave propagated in the direction specified by the unit vector* $s^{(i)}$ is completely determined when the time behaviour at one particular point in space is known. For if $\boldsymbol{F}(t)$ represents the time behaviour at any one point, the time behaviour at another point, whose position vector relative to the first point is $r$, is given by $F[t-(r.s)/v]$. At the boundary between the two media, the time variation of the secondary fields will be the same as that of the incident primary field. Hence, if $\boldsymbol{s^{(r)}}$ and $s^{(t)}$ denote unit vectors in the direction of propagation of the reflected and transmitted wave, one has, on equating the arguments of the three wave functions at a point $r$ on the boundary plane $z=0$ :

$$t-\frac{r\cdot s^{(i)}}{v_1}=t-\frac{r\cdot s^{(r)}}{v_1}=t-\frac{r\cdot s^{(t)}}{v_2}, \tag{1}$$

$v_1$ and $v_2$ being the velocities of propagation in the two media. Written out more explicitly, one has, with $r\equiv x,y,0$ :

$$\frac{xs_x^{(i)}+ys_y^{(i)}}{v_1}=\frac{xs_x^{(r)}+ys_y^{(r)}}{v_1}=\frac{xs_x^{(t)}+ys_y^{(t)}}{v_2}. \tag{2}$$

Since (2) must hold for all values $x$ and $y$ on the boundary,

$$\frac{s_x^{(i)}}{v_1}=\frac{s_x^{(r)}}{v_1}=\frac{s_x^{(t)}}{v_2},\quad \frac{s_y^{(i)}}{v_1}=\frac{s_y^{(r)}}{v_1}=\frac{s_y^{(t)}}{v_2}. \tag{3}$$

The plane specified by $s^{(i)}$ and the normal to the boundary is called the plane of incidence. (3) shows that both $s^{(t)}$ and $s^{(r)}$ lie in this plane.

Taking the plane of incidence as the $xz$-plane and denoting by $\theta_i,\theta_r$ and $\theta_t$ the angle which $s^{(i)},s^{(r)}$ and $s^{(t)}$ make with $Oz$, one has (see Fig. 1.10)

$$\left.\begin{array}{lll} s_x^{(i)}=\sin\theta_i, & s_y^{(i)}=0, & s_z^{(i)}=\cos\theta_i, \\ s_x^{(r)}=\sin\theta_r, & s_y^{(r)}=0, & s_z^{(r)}=\cos\theta_r, \\ s_x^{(t)}=\sin\theta_t, & s_y^{(t)}=0, & s_z^{(t)}=\cos\theta_t. \end{array}\right\} \tag{4}$$

For waves propagated from the first into the second medium, the $z$ components of the $s$ vectors are positive; for those propagated in the opposite sense, they are negative:

$$s_z{}^{(i)}=\cos\theta_i\geqslant 0,\quad s_z{}^{(r)}=\cos\theta_r\leqslant 0,\quad s_z{}^{(t)}=\cos\theta_t\geqslant 0. \tag{5}$$

The first set in (3) gives, on substituting from (4)

$$\frac{\sin\theta_i}{v_1}=\frac{\sin\theta_r}{v_1}=\frac{\sin\theta_t}{v_2}. \tag{6}$$

0

- General properties of time-harmonic electromagnetic waves of arbitrary form but with at least one of the field vectors linearly polarized have been investigated by A. Nisbet and E. Wolf, Proc. Camb. Phil. Soc., 50 (1954), 614.

Hence, $\sin\theta_r = \sin\theta_i$, and we find, also using (5), that $\cos\theta_r = -\cos\theta_i$, so that

$$\theta_r = \pi - \theta_i. \tag{7}$$

This relation, together with the statement that the reflected wave normal $s^{(r)}$ is in the plane of incidence, constitute the law of reflection.

Also from (6), using Maxwell's relation § 1.2 (14) connecting the refractive index and the dielectric constant,

$$\frac{\sin\theta_i}{\sin\theta_t} = \frac{v_1}{v_2} = \sqrt{\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}} = \frac{n_2}{n_1} = n_{12} \tag{8}$$

The relation $\sin\theta_i/\sin\theta_t = n_2/n_1$, together with the statement that the refracted wave normal $s^{(t)}$ is in the plane of incidence, constitute the law of refraction (or Snell's law).

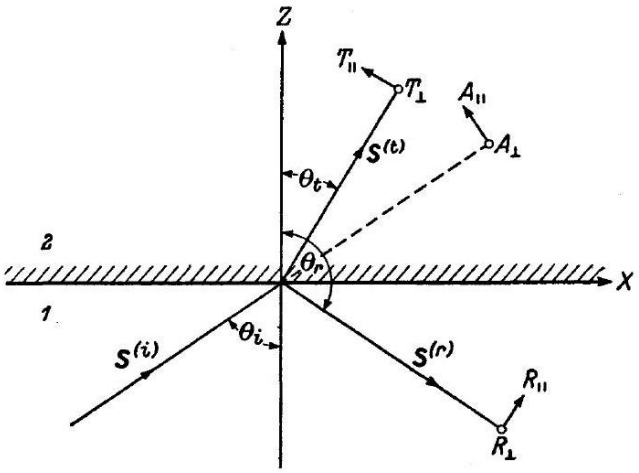



Fig. 1.10. Refraction and reflection of a plane wave. Plane of incidence.

When $n_2 > n_1$, then $n_{12} > 1$, and one says that the second medium is optically denser than the first medium. In this case, by (8),

$$\sin\theta_t = \frac{1}{n_{12}}\sin\theta_i < \sin\theta_i, \tag{9}$$

so that there is a real angle $\theta_t$ of refraction for every angle of incidence. If, however, the second medium is optically less dense than the first medium (i.e. if $n_{12} < 1$ ), one obtains a real value for $\theta_t$ only for those incident angles $\theta_i$ for which $\sin\theta_i \leqslant n_{12}$. For larger values of the angle of incidence, so-called total reflection takes place. It will be considered separately in § 1.5.4.

### 1.5.2 Fresnel formulae

Next we consider the amplitudes of the reflected and the transmitted waves. We shall assume that the two (homogeneous and isotropic) media are both of zero conductivity and consequently perfectly transparent; their magnetic permeabilities will then in fact differ from unity by negligible amounts, and accordingly we take $\mu_1 = \mu_2 = 1$.

0

- The suffixes $i, r$ and $t$ refer throughout to the incident, reflected and transmitted (refracted) waves respectively.

Let $A$ be the amplitude of the electric vector of the incident field. We take $A$ to be complex, with its phase equal to the constant part of the argument of the wave function; the variable part is

$$\tau_i = \omega\left(t - \frac{r \cdot s^{(i)}}{v_1}\right) = \omega\left(t - \frac{x\sin\theta_i + z\cos\theta_i}{v_1}\right). \tag{10}$$

We resolve each vector into components parallel (denoted by subscript $\parallel$ ) and perpendicular (subscript $\perp$ ) to the plane of incidence. The choice of the positive directions for the parallel components is indicated in Fig. 1.10. The perpendicular components must be visualized at right angles to the plane of the figure.

The components of the electric vector of the incident field then are

$$E_x^{(i)} = -A_{11}\cos\theta_2 e^{-i\tau_i}, \quad E_y^{(i)} = A_\perp e^{-i\tau_i}, \quad E_z^{(i)} = A_{11}\sin\theta_i e^{-i\tau_i} \tag{11}$$

The components of the magnetic vector are immediately obtained by using the relation 1.4 (4) (with $\mu = 1$ ):

$$H = \sqrt{\varepsilon}s \wedge E \tag{12}$$

This gives

$$H_x{}^{(i)} = -A_\perp\cos\theta_i\sqrt{\varepsilon_1}e^{-i\tau_i}, H_y{}^{(i)} = -A_{11}\sqrt{\varepsilon_1}e^{-i\tau_i}, H_z{}^{(i)} = A_\perp\sin\theta_i\sqrt{\varepsilon_1}e^{-i\tau_i} \tag{13}$$

Similarly if $T$ and $R$ are the complex amplitudes of the transmitted and reflected waves, the corresponding components of the electric and magnetic vectors are:

Transmitted field

$$\left.\begin{array}{lll} E_x{}^{(t)} = -T_1\cos\theta_t e^{-i\tau_t}, & E_y{}^{(t)} = T_\perp e^{-i\tau_t}, & E_z{}^{(t)} = T_{11}\sin\theta_t e^{-i\tau_t}, \\ H_x{}^{(t)} = -T_\perp\cos\theta_t\sqrt{\varepsilon_2}e^{-i\tau_t}, & H_y{}^{(t)} = -T_{11}\sqrt{\varepsilon_2}e^{-i\tau_t}, & H_z{}^{(t)} = T_\perp\sin\theta_t\sqrt{\varepsilon_2}e^{-i\tau_t}, \end{array}\right\} \tag{14}$$

with

$$\tau_t = \omega\left(t - \frac{r \cdot s^{(t)}}{v_2}\right) = \omega\left(t - \frac{x\sin\theta_t + z\cos\theta_t}{v_2}\right) \tag{15}$$

Reflected field

$$\left.\begin{array}{lll} E_x{}^{(r)} = -R_{11}\cos\theta_r e^{-i\tau_r}, & E_y{}^{(r)} = R_\perp e^{-i\tau_r}, & E_z{}^{(r)} = R_{11}\sin\theta_r e^{-i\tau_r}, \\ H_x{}^{(r)} = -R_\perp\cos\theta_r\sqrt{\varepsilon_1}e^{-i\tau_r}, & H_y{}^{(r)} = -R_{11}\sqrt{\varepsilon_1}e^{-i\tau_r}, & H_z{}^{(r)} = R_\perp\sin\theta_r\sqrt{\varepsilon_1}e^{-i\tau_r}, \end{array}\right\} \tag{16}$$

with

$$\tau_r = \omega\left(t - \frac{r \cdot s^{(r)}}{v_1}\right) = \omega\left(t - \frac{x\sin\theta_r + z\cos\theta_r}{v_1}\right) \tag{17}$$

The boundary conditions § 1.1 (23), (25) demand that across the boundary the tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$ should be continuous. Hence we must have

$$\left.\begin{array}{ll} E_x^{(i)} + E_x^{(r)} = E_x^{(t)}, & E_y^{(i)} + E_y^{(r)} = E_y^{(i)}, \\ H_x^{(i)} + H_x^{(r)} = H_x^{(i)}, & H_y^{(i)} + H_y^{(r)} = H_y^{(i)}, \end{array}\right\} \tag{18}$$

the conditions § 1.1 (15), and (19), being then automatically fulfilled for the normal components of $B$ and $D$. On substituting into (18) for all the components, and using the fact that $\cos\theta_r = \cos(\pi - \theta_i) = -\cos\theta_i$, we obtain the four relations

$$\left.\begin{aligned} \cos\theta_i\,(A_1 - R_1) &= \cos\theta_t T_1 \\ A_\perp + R_\perp &= T_\perp \\ \sqrt{\varepsilon_1}\cos\theta_i\,(A_\perp - R_\perp) &= \sqrt{\varepsilon_2}\cos\theta_t T_\perp, \\ \sqrt{\varepsilon_1}\,(A_1 + R_1) &= \sqrt{\varepsilon_2} T_1. \end{aligned}\right\} \tag{19}$$

We note that the equations (19) fall into two groups, one of which contains only the components parallel to the plane of incidence, whilst the other contains only those which are perpendicular to the plane of incidence, These two kinds of waves are, therefore, independent of one another.

We can solve (19) for the components of the reflected and transmitted waves in terms of those of the incident wave, giving (using again the Maxwell relation $n = \sqrt{\varepsilon}$ )

$$\left.\begin{aligned} T_\parallel &= \tfrac{2n_1\cos\theta_i}{n_2\cos\theta_i + n_1\cos\theta_t} A_\parallel \\ T_\perp &= \tfrac{2n_1\cos\theta_i}{n_1\cos\theta_i + n_2\cos\theta_t} A_\perp \\ R_\parallel &= \tfrac{n_2\cos\theta_i - n_1\cos\theta_t}{n_2\cos\theta_i + n_1\cos\theta_t} A_\parallel \\ R_\perp &= \tfrac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t} A_\perp \end{aligned}\right\} \tag{21}$$

Eqs. (20) and (21) are called Fresnel formulae, having first been derived in a slightly less general form by Fresnel in 1823, on the basis of his elastic theory of light. They are usually written in the following alternative form, which may be obtained from (20) and (21) by using the law of refraction (8):

$$\left.\begin{aligned} T_\mathrm{l} &= \tfrac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i+\theta_t)\cos(\theta_i-\theta_t)} A_\mathrm{l}, \\ T_\perp &= \tfrac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i+\theta_t)} A_\perp \\ R_\mathrm{lt} &= \tfrac{\tan(\theta_i-\theta_t)}{\tan(\theta_i+\theta_t)} A_\mathrm{l} \\ R_\perp &= -\tfrac{\sin(\theta_i-\theta_t)}{\sin(\theta_i+\theta_t)} A_\perp \end{aligned}\right\} \tag{21a}$$

Since $\theta_i$ and $\theta_t$ are real (the case of total reflection being excluded for the present), the trigonometrical factors on the right-hand side of (20a) and (21a) will also be real. Consequently the phase of each component of the reflected or transmitted wave is either equal to the phase of the corresponding component of the incident wave or differs from it by $\pi$. Since $T_\mathrm{l}$ and $T_\perp$ have the same signs as $A_\parallel$and $A_\perp$, the phase of the transmitted wave is actually equal to that of the incident wave. In the case of the reflected wave, the phase will, however, depend on the relative magnitudes of $\theta_i$ and $\theta_i$. For, if the second medium is optically denser than the first ( $\varepsilon_2 > \varepsilon_1$ ), then $\theta_t < \theta_i$; according to (21), the signs of $R_\perp$ and $A_\perp$ are different and the phases therefore differ* by $\pi$. Under the same circumstances $\tan(\theta_i - \theta_t)$ is positive, but the denominator $\tan(\theta_i + \theta_t)$ becomes negative for $\theta_i + \theta_t > \pi/2$, and the phase of $R_\parallel$and $A_\mathrm{l}$ then differ by $\pi$. Similar considerations apply when the second medium is optically less dense than the first.

For normal incidence, $\theta_i = 0$ and consequently $\theta_t = 0$, and (20) and (21) reduce to

- From (11) and (16) it follows that in the plane $z = 0$

$$\frac{E_y^{(r)}}{E_y^{(i)}} = \frac{R_\perp}{A_\perp}, \quad \frac{E_x^{(i)}}{E_x^{(i)}} = -\frac{R_\parallel}{A_\parallel}.$$

This result implies, that in the plane $z = 0$, the phases of ${E_y}^{(r)}$ and ${E_y}^{(i)}$ differ by $\pi$, whereas the phases of ${E_x}^{(r)}$ and ${E_x}^{(i)}$ are equal to each other. This difference in the behaviour of the phases of the $y$-and $x$-components is rather formal, arising from the way in which the angle of reflection $\theta_r$ was defined (see Fig. 1.10).

$$\left.\begin{array}{rl} T_{\parallel} & = \frac{2}{n+1} A_{\parallel} \\ T_{\perp} & = \frac{2}{n+1} A_{\perp} \end{array}\right\} \tag{23}$$

where $n = n_2/n_1$.

### 1.5.3 The reflectivity and transmissivity; polarization on reflection and refraction

We shall now examine how the energy of the incident field is divided between the two secondary fields.

According to § 1.4 (8), the light intensity is given (again assuming $\mu = 1$ ) by

$$S = \frac{c}{4\pi}\sqrt{\varepsilon}E^2 = \frac{cn}{4\pi}E^2 \tag{24}$$

The amount of energy in the primary wave which is incident on a unit area of the boundary per second is therefore

$$J^{(i)} = S^{(i)} \cos\theta_i = \frac{cn_1}{4\pi}|A|^2 \cos\theta_i \tag{25}$$

and the energies of the reflected and transmitted wave leaving a unit area of the boundary per second are given by similar expressions:

$$\left.\begin{array}{l} J^{(r)} = S^{(r)} \cos\theta_i = \frac{cn_1}{4\pi}|R|^2 \cos\theta_i, \\ J^{(t)} = S^{(t)} \cos\theta_t = \frac{cn_2}{4\pi}|T|^2 \cos\theta_t. \end{array}\right\} \tag{26}$$

The ratios

$$\mathscr{R} = \frac{J^{(r)}}{J^{(i)}} = \frac{|R|^2}{|A|^2} \quad \text{and} \quad \mathscr{T} = \frac{J^{(t)}}{J^{(i)}} = \frac{n_2}{n_1}\frac{\cos\theta_t}{\cos\theta_i}\frac{|T|^2}{|A|^2} \tag{27}$$

are called the reflectivity and transmissivity respectively.* It can easily be verified that, in agreement with the law of conservation of energy,

$$\mathscr{R} + \mathscr{T} = 1. \tag{28}$$

The reflectivity and transmissivity depend on the polarization of the incident wave. They may be expressed in terms of the reflectivity and transmissivity associated with polarizations in the parallel and perpendicular directions respectively.

Let $\alpha_i$ be the angle which the $\boldsymbol{E}$ vector of the incident wave makes with the plane of incidence. Then

$$A_{\text{ll}} = A\cos\alpha_i, \quad A_{\perp} = A\sin\alpha_i \tag{29}$$

Let

$$\left.\begin{array}{l} J_i^{(i)} = \frac{cn_1}{4\pi}\left|A_i\right|^2 \cos\theta_i = J^{(i)}\cos^2\alpha_i, \\ J_{\perp}^{(i)} = \frac{cn_1}{4\pi}\left|A_{\perp}\right|^2 \cos\theta_i = J^{(i)}\sin^2\alpha_i, \end{array}\right\} \tag{30}$$

and

0

- If $\mu \neq 1$, the factor $\frac{n_2}{n_1}$ in the expression for $\mathscr{T}$ must be replaced by $\sqrt{\frac{\varepsilon_2}{\mu_2}}/\sqrt{\frac{\varepsilon_1}{\mu_1}}$, as is immediately evident from § 1.4 (8).

$$\left.\begin{array}{l} J_1^{(r)} = \frac{cn_1}{4\pi} |R_1|^2 \cos\theta_i, \\ J_\perp^{(r)} = \frac{cn_1}{4\pi} |R_\perp|^2 \cos\theta_i. \end{array}\right\} \tag{31}$$

Then

$$\begin{aligned} \mathscr{R} = \frac{J^{(r)}}{J^{(i)}} &= \frac{J_1^{(r)} + J_\perp^{(r)}}{J^{(i)}} = \frac{J_1^{(r)}}{J_1^{(i)}} \cos^2\alpha_i + \frac{J_\perp^{(r)}}{J_\perp^{(i)}} \sin^2\alpha_i \\ &= \mathscr{R}_1 \cos^2\alpha_i + \mathscr{R}_\perp \sin^2\alpha_i \end{aligned} \tag{32}$$

where

$$\left.\begin{array}{l} \mathscr{R}_1 = \frac{J_\parallel^{(r)}}{J_\parallel^{(i)}} = \frac{\left|R_\parallel\right|^2}{\left|A_\parallel\right|^2} = \frac{\tan^2(\theta_i-\theta_t)}{\tan^2(\theta_i+\theta_t)} \\ \mathscr{R}_\perp = \frac{J_\perp^{(r)}}{J_\perp^{(i)}} = \frac{|R_\perp|^2}{|A_\perp|^2} = \frac{\sin^2(\theta_i-\theta_t)}{\sin^2(\theta_i+\theta_t)} \end{array}\right\} \tag{33}$$

In a similar way, we obtain

$$\mathscr{T} = \frac{J^{(t)}}{J^{(i)}} = \mathscr{T}_1 \cos^2\alpha_i + \mathscr{T}_\perp \sin^2\alpha_i, \tag{34}$$

with

$$\left.\begin{array}{l} \mathscr{T}_\parallel = \frac{J_\parallel{}^{(t)}}{J_\parallel{}^{(i)}} = \frac{n_2}{n_1}\frac{\cos\theta_t}{\cos\theta_i}\frac{|T_?|^2}{|A_{\text{il}}|^2} = \frac{\sin 2\theta_i \sin 2\theta_t}{\sin^2(\theta_i+\theta_t)\cos^2(\theta_i-\theta_t)}, \\ \mathscr{T}_\perp = \frac{J_\perp{}^{(t)}}{J_\perp{}^{(i)}} = \frac{n_2}{n_1}\frac{\cos\theta_t}{\cos\theta_i}\frac{|T_\perp|^2}{|A_\perp|^2} = \frac{\sin 2\theta_i \sin 2\theta_t}{\sin^2(\theta_i+\theta_t)}. \end{array}\right\} \tag{35}$$

Again we may verify that

$$\mathscr{R}_\parallel + \mathscr{T}_\parallel = 1, \quad \mathscr{R}_\perp + \mathscr{T}_\perp = 1. \tag{36}$$

For normal incidence the distinction between the parallel and perpendicular components disappears, and one has from (22), (23) and (27)

$$\left.\begin{array}{l} \mathscr{R} = \left(\frac{n-1}{n+1}\right)^2 \\ \mathscr{T} = \frac{4n}{(n+1)^2} \end{array}\right\} \tag{37}$$

It is seen from (37) that

$$\lim_{n\to 1} \mathscr{R} = 0, \quad \lim_{n\to 1} \mathscr{T} = 1 \tag{38}$$

Similar results also hold for the limiting values of $\mathscr{R}_\sharp$, $\mathscr{T}_\sharp$, and $\mathscr{R}_\perp$, $\mathscr{T}_\perp$, as can easily be seen from (33) and (35), making use of the fact that, according to the law of refraction, $\theta_i \to \theta_i$ as $n \to 1$. Hence the smaller the difference in the optical densities of the two media, the less energy is carried away by the reflected wave.

The denominators in (33) and (35) are finite, except when $\theta_i + \theta_t = \pi/2$. Then $\tan(\theta_i + \theta_t) = \infty$ and consequently $\mathscr{R}_1 = 0$. In this case (see Fig. 1.11) the reflected and transmitted rays are perpendicular to each other, and it follows from the law of refraction (since now $\sin\theta_t = \sin\{(\pi/2) - \theta_i\} = \cos\theta_i$ ) that

$$\tan\theta_i = n. \tag{39}$$

The angle $\theta_i$ given by (39) is called the polarizing or Brewster angle; its significance was noted first in 1815 by David Brewster (1781-1868): If light is incident under this angle, the electric vector of the reflected light has no component in the plane of incidence.

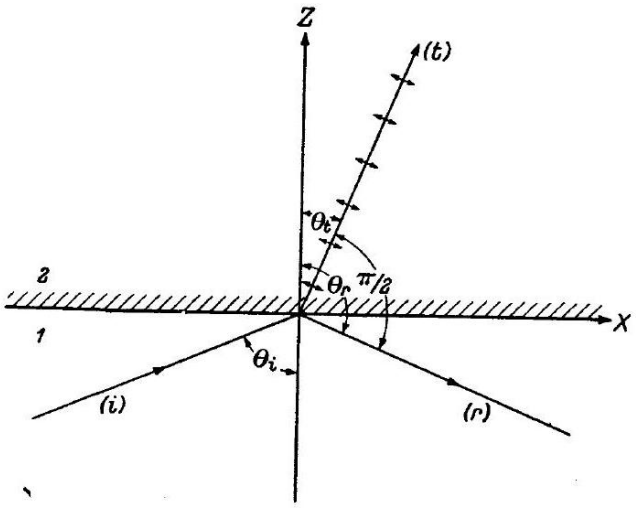



Fig. 1.11. Illustrating the polarizing (Brewster's) angle.

One usually says that the light is then polarized "in the plane of incidence". According to this traditional terminology the plane of polarization is therefore the plane which contains the magnetic vector and the direction of propagation. For reasons already explained in § 1.4.2 it is, however, better not to use this term.

The above result, often called Brewster's law, can also be explained by the following, more direct argument: The incident field gives rise to vibrations of electrons in the atoms of the second medium. These vibrations are in the direction of the electric vector of the transmitted wave. The vibrating electrons give rise to another wave, the reflected wave, which will be propagated back into the first medium. Now a linearly vibrating electron radiates transversally (cf. §2.2.3) so that there is no flux of radiant energy in the direction of vibration, and it follows that when the reflected and transmitted rays are at right angles to each other, the reflected ray does not receive any energy for oscillations in the plane of incidence.

In Fig. 1.12 the reflectivity for glass of refractive index 1.52 are plotted against the angle of incidence $\theta_i$. The numbers along the upper horizontal refer to the angle of refraction $\theta_t$. The zero value of the curve (a) (for $\mathscr{R}_\sharp$ ) corresponds to the polarizing angle $\tan^{-1} 1.52 = 56°40'$.

The refractive indices with respect to air are usually of the order of 1.5 at optical wavelengths, but at radio wavelengths they are much larger, there being a corresponding increase in the polarizing angle. For example, at optical wavelengths the refractive index of water is about 1.3 and the polarizing angle 53°. For radio wavelengths its value is about 9 , the polarizing angle then being in the neighbourhood of 84°.

According to (32), the curve (b) in Fig. 1.12 is seen to correspond to $\alpha = 45°$. The same curve, as will now be shown, represents also the reflectivity $\overline{\mathscr{R}}$ for natural light, i.e. for light obtained from a body which is made to glow, by raising its temperature. The directions of vibration of natural light vary rapidly in a random, irregular manner. The corresponding reflectivity $\overline{\mathscr{R}}$ may be obtained by averaging over all

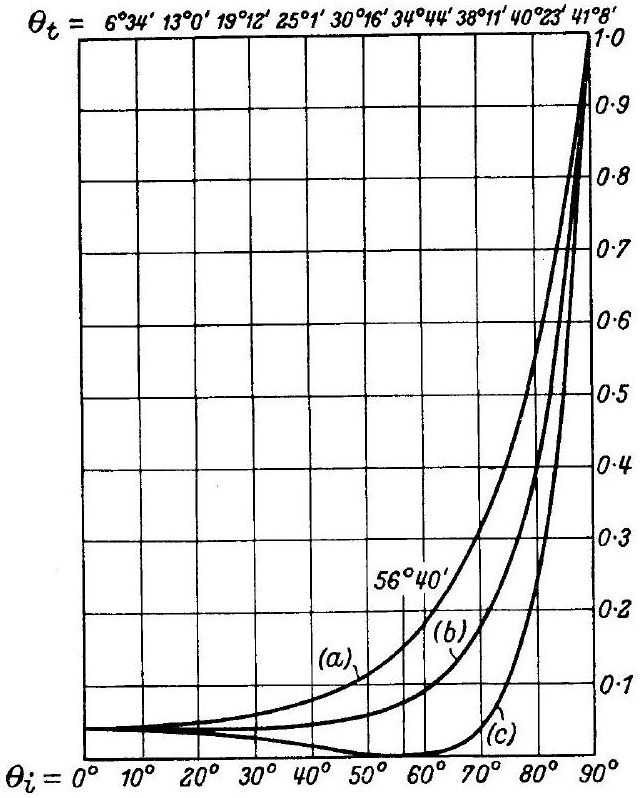



Fig. 1.12. Intensity of reflected light as a function of the angle of incidence.
(a) $\mathscr{R}_{\perp}$; (b) $\frac{1}{2}\left(\mathscr{R}_{\parallel}+\mathscr{R}_{\perp}\right)$; (c) $\mathscr{R}_{\parallel}$.

(After O. D. Chwolson, Lehrb. d. Physik (Braunschweig, Vieweg), Bd. II. 2, 2 Aufl. (1922), 716.)

directions. Since the averages of $\sin^2\alpha$ and $\cos^2\alpha$ are $\frac{1}{2}$, we obtain for the average values of ${J_{\parallel}}^{(i)}$ and ${J_{\perp}}^{(i)}$ the relations

$$\bar{J}_{\parallel}^{(i)} = \bar{J}_{\perp}^{(i)} = \frac{1}{2}J^{(i)}. \tag{40}$$

For the reflected light, however, the two components will in general differ from each other. For, using (40), one has

$$\left.\begin{aligned} \bar{J}_{\parallel}^{(r)} &= \tfrac{1}{2}\frac{\bar{J}_{\parallel}^{(r)}}{\bar{J}_{\parallel}^{(i)}}J^{(i)} = \tfrac{1}{2}\mathscr{R}_{\parallel}J^{(i)}, \\ \bar{J}_{\perp}^{(r)} &= \tfrac{1}{2}\frac{\bar{J}_{\perp}^{(r)}}{\bar{J}_{\perp}^{(i)}}J^{(i)} = \tfrac{1}{2}\mathscr{R}_{\perp}J^{(i)}. \end{aligned}\right\} \tag{41}$$

The reflected light is then said to be partially polarized and its degree of polarization $P$ may be defined to be*

$$P = \left|\frac{\mathscr{R}_{\parallel}-\mathscr{R}_{\perp}}{\mathscr{R}_{\parallel}+\mathscr{R}_{\perp}}\right|. \tag{42}$$

The reflectivity $\overline{\mathscr{R}}$ is now given by

0

- A more general definition of the degree of polarization is given in § 10.8.2, where its physical significance is also discussed.

$$\overline{\mathscr{R}} = \frac{\bar{J}^{(r)}}{\bar{J}^{(i)}} = \frac{\bar{J}_{\parallel}^{(r)} + \bar{J}_{\perp}^{(r)}}{\bar{J}^{(i)}} = \frac{1}{2}\left(\mathscr{R}_{\parallel} + \mathscr{R}_{\perp}\right) \tag{43}$$

and is therefore again represented by the curve (b) in Fig. 1.12. The degree of polarization may now be expressed in the form

$$P = \frac{1}{\overline{\mathscr{R}}}\frac{1}{2}\left\{\left|\mathscr{R}_{\parallel} - \mathscr{R}_{\perp}\right|\right\};$$

the quantity in the brace brackets is sometimes called the polarized proportion of the light reflected.

Similar results hold for the transmitted light. We also have for natural light

$$\overline{\mathscr{R}} + \overline{\mathscr{T}} = 1 \tag{44}$$

Returning to the case where the incident light is linearly polarized we see that this is true also for the reflected and for the transmitted light, since the phases change only by 0 or $\pi$. The directions of vibration in the reflected and transmitted light as compared with the incident light are, however, turned in opposite directions. This can be seen from the following:

The angle which we denoted by $\alpha$, i.e. the angle between the plane of vibration and the plane of incidence, may be called the azimuth of the vibration, and we shall regard it as positive when the plane of vibration turns clockwise around the direction of propagation (Fig. 1.13). It may be assumed that the azimuthal angle is in the range $-\pi/2$ to $\pi/2$. We have for the incident, reflected and transmitted electric wave

$$\tan\alpha_i = \frac{A_{\perp}}{A_{\parallel}}, \quad \tan\alpha_r = \frac{R_{\perp}}{R_{\parallel}}, \quad \tan\alpha_t = \frac{T_{\perp}}{T_{\parallel}} \tag{45}$$

Using the Fresnel formulae (20) and (21),

$$\tan\alpha_r = -\frac{\cos\left(\theta_i - \theta_t\right)}{\cos\left(\theta_i + \theta_t\right)}\tan\alpha_i \tag{46}$$

$$\tan\alpha_t = \cos\left(\theta_i - \theta_t\right)\tan\alpha_i \tag{47}$$

Since $0 \leqslant \theta_i \leqslant \pi/2, \quad 0 < \theta_t < \pi/2,$

$$|\tan\alpha_r| \geqslant |\tan\alpha_i| \tag{48}$$

$$|\tan\alpha_t| \leqslant |\tan\alpha_i| \tag{49}$$

In (48) the equality sign holds only for normal or tangential incidence $\theta_i = \theta_t = 0$ or $\theta_i = \pi/2$ ); in (49) it holds only for normal incidence. The two inequalities imply that on reflection the plane of vibration is turned away from the plane of incidence, whereas on refraction it is turned towards it. In Fig. 1.14 the behaviour of $\alpha_r$ and $\alpha_t$ is illustrated for $n = 1.52$ and $\alpha_i = 45°$. We see that, when $\theta_i$ is equal to the polarizing angle $56°40'$, $\alpha_r = 90°$. In fact, according to (46), $\tan\alpha_r = \infty$ (i.e. $\alpha_r = \pi/2$ ), for $\theta_i + \theta_t = \pi/2$, whatever the value of $\alpha_i$ may be.

It follows from Brewster's law that polarized light may be produced simply by allowing reflection to take place at the polarizing angle. One of the oldest instruments based on this principle is the so-called Nörrenberg's Reflecting Polariscope (after

Nörrenberg, 1787-1862). It consists essentially of two glassplates (Fig. 1.15) which the rays meet at the polarizing angle. The first plate plays the role of a polarizer, i.e. of an instrument producing linearly polarized light from unpolarized light; the second plays the role

of an analyzer, i.e. an arrangement which detects linearly polarized light. This instrument has however several disadvantages, chiefly that the fraction of the light reflected at the polarizing angle is comparatively small, and the

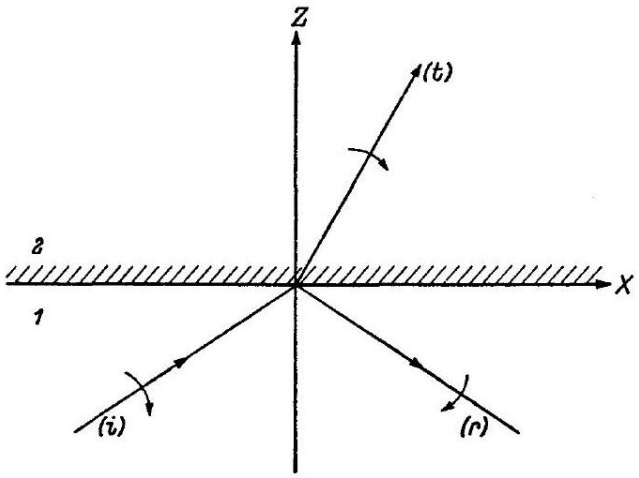



Fig. 1.13. Illustrating the signs of the azimuthal angles.

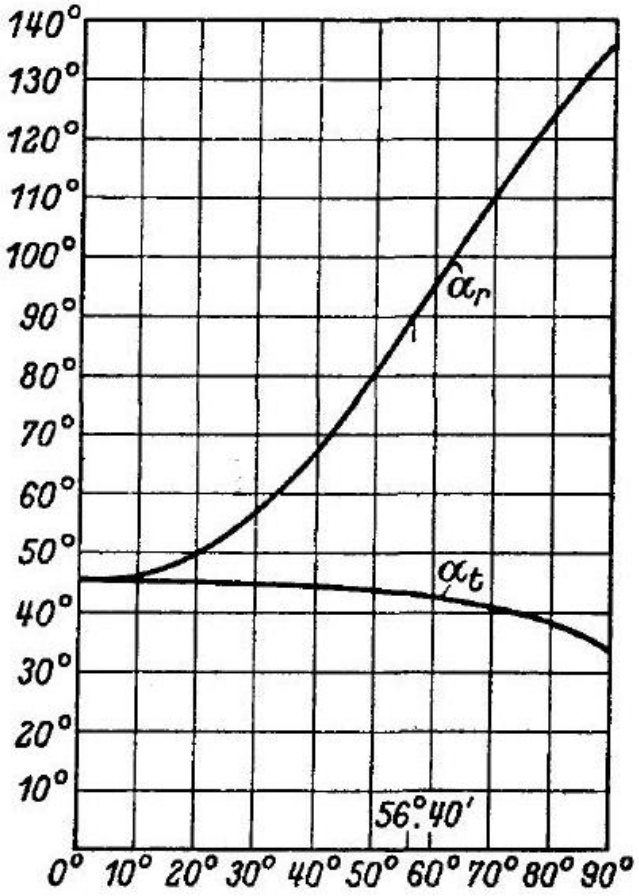



Fig. 1.14. Azimuthal angles as functions of the angle of incidence.
(After O. D. Chwolson, Lehrb. d. Physik, (Braunschweig, Vieweg) Bd. II. 2, 2 Aufl. (1922), 716.)

path of the ray through the instrument is rather complicated. It is preferable to employ

instruments which polarize the incident light without changing its direction of propagation. This can be done, for example, by using a pile of thin plane-parallel plates. If a beam of unpolarized light is incident on the pile, it is partially polarized on each refraction, and one can achieve a reasonably high degree of polarization even

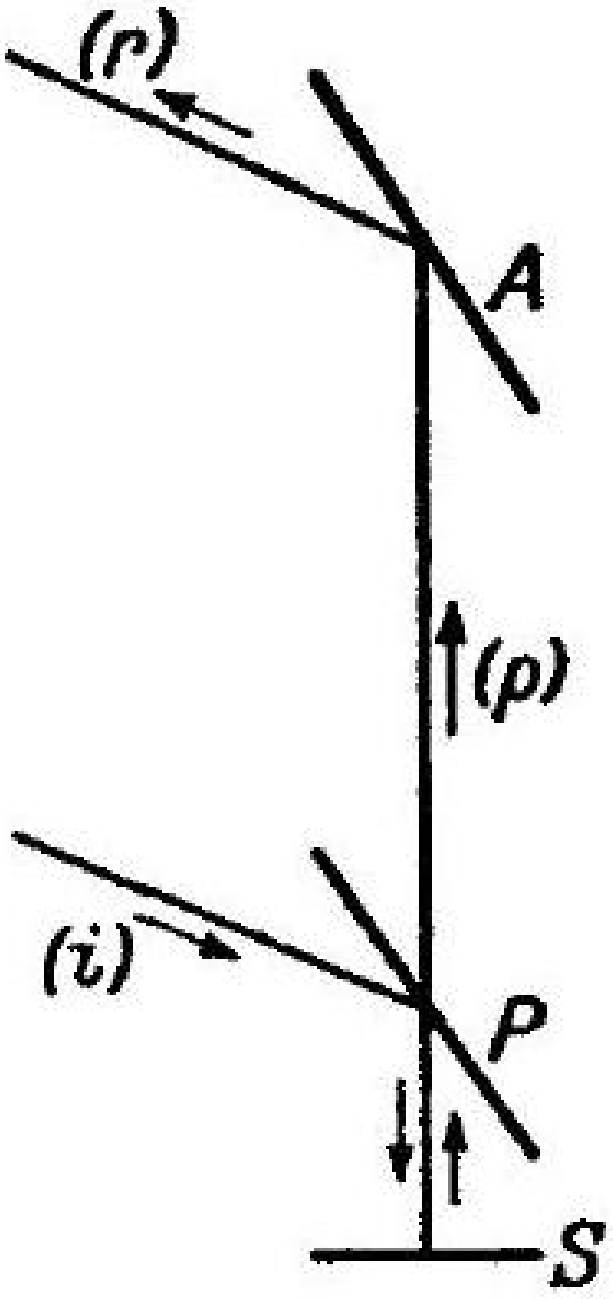



Fig. 1.15. Principle of the Nörrenberg's Reflecting Polariscope.

$P =$ polarizing glass plate
$S =$ reflecting mirror
$A =$ analyser
$i =$ incident beam
$p =$ polarized beam
$r =$ beam reflected from $A$.

with a small number of plates. After passing through both boundaries of a plate the intensities of the two components will be in the ratio

$$\left(\frac{\mathscr{I}_{\frac{1}{2}}}{\mathscr{I}_1}\right)^2 = \cos^4(\theta_i - \theta_t) < . \tag{50}$$

a result obtained by applying (35) twice in succession. This shows that, on emerging from the plate, the parallel component is stronger than the perpendicular component, the degree of polarization being greater, the greater $\theta_i$ is. If $\theta_i$ is equal to the polarizing angle, $\theta_i + \theta_t = \pi/2, \tan\theta_i = n$, and we then have

$$\left(\frac{\mathscr{I}_\perp}{\mathscr{I}_u}\right)^2 = \sin^4 2\theta_i = \left(\frac{2n}{1+n^2}\right)^4. \tag{51}$$

For $n = 1.5$ this expression has the value 0.73 . Hence, if the light passes through five plates, for example, we obtain the ratio $0.73^5 \simeq 0.2$.

In the past, polarized light was as a rule produced by double refraction in crystals of calcite or quartz, as described in § 14.4.1. To-day the most convenient method is by the use of so-called polaroid films. Their action is based on a property known as dichroism. By this we mean the property shown by certain materials of having different absorption coefficients for light polarized in different directions. Polyvinyl alcohol films impregnated by iodine, for example, can be made which transmit nearly 80 per cent of light polarized in one plane, and less than 1 per cent of light polarized at right angles to this plane. The theory of this effect will be briefly discussed in § 14.6.3.

### 1.5.4 Total reflection

So far we have excluded the case when the law of refraction

$$\sin\theta_t = \frac{\sin\theta_i}{n_{12}} \tag{52}$$

does not give a real value for the angle of refraction $\theta_t$. We shall now examine this case. It occurs when light is propagated from an optically denser medium into one which is optically less dense, i.e. when

$$n_{12} = \frac{n_2}{n_1} = \sqrt{\frac{\varepsilon_2\mu_2}{\varepsilon_1\mu_1}} < 1$$

provided that the angle of incidence $\theta_i$ exceeds the critical value $\bar{\theta}_i$ given by

$$\sin\bar{\theta}_i = n_{12} \tag{53}$$

When $\theta_i = \bar{\theta}_i, \sin\theta_t = 1$, i.e. $\theta_t = 90°$, so that the light emerges in a direction tangent to the boundary. If $\theta_i$ exceeds the limiting value $\bar{\theta}_i$, no light enters the second medium. All the incident light is then reflected back into the first medium and we speak of total reflection.

Nevertheless the electromagnetic field in the second medium does not disappear, only there is no longer a flow of energy across the boundary. For, if (omitting the subscript 12 on $n_{12}$ ) we set

$$\sin\theta_t = \frac{\sin\theta_i}{n}, \quad \cos\theta_t = \pm i\sqrt{\frac{\sin^2\theta_i}{n^2} - 1} \tag{54}$$

in the phase factor (15) of the transmitted wave, we have

$$e^{-i\tau_t} = e^{-i\omega\left(t-\frac{x\sin\theta_i}{nv_2}\right)} e^{\mp\frac{\omega z}{v_2}\sqrt{\frac{\sin^2\theta_i}{n_2}-1}}. \tag{55}$$

(55) represents a non-homogeneous wave which is propagated along the boundary in the plane of incidence (i.e. in the $x$ direction), and which varies exponentially with the distance $z$ from the boundary. Naturally only the negative sign in front of the square root in (55) corresponds to the physical situation, since otherwise the amplitude would tend to infinity with increasing distance. The amplitude is seen to decrease very rapidly with the depth $z$ of penetration, the effective depth of penetration being of the order of $v_2/\omega = \lambda_2/2\pi$, i.e. of the order of a wavelength. The wave is not transversal since, as will be shown below, the component of the electric vector in the direction of propagation does not vanish.

Experimental verification of the disturbance in the second (less dense) medium is somewhat troublesome, since any arrangement used for its detection will perturb the boundary conditions. A rough confirmation may be obtained by bringing up a second refracting medium within about a quarter of a wavelength of the interface at which total reflection is taking place, and observing the entry of the radiation into the second medium.*

To apply the Fresnel formulae (21a) to the case of total reflection, we rewrite them in the form

$$\left.\begin{aligned} R_{\parallel} &= \tfrac{\sin\theta_i\cos\theta_i-\sin\theta_t\cos\theta_t}{\sin\theta_i\cos\theta_i+\sin\theta_t\cos\theta_t}A_{\parallel} \\ R_{\perp} &= -\tfrac{\sin\theta_i\cos\theta_t-\sin\theta_t\cos\theta_i}{\sin\theta_i\cos\theta_t+\sin\theta_t\cos\theta_i}A_{\perp} \end{aligned}\right\} \tag{56}$$

and substitute into these expressions from (54), remembering that the upper sign is to be taken in front of the square root. We then obtain

$$\left.\begin{aligned} R_{\parallel} &= \tfrac{n^2\cos\theta_i-i\sqrt{\sin^2\theta_i-n^2}}{n^2\cos\theta_i+i\sqrt{\sin^2\theta_i-n^2}}A_{\parallel} \\ R_{\perp} &= \tfrac{\cos\theta_i-i\sqrt{\sin^2\theta_i-n^2}}{\cos\theta_i+i\sqrt{\sin^2\theta_i-n^2}}A_{\perp} \end{aligned}\right\} \tag{57}$$

Hence

$$\left|R_{\parallel}\right| = \left|A_{\parallel}\right|, \quad \left|R_{\perp}\right| = \left|A_{\perp}\right|, \tag{58}$$

i.e. for each component, the intensity of the light which is totally reflected is equal to the intensity of the incident light.

Although there is a field in the second medium, it is easy to see that no energy flows across the boundary. More precisely it will be shown that, although the component of the Poynting vector in the direction normal to the boundary is in general finite, its time average vanishes; this implies that the energy flows to and fro, but that there is no lasting flow into the second medium.

We write down the $x$ and $y$ components of the transmitted field, for $z = 0$, and make use of (54). (Real expressions must now be used for $\boldsymbol{E}$ and $\boldsymbol{H}$ since the energy flow is a quadratic function of the components). Denoting the conjugate of a complex quantity by an asterisk, we have from (14),

$$\begin{aligned} E_x{}^{(t)} &= -\frac{1}{2}\left[T_{\parallel}\cos\theta_t e^{-i\tau_t{}^o} + T_{\parallel}{}^{\star}\cos^{\star}\theta_t e^{+i\tau_t{}^o}\right] \\ &= -\frac{i}{2}\sqrt{\frac{\sin^2\theta_i}{n^2}-1}\left[T_{\parallel}e^{-i\tau_t{}^o} - T_{\parallel}{}^{\star}e^{+i\tau_t{}^o}\right] \\ E_y{}^{(t)} &= \frac{1}{2}\left[T_{\perp}e^{-i\tau_t{}^o} + T_{\perp}{}^{\star}e^{+i\tau_t{}^o}\right] \end{aligned}$$

0

$$H_x^{(t)} = -\frac{1}{2}\left[T_\perp \cos\theta_t \sqrt{\varepsilon_2} e^{-i\tau_t^0} + T_\perp^\star \cos^\star \theta_t \sqrt{\varepsilon_2} e^{+i\tau_t^0}\right]$$
$$= -\frac{i}{2}\sqrt{\varepsilon_2}\sqrt{\frac{\sin^2\theta_i}{n^2} - 1}\left[T_\perp e^{-i\tau_t^0 - T_\perp{}^\star e^{+i\tau_t^0}}\right]$$
$$H_y{}^{(t)} = -\frac{1}{2}\sqrt{\varepsilon_2}\left[T_1 e^{-i\tau_t^0} + T_1{}^\star e^{+i\tau_t^0}\right]$$

where

$$\tau_t^o = \omega\left[t - \frac{x\sin\theta_i}{n v_2}\right]$$

If we now form the time average of

$$S_z^{(t)} = \frac{c}{4\pi}\left[E_x^{(t)} H_y^{(t)} - E_y^{(t)} H_x^{(t)}\right]$$

over an interval $-t' \leqslant t \leqslant t'$, where $t'$ is large compared with the period $\mathrm{T} = 2\pi/\omega$, then both terms disappear for $z = 0$; for one contains the factor

$$\frac{1}{2t'}\int_{-t'}^{t'}\left[T_1{}^2 e^{-2i\tau_t{}^0} - T_1{}^{\star 2} e^{+2i\tau_t{}^0}\right] dt = \left[T_1{}^2 e^{+\frac{2i\omega x \sin 0_i}{n v_2}} - T_1{}^{\star 2} e^{-\frac{2i\omega x \sin\theta_i}{n v_2}}\right] O\left(\frac{\mathrm{T}}{t'}\right),$$

which is negligibly small when $t' \gg \mathrm{T}$; the other contains a similar factor with $T_\perp$ in place of $T_n$.

On the other hand, if the other two components of $\boldsymbol{S}^{(t)}$ for $z = 0$ are calculated, namely $S_x{}^{(t)}$ and $S_y{}^{(t)}$, their time averaged values are in general found to be finite. Energy therefore does not penetrate into the second medium, but flows along the boundary in the plane of incidence.

The preceding analysis applies to a stationary state, and is based on the assumption that the boundary surface and the wave-fronts are of infinite extent. It does not explain how the energy initially entered the second medium. In an actual experiment, the incident wave will be bounded both in space and time*; at the beginning of the process a small amount of energy will penetrate into the second medium and will give rise to a field there.

Finally, we determine the changes in the phases of the components of the reflected and the incident wave. On account of (58) we may set

$$\frac{R_{\mathrm{II}}}{A_{\mathrm{I}}} = e^{i\delta_{\mathrm{I}}}, \quad \frac{R_\perp}{A_\perp} = e^{i\delta_\perp}. \tag{59}$$

Now according to (57), $\frac{R_{\mathrm{I}}}{A_{\mathrm{I}}}$ and $\frac{R_\perp}{A_\perp}$ are each of the form $z\,(z^\star)^{-1}$. Hence if $\alpha$ is the argument of $z$ (i.e. $z = ae^{i\alpha}$, with $a, \alpha$ both real), then

$$e^{i\delta} = z\,(z^\star)^{-1} = e^{2i\alpha}, \quad \text{i.e.} \quad \tan\frac{\delta}{2} = \tan\alpha$$

and therefore

$$\left.\begin{aligned} \tan\tfrac{\delta_{\mathrm{I}}}{2} &= -\frac{\sqrt{\sin^2\theta_i - n^2}}{n^2\cos\theta_i} \\ \tan\tfrac{\delta_\perp}{2} &= -\frac{\sqrt{\sin^2\theta_i - n^2}}{\cos\theta_i} \end{aligned}\right\} \tag{60}$$

• An elegant way of doing this was described by W. Culshaw and D. S. Jones, Proc. Phys. Soc., B, 66 (1954), 859, using electrically generated waves of wavelength 1.25 cm .

0

The two components are seen to undergo phase jumps of different amounts. Linearly polarized light will in consequence become elliptically polarized on total reflection.

One can also immediately write down an expression for the relative phase difference $\delta = \delta_\perp - \delta_n$ :

$$\tan\frac{\delta}{2} = \frac{\tan\frac{\delta_\perp}{2} - \tan\frac{\delta_n}{2}}{1 + \tan\frac{\delta_\perp}{2}\tan\frac{\delta_n}{2}} = \frac{\left(\frac{1}{n^2} - 1\right)\frac{\sqrt{\sin^2\theta_i - n^2}}{\cos\theta_i}}{1 + \frac{\sin^2\theta_i - n^2}{n^2\cos^2\theta_i}},$$

i.e.

$$\tan\frac{\delta}{2} = \frac{\cos\theta_i\sqrt{\sin^2\theta_i - n^2}}{\sin^2\theta_i}. \tag{61}$$

This expression vanishes for grazing incidence ( $\theta_i = \pi/2$ ), and for incidence at the critical angle $\bar{\theta}_i$ $\left(\sin\bar{\theta}_i = n\right)$. Between these two values there lies the maximum value of the relative phase difference; it is determined from the equation

$$\frac{d}{d\theta_i}(\tan\delta/2) = \frac{2n^2 - (1+n^2)\sin^2\theta_i}{\sin^3\theta_i\sqrt{\sin^2\theta_i - n^2}} = 0.$$

This is satisfied when

$$\sin^2\theta_i = \frac{2n^2}{1+n^2}. \tag{62}$$

On substituting from (62) into (61), we obtain for the maximum $\delta_m$ of the relative phase difference $\delta$, the expression

$$\tan\frac{\delta_m}{2} = \frac{1-n^2}{2n}. \tag{63}$$

From (61) it is seen that, with $n$ given, there are two values of the angle $\theta_i$ of incidence for every value of $\delta$.

The phase change which takes place on total reflection may be used (as shown already by Fresnel) to produce circularly polarized light from light which is linearly polarized. The amplitude components of the incident light are made equal ( $\left|A_\parallel\right| = |A_\perp|$ ), by taking the incident wave to be polarized in a direction which makes an angle of 45° with the normal to the plane of incidence (i.e. $\alpha_i = 45°$ ). Then, by (58), $\left|R_\parallel\right| = |R_\perp|$. Further, $n$ and $\theta_i$ are chosen in such a way that the relative phase difference $\delta$ is equal to 90°. To attain this value of $\delta$ by a single reflection, it would be necessary, according to (63), that

$$\tan\frac{\pi}{4} = 1 < \frac{1-n^2}{2n}$$

i.e.

$$n = n_{12} < \sqrt{2} - 1 = 0.414$$

This means that the refractive index $n_{21} = 1/n$ of the denser with respect to the less dense medium would have to be at least $2 \cdot 41$. This value is rather large, although there are non-absorbing substances whose refractive index comes close to, and even exceeds, this value. Fresnel made use of two total reflections on glass. When $n_{21} = 1.51$, one obtains, according to

• The total reflection of a beam of light of finite cross-section has been treated by J. Picht, Ann. d. Physik (5), 3 (1929), 433.

(62) and (63), a maximum relative phase difference $\delta_m = 45°56'$ when the angle of incidence $\theta_i$ equals to $51°20'$. It is therefore just possible to attain the value $\delta = 45°$, namely, with either of the following angles of incidence:

$$\theta_i = 48°37', \quad \theta_i = 54°37'$$

A phase difference of 90° may therefore be obtained by means of two successive total reflections at either of these angles. For this purpose a glass block is used, of the form shown in Fig. 1.16, known as Fresnel's rhomb.

Fresnel's rhomb may, of course, be also used to produce elliptically polarized light; the azimuth of the incident (linearly polarized) light must then be taken

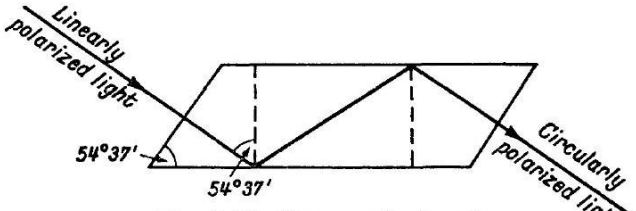



**Fig. 1.16. Fresnel's rhomb.**

different from 45°. One may also invert the procedure and produce, by means of Fresnel's rhomb, linearly polarized light from elliptically polarized light.

Measurement of the critical angle $\tilde{\theta}_i$ gives a convenient and accurate way of determining the index of refraction $n = \sin \bar{\theta}_i$. Instruments used for this purpose are called refractometers.

## 9.2.6 1.6 Wave Propagation in a Stratified Medium. Theory of Dielectric Films

A medium whose properties are constant throughout each plane perpendicular to a fixed direction is called a stratified medium. If the $z$-axis of a Cartesian reference system is taken along this special direction, then

$$\varepsilon = \varepsilon(z), \quad \mu = \mu(z). \tag{I}$$

We shall consider the propagation of a plane, time-harmonic electromagnetic wave through such a medium; this is a natural generalization of the simple case treated in the previous section.

The theory of stratified media is of considerable importance in optics, in connection with multilayers, i.e. a succession of thin plane-parallel films. Such films may be produced with the help of high-vacuum evaporation techniques, and their thickness may be controlled with very high accuracy. They have many useful applications. For example, as will be demonstrated later, they may be employed as antireflection films, i.e. as coatings which reduce the reflectivity of a given surface. On the other hand thin films will, under appropriate conditions, enhance the reflectivity so that when deposited on a glass surface they may be used as beam-splitters, i.e. arrangements employed in interferometry for the division of an incident beam into two parts. Under appropriate conditions a multilayer may also be employed as a filter which transmits (or reflects) only selected regions of the spectrum. Multilayers may also be used as polarizers.

The subject of dielectric and metallic films has been very extensively treated in the scientific literature and many schemes for the computation of the optical effects of multilayers have been proposed. We shall give an outline of the general theory as developed in elegant and important investigations by F. Abels,* and consider in

0

- F. Abelis, Ann. de Physique, 5 (1950), 596-640 and 706-782. For a detailed treatment of the subject of thin films see a more specialized treatise e.g. H. Mayer, Physik dnner Schichten (Stuttgart, Wis-

detail some special cases of particular interest. For the treatment of problems involving only a small number of films it is naturally not necessary to use the general theory, and accordingly we shall later (§7.6) describe an alternative and older method based on the concept of multiple reflections.

Only dielectric stratified media will be treated in this section. The extension of the analysis to conducting media will be described in § 13.4.

### 1.6.1 The basic differential equations

Consider a plane, time-harmonic electromagnetic wave propagated through a stratified medium. In the special case when the wave is linearly polarized with its electric vector perpendicular to the plane of incidence we shall speak of a transverse electric wave (denoted by $TE$ ); when it is linearly polarized with its magnetic vector perpendicular to the plane of incidence we shall speak of a transverse magnetic wave (denoted by $TM$ ).* Any arbitrarily polarized plane wave may be resolved into two waves, one of which is a $TE$ wave and the other a $TM$ wave. Since according to §1.5 the boundary conditions at a discontinuity surface for the perpendicular and parallel components are independent of each other, these two waves will also be mutually independent. Moreover, Maxwell's equations remain unchanged when $\boldsymbol{E}$ and $\boldsymbol{H}$ and simultaneously $\varepsilon$ and $-\mu$ are interchanged. Thus any theorem relating to $TM$ waves may immediately be deduced from the corresponding result for $TE$ waves by making this change. It will, therefore, be sufficient to study in detail the $TE$ waves only.

We take the plane of incidence to be the $yz$-plane, †$z$ being the direction of stratification. For a $TE$ wave, $E_y = E_z = 0$ and Maxwell's equations reduce to the following six scalar equations [time dependence $\exp(-i\omega t)$ being assumed]:

$$\begin{array}{llll} \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} + \frac{i\varepsilon\omega}{c} E_x = 0, & \text{(1a)} & \frac{i\omega\mu}{c} H_x = 0 \\ \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = 0, & \text{(1b)} & \frac{\partial E_x}{\partial z} - \frac{i\omega\mu}{c} H_y = 0 \\ \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = 0, & \text{(1c)} & \frac{\partial E_x}{\partial y} + \frac{i\omega\mu}{c} H_z = 0 \end{array}$$

These equations show that $H_y, H_z$ and $E_x$ are functions of $y$ and $z$ only. Eliminating $H_y$ and $H_z$ between (1a), (2b) and (2c) [or by taking the $x$ component of the wave equation § 1.2 (5) for $E$ ] it follows that

$$\frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} + n^2 k_0^2 E_x = \frac{d(\log \mu)}{dz} \frac{\partial E_x}{\partial z} \tag{3}$$

where

$$n^2 = \varepsilon\mu, \quad k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda_0} \tag{4}$$

To solve (3) we take, as a trial solution, a product of two functions, one involving $y$ only and the other involving $z$ only:

$$E_x(y, z) = Y(y)U(z) \tag{5}$$

---

senschaftliche Verlagsgesellschaft, 1950); S. Methfessel, Dnne Schichten (Halle (Saale), VEB Wilhelm Knapp Verlag, 1953); or O. S. Heavens, Optical Properties of Thin Solid Films (London, Butterworths Scientific Publications, 1955).

0

- The terms " $E$-polarized" and " $H$-polarized" are also used (cf. § 11.4.1). It should be mentioned that the terms "transverse electric wave" and "transverse magnetic wave" have different meanings in the theory of wave guides.

Eq. (3) then becomes

$$\frac{1}{Y}\frac{d^2Y}{dy^2} = -\frac{1}{U}\frac{d^2U}{dz^2} - n^2k_0^2 + \frac{d(\log\mu)}{dz}\frac{1}{U}\frac{dU}{dz} \tag{6}$$

Now the term on the left is a function of $y$ only whilst the terms on the right depend only on $z$. Hence (6) can only hold if each side is equal to a constant ( $-K^2$ say):

$$\frac{1}{Y}\frac{d^2Y}{dy^2} = -K^2 \tag{7}$$

$$\frac{d^2U}{dz^2} - \frac{d(\log\mu)}{dz}\frac{dU}{dz} + n^2k_0^2U = K^2U \tag{8}$$

It will be convenient to set

$$K^2 = {k_0}^2\alpha^2 \tag{9}$$

Then (7) gives

$$Y = \text{ const. } e^{ik_0\alpha y}$$

and consequently $E_x$ is of the form

$$E_x = U(z)e^{i(k_0\alpha y - \omega t)} \tag{10}$$

where $U(z)$ is a (possibly complex) function of $z$. From (2b) and (2c) we see that $H_v$ and $H_z$ are given by expressions of the same form:

$$H_y = V(z)e^{i(k_0\alpha y - \omega t)} \tag{11}$$

$$H_z = W(z)e^{i(k_0\alpha y - \omega t)} \tag{12}$$

On account of (1a), (2b) and (2c), the amplitude functions $U, V$ and $W$ are related by the following equations:

$$V' = ik_0[\alpha W + \varepsilon U], \tag{13a}$$

$$U' = ik_0\mu V, \tag{13b}$$

$$\alpha U + \mu W = 0, \tag{13c}$$

the prime denoting differentiation with respect to $z$. Substituting for $W$ from (13c) into (13a) we have, together with (13b), a pair of simultaneous first-order differential equations* for $U$ and $V$ :

$$\left.\begin{aligned} U' &= ik_0\mu V \\ V' &= ik_0\left(\varepsilon - \tfrac{\alpha^2}{\mu}\right)U. \end{aligned}\right\} \tag{14}$$

- Eqs. (14) are of the same form as the equations of an electric transmission line, i.e.

† Not the $xz$-plane as in the previous section.

$$\frac{dV}{dz} = -ZI, \quad \frac{dI}{dz} = -YV,$$

where $V$ is the voltage across the line, $I$ is the current in the line, $Z$ is the series impedance, and $Y$ the shunt admittance. The theory of stratified media may therefore be developed in a strict analogy with the theory of electric transmission lines, as has been done by several authors, for example, R. B. Muchmore, J. Opt. Soc. Amer., 38 (1948), 20; K. Schuster, Ann. d. Physik (6), 4 (1949), 352; R. Kronig, R. S. Blaisse and J. J. v.v. Sande, J. Appl. Sci. Res., B,1 (1947), 63.

Elimination between these equations finally gives the following second-order linear differential equations for $U$ and $V$ :

$$\frac{d^2U}{dz^2} - \frac{d(\log \mu)}{dz}\frac{dU}{dz} + {k_0}^2\left(n^2 - \alpha^2\right)U = 0, \tag{15}$$

$$\frac{d^2V}{dz^2} - \frac{d\left[\log\left(\varepsilon - \frac{\alpha^2}{\mu}\right)\right]}{dz}\frac{dV}{dz} + {k_0}^2\left(n^2 - \alpha^2\right)V = 0. \tag{16}$$

According to the substitution rule which is a consequence of the symmetry of Maxwelt's equations, it immediately follows that for the TM wave ( $H_y = H_z = 0$ ), the non-vanishing components of the field vectors are of the form:

$$H_x = U(z)e^{i(k_0\alpha y - \omega t)}, \tag{17}$$

$$E_y = -V(z)e^{i(k_0\alpha y - \omega t)}, \tag{18}$$

$$E_z = -W(z)e^{i(k_0\alpha y - \omega t)} \tag{19}$$

where

$$\left.\begin{aligned} U' &= ik_0\varepsilon V \\ V' &= ik_0\left(\mu - \tfrac{\alpha^2}{\varepsilon}\right)U, \end{aligned}\right\} \tag{20}$$

and $W$ is related to $U$ by means of the equation

$$\alpha U + \varepsilon W = 0. \tag{21}$$

$U$ and $V$ now satisfy the following second-order linear differential equations:

$$\frac{d^2U}{dz^2} - \frac{d\{\log \varepsilon\}}{dz}\frac{dU}{dz} + {k_0}^2\left(n^2 - \alpha^2\right)U = 0, \tag{22}$$

$$\frac{d^2V}{dz^2} - \frac{d\left\{\log\left(\mu - \frac{\alpha^2}{\varepsilon}\right)\right\}}{dz}\frac{dV}{dz} + {k_0}^2\left(n^2 - \alpha^2\right)V = 0. \tag{23}$$

$U, V$ and $W$ are in general complex functions of $z$. The surfaces of constant amplitude of $E_x$ are given by

$$|U(z)| = \text{ constant}$$

whilst the surfaces of constant phase have the equation

$$\phi(z) + k_0\alpha y = \text{ constant },$$

where $\phi(z)$ is the phase of $U$. The two sets of surfaces do not in general coincide so that $E_x$ (and similarly $H_y$ and $H_z$ ) is an inhomogeneous wave. For a small displacement ( $dy, dz$ ) along a co-phasal surface, $\phi'(z)dz + k_0\alpha dy = 0$; hence if $\theta$ denotes the angle which the normal to the co-phasal surface makes with $OZ$, then

$$\tan\theta = -\frac{dz}{dy} = \frac{k_0\alpha}{\phi'(z)}$$

In the special case when the wave is an homogeneous plane wave,

$$\phi(z) = k_0 n z\cos\theta, \quad \alpha = n\sin\theta. \tag{24}$$

Hence the relation

$$\alpha = \text{ constant}$$

imposed by (9) may be regarded as a generalization of Snell's law of refraction to stratified media.

### 1.6.2 The characteristic matrix of a stratified medium

The solutions, subject to appropriate boundary conditions, of the differential equations which we have just derived, and various theorems relating to stratified media, can most conveniently be expressed in terms of matrices. We shall therefore give a brief account of the main definitions relating to matrices before discussing the consequences of out equations.
I. By a matrix one understands a system of real or complex numbers, arranged in a rectangular or a square array:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix},$$

$a_{ij}$ denoting the element in the $i$ th row and the $j$ th column. The matrix is denoted symbolically by $\boldsymbol{A}$ or $[a_{ij}]$, and is said to be an $m$ by $n$ matrix (or $m \times n$ matrix), since it contains $m$ rows and $n$ columns. In the special case when $m = n, A$ is said to be a square matrix of order $m$. If $A$ is a square matrix, the determinant whose elements are the same, and are in the same positions as the elements of $\boldsymbol{A}$, is said to be the determinant of the matrix $\boldsymbol{A}$; it is denoted by $|\boldsymbol{A}|$ or $|a_{ij}|$. If $|\boldsymbol{A}| = 1, \boldsymbol{A}$ is said to be unimodular.

By definition two matrices are equal only if they have the same number of rows ( $m$ ) and the same number of columns ( $n$ ), and if their corresponding elements are equal. If $\boldsymbol{A} = [a_{ij}]$ and $\boldsymbol{B} = [b_{ij}]$ are two matrices with the same number of rows and the same number of columns, then their sum $\boldsymbol{A} + \boldsymbol{B}$ is defined as the matrix $\boldsymbol{C}$ whose elements are $c_{ij} = a_{ij} + b_{ij}$. Similarly their difference $\boldsymbol{A} - \boldsymbol{B}$ is defined as the matrix $D$ with elements $d_{ij} = a_{ij} - b_{ij}$.

A matrix having every element zero is called a null matrix. The square matrix with elements $a_{ij} = 0$ when $i \neq j$ and $a_{ii} = 1$ for every value of $i$ is called unit matrix and will be denoted by $I$.

The product of a matrix $A$ and a number $\lambda$ (real or complex) is defined as the matrix $B$ with elements $b_{ij} = \lambda a_{ij}$.

The product $AB$ of two matrices is defined only when the number of columns in $A$ is equal to the number of rows in $\mathbf{B}$. If $\mathbf{A}$ is a $m \times p$ matrix and $\mathbf{B}$ is a $p \times n$ matrix the product is then by definition the $m \times n$ matrix with elements

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}.$$

The process of multiplication of two matrices is thus analogous to the row-by-column rule for multiplication of determinants of equal orders. In general $AB \neq BA$. For example

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

whilst

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

In the special case when $\mathbf{AB} = \mathbf{BA}$, the matrices $\mathbf{A}$ and $\mathbf{B}$ are said to commute.

The above definitions and properties of matrices are the only ones necessary for our purposes, and we can, therefore, now return to our discussion of propagation of electromagnetic waves through a stratified medium.

II. Since the functions $U(z)$ and $V(z)$ of § 1.6.1 each satisfy a second-order linear differential equation [(15) and (16)], it follows that $U$ and $V$ may each be expressed as a linear combination of two particular solutions, say $U_1, U_2$ and $V_1, V_2$. These particular solutions cannot be arbitrary; they must be coupled by the first order differential equations (14):

$$\left.\begin{aligned} U_1' &= ik_0\mu V_1, \\ V_1' &= ik_0\left(\varepsilon - \tfrac{\alpha^2}{\mu}\right) U_1, \end{aligned}\right\} \qquad \left.\begin{aligned} U_2' &= ik_0\mu V_2 \\ V_2' &= ik_0\left(\varepsilon - \tfrac{\alpha^2}{\mu}\right) U_2. \end{aligned}\right\} \tag{25}$$

From these relations it follows that

$$V_1 U_2' - U_1' V_2 = 0, \quad U_1 V_2' - V_1' U_2 = 0,$$

so that

$$\frac{d}{dz}\left(U_1 V_2 - U_2 V_1\right) = 0.$$

This relation implies that the determinant

$$D = \begin{vmatrix} U_1 & V_1 \\ U_2 & V_2 \end{vmatrix} \tag{26}$$

associated with any two arbitrary solutions of (14) is a constant, i.e. that $D$ is an invariant of our system of equations.*

For our purposes the most convenient choice of the particular solutions is

$$\left.\begin{aligned} U_1 &= f(z), \quad U_2 = F(z), \\ V_1 &= g(z), \quad V_2 = G(z), \end{aligned}\right\} \tag{27}$$

such that

$$f(0) = G(0) = 0 \quad \text{and} \quad F(0) = g(0) = 1. \tag{28}$$

Then the solutions with

0

- This also follows from a well-known property of a Wronskian of second-order linear differential equations. Moreover it may also be shown that, if $U_1$ is known, $U_2$ may be obtained by integration from the relation

$$U(0) = U_0, \quad V(0) = V_0, \tag{29}$$

may be expressed in the form

$$\left.\begin{aligned} U &= FU_0 + fV_0, \\ V &= GU_0 + gV_0, \end{aligned}\right\}$$

or, in matrix notation,

$$\mathbf{Q} = \mathbf{N}\mathbf{Q_0}, \tag{30}$$

where

$$\mathbf{Q} = \begin{bmatrix} U(z) \\ V(z) \end{bmatrix}, \quad \mathbf{Q}_0 = \begin{bmatrix} U_0 \\ V_0 \end{bmatrix}, \quad \mathbf{N} = \begin{bmatrix} F(z) & f(z) \\ G(z) & g(z) \end{bmatrix}. \tag{31}$$

On account of the relation $D =$ constant, the determinant of the square matrix $N$ is a constant. The value of this constant may immediately be found by taking $z = 0$, giving

$$|N| = Fg - fG = 1.$$

It is usually more convenient to express $U_0$ and $V_0$ as functions of $U(z)$ and $V(z)$. Solving for $U_0$ and $V_0$, we obtain

$$\mathbf{Q_0} = \mathbf{M}\mathbf{Q}, \tag{32}$$

where

$$\mathbf{M} = \begin{bmatrix} g(z) & -f(z) \\ -G(z) & F(z) \end{bmatrix}. \tag{33}$$

This matrix is also unimodular,

$$|M| = 1. \tag{34}$$

The significance of $\boldsymbol{M}$ is clear: it relates the $x$ and $y$ components of the electric (or magnetic) vectors in the plane $z = 0$, to the components in an arbitrary plane $z =$ constant. Now we saw that the knowledge of $U$ and $V$ is sufficient for the complete specification of the field. Hence for the purposes of determining the propagation of a plane monochromatic wave through a stratified medium, the medium only need be specified by an appropriate two by two unimodular matrix M . For this reason we shall call M the characteristic matrix of the stratified medium. The constancy of the determinant $|\mathbf{M}|$ may be shown to imply the conservation of energy.*

We shall now consider the form of the characteristic matrix for cases of particular interest.
(a) A homogeneous dielectric film

In this case $\varepsilon, \mu$ and $n = \sqrt{\varepsilon\mu}$ are constants. If $\theta$ denotes the angle which the normal to the wave makes with the $z$-axis, we have by (24),

$$\alpha = n \sin\theta,$$

For a $TE$ wave, we have according to (15) and (16),

---

$$U_2 = ikDU_1 \int \frac{\mu}{U_1^2} dz$$

Cf. F. Abels, Ann. de Physique, 5 (1950), 603.

$$\left.\begin{array}{l} \frac{d^2U}{dz^2} + (k_0^2 n^2 \cos^2\theta)\, U = 0, \\ \frac{d^2V}{dz^2} + (k_0^2 n^2 \cos^2\theta)\, V = 0. \end{array}\right\} \tag{35}$$

The solutions of these equations, subject to the relations (14), are easily seen to be

$$\left.\begin{array}{l} U(z) = A\cos(k_0 nz\cos\theta) + B\sin(k_0 nz\cos\theta) \\ V(z) = \frac{1}{i}\sqrt{\frac{\varepsilon}{\mu}}\cos\theta\left\{B\cos(k_0 nz\cos\theta) - A\sin(k_0 nz\cos\theta)\right\}. \end{array}\right\} \tag{36}$$

Hence the particular solutions (27) which satisfy the boundary conditions (28) are

$$\left.\begin{array}{ll} U_1 & = f(z) = \frac{i}{\cos\theta}\sqrt{\frac{\mu}{\varepsilon}}\sin(k_0 nz\cos\theta), \\ V_1 & = g(z) = \cos(k_0 nz\cos\theta), \\ U_2 & = F(z) = \cos(k_0 nz\cos\theta), \\ V_2 & = G(z) = i\sqrt{\frac{\varepsilon}{\mu}}\cos\theta\sin(k_0 nz\cos\theta). \end{array}\right\} \tag{37}$$

If we set

$$p = \sqrt{\frac{\varepsilon}{\mu}}\cos\theta \tag{38}$$

the characteristic matrix is seen to be

$$M(z) = \left[\begin{array}{cc} \cos(k_0 nz\cos\theta) & -\frac{i}{p}\sin(k_0 nz\cos\theta) \\ -ip\sin(k_0 nz\cos\theta) & \cos(k_0 nz\cos\theta) \end{array}\right]. \tag{39}$$

For a $TM$ wave, the same equations hold, with $p$ replaced by

$$q = \sqrt{\frac{\mu}{\varepsilon}}\cos\theta \tag{40}$$

(b) A stratified medium as a pile of thin homogeneous films

Consider two adjacent stratified media, the first one extending from $z = 0$ to $z = z_1$, and the second from $z = z_1$ to $z = z_2$. If $\mathbf{M}_1(z)$ and $\mathbf{M}_2(z)$ are the characteristic matrices of the two media, then

$$\mathbf{Q}_0 = \mathbf{M}_1(z_1)\,\mathbf{Q}(z_1), \quad \mathbf{Q}(z_1) = \mathbf{M}_2(z_2 - z_1)\,\mathbf{Q}(z_2)$$

so that

$$\mathbf{Q}_0 = \mathbf{M}(z_2)\,\mathbf{Q}(z_2),$$

where

$$\mathbf{M}(z_2) = \mathbf{M}_1(z_1)\,\mathbf{M}_2(z_2 - z_1).$$

This result may immediately be generalized to the case of a succession of stratified media extending from $0 \leqslant z \leqslant z_1, z_1 \leqslant z \leqslant z_2, \ldots z_{N-1} \leqslant z \leqslant z_N$. If the characteristic matrices are $\mathbf{M}_1, \mathbf{M}_2, \ldots \mathbf{M}_N$, then
where

0

- To show this, one evaluates the reflectivity and transmissivity (51) in terms of the matrix elements. If further one uses the fact that for a non-absorbing medium the characteristic matrix is of the form indicated by (45), it follows that the conservation law $\mathscr{R} + \mathscr{T} = 1$ will be satisfied provided that $|\mathrm{M}| = 1$.

$$\left.\begin{array}{c} \mathbf{Q}_0 = \mathbf{M}(z_N)\,\mathbf{Q}(z_N) \\ \mathbf{M}(z_N) = \mathbf{M}_1(z_1)\,\mathbf{M}_2(z_2 - z_1)\ldots\mathbf{M}_N(z_N - z_{N-1}) \end{array}\right\} \tag{41}$$

With the help of (41) an approximate expression for the characteristic matrix of any stratified medium may easily be derived*: we regard the medium as consisting of a very large number of thin films of thickness $\delta z_1, \delta z_2, \delta z_3, \ldots \delta z_n$. If the maximum

thickness is sufficiently small, it is permissible to regard $\varepsilon, \mu$ and $n$ to be constant throughout each film. From (39) it is seen that the characteristic matrix of the $j$ th film is then approximately given by

$$M_j = \begin{bmatrix} 1 & -\frac{i}{p_j}k_0 n_j \delta z_j \cos\theta_j \\ -ip_j k_0 n_j \delta z_j \cos\theta_j & 1 \end{bmatrix}$$

Hence the characteristic matrix of the whole medium, considered as a pile of thin films, is approximately equal to (again retaining terms up to the first power in $\delta z$ only):

$$\mathbf{M} = \prod_{j=1}^{N} \mathbf{M}_j = \begin{bmatrix} 1 & -ik_0 B \\ -ik_0 A & 1 \end{bmatrix} \tag{42}$$

where

$$A = \sum_{j=1}^{N} p_j n_j \delta z_j \cos\theta_j = \sum_{j=1}^{N} \left(\varepsilon_j - \frac{\alpha^2}{\mu_j}\right)\delta z_j$$

$$B = \sum_{j=1}^{N} \frac{n_j}{p_j}\delta z_j \cos\theta_j = \sum_{j=1}^{N} \mu_j \delta z_j.$$

Proceeding to the limit as $N \to \infty$ in such a way that $\max|\delta z_j| \to 0$, we obtain

$$M = \begin{bmatrix} 1 & -ik_0 \mathscr{B} \\ -ik_0 \mathscr{A} & 1 \end{bmatrix}, \tag{43}$$

where

$$\mathscr{A} = \int \left(\varepsilon - \frac{\alpha^2}{\mu}\right) dz, \quad \mathscr{B} = \int \mu dz, \tag{44}$$

the integration being taken throughout the whole $z$ range. Eq. (43) gives a first approximation to the characteristic matrix of an arbitrary stratified medium. Improved approximations can be obtained by retaining higher-order terms* in the expansions of $\cos(k_0 n \delta z \cos\theta)$ and $\sin(k_0 n \delta z \cos\theta)$ and in the product (42).

Since, for a non-absorbing medium, $\varepsilon$ and $\mu$ are real, it is also seen that the characteristic matrix of a non-absorbing stratified medium has the form

$$\mathbf{M} = \begin{bmatrix} a & ib \\ ic & d \end{bmatrix} \tag{45}$$

where $a, b, c$ and $d$ are real.

0

- For a fuller treatment of stratified media of continuously varying refractive index see R. Jacobsson, Progress in Optics, Vol. 5, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 247.

### 1.6.3 The reflection and transmission coefficients

Consider a plane wave incident upon a stratified medium that extends from $z = 0$ to $z = z_1$ and that is bounded on each side by a homogeneous, semi-infinite medium. We shall derive expressions for the amplitudes and intensities of the reflected and transmitted waves. †

Let $A, R$ and $T$ denote as before the amplitudes (possibly complex) of the electric vectors of the incident, reflected and transmitted waves. Further, let $\varepsilon_1, \mu_1$ and $\varepsilon_l, \mu_l$ be the dielectric constant and the magnetic permeability of the first and the last

medium, and let $\theta_1$ and $\theta_l$ be the angles which the normals to the incident and the transmitted waves make with the $z$-direction (direction of stratification).

The boundary conditions of § 1.1 demand that the tangential components of $\boldsymbol{E}$ and $H$ shall be continuous across each of the two boundaries of the stratified medium. This gives, if the relation § 1.4 (4)

$$\boldsymbol{H} = \sqrt{\frac{\varepsilon}{\mu}} s \wedge \boldsymbol{E}$$

is also used, the following relations for a $TE$ wave:

$$\left.\begin{array}{ll} U_0 = A + R, & U(z_1) = T \\ V_0 = p_1(A - R), & V(z_1) = p_l T, \end{array}\right\} \tag{46}$$

where

$$p_1 = \sqrt{\frac{\varepsilon_1}{\mu_1}} \cos\theta_1, \quad p_l = \sqrt{\frac{\varepsilon_l}{\mu_l}} \cos\theta_l \tag{47}$$

The four quantities $U_0, V_0, U$ and $V$ given by (46) are connected by the basic relation (32); hence

$$\left.\begin{array}{r} A + R = (m'_{11} + m'_{12} p_l)\, T \\ p_1(A - R) = (m'_{21} + m'_{22} p_l)\, T, \end{array}\right\} \tag{48}$$

$m'_{ij}$ being the elements of the characteristic matrix of the medium, evaluated for $z = z_1$. From (48) we obtain the reflection and transmission coefficients of the film:

$$r = \frac{R}{A} = \frac{(m'_{11} + m'_{12} p_l)\, p_1 - (m'_{21} + m'_{22} p_l)}{(m'_{11} + m'_{12} p_l)\, p_1 + (m'_{21} + m'_{22} p_l)} \tag{49}$$

$$t = \frac{T}{A} = \frac{2p_1}{(m'_{11} + m'_{12} p_l)\, p_1 + (m'_{21} + m'_{22} p_l)} \tag{50}$$

In terms of $r$ and $t$, the reflectivity and transmissivity are

$$\mathscr{R} = |r|^2, \quad \mathscr{T} = \frac{p_l}{p_1} |t|^2 \tag{51}$$

The phase $\delta_r$ of $r$ may be called the phase change on reflection and the phase $\delta_t$ of $t$ the phase change on transmission. The phase change $\delta_r$ is referred to the first surface of discontinuity, whilst the phase change $\delta_i$ is referred to the plane boundary between the stratified medium and the last semi-infinite medium.

0

- This is discussed fully in the paper by F. Abels, Ann. d Physique, 5 (1950), p. 611.
  † We consider the amplitudes of the electric vectors when studying a $TE$ wave and those of the magnetic vectors when studying a TM wave.

The corresponding formulae for a $TM$ wave are immediately obtained from (49)-(51) on replacing the quantities $p_1$ and $p_l$ by

$$q_1 = \sqrt{\frac{\mu_1}{\varepsilon_1}}\cos\theta_1, \quad q_l = \sqrt{\frac{\mu_l}{\varepsilon_l}}\cos\theta_l. \tag{52}$$

$r$ and $t$ are then the ratios of the amplitudes of the magnetic and not the electric vectors.

### 1.6.4 A homogeneous dielectric film*

The properties of an homogeneous dielectric film situated between two homogeneous media is of particular interest in optics, and we shall, therefore, study this case more fully. We assume all the media to be non-magnetic ( $\mu = 1$ ).

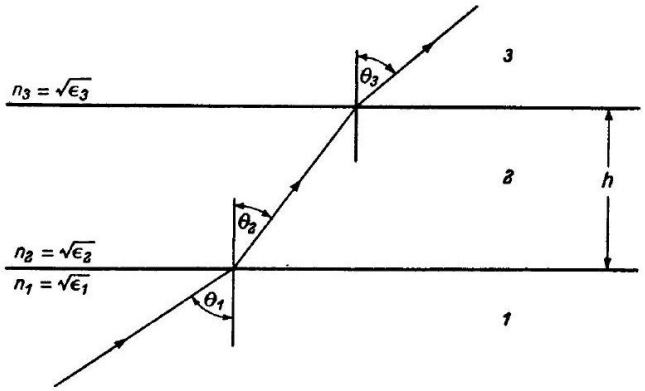



Fig. 1.17. Propagation of an electromagnetic wave through a homogeneous film.

The characteristic matrix of a homogeneous dielectric film is given by (39). Denoting by subscripts 1, 2 and 3 quantities which refer to the three media (see Fig. 1.17), and by $h$ the thickness of the film, we have

$$m'_{11} = m'_{22} = \cos\beta, \quad m'_{12} = -\frac{\imath}{p_2}\sin\beta, \quad m'_{21} = -ip_2\sin\beta, \tag{53}$$

where

$$\beta = \frac{2\pi}{\lambda_0} n_2 h\cos\theta_2$$

and

$$p_j = n_j\cos\theta_j, \quad (j = 1, 2, 3). \tag{54}$$

The reflection and transmission coefficients $r$ and $t$ may be obtained by substituting these expressions into (49) and (50), with $l = 3$. The resulting formulae may be conveniently expressed in terms of the corresponding coefficients $r_{12}, t_{12}$ and $r_{23}, t_{23}$ associated with the reflection and transmission at the first and the second surface respectively. According to the Fresnel formulae § 1.5 (20) and (21) we have for a $TE$ wave,

$$r_{12} = \frac{n_1\cos\theta_1 - n_2\cos\theta_2}{n_1\cos\theta_1 + n_2\cos\theta_2} = \frac{p_1 - p_2}{p_1 + p_2} \tag{55}$$

$$t_{12} = \frac{2n_1\cos\theta_1}{n_1\cos\theta_1 + n_2\cos\theta_2} = \frac{2p_1}{p_1 + p_2} \tag{56}$$

with analogous expressions for $r_{23}$ and $t_{23}$. In terms of these expressions, the formulae for $r$ and $t$ become*

$$r = \frac{r_{12} + r_{23}e^{2i\beta}}{1 + r_{12}r_{23}e^{2i\beta}} \tag{57}$$

$$t = \frac{t_{12}t_{23}e^{i\beta}}{1 + r_{12}r_{23}e^{2i\beta}} \tag{58}$$

the reflectivity and transmissivity are therefore given by

$$\mathscr{R} = |r|^2 = \frac{r_{12}^2 + r_{23}^2 + 2r_{12}r_{23}\cos 2\beta}{1 + r_{12}^2 r_{23}^2 + 2r_{12}r_{23}\cos 2\beta}; \tag{59}$$

and

$$\mathscr{T} = \frac{p_3}{p_1}|t|^2 = \frac{n_3\cos\theta_3}{n_1\cos\theta_1}\frac{{t_{12}}^2{t_{23}}^2}{1 + {r_{12}}^2{r_{23}}^2 + 2r_{12}r_{23}\cos 2\beta} \tag{60}$$

A straightforward calculation gives, as expected,

$$\mathscr{R} + \mathscr{T} = 1.$$

The phase changes can also easily be calculated from (57) and (58), and are found to be given by

$$\tan\delta_r = \tan(\arg r) = \frac{r_{23}\,(1 - r_{12}^2)\sin 2\beta}{r_{12}\,(1 + r_{23}^2) + r_{23}\,(1 + r_{12}^2)\cos 2\beta} \tag{61}$$

$$\tan\delta_t = \tan(\arg t) = \frac{1 - r_{12}r_{23}}{1 + r_{12}r_{23}}\tan\beta \tag{62}$$

Let us now briefly consider the implications of these formulae. We first note that (59) and (60) remain unchanged when $\beta$ is replaced by $\beta + \pi$, i.e. when $h$ is replaced by $h + \Delta h$, where

$$\Delta h = \frac{\lambda_0}{2n_2\cos\theta_2}. \tag{63}$$

Hence the reflectivity and transmissivity of dielectric films which differ in thickness by an integral multiple of $\lambda_0/2n_2\cos\theta_2$ are the same.

Next we determine the optical thickness for which the reflection coefficient has a maximum or a minimum. If we set

$$H = n_2 h, \tag{64}$$

we find from (59) that

$$\frac{d\mathscr{R}}{dH} = 0 \quad \text{when} \quad \sin 2\beta = 0$$

i.e. when

0

- An alternative derivation of the main formulae relating to the properties of a single dielectric film will be found in § 7.6.1. The formulae may, of course, also be derived directly by applying the boundary conditions of § 1.1.3 at each boundary of the film (cf. M. Born, Optik (Berlin, Springer, 1933), p. 125; or H. Mayer, Physik dnner Schichten (Stuttgart, Wissenschaftliche Verlagsgesellschaft 1950), p. 145.

$$H = \frac{m\lambda_0}{4\cos\theta_2}, \quad (m = 0, 1, 2, \ldots)$$

We must distinguish two cases:

(1) When $m$ is odd, i.e. when $H$ has any of the values

$$H = \frac{\lambda_0}{4\cos\theta_2}, \quad \frac{3\lambda_0}{4\cos\theta_2}, \quad \frac{5\lambda_0}{4\cos\theta_2}, \ldots$$

then $\cos 2\beta = -1$ and (59) reduces to

$$\mathscr{R} = \left(\frac{r_{12} - r_{23}}{1 - r_{12}r_{23}}\right)^2. \tag{65}$$

In particular for normal incidence, one has from (55)

$$r_{12} = \frac{n_1 - n_2}{n_1 + n_2}, \quad r_{23} = \frac{n_2 - n_3}{n_2 + n_3} \tag{66}$$

and (65) becomes

$$\mathscr{R} = \left(\frac{n_1 n_3 - n_2^2}{n_1 n_3 + n_2^2}\right)^2 \tag{67}$$

(2) When $m$ is even, i.e. when the optical thickness has any of the values

$$H = \frac{\lambda_0}{2\cos\theta_2}, \quad \frac{2\lambda_0}{2\cos\theta_2}, \quad \frac{3\lambda_0}{2\cos\theta_2}, \ldots$$

then $\cos 2\beta = 1$ and (59) reduces to

$$\mathscr{R} = \left(\frac{r_{12} + r_{23}}{\mathrm{I} + r_{12}r_{23}}\right)^2 \tag{68}$$

In particular, for normal incidence, this becomes

$$\mathscr{R} = \left(\frac{n_1 - n_3}{n_1 + n_3}\right)^2 \tag{69}$$

and is seen to be independent of $n_2$. Now the only difference in the case of oblique incidence is the replacement of $n_j$ by $n_j \cos\theta_j (j = 1, 2, 3)$ in all the formulae; hence a plate whose optical thickness is $m\lambda_0/2\cos\theta_2 (m = 1, 2, 3 \ldots)$ has no influence on the intensity of the reflected (or transmitted) radiation.

Next we must determine the nature of these extreme values. After a straightforward calculation we find that when $H = m\lambda_0/4\cos\theta_2 (m = 1, 2, \ldots)$
according as

$$\left.\begin{array}{c} \left(\frac{d^2\mathscr{R}}{dH^2}\right) \gtrless 0 \\ (-1)^m r_{12}r_{23}\left[1 + {r_{12}}^2{r_{23}}^2 - {r_{12}}^2 - {r_{23}}^2\right] \lesseqgtr 0, \end{array}\right\} \tag{70}$$

so that with the upper sign there is a minimum and with the lower sign a maximum. In particular, for normal incidence, $r_{12}$ and $r_{23}$ are given by (66) and we have

---

0

- These formulae were first derived in a different manner by G. B. Arry, Phil. Mag., 2 (1833), 20; also Ann. Phys. und Chem. (Ed. Poggendorf), 41 (1837), 512.

$$\left.\begin{array}{ll} \text{maximum, if} & (-1)^m (n_1 - n_2)(n_2 - n_3) > 0, \\ \text{minimum, if} & (-1)^m (n_1 - n_2)(n_2 - n_3) < 0. \end{array}\right\} \tag{71}$$

Usually the first medium is air ( $n_1 \sim 1$ ) and we see that with a film whose optical thickness has any of the values $\lambda_0/4, 3\lambda_0/4, 5\lambda_0/4, \ldots$ the reflectivity is then a maximum or a minimum according to whether the refractive index of the film is greater or smaller than the refractive index of the last medium; for a film whose optical thickness has any of the values $\lambda_0/2, 2\lambda_0/2, 3\lambda_0/2, \ldots$ the opposite is the case.

These results, which are illustrated in Fig. 1.18, are found to be in good agreement with experiment.*

It is evident from the preceding analysis that a plate whose optical thickness is a quarter of the wavelength and whose refractive index is low enough may be used as an antireflection film, i.e. a film by means of which the reflectivity of a surface is reduced. (The surface is then said to be "bloomed".) The two substances most commonly used for this purpese are cryolite ( $n \sim 1.35$ ) and magnesium fluoride ( $MgF_2, n \sim 1.38$ ).

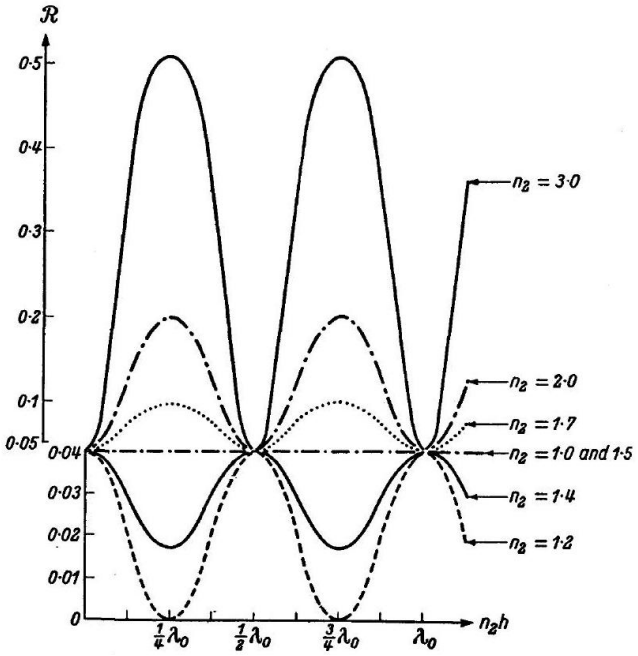



Fig. 1.18. The reflectivity of a dielectric film of refractive index $n_{2}$ as a function of its optical thickness. ($\theta_1 = 0, n_1 = 1, n_3 = 1.5$).

(After R. Messner, Zeiss Nachr., 4 (H9) (1943), 253.)

According to (67), the reflectivity at normal incidence would be strictly zero if

$$n_2 = \sqrt{n_1 n_3}. \tag{72}$$

With $n_1 = 1, n_3 = 1.5$ this demands $n_2 \sim 1.22$, a condition which cannot be satisfied in practice. A fuller analysis of (59) shows, however, that with oblique incidence it is possible to have zero reflectivity for a $TM$ wave (electric vector parallel to the plane of incidence) but not for a $TE$ wave (electric vector perpendicular to the plane of incidence), i.e. under favourable conditions one can have simultaneously $\mathscr{R}_{11} = 0, \mathscr{R}_{\perp} \neq 0$. Hence a thin film of a suitable dielectric material may also be used as a polarizer, working by reflection. Such a polarizer may be regarded as a generalization of the simple arrangement discussed earlier in connection with

Brewster's angle. To obtain a large value for $\mathscr{R}_\perp$ (with $\mathscr{R}_\parallel = 0$ ), the refractive index $n_2$ of the film must

be as large as possible.* For example, with $n_1 = 1, n_2 = 2.5, n_3 = 1.53$ one obtains $\mathscr{R}_\parallel = 0, \mathscr{R}_\perp = 0.79$ when $\theta_1 = 74°30'$.

If a glass surface is coated with a material of sufficiently high refractive index, the reflectivity of the surface will, according to the preceding analysis, be greatly enhanced (see Figs. 1.18 and 1.19). The surface will then act as a good beam splitter. Coatings of titanium dioxide ( $TiO_2, n \sim 2 \cdot 45$ ) or zinc sulphide ( $ZnS, n \sim 2 \cdot 3$ ) are very suitable for this purpose, giving a maximum reflectivity of about $0 \cdot 3$. There are other substances which have high refractive indices, but they absorb some of the incident light. For example, with a coating of stibnite ($Sb_2\ S_3, n \sim 2.8$) one can attain the values $\mathscr{R} = \mathscr{T} = 0 \cdot 46$, but 8 per cent of the incident light is then absorbed by the film.

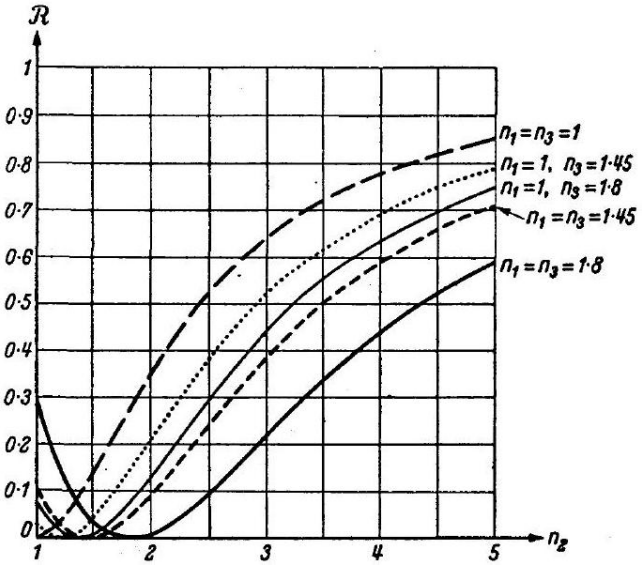



Fig. 1.19. The reflectivity at normal incidence of a quarter-wave film ( $n_{2} h=\lambda_{0} / 4$ ) as function of the refractive index $n_2$ of the film.

(After K. Hammer, Z. tech. Phys., 24 (1943), 169.)

It is also of interest to examine the case when total reflection takes place on the first boundary. In this case

$$n_1 \sin\theta_1 > n_2, \quad n_1 \sin\theta_1 < n_3$$

and (cf. § 1.5, eq. 54),

$$n_2 \cos\theta_2 = -i\sqrt{n_1^2 \sin^2\theta_1 - n_2^2} \tag{73}$$

The coefficients for reflection at the two boundaries now are (cf. §1.5 (21))

$$\left.\begin{aligned} r_{12} &= \frac{n_1\cos\theta_1 - i\sqrt{n_1^2\sin^2\theta_1 - n_2^2}}{n_1\cos\theta_1 + i\sqrt{n_1^2\sin^2\theta_1 - n_2^2}} \\ r_{23} &= \frac{-i\sqrt{n_1^2\sin^2\theta_1 - n_2^2} - n_3\cos\theta_3}{-i\sqrt{n_1^2\sin^2\theta_1 - n_2^2} + n_3\cos\theta_3} \end{aligned}\right\} \tag{74}$$

0

• See, for example, K. Hammer, Z. tech. Phys., 24 (1943), 169.

If we set

$$k_0 n_2 h \cos\theta_2 = ib \tag{75}$$

where, according to (73),

$$b = -\frac{2\pi}{\lambda_0} h \sqrt{n_1^2 \sin^2\theta_1 - n_2^2} \tag{76}$$

we obtain for the reflection coefficients the following expression, in place of (57):

$$r = \frac{r_{12} + r_{23} e^{-2b}}{1 + r_{12} r_{23} e^{-2b}} \tag{77}$$

Since $|r_{12}| = |r_{23}| = 1, r_{12}$ and $r_{23}$ are of the form

$$r_{12} = e^{i\phi_{12}}, \quad r_{23} = e^{i\phi_{23}}, \tag{78}$$

where the $\phi$ 's are real; hence the reflectivity now is

$$\mathscr{R} = |r|^2 = \frac{e^{2b} + e^{-2b} + 2\cos(\phi_{12} - \phi_{23})}{e^{2b} + e^{-2b} + 2\cos(\phi_{12} + \phi_{23})}. \tag{79}$$

In contrast with the previous case, $\mathscr{R}$ is now no longer a periodic function of the thickness of the film. (76) shows that if the dependence of the refractive index on the wavelength is neglected, $b$ is inversely proportional to wavelength. Since for sufficiently large values of $b, \mathscr{R}$ will be practically unity, the shorter wavelengths will not be transmitted; the film then acts as a low-pass filter, i.e. one which transmits the long wavelengths only.

We have seen that by the use of dielectric films of suitable material, many useful effects can be attained. It will be apparent that, with a number of such films arranged in succession, the desired features may be still further enhanced. The characteristic matrix of such a multilayer may be obtained with the help of the theorem expressed by (41).* We shall discuss in detail only the case when the multilayer is periodic.

### 1.6.5 Periodically stratified media

A stratified periodic medium with period $h$ is characterized by a dielectric constant $\varepsilon$ and a magnetic permeability $\mu$ which are functions of $z$ only and are such that

$$\varepsilon(z + jh) = \varepsilon(z), \quad \mu(z + jh) = \mu(z)$$

being any integer in some fixed range $1 \leqslant j \leqslant N$.
Let $\mathbf{M}(h)$ be the characteristic matrix corresponding to one period and write †

$$\mathbf{M}(h) = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}. \tag{80}$$

According to (41) we then have, on account of the periodicity,

$$\mathbf{M}(Nh) = \underbrace{\mathbf{M}(h) \cdot \mathbf{M}(h) \ldots \mathbf{M}(h)}_{N \text{ times}} = (\mathbf{M}(h))^N. \tag{81}$$

0

• Cf. H. Schröder, Optik, 3 (1948), 499.

0

To evaluate the elements of the matrix $\boldsymbol{M}(Nh)$ we use a result from the theory of matrices, according to which the $N$ th power of a unimodular matrix $\mathbf{M}(h)$ is*

$$[\mathbf{M}(h)]^N = \begin{bmatrix} m_{11}\mathscr{U}_{N-1}(a) - \mathscr{U}_{N-2}(a) & m_{12}\mathscr{U}_{N-1}(a) \\ m_{21}\mathscr{U}_{N-1}(a) & m_{22}\mathscr{U}_{N-1}(a) - \mathscr{U}_{N-2}(a) \end{bmatrix}, \tag{82}$$

where

$$a = \frac{1}{2}(m_{11} + m_{22}) \tag{83}$$

and $\mathscr{U}_N$ are the Chebyshev Polynomials of the second kind † :

$$\mathscr{U}_N(x) = \frac{\sin[(N+1)\cos^{-1}x]}{\sqrt{1-x^2}} \tag{84}$$

A multilayer usually consists of a succession of homogeneous layers of alternately low and high refractive indices $n_2$ and $n_3$ and of thickness $h_2$ and $h_3$, placed between two homogeneous media of refractive indices $n_1$ and $n_l$. (See Fig. 1.20.) We again assume the media to be non-magnetic ( $\mu = 1$ ) and set

$$\left.\begin{array}{cc} \beta_2 = \frac{2\pi}{\lambda_0} n_2 h_2 \cos\theta_2, & \beta_3 = \frac{2\pi}{\lambda_0} n_3 h_3 \cos\theta_3, \\ p_2 = n_2 \cos\theta_2, & p_3 = n_3 \cos\theta_3. \\ h = h_2 + h_3. & \end{array}\right\} \tag{85}$$

The characteristic matrix $M_2(h)$ of one period then is, according to (39) and (41),

$$\begin{aligned} \boldsymbol{M}_2(h) &= \begin{bmatrix} \cos\beta_2 & -\frac{i}{p_2}\sin\beta_2 \\ -ip_2\sin\beta_2 & \cos\beta_2 \end{bmatrix} \begin{bmatrix} \cos\beta_3 & -\frac{i}{p_3}\sin\beta_3 \\ -ip_3\sin\beta_3 & \cos\beta_3 \end{bmatrix} \\ &= \begin{bmatrix} \cos\beta_2\cos\beta_3 - \frac{p_3}{p_2}\sin\beta_2\sin\beta_3 & -\frac{i}{p_3}\cos\beta_2\sin\beta_3 - \frac{i}{p_2}\sin\beta_2\cos\beta_3 \\ -ip_2\sin\beta_2\cos\beta_3 - ip_3\cos\beta_2\sin\beta_3 & \cos\beta_2\cos\beta_3 - \frac{p_2}{p_3}\sin\beta_2\sin\beta_3 \end{bmatrix}. \end{aligned} \tag{86}$$

- The correctness of this result may be verified by induction, using the recurrence relation

$$\mathscr{U}_j(x) = 2x\mathscr{U}_{j-1}(x) - \mathscr{U}_{j-2}(x),$$

which follows as an identity from the definition of the Chebyshev Polynomials.
A direct proof based on the theory of matrices was given by F. Abels, Ann. de Physique, 5 (1950), 777.
† These polynomials satisfy the following orthogonality and normalizing conditions:

$$\int_{-1}^{+1} \mathscr{U}_m(x)\mathscr{U}_n(x)\sqrt{1-x^2}\,dx = 0 \text{ when } n \neq m$$
$$= \frac{\pi}{2} \text{ when } n = m$$

For convenience we note the explicit expressions of the first six polynomials:

$$\begin{array}{ll} \mathscr{U}_0(x) = 1, & \mathscr{U}_3(x) = 8x^3 - 4x \\ \mathscr{U}_1(x) = 2x, & \mathscr{U}_4(x) = 16x^4 - 12x^2 + 1 \\ \mathscr{U}_2(x) = 4x^2 - 1, & \mathscr{U}_5(x) = 32x^5 - 32x^3 + 6x \end{array}$$

- Formulae relating to multilayers have been given by many writers, e.g. R. L. Mooney, J. Opt. Soc. Amer., 36 (1946), 256; W. Weinstein, ibid, 37 (1947), 576, † We now omit the prime on the matrix elements,

Tables of Chebyshev polynomials have been published by the National Bureau of Standards Washington (Applied Mathematics Series 9 (1952)), where the main properties of the polynomials are also summarized. See also Higher Transcendental Functions (Bateman Manuscript Project, New York, McGraw-Hill, Vol. 2 (1953), p. 183).

Hence according to (81), the characteristic matrix $\mathbf{M}_{2N}(Nh)$ of the multilayer (with $2N$ films in all) is given by the following formula due to Abels:

$$\mathbf{M}_{2N}(Nh) = \begin{bmatrix} \mathscr{M}_{11} & \mathscr{M}_{12} \\ \mathscr{M}_{21} & \mathscr{M}_{22} \end{bmatrix}, \tag{87}$$

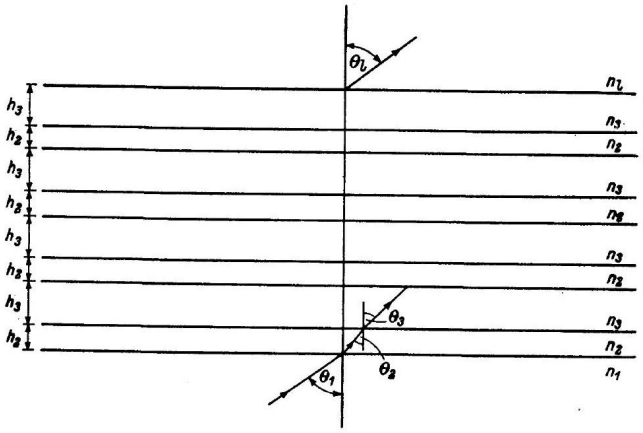



Fig. 1.20. A periodic multilayer.

where

$$\left.\begin{aligned} \mathscr{M}_{11} &= \left[\cos\beta_2\cos\beta_3 - \tfrac{p_3}{p_2}\sin\beta_2\sin\beta_3\right]\mathscr{U}_{N-1}(a) - \mathscr{U}_{N-2}(a), \\ \mathscr{M}_{12} &= -i\left[\tfrac{1}{p_3}\cos\beta_2\sin\beta_3 + \tfrac{1}{p_2}\sin\beta_2\cos\beta_3\right]\mathscr{U}_{N-1}(a), \\ \mathscr{M}_{21} &= -i\left[p_2\sin\beta_2\cos\beta_3 + p_3\cos\beta_2\sin\beta_3\right]\mathscr{U}_{N-1}(a), \\ \mathscr{M}_{22} &= \left[\cos\beta_2\cos\beta_3 - \tfrac{p_2}{p_3}\sin\beta_2\sin\beta_3\right]\mathscr{U}_{N-1}(a) - \mathscr{U}_{N-2}(a), \end{aligned}\right\} \tag{88}$$

and

$$a = \cos\beta_2\cos\beta_3 - \frac{1}{2}\left(\frac{p_2}{p_3} + \frac{p_3}{p_2}\right)\sin\beta_2\sin\beta_3. \tag{89}$$

The reflection and transmission coefficients of the multilayer are immediately obtained by substituting these expressions into (49) and (50).

Of particular interest is the case when the two basic layers are of the same optical thickness (usually $\lambda_0/4$ ), i.e. when

$$n_2h_2 = n_3h_3, \tag{90}$$

and the incidence is normal ( $\theta_1 = 0$ ). Then

$$\beta_2 = \beta_3 = \frac{2\pi}{\lambda_0}n_2h_2 = \frac{2\pi}{\lambda_0}n_3h_3, \tag{91}$$

and if we denote this common value by $\beta$, the argument of the Chebyshev polynomials reduces to

$$a = \cos^2\beta - \frac{1}{2}\left(\frac{n_2}{n_3} + \frac{n_3}{n_2}\right)\sin^2\beta. \tag{92}$$

It is seen that $a$ cannot exceed unity, but that for some values of $\beta$ it may become smaller than -1 . Then $\cos^{-1} a$ will be imaginary and consequently, since for any $\chi$,

$$\sin i\chi = i\sinh\chi = i\frac{e^{\chi} - e^{-\chi}}{2}$$

$\mathscr{U}_N$ will have exponential behaviour. It follows that the reflectivity of such a multilayer will increase rapidly with the number of the periods.

With quarter-wave films ( $n_2h_2 = n_3h_3 = \lambda_0/4$ ) at normal incidence (again assuming non-magnetic media),

$$\beta = \pi/2, \quad p_2 = n_2, \quad p_3 = n_3 \tag{93}$$

and (86) reduces to

$$\mathbf{M}_2(h) = \begin{bmatrix} -\frac{n_3}{n_2} & 0 \\ 0 & -\frac{n_2}{n_3} \end{bmatrix} \tag{94}$$

The characteristic matrix (87) of the multilayer whose basic period is such a double layer is, as can be directly verified by multiplying (94) $N$ times by itself,

$$M_{2N}(Nh) = \begin{bmatrix} \left(-\frac{n_3}{n_2}\right)^N & 0 \\ 0 & \left(-\frac{n_2}{n_3}\right)^N \end{bmatrix}. \tag{95}$$

Aocording to (49) and (51) the reflectivity is

$$\mathscr{R}_{2N} = \left(\frac{1 - \frac{n_l}{n_1}\left(\frac{n_2}{n_3}\right)^{2N}}{1 + \frac{n_l}{n_1}\left(\frac{n_2}{n_3}\right)^{2N}}\right)^2 \tag{96}$$

This shows that for a fixed number $N$ of the double layers, $\mathscr{R}_{2N}$ increases when the ratio $n_2/n_3$ is increased, and that if this ratio is fixed $\mathscr{R}_{2N}$ increases with $N$.

Sometimes, for example, for plate coatings of the Fabry-Perot Interferometer (cf. §7.6) the layers are arranged in succession characterized by the sequence $n_2, n_3,\ n_2, n_3, \ldots n_2, n_3, n_2$ of refractive indices. The characteristic matrix of this multilayer is

$$\mathbf{M}_{2N+1} = \mathbf{M}_{2N} \cdot \mathbf{M} \tag{97}$$

Reflectivity $\mathscr{R}_{2N+1}$ of multilayers formed by a periodic succession of quarter-wave films of zinc sulphide and cryolite at normal incidence ( $n_1 = 1, n_2 = 2 \cdot 3$, $n_3 = 1 \cdot 35, n_l = 1 \cdot 52, n_2h_2 = n_3h_3 = \lambda_0/4, \lambda_0 = 5460A, \theta_1 = 0$). The values in brackets are experimental results obtained by P. Glacomo, Compt. Rend. Acad. Sci., Paris, 235 (1952), 1627.

0

- A thorough discussion of the properties of a system of double layers will be found in a paper by C. Dufour and A. Herpin, Rev. Opt., 32 (1953), 321.

| $N$ | $\mathscr{R}_{2N+1}$ |
|---|---|
| 0 | 0.306 |
| 1 | 0.672 |
| 2 | 0.872(0.865) |
| 3 | 0.954(0.945) |
| 4 | 0.984(0.97) |

where $M_{2N}$ is given by (87) and $M$ is the characteristic matrix of the last film in the sequence. In particular with quarter-wave films at normal incidence, $\mathbf{M}_{2N}$ reduces to ( 95 ), $\beta_2 = \pi/2$, and ( 97 ) then becomes

$$\mathbf{M}_{2N+1} = \begin{bmatrix} 0 & -\frac{i}{n_2}\left(-\frac{n_3}{n_2}\right)^N \\ -in_2\left(-\frac{n_2}{n_3}\right)^N & 0 \end{bmatrix}. \tag{98}$$

Substitution into (49) and (51) gives the required reflectivity:

$$\mathscr{R}_{2N+1} = \left( \frac{1 - \left(\frac{n_2}{n_1}\right)\left(\frac{n_2}{n_l}\right)\left(\frac{n_2}{n_3}\right)^{2N}}{1 + \left(\frac{n_2}{n_1}\right)\left(\frac{n_2}{n_l}\right)\left(\frac{n_2}{n_3}\right)^{2N}} \right)^2. \tag{99}$$

The reflectivity is seen to increase rapidly with the ratio $n_2/n_3$ and with $N$ (see Table III).

## 9.3 Electromagnetic Potentials and Polarization

In the preceding chapter the effect of matter on an electromagnetic field was expressed in terms of a number of macroscopic constants. These have only a limited range of validity and are in fact inadequate to describe certain processes, such as the emission, absorption and dispersion of light. A full account of these phenomena would involve an extensive study of the atomistic theory and lies therefore outside the scope of this book.

It is possible, however, to describe the interaction of field and matter by means of a simple model which is entirely adequate for most branches of optics. For this purpose each of the vectors $D$ and $B$ is expressed as the sum of two terms.* Of these one is taken to be the vacuum field and the other is regarded as arising from the influence of matter. Thus one is led to the introduction of two new vectors for describing the effects of matter: the electric polarization ( $\boldsymbol{P}$ ) and the magnetic polarization or magnetization ( $M$ ). Instead of the material equations (10) and (11) in § 1.1 connecting $\boldsymbol{D}$ and $\boldsymbol{B}$ with $\boldsymbol{E}$ and $\boldsymbol{H}$, we now have equations connecting $\boldsymbol{P}$ and $\boldsymbol{M}$ with $\boldsymbol{E}$ and $H$. These new equations have a more direct physical meaning and lead to the following conception of the propagation of an electromagnetic field in matter:
An electromagnetic field produces at a given volume element certain amounts of polarization $\boldsymbol{P}$ and $\boldsymbol{M}$ which, in the first approximation, are proportional to the field, the constant of proportionality being a measure of the reaction of the field. Each volume element then becomes a source of a new secondary or scattered wavelet, whose strength is related in a simple way to $\boldsymbol{P}$ and $\boldsymbol{M}$. All the secondary wavelets combine with each other and with the incident field and form the total field, this being the one from which the whole consideration proceeds. By expressing this identity formally, we obtain two integral equations, † which may be shown to be equivalent to Maxwell's differential equations, but which describe the propagation of the electromagnetic field in a manner more clearly related to the atomic constitution of matter.

The two main results which will be obtained from the theory are: (1) The LorentzLorenz formula, which relates the macroscopic-optical properties of the medium to the number and the properties of the scattering particles, and (2) the so-called extinction theorem of Ewald and

Oseen, which shows how an external electromagnetic disturbance travelling with the velocity of the light in vacuum is exactly cancelled out and replaced in the substance by the secondary disturbance travelling with an appropriately smaller velocity.

The theory provides an alternative mathematical approach to the treatment of some problems of electromagnetic theory. This will be illustrated by the derivation of the laws of refraction and reflection and the Fresnel formulae, and will be further illustrated by the treatment of a more complex problem in Chapter XII.

In the derivation of these results some new mathematical apparatus will be required. Accordingly, we shall first discuss the representation of an electromagnetic field in terms of so-called retarded potentials, these being generalizations of the wellknown static potentials. The expressions of the potentials in terms of the polarization vectors lead to another set of auxiliary quantities, known as Hertz vectors. In § 2.1 and § 2.2 we discuss these mathematical preliminaries and in § 2.3 we briefly explain the underlying physical concepts. The integral equations and the two basic theorems already referred to will be derived in § 2.4.

## 9.3.1 2.1 The Electrodynamic Potentials in the Vacuum

### 2.1.1 The vector and scalar potentials

Let us consider an electromagnetic field in vacuum due to a given distribution of charges $\rho(r,t)$ and currents $j(r,t)$. It obeys Maxwell's equations § 1.1 (1)-(4) which may be written (since in vacuum $\boldsymbol{D} = \boldsymbol{E}$ and $\boldsymbol{B} = \boldsymbol{H}$ ) in the form:

$$\operatorname{curl} B - \frac{1}{c}\dot{E} = \frac{4\pi}{c} j \tag{1}$$

$$\operatorname{curl} E + \frac{1}{c}\dot{B} = 0 \tag{2}$$

$$\operatorname{div} E = 4\pi\rho \tag{3}$$

$$\operatorname{div} B = 0 \tag{4}$$

Since the divergence of the curl of any vector is zero, (4) will be satisfied if we set

$$\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}, \tag{5}$$

$A$ being an arbitrary vector function of position and time. If we substitute from (5) into Maxwell's second equation, we have

$$\operatorname{curl}\left(E + \frac{1}{c}\dot{A}\right) = 0 \tag{6}$$

(6) will hold if

$$E = -\frac{1}{c}\dot{A} - \operatorname{grad}\phi \tag{7}$$

0

- According to the remarks on p. 1, it would be more appropriate to write $\boldsymbol{H}$ in place of $\boldsymbol{B}$. This departure from the general theory is made here for the sake of the trivial convenience of having the positive signs on the right-hand side of eq. (2) in § 2.2, as is customary.
  † In the case of non-magnetic substances there is only one integral equation. The second integral equation reduces to a relatively simple expression for the magnetic field, and may be evaluated when the solution of the first equation has been determined.

The "material" integral equations discussed here must be distinguished from the "geometrical" integral equations used in the treatment of certain diffraction problems (cf. Chapter XI).

$\phi$ being an arbitrary scalar function. $A$ and $\phi$ must now be determined in such a way as to satisfy the remaining Maxwell equations.

Substituting from (5) and (7) into (1) and (3), and using the identities curl curl $\equiv$ grad div $-\nabla^2$ and div grad $\equiv \nabla^2$, we obtain

$$\nabla^2 A - \frac{1}{c^2}\ddot{A} - \operatorname{grad}\left(\operatorname{div} A + \frac{1}{c}\dot{\phi}\right) = -\frac{4\pi}{c} j \tag{8}$$

and

$$\nabla^2 \phi - \frac{1}{c^2}\ddot{\phi} + \frac{1}{c}\frac{\partial}{\partial t}\left(\operatorname{div} A + \frac{1}{c}\dot{\phi}\right) = -4\pi\rho \tag{9}$$

If we prescribe between $A$ and $\phi$ the relation

$$\operatorname{div} A + \frac{1}{c}\dot{\phi} = 0 \tag{10}$$

then (8) and (9) reduce to the inhomogeneous wave equations

$$\nabla^2 A - \frac{1}{c^2}\ddot{A} = -\frac{4\pi}{c} j \tag{11}$$

and

$$\nabla^2 \phi - \frac{1}{c^2}\ddot{\phi} = -4\pi\rho \tag{12}$$

The two functions $\boldsymbol{A}$ and $\phi$, from which $\boldsymbol{B}$ and $\boldsymbol{E}$ may be determined by means of the relations (5) and (7), are known as the magnetic vector potential and the electric scalar potential respectively. The relation (10), which couples the two potentials, is called the Lorentz condition. We note that it is consistent with the continuity equation § 1.1 (5)

$$\dot{\rho} + \operatorname{div} j = 0 \tag{13}$$

It is to be observed that (11), (12) and (10) do not define the two potentials uniquely. For if we add to $A$ the vector grad $\chi$, where $\chi$ is arbitrary, $B$ will remain unchanged, and if, in addition, $\phi$ is replaced by $\phi - \dot{\chi}/c$, $\boldsymbol{E}$ also will remain unchanged. In other words $\boldsymbol{B}$ and $\boldsymbol{E}$ are invariant under the transformation

$$A' = A + \operatorname{grad}\chi \tag{14a}$$

$$\phi' = \phi - \frac{1}{c}\dot{\chi} \tag{14b}$$

Now from (10) and (14),

$$\operatorname{div} A' + \frac{1}{c}\dot{\phi} + \left(\nabla^2\chi - \frac{1}{c^2}\ddot{\chi}\right) = 0 \tag{15}$$

Hence $A'$ and $\phi'$ will satisfy the Lorentz relation, if one imposes on $\chi$ the condition

$$\nabla^2\chi - \frac{1}{c^2}\ddot{\chi} = 0. \tag{16}$$

Eqs. (14), subject to the condition (16), express the so-called gauge transformation. The gauge transformation may be used to simplify the expressions for the field vectors. For example in a region where the charge density $\rho$ is zero, $\phi$ satisfies the homogeneous wave equation

$$\nabla^2\phi - \frac{1}{c^2}\ddot{\phi} = 0 \tag{17}$$

and $\chi$ may then be so chosen, that the scalar potential vanishes. According to (14b) and (16) it is only necessary to take

$$\chi = c\int \phi dt \tag{18}$$

The field can then be derived from the vector potential alone* by means of the relations (dropping the prime on $A$ )

$$B = \operatorname{curl} A, \quad E = -\frac{1}{c}\dot{A}, \tag{19}$$

whilst the Lorentz condition reduces to

$$\operatorname{div} A = 0. \tag{20}$$

### 2.1.2 Retarded potentials

We now consider the solution (subject to the relation (10)) of the inhomogeneous wave equations (11) and (12) for the vector and scalar potentials, and show first that these equations are satisfied by the following functions:

$$A(r,t) = \frac{1}{c}\int \frac{j\,(r', t - R/c)}{R} dV' \tag{21}$$

$$\phi(r,t) = \int \frac{\rho\,(r', t - R/c)}{R} dV' \tag{22}$$

Here

$$R = |r - r'| = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2} \tag{23}$$

is the distance from the point $r(x,y,z)$ to the volume element $dV'$ at $r'\,(x',y',z')$, the integration being carried throughout the whole space.

To verify that (22) satisfies the inhomogeneous wave equation for the scalar potential, we imagine that the point $r$ is surrounded by a sphere of radius $a$, centred on that point, and divide (22) into two parts:

$$\phi = \phi_1 + \phi_2, \tag{24}$$

where $\phi_1$ represents the contribution to the integral from the interior of the sphere and $\phi_2$ the contribution from the rest of space. Since $R \neq 0$ for every point $r'\,(x',y',z')$ outside the sphere it will be permissible to differentiate $\phi_2$ under the integral sign. We can then verify by a straightforward calculation that $\phi_2$ satisfies the homogeneous wave equation

$$\nabla^2\phi_2 - \frac{1}{c^2}\ddot{\phi}_2 = 0 \tag{25}$$

consisting, simply as it does, of a set of superposed spherical waves (cf. eq. (12), § 1.3). In the case of $\phi_1$ we must, however, proceed differently, since the integrand has a singularity at the centre $R = 0$.

We note that by making the radius of the sphere sufficiently small, we can secure (assuming $\rho$ to be a continuous function of $r$ and $t$ ) that for all points $r'$ inside the sphere $\rho\,(r', t - R/c)$

differs from $\rho(r,t)$ by an amount smaller than any prescribed value. Hence, as the radius $a$ tends to zero, $\nabla^2\phi_1$ will approach more and more closely the value of the electrostatic potential of a homogeneously charged sphere, with charge density $\rho$; i.e. for a sufficiently small $a$,

$$\nabla^2\phi_1 = -4\pi\rho(r,t) \tag{26}$$

Also, as $a \to 0, \ddot{\phi}_1 \to 0$. For if $a$ is sufficiently small we may write

$$\begin{aligned}\ddot{\phi}_1 &= \ddot{\rho}(r,t)\int_{R\leqslant a}\frac{dV'}{R}\\ &= 4\pi\ddot{\rho}\int_0^a Rdr\\ &= 2\pi a^2\ddot{\rho}\end{aligned} \tag{27}$$

and this tends to zero with $a$. Hence it follows from (24)-(26) that as $a \to 0$,

$$\nabla^2\phi - \frac{1}{c^2}\ddot{\phi} = \nabla^2(\phi_1+\phi_2) - \frac{1}{c^2}\left(\ddot{\phi}_1+\ddot{\phi}_2\right) \to -4\pi\rho(r,t), \tag{28}$$

so that (22) satisfies the inhomogeneous wave equation for the scalar potential. In a strictly analogous manner we can show that each Cartesian component of (21) is a solution of the corresponding scalar wave equation, whose inhomogeneous term contains, in place of $\rho$, the appropriate component of $j/c$. Hence (21) satisfies the inhomogeneous wave equation for the vector potential. Moreover, on account of the continuity relation (13), these solutions also satisfy the Lorentz relation (10).

The expressions (21) and (22) have a simple physical interpretation. They show that one may regard $\boldsymbol{A}$ and $\phi$ as arising from contributions from each volume element of space, a typical element $dV'$ contributing the amounts

$$\frac{1}{c}\frac{j(r',t-R/c)}{R} \quad \text{and} \quad \frac{\rho(r',t-R/c)}{R}$$

to $A$ and $\phi$ respectively. Now $R/c$ is precisely the time needed for light to travel from the point $r'$ to $r$, so that each contribution has to leave the element at such preceding time as to reach the point of observation at the required time $t$. For this reason (21) and (22) are called retarded potentials.*

(21) and (22) represent a special solution of the wave equations (11) and (12), namely that which arises from the given charges and currents. The general solution is obtained by adding to this the general solutions (again subject to the Lorentz condition) of the homogeneous wave equations

$$\nabla^2 A - \frac{1}{c^2}\ddot{A} = 0 \tag{29}$$

$$\nabla^2\phi - \frac{1}{c^2}\ddot{\phi} = 0 \tag{30}$$

0

- Eq. (19) expresses the field in a charge-free region in vacuum in terms of three scalar wave functions (Cartesian components of $A$ ). On account of (20) they are, however, not independent. It may be shown that in such a region the field can in fact be derived from two real scalar wave functions alone (see, for example, S. A. Schelkunoff, Electromagnetic Waves, New York, Van Nostrand, 1943, 382; H. S. Green and E. Wolf, Proc. Phys. Soc., A, 66 (1953), 1129; A. Nisbet, Proc. Roy. Soc., A, 231 (1955) 251).

0

## 9.3.2 2.2 Polarization and Magnetization

### 2.2.1 The potentials in terms of polarization and magnetization

In Chapter I the field was studied with the help of the material relations $\boldsymbol{D} = \varepsilon \boldsymbol{E}$, $\boldsymbol{B} = \mu \boldsymbol{H}$; these imply that the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ are regarded as producing at each point of the substance certain displacements $\boldsymbol{D}$ and $\boldsymbol{B}$ which are proportional to them. Instead of these "multiplicative relations" we shall now describe the interaction of matter and field by means of "additive relations",

$$D = E + 4\pi P \tag{1}$$
$$B = H + 4\pi M \tag{2}$$

Here $\boldsymbol{P}$ is called the electric polarization and $\boldsymbol{M}$ the magnetic polarization or magnetization. The physical meaning of these quantities will become clear later; here we only observe that $P$ and $M$ both vanish in vacuum so that these quantities represent the influence of matter on the field in a simple intuitive way.

Assume the material to be nonconducting ( $\sigma = 0$ ) and consider the field in a region where the current and charge densities are zero ($\rho = j = 0$). Eliminating $D$ and $H$ from Maxwell's equations § 1.1 (1)-(4) by means of (1) and (2), we obtain

$$\operatorname{curl} B - \frac{1}{c}\dot{E} = \frac{4\pi}{c}\tilde{j}, \tag{3}$$
$$\operatorname{curl} E + \frac{1}{c}\dot{B} = 0, \tag{4}$$
$$\operatorname{div} E = 4\pi\tilde{\rho}, \tag{5}$$
$$\operatorname{div} B = 0, \tag{6}$$

where the free current density $\tilde{\boldsymbol{j}}$ and the free charge density $\tilde{\boldsymbol{\rho}}$ are defined by the relations

$$\tilde{\boldsymbol{j}} = \dot{\boldsymbol{P}} + c \operatorname{curl} \boldsymbol{M}, \tag{7}$$
$$\tilde{\boldsymbol{\rho}} = -\operatorname{div} \boldsymbol{P}. \tag{8}$$

Eqs. (3)-(6) are formally identical with the equations for the field in vacuum considered in the previous section. Hence, as in the vacuum case, we may introduce a vector potential $A$ and a scalar potential $\phi$ such that

$$B = \operatorname{curl} A \tag{9}$$
$$E = -\frac{1}{c}\dot{A} - \operatorname{grad} \phi \tag{10}$$

Moreover, if we impose, as before, the Lorentz condition

- It is also possible to construct solutions in the form of advanced potentials (with $t + R/c$ in place of $t - R/c$). These represent the effect of incoming spherical waves, whereas the retarded potentials represent the effect of outgoing spherical waves.

At the cost of some formal asymmetry, we may restrict our attention to solutions in terms of retarded potentials only. When the interaction of radiation and matter is considered thermodynamically, then there is an asymmetry in the physical situation which reinforces this choice (cf. M. Born, Natural Philosophy of Cause and Chance, Oxford, Clarendon Press, 1949; Dover Publications, New York, 1964; especially p. 26).

$$\operatorname{div} A + \frac{1}{c}\dot{\phi} = 0, \tag{11}$$

we have, by analogy with (11) and (12) in § 2.1, the following equations for $\boldsymbol{A}$ and $\phi$ :

$$\nabla^2 A - \frac{1}{c^2}\ddot{A} = -\frac{4\pi}{c}\tilde{j}, \tag{12}$$

$$\nabla^2 \phi - \frac{1}{c^2}\phi = -4\pi\tilde{\rho}. \tag{13}$$

Now condition (11) is consistent with (12) and (13) provided that

$$\frac{\partial\tilde{\rho}}{\partial i} + \operatorname{div}\tilde{j} = 0; \tag{14}$$

this relation is, in fact, satisfied identically, as can be immediately seen from (7) and (8), on using the vector identity div curl $\equiv 0$.

In terms of the polarization and magnetization, the solutions of (12) and (13) may, according to (21) and (22) in § 2.1, and by (7) and (8), be expressed in the form

$$\phi = -\int \frac{1}{R}\left[\operatorname{div}' P\right] dV' \tag{15}$$

$$A = \int \frac{1}{R}\left[\operatorname{curl}' M + \frac{1}{c}\dot{P}\right] dV' \tag{16}$$

Here the differential operators div' and curl' are taken with respect to the coordinates ( $x', y', z'$ ) of the variable point of integration $r'$ at which the volume element $dV'$ is situated, and the square brackets denote retarded values, i.e. they imply that the argument $t$ is replaced by $t - R/c$ inside each bracket.

By a straightforward calculation the following identities can be verified:

$$\operatorname{div}'[P] = [\operatorname{div}' P] + \frac{1}{cR} R \cdot [\dot{P}] \tag{17a}$$

$$\operatorname{curl}'[M] = [\operatorname{curl}' M] + \frac{1}{cR} R \wedge [\dot{M}] \tag{17b}$$

where

$$\boldsymbol{R} = \boldsymbol{r} - \boldsymbol{r}' \tag{18}$$

Hence (15) and (16) may be written as

$$\phi = -\int \left\{ \frac{1}{R}\operatorname{div}'[P] - \frac{1}{cR^2} R \cdot [\dot{P}] \right\} dV' \tag{19}$$

$$A = \int \left\{ \frac{1}{R}\operatorname{curl}'[M] - \frac{1}{cR^2} R \wedge [\dot{M}] + \frac{1}{cR}[\dot{P}] \right\} dV' \tag{20}$$

From the first term in each of these integrals we can separate an integral over the boundary surface. To do this we use the vector identities

$$\operatorname{div}\left(\frac{1}{R}P\right) = \frac{1}{R}\operatorname{div}P + P\cdot\operatorname{grad}\frac{1}{R} \tag{21a}$$

$$\operatorname{curl}\left(\frac{1}{R}M\right) = \frac{1}{R}\operatorname{curl}M - M\wedge\operatorname{grad}\frac{1}{R} \tag{21b}$$

Integrating these relations throughout an arbitrary finite domain, we obtain, on using the Gauss theorem,

$$\int\frac{1}{R}(n\cdot P)dS = \int\left\{\frac{1}{R}\operatorname{div}P + P\cdot\operatorname{grad}\frac{1}{R}\right\}dV \tag{22a}$$

$$\int\frac{1}{R}(n\wedge M)dS = \int\left\{\frac{1}{R}\operatorname{curl}M - M\wedge\operatorname{grad}\frac{1}{R}\right\}dV \tag{22b}$$

the integrals on the left extending over the surface bounding the domain, $n$ being the unit outward normal to the surface.

Assume now that the substance (i.e. the region of space where $\boldsymbol{P}$ and $\boldsymbol{M}$ differ from zero) remains within a finite closed surface. If then the integrals (19) and (20) are extended over the volume inside this surface, the surface integrals in (22) vanish and (19) and (20) may be expressed in the form

$$\phi = \int\left\{[\boldsymbol{P}]\cdot\operatorname{grad}'\frac{1}{R} + \frac{1}{cR^2}R\cdot[\dot{\boldsymbol{P}}]\right\}dV' \tag{23}$$

$$A = \int\left\{[M]\wedge\operatorname{grad}'\frac{1}{R} - \frac{1}{cR^2}R\wedge[\dot{\boldsymbol{M}}] + \frac{1}{cR}[\dot{\boldsymbol{P}}]\right\}dV' \tag{24}$$

From (23) and (24) the physical significance of $\boldsymbol{P}$ and $\boldsymbol{M}$ can be appreciated. For this purpose consider the case when $\boldsymbol{P}$ and $\boldsymbol{M}$ are zero everywhere except within a

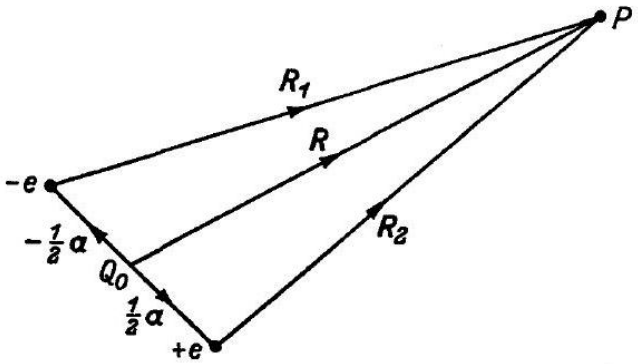



Fig. 2.1. Notation used in calculating the potentials of an electric dipole.

vanishingly small volume element at $r_0\,(x_0, y_0, z_0)$. We may express this assumption formally with the help of Dirac's delta function (see Appendix IV):

$$\boldsymbol{P}\,(\boldsymbol{r}', t) = \boldsymbol{p}(t)\delta\,(\boldsymbol{r}' - \boldsymbol{r}_0) \tag{25}$$

$$\boldsymbol{M}\,(\boldsymbol{r}', t) = \boldsymbol{m}(t)\delta\,(\boldsymbol{r}' - \boldsymbol{r}_0) \tag{26}$$

(23) and (24) then become

$$\phi = [p] \cdot \operatorname{grad}_0 \frac{1}{R} + \frac{1}{cR^2} R \cdot [\dot{p}], \tag{27}$$

$$A = [m] \wedge \operatorname{grad}_0 \frac{1}{R} - \frac{1}{cR^2} R \wedge [\dot{m}] + \frac{1}{cR} [\dot{p}], \tag{28}$$

where now

$$\boldsymbol{R} = r - r_0, \quad R = |r - r_0|, \tag{29}$$

and $\operatorname{grad}_0$ implies that the operator is taken with respect to the coordinates $x_0, y_0, z_0$ of the point $\boldsymbol{r}_0$.

Eqs. (27) and (28) have a simple interpretation. Consider the electrostatic potential of two fixed electric charges $-e$ and $+e$ situated at points whose position vectors with respect to a fixed point $Q_0\,(r_0)$ are $-\frac{1}{2}a$ and $\frac{1}{2}a$ respectively. The Coulomb potential $\phi(\boldsymbol{R})$ due to the charges is (see Fig. 2.1):

$$\phi = \frac{e}{R_2} - \frac{e}{h_1}. \tag{30}$$

If $a$ is sufficiently small, we may expand $1/R_2$ in powers of the components of $a$ and obtain

$$\frac{1}{R_2} = \frac{1}{R_1} + a \cdot \operatorname{grad} \frac{1}{R_1} + \ldots, \tag{31}$$

where the operator grad is taken with respect to the coordinates of the point at which the charge $-e$ is situated. If we neglect the higher-order terms in (31), (30) becomes

$$\phi = ea \cdot \operatorname{grad} \frac{1}{R_1} = e \frac{a \cdot R_1}{{R_1}^3}. \tag{32}$$

Suppose now that $a$ is gradually decreased whilst $e$ is increased without limits in such a way that $ea$ approaches a finite value $p$ :

$$\lim_{\boldsymbol{a} \to \boldsymbol{0}} \mathbf{ea} = \boldsymbol{p}. \tag{33}$$

Then $R_1$ approaches $R$ and (32) becomes in the limit

$$\phi = p \cdot \operatorname{grad}_0 \frac{1}{R}. \tag{34}$$

This expression is identical with (27) in the special case when $p$ is independent of time.
If the charges depend on time but at each instant of time differ only in sign, we bave, in place of (30),

$$\phi = \frac{e\,(t - R_2/c)}{R_2} - \frac{e\,(t - R_1/c)}{R_1} \tag{35}$$

and obtain by the same limiting process

$$\phi = [p] \cdot \operatorname{grad}_0 \frac{1}{R} + \frac{1}{cR^2} R \cdot [\dot{\boldsymbol{p}}],$$

which is (27). Thus (23) may be interpreted as the scalar potential of a distribution of electric dipoles of moment $\boldsymbol{P}$ per unit volume. The last term in (28) may be shown to be the magnetic potential arising from these dipoles.

In a similar way (28) may be shown to be the vector potential of a magnetic dipole of moment $m(t)$, such a dipole being equivalent, of course, to an infinitesimal closed circuit of area $A$ normal to $m$, carrying the current* $cm/A$. Hence the first two terms in (24) may be interpreted as the vector potential of a distribution of magnetic dipoles of moment $M$ per unit volume.

### 2.2.2 Hertz vectors

Instead of using the potentials $A$ and $\phi$ one can represent the field in terms of another pair of potential functions $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ which depend on $\boldsymbol{P}$ and $\boldsymbol{M}$ in a much simpler way. They are known as the Hertz vectors or Polarization Potentials † and may be introduced by means of the relations

$$A = \frac{1}{c}\dot{\Pi}_e + \operatorname{curl}\Pi_m \tag{36}$$

$$\phi = -\operatorname{div}\Pi_e \tag{37}$$

$\Pi_e$ and $\Pi_m$ are seen to bear the same mathematical relationship to $A$ and $\phi$ as the polarizations $\boldsymbol{P}$ and $\boldsymbol{M}$ bear to $\tilde{j}/c$ and $\tilde{\rho}$.

The Lorentz condition is now automatically satisfied and the eqs. (12) and (13) for $A$ and $\phi$ will also be satisfied if $\Pi_e$ and $\Pi_m$ are solutions of the inhomogeneous wave equations

$$\nabla^2\Pi_e - \frac{1}{c^2}\ddot{\Pi}_e = -4\pi P \tag{38}$$

$$\nabla^2\Pi_m - \frac{1}{c^2}\ddot{\Pi}_m = -4\pi M \tag{39}$$

According to § 2.1, particular solutions of these equations may be expressed in the form of retarded potentials

$$\mathbf{\Pi}_e = \int \frac{[\boldsymbol{P}]}{R}dV' \tag{40}$$

$$\mathbf{\Pi}_m = \int \frac{[\boldsymbol{M}]}{R}dV' \tag{41}$$

The field vectors may be expressed directly in terms of $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ by simple differentiation. We have from (9) and (36),

$$\boldsymbol{B} = \operatorname{curl}\left(\frac{1}{c}\dot{\mathbf{\Pi}}_e + \operatorname{curl}\mathbf{\Pi}_m\right) \tag{42}$$

0

- For a discussion of Ampre's infinitesimal currents see any standard textbook on electricity, or Max Born, Optik (Berlin, Springer, 1933), pp. 305-306.
†$\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ have similar transformation properties as $\boldsymbol{E}$ and $\boldsymbol{B}$ (or $\boldsymbol{D}$ and $\boldsymbol{H}$ ), i.e. they form a six vector (antisymmetric tensor of the second order). They are generalizations of a certain potential function introduced for the electromagnetic field of an oscillating dipole by H. Hertz, Ann. d. Physik., 36 (1889), 1. The vector character of the Hertz potential was noted by A. Righi, Nuovo Cimento, 2 (1901), 104, who introduced the corresponding magnetic potential. A general theory of the Hertz vectors and the associated gauge transformation is due to F. Kottler, Ann. d. Physik, (4), 71 (1923), 462 and A. Nisbet, Proc. Roy. Soc., A 231 (1955), 250. See also W. H. McCrea, ibid., 240 (1957), 447.

and from (10), (36) and (37),

$$\begin{aligned}\boldsymbol{E} &= -\frac{1}{c}\frac{\partial}{\partial t}\left(\frac{1}{c}\dot{\Pi}_e + \operatorname{curl}\Pi_m\right) + \operatorname{grad}\operatorname{div}\Pi_e \\ &= \operatorname{curl}\left(-\frac{1}{c}\dot{\Pi}_m + \operatorname{curl}\Pi_e\right) + \left(\nabla^2\Pi_e - \frac{1}{c^2}\ddot{\Pi}_e\right),\end{aligned} \tag{43}$$

where the identity grad div $\equiv$ curl curl $+\nabla^2$ has been used. According to (38) the term in the second bracket is equal to $-4\pi P$; hence, using (1),

$$D = \operatorname{curl}\left(-\frac{1}{c}\dot{\boldsymbol{\Lambda}}_m + \operatorname{curl}\boldsymbol{\Pi}_e\right) \tag{44}$$

Finally, from (2), (42) and (39)

$$\begin{aligned}\boldsymbol{H} &= \operatorname{curl}\left(\frac{1}{c}\dot{\Pi}_e + \operatorname{curl}\Pi_m\right) + \left(\nabla^2\Pi_m - \frac{1}{c^2}\ddot{\Pi}_m\right) \\ &= \frac{1}{c}\frac{\partial}{\partial t}\left(-\frac{1}{c}\dot{\Pi}_m + \operatorname{curl}\Pi_e\right) + \operatorname{grad}\operatorname{div}\Pi_m\end{aligned} \tag{45}$$

It was shown in § 2.1 that the potentials $A$ and $\phi$ associated with a given field are not unique; any other pair ( $A', \phi'$ ) which can be derived from a possible ( $A, \phi$ ) by means of a gauge transformation § 2.1 (14) will represent the same field. The Hertz vectors too are not unique. They may be subjected to the following transformation, which leaves the field vectors invariant:

$$\left.\begin{aligned}\boldsymbol{\Pi}'_e &= \boldsymbol{\Pi}_e + \operatorname{curl}\boldsymbol{F} - \operatorname{grad} G, \\ \boldsymbol{\Pi}'_m &= \boldsymbol{\Pi}_m - \tfrac{1}{c}\dot{\boldsymbol{F}},\end{aligned}\right\} \tag{46}$$

where the vector function $F$ and the scalar function $G$ are any solutions of the homogeneous wave equations

$$\left.\begin{aligned}\nabla^2 F - \tfrac{1}{c^2}\ddot{F} = 0 \\ \nabla^2 G - \tfrac{1}{c^2}\ddot{G} = 0\end{aligned}\right\} \tag{47}$$

The invariance under this transformation is readily verified by substituting from (46) into (36) and (37). $A$ and $\phi$ are then transformed into $A'$ and $\phi'$ according to (14), § 2.1, with

$$\chi = -\frac{1}{c}\dot{G}$$

Moreover, substitution from (46) into (38) and (39) shows that $\Pi'_e$ and $\Pi'_m$ also satisfy the inhomogeneous wave equations for the Hertz vectors.

### 2.2.3 The field of a linear electric dipole

For the purpose of later applications it will be useful to write down the full explicit solutions of the field equations for a field in vacuum produced by a linear electric dipole, situated at a point $r_0$ and vibrating in a fixed direction specified by a unit vector $n$. Such a dipole is characterized by an electric polarization

$$\boldsymbol{P}(\boldsymbol{r}, t) = p(t)\delta\left(\boldsymbol{r} - \boldsymbol{r}_0\right)\boldsymbol{n} \tag{48}$$

where $p$ is a function of time and $\delta$ the Dirac delta function. A general dipole whose orientation $n$ is also a function of time is equivalent to three linear dipoles, with their moment vectors oriented along three mutually perpendicular directions.

According to (40) the electric Hertz vector associated with (48) is

$$\mathbf{\Pi}_e = \frac{p(t - R/c)}{R}\boldsymbol{n} \tag{49}$$

where $R$ is the distance of the point $r$ from the point $r_0$.
Since $\Pi_m = 0$ and $\Pi_e$ satisfies (except at the origin) the homogeneous wave equation for vacuum, the eqs. (42)-(45) become

$$\boldsymbol{E} = \boldsymbol{D} = \text{ curl curl } \mathbf{\Pi}_e \tag{50}$$

$$\boldsymbol{B} = \boldsymbol{H} = \frac{1}{c}\,\text{curl}\,\dot{\mathbf{\Pi}}_e \tag{51}$$

Using the identity curl curl $\equiv$ grad div $-\nabla^2$ and the relation $\nabla^2\Pi_e = \frac{1}{c^2}\ddot{\Pi}_e$, (50) may be re-written in the form

$$\boldsymbol{E} = \boldsymbol{D} = \text{grad div}\,\mathbf{\Pi}_e - \frac{\mathbf{1}}{c^2}\ddot{\mathbf{\Pi}}_e \tag{52}$$

From (49) we obtain, after a simple calculation,

$$\begin{aligned}
\text{div}\,\Pi_e &= -\left\{\frac{[p]}{R^3} + \frac{[\dot{p}]}{cR^2}\right\}(\boldsymbol{n}\cdot\boldsymbol{R}) \\
\text{grad div}\,\Pi_e &= \left\{\frac{3[p]}{R^5} + \frac{3[\dot{p}]}{cR^4} + \frac{[\ddot{p}]}{c^2R^3}\right\}(\boldsymbol{n}\cdot\boldsymbol{R})\boldsymbol{R} - \left\{\frac{[p]}{R^3} + \frac{[\dot{p}]}{cR^2}\right\}\boldsymbol{n} \\
\text{curl}\,\Pi_e &= \left\{\frac{[p]}{R^3} + \frac{[\dot{p}]}{cR^2}\right\}(\boldsymbol{n}\wedge\boldsymbol{R})
\end{aligned}$$

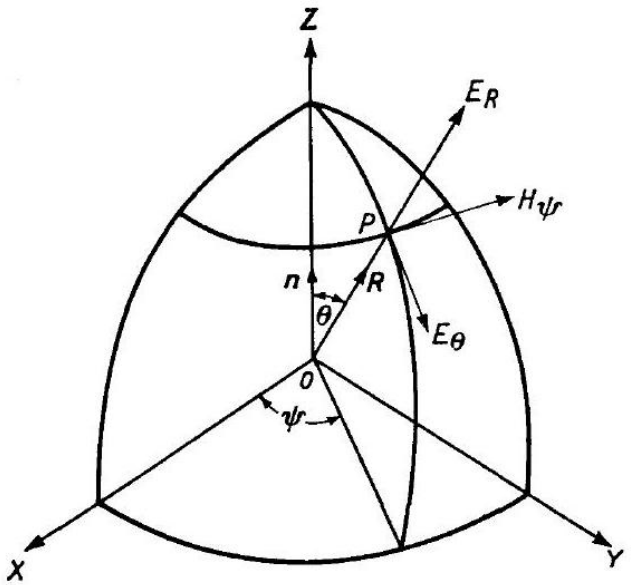



Fig. 2.2. Calculation of the field of a linear electric dipole with moment-vector along the $z$-axis.

where, as before, the square brackets denote retarded values. Substituting into (52) and (51) we obtain the required formulae for the field vectors:

$$\boldsymbol{E} = \boldsymbol{D} = \left\{ \frac{3[p]}{R^5} + \frac{3[\dot{p}]}{cR^4} + \frac{[\ddot{p}]}{c^2R^3} \right\} (\boldsymbol{n} \cdot \boldsymbol{R})\boldsymbol{R} - \left\{ \frac{[p]}{R^3} + \frac{[\dot{p}]}{cR^2} + \frac{[\ddot{p}]}{c^2R} \right\} \boldsymbol{n} \tag{53}$$

$$\boldsymbol{B} = \boldsymbol{H} = \left\{ \frac{[\dot{p}]}{cR^3} + \frac{[\ddot{p}]}{c^2R^2} \right\} (\boldsymbol{n} \wedge \boldsymbol{R}) \tag{54}$$

We shall also later need expressions for $\boldsymbol{E}$ and $\boldsymbol{H}$ in terms of spherical polar coordinates. Taking the $n$ direction as the $z$-axis (see Fig. 2.2), and denoting by $i_R, i_\theta$ and $\boldsymbol{i}_\psi$ the unit vectors in the direction of increasing $R, \theta$ and $\psi$ we have

$$\begin{aligned} \boldsymbol{R} = R\boldsymbol{i}_{\boldsymbol{R}}, \quad \boldsymbol{n} = (\cos\theta)\boldsymbol{i}_{\boldsymbol{R}} - (\sin\theta)\boldsymbol{i}_\theta \\ (\boldsymbol{n} \cdot \boldsymbol{R}) = R\cos\theta, \quad (\boldsymbol{n} \wedge \boldsymbol{R}) = (R\sin\theta)\boldsymbol{i}_{\boldsymbol{\psi}} \end{aligned} \tag{55}$$

and (53) and (54) give

$$E = E_R i_R + E_\theta i_\theta, \quad H = H_\psi i_\psi, \tag{56}$$

where

$$\left.\begin{aligned} E_R &= 2\left(\tfrac{[p]}{R^3} + \tfrac{[\dot{p}]}{cR^2}\right)\cos\theta \\ E_\theta &= \left(\tfrac{[p]}{R^3} + \tfrac{[\dot{p}]}{cR^2} + \tfrac{[\ddot{p}]}{c^2R}\right)\sin\theta \\ H_v &= \left(\tfrac{[\dot{p}]}{cR^2} + \tfrac{[\ddot{p}]}{c^2R}\right)\sin\theta \end{aligned}\right\} \tag{57}$$

Thus the electric vector lies in the meridional plane through the axis of the dipole, and the magnetic vector is perpendicular to this plane.

Of particular interest is the field in regions which are so far away from the dipole that all terms but those in $1/R$ may be neglected in the above expressions. This wave- (or radiation-) zone is characterized by

$$R \gg c\left|\frac{p}{\dot{p}}\right|, \quad R \gg c\left|\frac{\dot{p}}{\ddot{p}}\right|. \tag{58}$$

In this region

$$E_\theta \sim H_\varphi \sim \frac{[\ddot{p}]}{c^2R}\sin\theta \tag{59}$$

and the other components are negligible. Hence, in the wave zone, $\boldsymbol{E}$ and $\boldsymbol{H}$ are of equal magnitude and perpendicular to each other and to the radius vector $R$, which now coincides with the direction of the Poynting vector. In this zone the structure of the field of a linear electric dipole is therefore similar to that of a plane wave. However, on each "wave surface" (sphere centred on $r_0$ ) the field vectors now vary from point to point, diminishing in magnitude from the equator to the poles, being zero on the axis of the oscillator. The dipole therefore does not radiate energy in the direction of its axis.

Let us now calculate the amount of energy radiated per second across each of the spherical wave surfaces. It is given by the integral of the magnitude $S$ of the Poynting vector taken over this surface. In the wave zone,

$$S = \frac{c}{4\pi}|E \wedge H| = \frac{c}{4\pi}\,|E_\theta H_\psi| = \frac{[\ddot{p}]^2}{4\pi c^3 R^2}\sin^2\theta \tag{60}$$

Hence the total amount of energy which is radiated across the spherical surface per second is ( $d\sigma$ denoting element of the surface)

$$\int S d\sigma = \frac{[\ddot{p}]^2}{4\pi c^3 R^2} \int_0^\pi \sin^2\theta \cdot 2\pi R^2 \sin\theta d\theta$$
$$= \frac{2[\ddot{p}]^2}{3c^3} \tag{61}$$

Consider the special case when $p(t)$ is periodic with angular frequency $\omega$ :

$$p(t) = p_0 e^{-i\omega t} \tag{62}$$

where $p_0$ is a complex constant, and the real part of (62) is understood to represent $p$. The two conditions (58) now reduce to the single condition

$$R \gg \lambda/2\pi, \quad (\lambda = 2\pi c/\omega) \tag{63}$$

and (59) gives, for the non-vanishing components of the field in the wave zone,

$$E_\theta = H_\psi \sim -\left(\frac{\omega}{c}\right)^2 p_0 \sin\theta \frac{e^{-i\omega(t-R/c)}}{R} \tag{64}$$

it being understood that the real part of the expression on the right-hand side of (64) is taken. The amount of energy crossing per second a unit area of the spherical surface in the wave zone is

$$S = \frac{c}{4\pi}\left(\frac{\omega}{c}\right)^4 \frac{|p_0|^2}{R^2} \sin^2\theta \cos^2[\omega(t - R/c) - \alpha] \tag{65}$$

$\alpha$ being the phase of $p_0$.

On taking the time average over an interval of time which is large compared with the period $T = 2\pi/\omega$, (65) reduces to

$$\langle S\rangle = \frac{|p_0|^2}{8\pi R^2}\frac{\omega^4}{c^3}\sin^2\theta \tag{66}$$

The (time averaged) energy crossing the whole surface per second is therefore

$$\int \langle S\rangle \delta\sigma = \frac{1}{3}\frac{|p_0|^2\omega^4}{c^3} \tag{67}$$

## 9.3.3 2.3 The Lorentz-Lorenz Formula and Elementary Dispersion Theory*

### 2.3.1 The dielectric and magnetic susceptibilities

The interpretation of the potentials in terms of polarization and magnetization is of fundamental importance in the atomic theory of matter. From the standpoint of such a theory the additive relations § 2.2 (1) and (2) connecting $\boldsymbol{E}$ with $\boldsymbol{D}$ and $\boldsymbol{P}$ and $\boldsymbol{H}$ with $\boldsymbol{B}$ and $\boldsymbol{M}$ have a more direct physical meaning than the multiplicative relations (the material equations (10) and (11), in § 1.1).

In the atomic theory one regards matter as composed of interacting particles (atoms and molecules) embedded in the vacuum. These entities produce a field which has large local variations in the interior of the matter. This internal field is modified by any field which is applied externally; the properties of the matter are then derived by averaging over the total field within it. As long as the region over which the average is taken is large compared with the linear

dimensions of the particles, the electromagnetic properties of each can be simply described by an electric and a magnetic dipole; the secondary field is then the field due to these dipoles (with retardation). This is, in fact, exactly what we have just described by regarding matter as a continuous distribution interacting with the field: it corresponds to the first approximation (for slow variation in space) of the atomic theory. In this approximation for sufficiently weak fields † one can assume $\boldsymbol{P}$ and $\boldsymbol{M}$ to be proportional to $\boldsymbol{E}$ and $\boldsymbol{H}$ respectively:

$$\boldsymbol{P} = \eta \boldsymbol{E}, \quad \boldsymbol{M} = \chi \boldsymbol{H}. \tag{1}$$

The factor $\eta$ is called the dielectric susceptibility and $\chi$ the magnetic susceptibility, the latter being a concept current in practical magnetism. From (1) and from eq. (1) and (2) in § 2.2 it follows, on comparison with the material equations (10) and (11) in § 1.1, that the dielectric constant $\varepsilon$ and the magnetic permeability $\mu$ are related to the dielectric and magnetic susceptibilities by the formulae

$$\varepsilon = 1 + 4\pi\eta, \quad \mu = 1 + 4\pi\chi \tag{2}$$

By means of (2) the complete formal equivalence of the "multiplicative" and "additive" treatments is secured. But such formal comparison of the constants makes no use of any of the consequences of the atomistic structure. While a detailed discussion involving the results of the atomistic theory is outside the scope of this book, it will be useful to introduce some of the relevant concepts and formulae. With their help we shall obtain a clearer understanding of the physical content of the integral equations which will be introduced in § 2.4 in place of the usual differential equations of Maxwell's theory.

The simplest assumption which leads one step into the atomistic theory is to regard matter as being made up of certain physical entities-the molecules-which are polarizable, so that under the influence of an external field they show electric and magnetic moments. In the first approximation it may be assumed that the components of these moments are linear functions of the components of the field; in general the direction of the moment vector does not coincide with the direction of the field. This assumption has many consequences, which can only be briefly studied here. We shall confine our attention to isotropic non-magnetic substances, and examine first the dependence of the electric constants on the density in a substance composed of similar molecules. We shall also study the dependence of the refractive index on frequency. Only somewhat oversimplified arguments can be put forward here, but a rigorous, though a more formal, derivation of the main result (the Lorentz-Lorenz formula) will be given in § 2.4.

### 2.3.2 The effective field

To begin with we must distinguish between the effective field $\boldsymbol{E}', \boldsymbol{H}'$ acting on a molecule and the mean or observed field $\boldsymbol{E}, \boldsymbol{H}$ obtained by averaging over a region which contains a great number of molecules. The difference between the two fields is due to the gaps between

---

0

- In this section we depart from our deductive treatment in order to obtain a clearer understanding of the phenomena before formulating the integral equations which will be introduced in §2.4 in place of Maxwell's differential equations. The considerations of this section are more illustrative than rigorous, since a rigorous treatment would involve most of the atomistic theory of matter and take us far beyond the reasonable bounds of this book.
  † A field which is considered strong on the atomic level is one in which the electrostatic energy of an electron changes by an appreciable fraction of the ionization potential of the atom in a distance equal to the atomic diameter, i.e. a field of millions of volts per centimetre. Electrical breakdowns in real materials actually occur at much smaller fields.

the molecules and depends on the number of molecules per unit volume (their number density $N$ ).

To estimate the difference $\boldsymbol{E}' - \boldsymbol{E}$, consider a particular molecule and imagine it to be surrounded by a small sphere whose radius is nevertheless large compared with its linear dimensions. We shall consider separately the effects on the central molecule produced by the matter outside and inside this sphere.

In determining the effect of the matter outside the sphere we may clearly neglect the molecular structure and treat the substance as continuous. We may then assume that, outside the sphere, the polarization $\boldsymbol{P}$ produced by the mean electric field is constant. Concerning the effect of the molecules inside the sphere we shall assume, as may be shown for a number of important special cases, including that of random distribution, that they do not produce any resulting field at the central molecule. Hence we may regard the molecule as being situated in a spherical region, inside which there is vacuum and outside which there is a homogeneously polarized medium. We must then determine the potential $\phi$ of this configuration, i.e. of the free charges on the spherical discontinuity surface at which $\boldsymbol{P}$ changes from the interior value zero to a constant exterior value.

For this purpose consider the potential $\tilde{\phi}$ of the "complementary" configuration, namely of a homogeneously polarized sphere surrounded by vacuum. The superposition of these two configurations is a homogeneously polarized substance with no boundary. Hence the potential due to the boundary is zero and we have

$$\phi + \tilde{\phi} = 0. \tag{3}$$

Now the potential $\tilde{\phi}$ can be immediately obtained from § 2.2 (23) by taking $P$ to be constant. This gives

$$\phi = -\tilde{\phi} = -\boldsymbol{P} \cdot \int \operatorname{grad}' \frac{1}{R} dV' \tag{4}$$

Since

$$R = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2},$$

we may replace grad' by - grad in (4), and write

$$\phi = -\tilde{\phi} = \boldsymbol{P} \cdot \operatorname{grad} \int \frac{dV'}{R} = -\boldsymbol{P} \cdot \operatorname{grad} \phi_0, \tag{5}$$

where

$$\phi_0 = -\int \frac{dV'}{R}. \tag{6}$$

(6) may be interpreted as the potential of a uniformly charged sphere of charge density - 1. Hence it satisfies the Poisson equation

$$\nabla^2 \phi_0 = 4\pi. \tag{7}$$

Now the components of the field strengths associated with the potential (5) are

$$\begin{aligned} -\frac{\partial \phi}{\partial x} &= \frac{\partial}{\partial x}\left[P_x \frac{\partial \phi_0}{\partial x} + P_y \frac{\partial \phi_0}{\partial y} + P_z \frac{\partial \phi_0}{\partial z}\right] \\ &= P_x \frac{\partial^2 \phi_0}{\partial x^2} + P_y \frac{\partial^2 \phi_0}{\partial x \partial y} + P_z \frac{\partial^2 \phi_0}{\partial x \partial z} \end{aligned} \tag{8}$$

etc. At the centre of the sphere we have, by symmetry,

$$\frac{\partial^2\phi_0}{\partial x\partial y} = \frac{\partial^2\phi_0}{\partial y\partial z} = \frac{\partial^2\phi_0}{\partial z\partial x} = 0, \tag{9}$$

and

$$\frac{\partial^2\phi_0}{\partial x^2} = \frac{\partial^2\phi_0}{\partial y^2} = \frac{\partial^2\phi_0}{\partial z^2} \tag{10}$$

Using (7) it follows that each term in (10) is equal to $4\pi/3$; (8) then shows that the required contribution to the effective field strength is

$$-\nabla\phi = \frac{4\pi}{3}\boldsymbol{P} \tag{11}$$

The total field inside the sphere, which is the effective field acting on the central molecule, is obtained by adding to this the mean field $\boldsymbol{E}$, giving

$$E' = E + \frac{4\pi}{3}P. \tag{12}$$

For magnetizable substances there is a corresponding relation between $\boldsymbol{H}', \boldsymbol{H}$ and $\boldsymbol{M}$. As we shall however treat only non-magnetic substances we shall always have $\boldsymbol{H}' = \boldsymbol{H}$.

Next we must connect the polarization $P$ with the density.

### 2.3.3 The mean polarizability: the Lorentz-Lorenz formula

As already explained, it will be assumed that, for each molecule, the electric dipole moment $\boldsymbol{p}$ established under the influence of the field is proportional to the effective field* $\boldsymbol{E}'$

$$\boldsymbol{p} = \alpha\boldsymbol{E}'. \tag{13}$$

The molecule is here assumed to be isotropic; † but as we are interested only in the average effect over all possible orientations of the molecule, it will not be necessary to assume that each individual molecule is isotropic. $\alpha$ will then be regarded as representing the mean polarizability, and since $\boldsymbol{p}$ is of dimensions [ $el$ ] and $\boldsymbol{E}'$ of dimensions $[e^{-2}]$ ( $e$ = charge, $l$ = length), $\alpha$ is seen to have the dimensions [ $l^3$ ], i.e. those of a volume.

If, as before, $N$ is the number of molecules per unit volume, the total electric moment $\boldsymbol{P}$ per unit volume is given by

$$\boldsymbol{P} = N\boldsymbol{p} = N\alpha\boldsymbol{E}'. \tag{14}$$

On eliminating $\boldsymbol{E}'$ between (12) and (14) we obtain the first formulae (1), with the following explicit expression for the dielectric susceptibility:

$$\eta = \frac{N\alpha}{1 - \frac{4\pi}{3}N\alpha} \tag{15}$$

If now we substitute for $\eta$ into the first formula in (2) we obtain the following expression for the dielectric constant:

$$\varepsilon = \frac{1 + \frac{8\pi}{3}N\alpha}{1 - \frac{4\pi}{3}N\alpha} \tag{16}$$

Conversely (16) gives information about the dependence of the mean polarizability on $\varepsilon$ and $N$, or (if Maxwell's relation $\varepsilon = n^2$ is used) on $n$ and $N$ :

$$\alpha = \frac{3}{4\pi N}\frac{\varepsilon - 1}{\varepsilon + 2} = \frac{3}{4\pi N}\frac{n^2 - 1}{n^2 + 2}. \tag{17}$$

By a remarkable coincidence, the relation (17) was discovered independently and practically at the same time by two scientists of almost identical names, ‡ Lorentz and Lorenz, and is accordingly called the Lorentz-Lorenz formula. It is seen to be a

bridge which connects Maxwell's phenomenological theory with the atomistic theory of matter.

Instead of the mean polarizability $\alpha$ one often uses another quantity $A$, called molar refractivity. (In the case of monatomic substances it is also called atomic refractivity). This is essentially the total polarizability of a mole of the substance,* being defined as

$$A = \frac{4\pi}{3} N_m \alpha \tag{18}$$

Here $N_m = 6 \cdot 02.10^{23}$ is the Avogadro number, † namely the number of molecules in a mole. If $W$ is the molecular weight, $\rho$ the density, $p$ the pressure and $T$ the absolute temperature, then the molar volume is

$$\frac{N_m}{N} = \frac{W}{\rho} = \frac{RT}{p}, \tag{19}$$

where Boyle's law was used. Hence the molar refractivity may be written in the alternative forms

$$A = \frac{W}{\rho}\frac{n^2 - 1}{n^2 + 2} = \frac{RT}{p}\frac{n^2 - 1}{n^2 + 2}. \tag{20}$$

(20) shows that $A$ has the dimensions and the order of magnitude of a molar volume.

For a gas, $n^2$ is nearly unity, so that

$$\alpha \sim \frac{n^2 - 1}{4\pi N}, \quad A \sim \frac{W}{\rho}\frac{n^2 - 1}{3} = \frac{RT}{p}\frac{n^2 - 1}{3}. \tag{21}$$

The formula (20) gives the explicit dependence of the refractive index, for light of any one particular colour, on the density, and it should hold when the density is changed provided that isotropy is preserved. In the case of gases we actually find that $n^2 - 1$ and $\rho$ are very nearly proportional to each other as demanded by the second formula in (21). But even at high pressure, when $n$ will differ appreciably from unity, the molar refractivity remains substantially constant, as illustrated in Table IV. The molar refractivity is also found to remain practically constant when the gas is condensed into a liquid (Table V).

0

- We are restricting our discussion to so-called non-polar molecules, having no permanent dipole moment in the absence of the field.
  † In general the polarizability $\alpha$ is a tensor of the second rank.
  ‡ H. A. Lorentz, Wiedem. Ann., 9 (1880), 641; L. Lorenz, Wiedem. Ann., 11 (1881), 70.
  An analogous formula for statical fields was derived earlier by R. Clausius, Mechanische Wrmetheorie, 2 (Braunschweig, 2nd ed. 1879), p. 62), and by O. F. Mossotti, Mem. Soc. Sci. Modena, 14 (1850), p. 49. These authors tried to explain the dielectric properties of insulators on the assumption that the atoms are small conducting spheres whose mutual distances are large compared to their diameters, and derived the following expression for the number density $g$ of the spheres in terms of the dielectric constant:

$$g = \frac{\varepsilon - 1}{\varepsilon + 2}.$$

0

Table IV
The molar refractivity $A$ of air at different pressures and at about \$14 \cdot 5\hat{}\{\}\{\circ\} \mathrm{C}\$, for sodium $D$ light

| Pressure in atm | $n$ | $A$ |
|---|---|---|
| 1.00 | 1.0002929 | 4.606 |
| 42.13 | 1.01241 | 4.661 |
| 96.16 | 1.02842 | 4.713 |
| 136.21 | 1.04027 | 4.743 |
| 176.27 | 1.05213 | 4.772 |

Table V
Molar refractivity $A$ of different compounds for sodium $D$ light; in the columns giving $n$ and $A$ the first entry refers to liquid state, the second to vapour. The densities of the vapours are not stated but were calculated assuming perfect gas law from the known density of hydrogen

| Substance | Formula | $\boldsymbol{W}$ | $n$ | $\rho$ | $\boldsymbol{A}$ |
|---|---|---|---|---|---|
| Oxygen . . | $O_2$ | 32 | 1.221<br>1 · 000271 | 1.124 | 4 · 00<br>4.05 |
| Hydrochloric acid | HCl | 36 · 5 | 1 · 245<br>1 · 000447 | 0.95 | 5 · 95<br>6.68 |
| Water . . | $H_2O$ | 18 | 1.334<br>1.000249 | 1 · 00 | 3 · 71<br>3.72 |
| Carbon disulphide | $CS_2$ | 76 | 1.628<br>1 · 00147 | 1 · 264 | 21 · 33<br>21 · 99 |
| Acetone . . | $C_3H_6O$ | 58 | 1.3589<br>1 · 00108 | 0.791 | 16 · 14<br>16-16 |

From (20) it can be shown that, to a good approximation, the molar refractivity of a mixture of two substances is equal to the sum of the contributions due to each substance. Thus, if two liquids of refractivities $A_1$ and $A_2$ are mixed, and if a unit volume of the first liquid contains $N_1$ molecules and of the second $N_2$ molecules, then the molar refractivity of the mixture will be

$$A = \frac{N_1 A_1 + N_2 A_2}{N_1 + N_2}. \tag{22}$$

This result is also found to be in reasonably good agreement with experiment [cf. Table VI].

Finally, we consider the dependence of the molar refractivity $A$ of a compound on the atomic refractivities of its constituent atoms. If the molecule consists of $\mathscr{N}_1$ atoms of refractivity $A_1$ and of $\mathscr{N}_2$ atoms of refractivity $A_2$, etc., then obviously

$$A = \mathscr{N}_1 A_1 + \mathscr{N}_2 A_2. \tag{23}$$

• A mole or a gram-molecule is that amount of the substance whose weight, in grams, is equal to the molecular weight of the substance.
† This is the name internationally accepted; however, in the German literature the term Loschmidt's number is used, after the physicist who first gave a numerical estimate for it from experimental data.

Table VI
Molar refractivity for sodium $D$ light of mixtures of water and sulphuric acid at \$15\hat{}\{\}\{\circ\}\mathrm\{C\}\$

| Weight per cent | $\frac{N_1}{N_1+N_2}$ | $W$ | $n$ | $\rho$ | $A$ (interpolated from (22)) | $A$ (calculated from |
|---|---|---|---|---|---|---|
| 0 | 0 | 18.00 | 1.3336 | 0.9991 | 3.72 | 3.72 |
| $19 \cdot 98$ | 0.044 | $21 \cdot 52$ | $1 \cdot 3578$ | $1 \cdot 1381$ | $4 \cdot 19$ | 4.15 |
| $39 \cdot 76$ | $0 \cdot 109$ | 26.72 | 1.3817 | 1.2936 | 4.80 | 4.81 |
| 59.98 | 0.216 | $35 \cdot 28$ | $1 \cdot 4065$ | 1.4803 | $5 \cdot 86$ | 5.86 |
| $80 \cdot 10$ | 0.425 | 52.00 | $1 \cdot 4308$ | 1.6955 | 7.95 | 7.93 |
| 100 | 1 | 98 | $1 \cdot 4277$ | 1.8417 | 13.68 | 13.68 |

This formula too agrees fairly well with experiment, as is seen from Table VIII. The refractivities of the constituent atoms are given in Table VII.

Table VII
Atomic refractivities of various elements for sodium $D$ light

| Element | Formula | Atomic Weight | $\boldsymbol{A}$ |
|---|---|---|---|
| Hydrogen | H | 1 | 1.02 |
| Carbon . | C | 12 | $2 \cdot 11$ |
| Oxygen . | O | 16 | 2.01 |
| Sulphur | S | 32 | 8.23 |
| Chlorine | Cl | $35 \cdot 5$ | $5 \cdot 72$ |

Table VIII
The molar refractivities of various compounds for sodium $D$ light

| Substance | Formula | $A$ (calculated from (20)) | $A$ (calculated from (23)) |
|---|---|---|---|
| Hydrochloric acid | HCl | 6.68 | 6.74 |
| Water | $H_2O$ | $3 \cdot 72$ | $4 \cdot 05$ |
| Carbon disulphide | $CS_2$ | 21.97 | 18.57 |
| Acetone | $C_3H_6O$ | 16.14 | $14 \cdot 46$ |

The good agreement obtained in this case is really better than one could reasonably expect, on the basis of our simple considerations. Refinements are in fact necessary in some cases; for example in the case of nitrogen one has to attribute different atomic polarizabilities to the nitrogen atom according to the compounds which it enters. With the necessary modifications, the polarizability $\alpha$ is quite a well defined atomic parameter which may be used also in the study of other phenomena, such as the van der Waals cohesion forces, adsorption phenomena on surfaces, etc. These can only be satisfactorily treated, however, with the help of quantum mechanics.

### 2.3.4 Elementary theory of dispersion

In § 1.2 it was pointed out that the phase velocity and consequently the refractive index cannot be constants of the medium as suggested by our earlier formal considerations, but that

these quantities must in fact depend on frequency. The variation of the refractive index with frequency constitutes the phenomenon of dispersion. For an adequate treatment of dispersion, it would be necessary to go deeply into the atomic
theory of matter, but it is possible to give a simplified model of a dispersing medium by making use of one or two basic results concerning the structure of molecules.

A molecule consists of a number of heavy particles (the atomic nuclei of the atoms forming the molecule) around which light particles (electrons) revolve. The electrons carry negative and the nuclei positive charges. In neutral molecules the charges of the electrons just compensate those of the nuclei. However, the centres of the positive (nuclear) charges and that of the negative (electronic) charges may not coincide; this system has then an electric dipole and is called polar. For the sake of simplicity we shall here exclude polar molecules, although these play a great part in many physical and chemical phenomena.

If a non-polar molecule is subjected to an electric field, the electrons and nuclei are displaced and a dipole moment is generated. The vectorial sum of all the dipole moments of the molecules in the unit of volume is essentially the polarization vector $\boldsymbol{P}$ introduced formally in the preceding section.

In order to determine the dependence of the polarization and of the refractive index on the frequency of the field we must first find the displacement $r$ of each charged particle from its equilibrium position. We may assume that each electron is acted on by the Lorentz force $F$ (see eq. (34) in § 1.1):

$$F = e\left(E' + \frac{v}{c} \wedge B'\right),$$

where $e$ is the charge of the electron and $v$ its velocity. It will be assumed that the velocity of the electron is small in comparison with the velocity $c$ of light in vacuum, so that the contribution of the magnetic field may be neglected in the expression for the Lorentz force. The rigorous determination of the effective displacement of the nuclei and electrons under the action of the electric force is a complicated problem of quantum mechanics. It is however plausible, and actually confirmed by the rigorous theory, that the electron behaves, to a good approximation, as if it were bound to an equilibrium position by a quasi-elastic restoring force

$$Q = -qr.$$

Hence if $m$ denotes the mass of the electron, its equation of motion is

$$m\ddot{\boldsymbol{r}} + q\boldsymbol{r} = e\boldsymbol{E'} \tag{24}$$

Let $\omega$ be the angular frequency of the incident field,

$$\boldsymbol{E'} = \boldsymbol{E'}_0 e^{-i\omega t} \tag{25}$$

and take as trial solution of (24) the expression

$$r = r_0 e^{-i\omega t}. \tag{26}$$

(24) then gives the required stationary solution

$$r = \frac{eE'}{m\left(\omega_0^2 - \omega^2\right)} \tag{27}$$

where

$$\omega_0 = \sqrt{\frac{q}{m}} \tag{28}$$

is called the resonance (or absorption) frequency. According to (27) the electron oscillates with the frequency of the incident field.

Each electron contributes to the polarization a moment $p = er$. There will also be contributions from the nuclei; but since the nuclear masses are heavy in comparison with the masses of the electrons, their contributions may in the first approximation be neglected. Also, assuming for the moment that there is only one effective electron in a molecule with resonance frequency $\omega_0$, we obtain for the total polarization $P$ the expression

$$\boldsymbol{P} = N\boldsymbol{p} = Ne\boldsymbol{r} = N\frac{e^2}{m}\frac{\boldsymbol{E}'}{(\omega_0^2 - \omega^2)} \tag{29}$$

Comparison of (29) with (14) gives

$$N\alpha = N\frac{e^2}{m(\omega_0^2 - \omega^2)} \tag{30}$$

expressing the "density of polarizability" in terms of atomic parameters. Thus it turns out that the quantity $\alpha$ is not a constant, as it ought to be if the $\varepsilon$ in (17) denotes the static dielectric constant. It is convenient to introduce the notion of a frequency-dependent dielectric constant $\varepsilon(\omega)$ defined by Maxwell's relation $\varepsilon = n^2$, where $n$ is the refractive index, a function $n(\omega)$ of $\omega$. The static dielectric constant is then the value $\varepsilon(0) = n^2(0)$; it corresponds, according to (17), to the limiting value $N\alpha(0)$ which is found from (30) to be

$$N\alpha(0) = \frac{Ne^2}{m\omega_0^2}$$

if this is introduced into (16), one obtains the static constant $\varepsilon(0)$.
For $\omega \neq 0$, the function $N\alpha(\omega)$ is, according to (30), monotonically increasing with $\omega$, but has an infinity (resonance point) for $\omega = \omega_0$; for $\omega > \omega_0$, it approaches with increasing $\omega$ the value 0 from the negative side. Substituting from (30) into (17), we find the explicit dependence of the refractive index on frequency:

$$\frac{n^2 - 1}{n^2 + 2} = \frac{4\pi}{3}\frac{Ne^2}{m(\omega_0^2 - \omega^2)} \tag{31}$$

For a gas, $n$ is close to unity so that we may set $n^2 + 2 \sim 3$ in the denominator on the left-hand side and obtain

$$n^2 - 1 \sim 4\pi N\alpha = \frac{4\pi Ne^2}{m(\omega_0^2 - \omega^2)} \tag{32}$$

It is seen that $n$ is an increasing function of the frequency. The dispersion is then said to be normal. Further $n \gtrless 1$ according as $\omega \lessgtr \omega_0$, and $n$ approaches unity with increasing $\omega$ (see Fig. 2.3).

At the resonance frequency $(\omega = \omega_0)$ $n$ and $\alpha$ are actually not infinite as our formula suggests. The singularity arises only formally, because we neglected the effect of damping. Damping is in fact an essential factor in the whole process as the vibrating electrons emit electromagnetic waves which carry away energy. But there are also other reasons for the dissipation of energy (e.g. because of collisions between the atoms). Formally damping may be taken into account by adding in the equation of motion (24) a term $g\dot{r}$ representing a resisting force:

$$m\ddot{r} + g\dot{r} + qr = eE'. \tag{33}$$

Instead of (27) we now have

$$r = \frac{eE'}{m\left(\omega_0^2 - \omega^2\right) - i\omega g}. \tag{34}$$

The polarization, and hence also $N\alpha$, becomes a complex quantity. It can be shown (cf. the similar situation in the theory of metals where a complex refractive index is used-Chapter XIII) that the modulus of this complex function is-apart from a small factor, due to absorption-the actual "density of polarizability" and this is shown by the dotted curve in Fig. 2.3. The curve has a sharp maximum at a value of $\omega$ which is slightly smaller than $\omega_0$, and a sharp minimum at a value slightly greater than $\omega_0$. Between the maximum and minimum the function decreases with increasing frequency

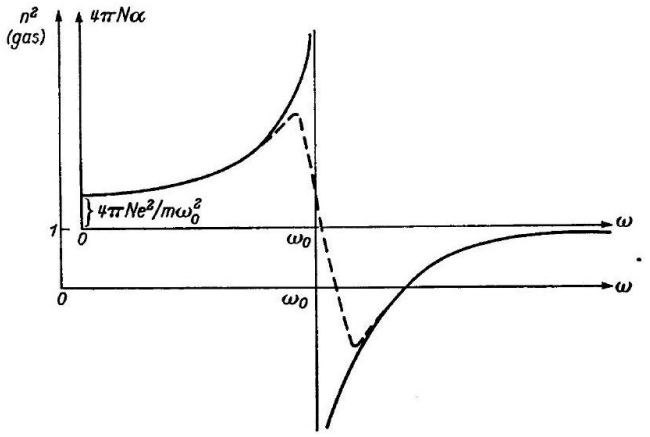



Fig. 2.3. Dispersion curves.
Full curve: effect of damping neglected. Broken curve: effect of damping taken into account. (The ordinates now represent the real parts of $n\hat{}\{\}\{2\}$ and of $4\pi N\alpha$.)

and we speak of a region of anomalous dispersion. The rays of shorter wavelength are then refracted less than those of longer wavelength, and this results in the reversal of the usual sequence of prismatic colours. For our purposes the region of anomalous dispersion is of little importance, since the absorption frequencies of free atoms lie almost exclusively in the ultra-violet region of the spectrum. The refractive index for visible light is then always greater than unity.

So far we have assumed that the system has only one resonance frequency. In general there will be many such frequencies, even in systems with the same kind of molecules; ( 31 ) and ( 32 ) must then be replaced by more general expressions. Neglecting for the moment again the motion of the nuclei, one has, in place of (31),

$$\frac{4\pi}{3}N\alpha = \frac{n^2-1}{n^2+2} = \frac{4\pi}{3}N\frac{e^2}{m}\sum_k \frac{f_k}{\omega_k^2 - \omega^2}, \tag{35}$$

where $Nf_k$ is the number of electrons corresponding to the resonance frequency $\omega_k$. For gases ( $n \sim 1$ ) we may rewrite (35) in the form

$$n^2 - 1 = 4\pi N\alpha = \sum_k \frac{\rho_k}{\nu_k^2 - \nu^2} = \sum_k \frac{\rho_k}{c^2}\frac{\lambda^2\lambda_k^2}{\lambda^2 - \lambda_k^2}, \tag{36}$$

where

$$\left.\begin{aligned} &\rho_k = N\tfrac{e^2}{\pi m}f_k, \\ &v_k = \tfrac{\omega_k}{2\pi} = \tfrac{c}{\lambda_k}, \quad v = \tfrac{\omega}{2\pi} = \tfrac{c}{\lambda}. \end{aligned}\right\} \tag{37}$$

Using the identity

$$\frac{\lambda^2}{\lambda^2-\lambda_k^2}=1+\frac{\lambda_k^2}{\lambda^2-\lambda_k^2},$$

(36) becomes

$$n^2-1=a+\sum_k\frac{b_k}{\lambda^2-\lambda_k^2}, \tag{38}$$

with

$$\left.\begin{array}{ll} a & =\sum_k\frac{\rho_k}{v_k^2}=\frac{1}{c^2}\sum_k\rho_k\lambda_k^2, \\ b_k & =\frac{c^2\rho_k}{v_k^4}=\frac{1}{c^2}\rho_k\lambda_k^4. \end{array}\right\} \tag{39}$$

For the representation of the refractive index over the whole visible range it is usually sufficient to take account of only one or two resonance frequencies in the ultra-violet region. Косн* for example found that for the range from $\lambda=0.436\mu$ to $\lambda=8.68\mu$, the following formula holds for hydrogen, oxygen and air:

$$n^2-1=a+\frac{b}{\lambda^2-\lambda_0^2}, \tag{40}$$

where the constants $a, b$ and $\lambda_0$ are shown in Table IX.

Table IX
Dispersion constants for oxygen, hydrogen and air between $\lambda=0.436\mu$ and $\lambda=8.68\mu$ at \$0\hat{}\{\}\{\circ\} \mathrm{C}\$ and 760 mm Hg

| | Gas | $a.10^8$ | $b.10^8$ | $\lambda_0^2$ in $10^{-8}$ cm$^2$ | $v_0$ in $10^{15}$sec$^{-1}$ |
|---|---|---|---|---|---|
| Hydrogen | • | 27,216 | 211.2 | 0.007760 | 3.40 |
| Oxygen | - | 52,842 | 369.9 | 0.007000 | $3\cdot 55$ |
| Air | • | 57,642 | $327\cdot 7$ | 0.005685 | 3.98 |

In a spectral range which contains no resonance frequencies, (38) may with a good accuracy be replaced by a simpler formula. For all substances which appear transparent to the eye, the visible domain of the spectrum is such a range. Denoting by $v_v$ the absorption frequencies which are on the short wavelength side (violet), and by $v_r$

those on the long wavelength side (red), the dispersion formula (36) becomes, on expanding into power series with respect to $\nu$ and $\lambda$ respectively:

$$\left.\begin{array}{rll} n^2-1= & A & +Bv^2+Cv^4+\dots \\ & & -\frac{B'}{v^2}-\frac{C'}{v^4}-\dots \\ = & A & +\frac{Bc^2}{\lambda^2}+\frac{Cc^4}{\lambda^4}+\dots \\ & & -\frac{B'\lambda^2}{c^2}-\frac{C'\lambda^4}{c^4}-\dots \end{array}\right\} \tag{41}$$

where

$$\left.\begin{array}{ll} A=\sum_v\frac{\rho_v}{v_v^2}, & B=\sum_v\frac{\rho_v}{v_v^4}, \quad C=\sum_v\frac{\rho_v}{v_v^6},\dots \\ B'=\sum_r\rho_r, & C'=\sum_r\rho_r v_r^2,\dots \end{array}\right\} \tag{42}$$

0

• J. Koch, Nov. Act. Soc. Ups. (4), 2 (1909), p. 61.

In such an absorption-free range the value of $n$ for gases differs so little from unity that one may replace $n^2 - 1$ by $2(n - 1)$. Moreover the terms $B', C', \ldots$ arising from the ultra-violet resonance frequencies are usually without appreciable influence. Then, if only terms down to $1/\lambda^2$ are retained, (41) reduces to Cauchy's formula*

$$n - 1 = A_1 \left(1 + \frac{B_1}{\lambda^2}\right). \tag{43}$$

where

$$A_1 = \frac{A}{2}, \quad B_1 = \frac{Bc^2}{A} \tag{44}$$

In Table X the values of $A_1$ and $B_1$ for the more important gases are given. Also, to illustrate the accuracy of the formula, the values of the refractive index for air as given by Cauchy's formula and the observed values are compared with each other in Table XI.

Table X
The constants $A_{1}$ and $B_1$ of Cauchy's dispersion formula for various gases

| Gas | $A_1 \cdot 10^5$ | $B_1 \cdot 10^5$ |
|---|---|---|
| Argon | 27.92 | $5 \cdot 6$ |
| Nitrogen | $29 \cdot 19$ | $7 \cdot 7$ |
| Helium | $3 \cdot 48$ | $2 \cdot 3$ |
| Hydrogen | 13.6 | $7 \cdot 7$ |
| Oxygen | 26.63 | $5 \cdot 07$ |
| Air . | $28 \cdot 79$ | $5 \cdot 67$ |
| Ethane | 73.65 | 9.08 |
| Methane | $42 \cdot 6$ | $14 \cdot 41$ |

In the case of substances of high density, i.e. liquids or solids, it is no longer permissible to replace $n$ by unity in the denominator of the second term in (35). Nevertheless, it is possible to reduce (35) to the same form as before. Since

$$\frac{n^2 - 1}{n^2 + 2} = \frac{4\pi}{3} N\alpha,$$

it follows that

$$n^2 - 1 = \frac{12\pi N\alpha}{3 - 4\pi N\alpha} \tag{45}$$

where

$$4\pi N\alpha = \frac{Ne^2}{\pi m} \sum_k \frac{f_k}{v_k^2 - v^2} = \sum_k \frac{\rho_k}{v_k^2 - v^2}. \tag{46}$$

It will, as a rule, be sufficient to take into account only a finite number of the absorption frequencies, and it follows that $n^2 - 1$, given by (45), represents a rational function in $\nu^2$, and

0

- L. Cauchy, Bull. des. sc. math., 14 (1830), 9; Sur la dispersion de la lumire (Nouv. exerc. de math., 1836).

Table XI
Observed values of the refractive index for air and the values given by Cauchy's dispersion formula (43)

| $\lambda.10^5$ cm | $(n-1)\cdot 10^4$ observed | $(n-1)\cdot 10^4$ calculated | Difference |
|---|---|---|---|
| 7.594 | 2.905 | 2.907 | 0.002 |
| 6.563 | 2.916 | $2\cdot 917$ | $0\cdot 001$ |
| $5\cdot 896$ | 2.926 | 2.926 | $0\cdot 000$ |
| 5.378 | $2\cdot 935$ | 2.935 | $0\cdot 000$ |
| 5.184 | $2\cdot 940$ | 2.940 | $0\cdot 000$ |
| $4\cdot 861$ | $2\cdot 948$ | 2.948 | $0\cdot 000$ |
| $4\cdot 677$ | 2.951 | 2.954 | $0\cdot 003$ |
| $4\cdot 308$ | 2.966 | 2.967 | 0.001 |
| $3\cdot 969$ | $2\cdot 983$ | 2.983 | $0\cdot 000$ |
| 3.728 | 2.995 | $2\cdot 996$ | $0\cdot 001$ |
| 3.441 | 3.016 | $3\cdot 017$ | 0.001 |
| $3\cdot 180$ | $3\cdot 040$ | $3\cdot 041$ | $0\cdot 001$ |
| 3.021 | $3\cdot 056$ | 3.058 | $0\cdot 002$ |
| 2.948 | 3.065 | 3.067 | $0\cdot 002$ |

may therefore be decomposed into partial fractions. For this purpose one must find the zero values of the denominator, i.e. the roots of the equation

$$3 - 4\pi N\alpha = 3 - \sum_k \frac{\rho_k}{\nu_k^2 - \nu^2} = 0. \tag{47}$$

If the roots are denoted by $\nu_k$, then (45) may be reduced to the form

$$n^2 - 1 = \sum_k \frac{\bar{\rho}_k}{\bar{\nu}_k^2 - \nu^2}, \tag{48}$$

which is identical in form with the formula (36) for gases.

For example if there is only a single absorption frequency, $\bar{\nu}_1$, this frequency will be the root of the equation

$$3 - \frac{\rho_1}{v_1^2 - v^2} = 0 \tag{49}$$

whence

$$\left.\begin{aligned} \bar{\nu}_1^2 &= \nu_1^2 - \tfrac{1}{3}\rho_1, \\ \bar{\rho}_1 &= \rho_1. \end{aligned}\right\} \tag{50}$$

Eq. (48), which may also be written in the form (38), is known as Sellmeir's dispersion formula.

So far the effect of the motion of the nuclei has been neglected. It is, in fact, only of importance at very long wavelengths (infra-red). The reason for this can easily be understood. The electric moment and the mean polarizability may be approximately divided into two parts, one associated with the electrons and the other with the nuclei,

$$\boldsymbol{p} = \boldsymbol{p}_e + \boldsymbol{p}_n \tag{51}$$

where

$$\boldsymbol{p}_e = \alpha_e \boldsymbol{E}', \quad \boldsymbol{p}_n = \alpha_n \boldsymbol{E}' \tag{52}$$

The electrons will follow the field almost instantly, up to rather high frequencies, including in many cases those of the whole visible spectrum. The nuclear masses, on the other hand, are so heavy that they cannot follow the field in the high-frequency region, i.e. $\alpha_n \sim 0$ for visible light. This can also be seen from the dispersion formula (41) where $n^2 - 1$ is represented by two sets of terms, one due to the high-frequency (violet) vibration $\nu_v$, the other due to the low-frequency (red) vibrations $\nu_r$. It is reasonable to assume, and it follows rigorously from quantum mechanics, that the quasi-elastic forces binding the nuclei and electrons are all of the same order of magnitude; hence the order of magnitude of the frequencies will be related by

$$m v_e^2 \sim M v_n^2 \tag{53}$$

where $m$ is the electronic and $M$ the nuclear mass, and $\nu_e$ and $\nu_n$ are the frequencies of vibrations of the electrons and nuclei respectively. In order to extend the formulae (41) and (42) to take into account also the motion of the nuclei, we may simply assume that the violet vibrations are due to the electrons, and the red vibrations due to the nuclei. Then we have also to distinguish two types of factors $\rho_k$ as given by (37), those due to the electronic and those due to the nuclear vibrations, and we have the order of magnitude relation

$$m\rho_e \sim M\rho_n \tag{54}$$

The coefficients (42) of the dispersion formulae are therefore of the following orders of magnitude:

$$\left.\begin{array}{llll} B & \sim \frac{A}{v_e^2}, & C & \sim \frac{A}{v_e^4} \\ B' & \sim \frac{m}{M} A v_e^2, & C' & \sim \frac{m}{M} A v_e^4, \end{array}\right\} \tag{55}$$

and we have

$$\begin{aligned} n^2 - 1 \sim A\{1 + a\frac{v^2}{v_e^2} + b\frac{v^4}{v_e^4} + \ldots \\ - \frac{m}{M}\left[a'\frac{v_e^2}{v^2} + b'\frac{v_e^4}{v^4} + \ldots\right]\Big\} \end{aligned} \tag{56}$$

where $a, b, \ldots a', b', \ldots$ are numerical constants of the order of unity. The terms in the first line correspond to the electronic part $\alpha_e$ of the polarizability, in the second line to its nuclear part $\alpha_n$; since $m/M$ is small ($\sim 1/1840$) we see that the latter terms will be negligible, provided that the frequency $v$ is not a very small fraction of the violet resonance frequencies $\nu_e$.

Hence for an optical frequency (i.e. one corresponding to visible light), the polarization is given essentially by $\alpha_e$ alone (as assumed in our preceding calculations), while for a static field it is given by $\alpha_e + \alpha_n$.

The considerations of the present section were entirely based on classical mechanics. When similar calculations are carried on the basis of quantum mechanics, the interaction of the field and matter may still be described in terms of virtual oscillators, but the number of the oscillators, even when only a single electron is present, is found to be infinite. Eq. (35) still applies, but the strength factors $f_k$ no longer specify the number of electrons of a particular type, but rather the number of virtual oscillators belonging to one electron or to a set of electrons. In most cases only a finite number of the $f_k$ factors have an appreciable value, the others may be neglected. In fact the whole formal theory is hardly affected by the introduction of quantum mechanics, but with its help one can calculate the $f_k$ factors for a given electronic system.

## 9.3.4 2.4 Propagation of Electromagnetic Waves Treated by Integral Equations*

At the beginning of this chapter a method was outlined for the treatment of propagation of electromagnetic waves in terms of integral equations, and in the sections which followed, auxiliary quantities needed for this purpose were introduced. We shall now formulate these integral equations and discuss some of their consequences. This may be done entirely within the framework of the macroscopic theory as dealt with throughout this book. Within this framework the integral equations are equivalent to Maxwell's equations and represent the mathematical description of the electromagnetic phenomena in terms of mutual interactions over finite distances (which of course need time for transfer). A definite advantage of this treatment, which in some cases is more powerful than the more customary treatment based on differential equations, is that it connects macroscopic phenomena with the molecular ones, discussed in the previous section.

As already explained, a substance may be considered to consist of molecules which react to the incident waves like dipoles. All the emitted dipole waves then give rise to the effective force acting on any other dipole and also determine the average measurable field. It will be assumed that the dipoles are evenly distributed throughout the medium, and the average value $\boldsymbol{P}$ of their electric moment per unit volume will be considered as the basic quantity. Actually the molecules are never quite evenly distributed throughout a medium (i.e. there are density fluctuations) and consequently the electric moment of the individual particles exhibits fluctuations about the average

value. The resulting phenomena may be accounted for in the present theory by carrying the calculations one step further, i.e. one calculates not only the averages but also their mean square deviations. Such calculations are of importance in some problems, e.g. for the explanation of the blue colour of the sky, as first given by Rayleigh.* But this extension of the theory cannot be given here. †

### 2.4.1 The basic integral equation

We consider the propagation of an electromagnetic wave in an homogeneous, isotropic non-magnetic medium. The electric and magnetic fields $\boldsymbol{E}'_j$ and $\boldsymbol{H}'_j$ which act on the $j$ th dipole in the interior of the medium can be divided into the incident fields $E^{(i)}, H^{(i)}$ (propagated with the vacuum velocity $c$ ) and the contribution arising from all the dipoles:

$$\left.\begin{aligned} \boldsymbol{E}'_j &= \boldsymbol{E}^{(i)} + \textstyle\sum'_l \boldsymbol{E}_{jl}, \\ \boldsymbol{H}'_j &= \boldsymbol{H}^{(i)} + \textstyle\sum'_l \boldsymbol{H}_{jl}, \end{aligned}\right\} \tag{1}$$

the summation extending over all the dipoles except the $j$ th one. At the point $r_1$ where the $j$ th dipole is situated, the field of the $l$ th dipole is obtained from § 2.2 (49)-(51):

0

- The theory here described was developed chiefly by W. Esmarch, Ann. d. Physik, 42 (1913), 1257 ; C. W. Oseen, ibid., 48 (1915), 1; and W. Boethe, Beitrge zur Theorie der Brechung und Reflexion (Dissertation, Berlin, 1914), and Ann. d. Physik, 64 (1921), 693. A full account of the theory was given by R. Lundblad, Univ. Arsskrift, Upsala (1920). See also C. G. Darwin, Trans. Camb. Phil. Soc., 23 (1924), 137, and H. Hoek, Algemeene theorie der optische activiteit van isotrope media (Thesis, Leiden, 1939).

Some generalizations of this theory which take in to account non-linear behaviour of the medium (arising when the fields are exceptionally strong) were discussed by N. Bloembergen and P. S. Pershan, Phys. Rev., 128 (1962), 619.

$$\left.\begin{aligned} E_{jl} &= \operatorname{curl}\operatorname{curl}\frac{p_l\left(t-R_{jl}/c\right)}{R_{jl}}, \\ H_{jl} &= \tfrac{1}{c}\operatorname{curl}\frac{\dot{p}_l\left(t-R_{jl}/c\right)}{R_{jl}}, \end{aligned}\right\} \tag{2}$$

$\boldsymbol{p}_l(t)$ denoting the moment of the $l$ th dipole, $R_{jl} = |\boldsymbol{r}_j - \boldsymbol{r}_l|$, and the operation curl is taken with respect to the coordinates $x_j, y_j, z$, of the $j$ th dipole.

As already explained, the distribution may to a good approximation be treated as continuous, i.e. the moment of the dipoles may be considered to be a continuous function of position (and of time): $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{r}, \boldsymbol{t})$. Likewise the number density $N$ will be assumed to be a continuous function $N(r)$ of position. The total electric dipole $P$ per unit volume is then given by (14), §2.3:

$$\boldsymbol{P} = N\boldsymbol{p} = N\alpha\boldsymbol{E}' \tag{3}$$

For reasons explained in § 2.3, the contribution of the magnetic force has been neglected in (3). Since the material is also assumed to be non-magnetic (i.e. $M = 0$ ), the effective field $\boldsymbol{H}'$ will not enter into the conditions for the dynamical equilibrium to be established.

If we substitute from (2) into (1) and go over to a continuous distribution, using (3), we obtain ‡

$$\boldsymbol{E}'(\boldsymbol{r}, t) = \boldsymbol{E}^{(i)} + \int \operatorname{curl}\operatorname{curl} N\alpha\frac{\boldsymbol{E}'\left(\boldsymbol{r}', t - R/c\right)}{R}dV' \tag{4}$$

$$\boldsymbol{H}'(\boldsymbol{r}, t) = \boldsymbol{H}^{(i)} + \frac{1}{c}\int \operatorname{curl} N\alpha\frac{\dot{\boldsymbol{E}}'\left(\boldsymbol{r}', t - R/c\right)}{R}dV' \tag{5}$$

where

$$R = |\boldsymbol{r} - \boldsymbol{r}'|\,.$$

If the point of observation $r$ is outside the medium, the integral is taken throughout the whole medium. If it is inside the medium, a small domain occupied by the atom must be first excluded; we consider it to be a small sphere $\sigma$ of radius $a$. Eventually we will proceed to the limit $a \to 0$ in the usual way.

Eq. (4) is an integro-differential equation for $\boldsymbol{E}'$. When it is solved, $\boldsymbol{H}'$ may be obtained from (5). The two equations are essentially equivalent to Maxwell's equations for isotropic non-magnetic substances. Generalization to magnetic substances may be obtained with the help of the second Hertz vector.

---

0

- We shall refer to Rayleigh's theory again in a later section (§ 13.5) which deals with the scattering of spherical particles embedded in vacuum. If the particles are small, the dependence of the intensity of the scattered radiation on the wavelength is the same as in the present case of spontaneous fluctuations of density in a homogeneous medium (namely inversely proportional to the fourth power of the wavelength). The fluctuation theory, however, gives in addition to this result also the dependence of the scattered intensity on the density fluctuation.

† It is discussed, for example, in M. Born, Optik (Berlin, Springer, 1933), § 81.

‡ One must sum the fields and not the potentials, i.e. in (4) the operator curl curl must be under the integration sign and not outside it; for the field acting on the molecule as expressed by (3) is obtained by summing all the individual fields. The singularities of the fields at the sources lead to a difference between the sum of the fields and the field associated with the total potential (sum of potentials associated with the individual fields). Cf. (15) below.

### 2.4.2 The Ewald-Oseen extinction theorem and a rigorous derivation of the Lorentz-Lorenz formula

Eq. (4) connects the effective electric field with the incident electric field in a rather complicated manner and only in special cases can this equation be solved explicitly. Nevertheless it is possible to derive from it some basic results, such as the LorentzLorenz formula, the laws of refraction and reflection, and the Fresnel formulae. Before showing this we shall establish an important general consequence of the solution.

Let $\Sigma$ denote the boundary of the medium. For points of observation inside the medium the basic equation (4) may be written in the form

$$\boldsymbol{P} = N\alpha \left(E^{(i)} + E^{(d)}\right), \tag{6}$$

where $\boldsymbol{E}^{(d)}$ denotes the contribution from the dipoles:

$$E^{(d)} = \int_{\sigma}^{\Sigma} \operatorname{curl}\operatorname{curl} \frac{P\left(r', t - R/c\right)}{R} dV'. \tag{7}$$

Here we have indicated explicitly the boundary of the volume over which the integration is carried out.

The incident field $\boldsymbol{E}^{(i)}$ will be assumed to be monochromatic, with angular frequency $\omega$,

$$E^{(i)} = A^{(i)}(r)e^{-i\omega t} \tag{8}$$

and we take, as a trial solution for $\boldsymbol{P}$, a wave which is also monochromatic and of the same frequency, but which has a different velocity of propagation (say $c/n$ ),

$$P = \left(n^2 - 1\right) k_0^2 Q(r) e^{-i\omega t} \tag{9}$$

where, as before, $k_0 = \omega/c$ and

$$\nabla^2 Q + n^2 k_0^2 Q = 0. \tag{10}$$

In (9) the constant factor $(n^2 - 1)\, {k_0}^2$ has been introduced to simplify later formulae. The constant $n$ must be regarded as an unknown quantity, its determination being one of the chief aims of the present analysis.

We also assume that $Q$ has no sources in the medium, i.e. that inside the medium

$$\operatorname{div} Q = 0. \tag{11}$$

The possibility of solving the basic integro-differential equation (6) in the form represented by our trial solution may appear somewhat strange at first sight, since $E^{(i)}$ represents a wave propagated with the vacuum velocity $c$, whereas $P$ is assumed to be propagated with the velocity $c/n$. It will however be shown that the dipole field $E^{(d)}$ may be expressed as the sum of two terms, one of which obeys the wave equation in the vacuum and cancels out exactly the incident wave, whereas the other satisfies the wave equation for propagation with velocity $c/n$. The incident wave may therefore be regarded as extinguished at any point within the medium by interference with the dipole field and replaced by another wave with a different velocity (and generally also with a different direction) of propagation. This result is known as the extinction theorem and was established first for crystalline media by Ewald* and for isotropic media by Oseen. †

To prove the extinction theorem we rewrite (6) in its time-free form

$$Q = N\alpha \left\{ \frac{1}{\left(n^2 - 1\right) k_0^2} A^{(i)} + A^{(d)} \right\} \tag{12}$$

where

$$\left.\begin{aligned} E^{(d)} &= (n^2-1)\,k_0^2 A^{(d)} e^{-i\omega t} \\ A^{(d)} &= \int_\sigma^\Sigma \operatorname{curl}\operatorname{curl} Q\,(r')\,G(R)dV', \end{aligned}\right\} \tag{13}$$

and

$$G = \frac{e^{ik_0 R}}{R} \tag{14}$$

It is shown in Appendix $V$ that, when the radius of the sphere $\sigma$ is sufficiently small,

$$\int_\sigma^\Sigma \text{ curl curl } Q\,(r')\,G(R)dV' = \operatorname{curl}\operatorname{curl}\int_\sigma^\Sigma Q\,(r')\,G(R)dV' - \frac{8\pi}{3}Q(r) \tag{15}$$

Now the function $G$ represents a spherical wave in the vacuum and therefore satisfies the wave equation

$$\nabla^2 G + k_0^2 G = 0 \tag{16}$$

Hence it follows from (10) and (16) that the term under the integral sign on the right of (15) may be expressed in the form

$$QG = \frac{1}{(n^2-1)\,k_0^2}\left(Q\nabla^2 G - G\nabla^2 Q\right) \tag{17}$$

giving on integration with the help of Green's theorem,

$$\int_\sigma^\Sigma Q\,(r')\,G(R)dV' = \frac{1}{(n^2-1)\,k_0^2}\left[\int_\Sigma\left\{Q\frac{\partial G}{\partial \nu'} - G\frac{\partial Q}{\partial \nu'}\right\}dS' - \int_\sigma\left\{Q\frac{dG}{dR} - G\frac{\partial Q}{\partial R}\right\}dS'\right] \tag{18}$$

where $\partial/\partial\nu'$ denotes differentiation along the outward normal to the boundary $\Sigma$. A straightforward calculation gives (cf. § 8.3.1) for the limit of the surface integral over $\sigma$ as $a \to 0$ the value $-4\pi Q(r)$, so that

$$\int_\sigma^\Sigma Q\,(r')\,G(R)dV' \to \frac{1}{(n^2-1)\,k_0^2}\left\{\int_\Sigma\left(Q\frac{\partial G}{\partial v'} - G\frac{\partial Q}{\partial v'}\right)dS' + 4\pi Q\right\} \tag{19}$$

From (13), (15) and (19) we obtain, in the limit,

$$\begin{aligned} A^{(d)} &= \int_\sigma^\Sigma \text{ curl curl } Q\,(r')\,G(R)dV' = \\ &\frac{1}{(n^2-1)\,k_0^2}\left[4\pi \text{ curl curl } Q + \text{ curl curl } \int_\Sigma\left\{Q\frac{\partial G}{\partial v'} - G\frac{\partial Q}{\partial v'}\right\}dS'\right] - \frac{8\pi}{3}Q(r) \end{aligned} \tag{20}$$

Next we use the identity

$$\operatorname{curl}\operatorname{curl} Q \equiv \operatorname{grad}\operatorname{div} Q - \nabla^2 Q.$$

0

- P. P. Ewald (Dissertation, Mnchen, 1912); Ann. d. Physik, 49 (1916), 1.
† C. W. Oseen, Ann. d. Physik, 48 (1915), 1.

The first term on the right vanishes on account of (11), and the second is by (10) equal to $n^2 k_0{}^2 Q$. Hence, on substituting into (20), we finally obtain the following expression for the dipole field:

$$A^{(d)} = \frac{4\pi}{3}\left(\frac{n^2+2}{n^2-1}\right)\cdot Q + \frac{1}{(n^2-1)\,k_0^2}\ \text{curl curl}\ \int_\Sigma \left\{Q\frac{\partial G}{\partial v'} - G\frac{\partial Q}{\partial v'}\right\} dS' \tag{21}$$

According to (10) the first term on the right represents a wave propagated with the velocity $c/n$, whereas the second is seen to represent a wave which, like $G$, is propagated with the vacuum velocity $c$. Hence, as anticipated, the basic equation (12) splits into two groups of terms, each of which represents a wave propagated with a different velocity of propagation. This is only possible if each group vanishes separately, so that

$$\frac{4\pi}{3}N\alpha = \frac{n^2-1}{n^2+2} \tag{22}$$

and

$$A^{(i)} + \text{curl}\,\text{curl}\int_\Sigma \left\{Q\frac{\partial G}{\partial v'} - G\frac{\partial Q}{\partial v'}\right\} dS' = 0 \tag{23}$$

The relation (23) expresses the extinction of the incident wave

$$E^{(i)} = A^{(i)} e^{-i\omega t}$$

at any point inside the medium, by interference with part of the dipole field. The incident wave is replaced by another wave, namely by

$$E' = \frac{1}{N\alpha}P = \frac{1}{N\alpha}\left(n^2-1\right)k_0^2 Q e^{-i\omega t} = \frac{4\pi}{3}\left(n^2+2\right)k_0^2 Q e^{-i\omega t} \tag{24}$$

which is propagated inside the medium with velocity $c/n.n$ is given by (22) in terms of the number density $N$ and the polarizability $\alpha$; this relation will be recognized as the Lorentz-Lorenz formula, tentatively introduced in Section 2.3.3. It is to be noted that the extinction of the incident wave is brought about entirely by dipoles on the boundary of the medium and the polarization $\boldsymbol{P}$ is set up entirely by mutual interaction of neighbouring dipoles.

The general solution of the integro-differential equation (23) may, in principle, be obtained by reducing it to a system of two coupled Fredholm integral equations for $Q$ and $\partial Q/\partial v'$ on the surface. $Q$ at any point inside the medium may then be obtained by solving (10) subject to the condition that it takes these values on the boundary. This may be done by standard methods, using Green's functions, or by applying the Helmholtz-Kirchhoff integral formula (§ 8.3, eq. (7)).

We must now connect the effective field $\boldsymbol{E}'$ with the ordinary field $\boldsymbol{E}$ of Maxwell's theory. To do this, we write down the two definitions of the electric displacement $D$ :

$$D = \varepsilon E = n^2 E \tag{25a}$$

$$D = E + 4\pi P \tag{25b}$$

and

Elimination of $\boldsymbol{D}$ between these two equations gives

$$E = \frac{4\pi}{n^2-1}P \tag{26}$$

Now from (3) and (22),

$$E' = \frac{1}{N\alpha}P = \frac{4\pi}{3}\frac{n^2+2}{n^2-1}P$$

so that, using (26), we obtain

$$E' = E + \frac{4\pi}{3}P \tag{27}$$

in agreement with (12), § 2.3. Finally from (26) and (9) it is seen that, in terms of $Q$, the field $\boldsymbol{E}$ inside the medium ( $n \neq 1$ ) is given by

$$E = 4\pi k_0^2 Q e^{-i\omega t} \tag{28}$$

To determine the electric field outside the medium we must return to (4). The integrals occurring in these equations now extend over the whole medium, so that the operator curl curl may be taken outside the respective integrals. Further, since outside the medium $D = E$, we have $P = 0$ there, and consequently, by (27), we have, formally, $\boldsymbol{E}' = \boldsymbol{E}$. In place of (6) we now obtain
where

$$\left.\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}^{(i)} + \boldsymbol{E}^{(r)} \\ \boldsymbol{E}^{(r)} &= \text{curl curl } \int \tfrac{\boldsymbol{P}(\boldsymbol{r}', t-R/c)}{R} dV'. \end{aligned}\right\} \tag{29}$$

For the time harmonic case here considered, we have, on substituting from (9), using (17) and Green's theorem,
where

$$\left.\begin{aligned} E^{(r)} &= A^{(r)} e^{-i\omega t} \\ A^{(r)} &= \text{curlcurl} \int_\Sigma \left\{ Q \tfrac{\partial G}{\partial \nu'} - G \tfrac{\partial Q}{\partial \nu'} \right\} dS', \end{aligned}\right\} \tag{30}$$

and $Q$ is the same function on the surface $\Sigma$ as before.
So far we have considered the electric field only. To determine the magnetic field we have only to substitute for $\boldsymbol{E}'$ into (5) and evaluate the integral. This can be done by a similar calculation as above, but the evaluation is somewhat simpler now, since the operator curl in (5) may be taken outside the integral sign, as is evident from discussion of Appendix V. On comparing the resulting expressions for $H$ with those for $\boldsymbol{E}$ we find that the solutions are consistent with Maxwell's equations. Since these calculations are straightforward and do not bring out any important new features, they will not be given here.
It is of very great interest that an approach based on the physical remark that the field in matter is more naturally characterized by the polarization than by the displacement vector, should lead so elegantly, via the integro-differential equation (4), to
a strict derivation of the Lorentz-Lorenz formula and the extinction theorem. This powerful method has so far been little used in treating more specific problems,* but an example of its application will be given in Chapter XIII.

### 2.4.3 Refraction and reflection of a plane wave, treated with the help of the Ewald-Oseen extinction theorem

The Ewald-Oseen extinction theorem, expressed by the formula (23), will now be applied to the case of a plane monochromatic wave entering a homogeneous medium which fills the half-space $z < 0$. It will be shown that it leads to the laws of refraction and reflection and to the Fresnel formulae.

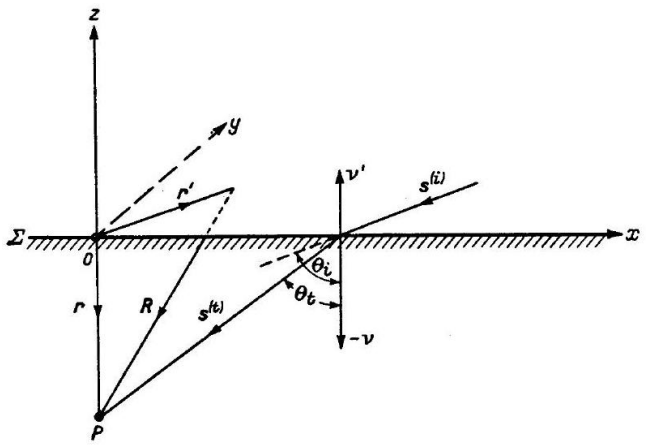



Fig. 2.4. Penetration of a wave into a homogeneous medium which is considered as a system of dipoles.

The incident wave will be written in the form

$$E^{(i)} = A^{(i)}(r)e^{-i\omega t} = A_0^{(i)} e^{i\left[k_0\left(r.s^{(i)}\right)-\omega t\right]}, \tag{31}$$

where $k_0 = \omega/c$, $A_0{}^{(i)}$ is a constant vector and $s^{(i)}$ the unit vector in the direction of propagation.

We choose the $z$-axis so that it passes through the point of observation $P$, assumed first to be inside the medium at a distance $r$ from the boundary $\Sigma$ (see Fig. 2.4). The $x$-axis will be chosen so that $s^{(i)}$ lies in the $xz$-plane. Hence, if $\theta_i$ denotes the angle of incidence, we have

$$s_x^{(i)} = -\sin\theta_i, \quad \dot{s}_y^{(i)} = 0, \quad s_z^{(i)} = -\cos\theta_i. \tag{32}$$

In accordance with the analysis of the preceding section, the transmitted wave will be assumed to have the same frequency as the incident wave, but a different velocity $c/n, n$ being given in terms of the polarizability and the density by the Lorentz-Lorenz formula. As our trial solution for the transmitted wave we take a plane wave propagated in the direction of a unit vector $\boldsymbol{s}^{(t)}$ which is assumed to lie in the $xz$-plane:

$$s_x{}^{(t)} = -\sin\theta_t, \quad s_y{}^{(t)} = 0, \quad s_z{}^{(t)} = -\cos\theta_t. \tag{33}$$

(28) then becomes

$$E = 4\pi k_0^2 Q e^{-i\omega t} = 4\pi k_0^2 Q_0 e^{i\left[k_0 n\left(r.s^{(t)}\right)-\omega t\right]}, \tag{34}$$

$Q_0$ being a constant vector which, on account of (11), is orthogonal to $s^{(t)}$.
The solution of the integro-differential equation can easily be obtained if we only consider points of observation $P$ whose distance from the boundary $z = 0$ is large compared to the wavelength, i.e. if

$$\frac{2\pi r}{\lambda} = k_0 r \gg 1 \tag{35}$$

This condition is merely imposed to simplify the calculations and does not correspond to a limitation on the physical validity of the integro-differential equation.

0

- The method has been applied to the propagation of electromagnetic waves through a stratified medium by D. R. Hartree, Proc. Cambr. Phil. Soc., 25 (1929), 97.

The derivatives occurring in the integral over $\Sigma$ are now

$$\frac{\partial Q\left(r'\right)}{\partial v'} = \frac{\partial}{\partial v'} Q_0 e^{ink_0\left(r'\cdot s^{(t)}\right)} = ink_0 \left(\frac{\partial r'}{\partial v'} \cdot s^{(t)}\right) Q_0 e^{ink_0\left(r'\cdot s^{(t)}\right)}, \tag{36a}$$

$$\frac{\partial G(R)}{\partial v'} = \frac{\partial}{\partial v'} \frac{e^{ik_0 R}}{R} = ik_0 \frac{\partial R}{\partial v'} G \left(1 + \frac{i}{k_0 R}\right). \tag{36b}$$

The last term in (36b) may be neglected because of (35), and the integral occurring in (23) becomes

$$J = \int_\Sigma \left(Q\frac{\partial G}{\partial v'} - G\frac{\partial Q}{\partial v'}\right) dS' = ik_0 Q_0 \int_\Sigma \left\{\frac{\partial R}{\partial v'} - n\left(\frac{\partial r'}{\partial v'} \cdot s^{(t)}\right)\right\} \frac{e^{ik_0\left[R + n\left(r'\cdot s^{(t)}\right)\right]}}{R} dS' \tag{37}$$

The components of the vectors $r, r'$ and $R$ are (see Fig. 2.4),

$$\begin{array}{rrrr} \boldsymbol{r}: & 0, & 0, & -r, \\ \boldsymbol{r'}: & x', & y', & 0, \\ \boldsymbol{R}: & -x', & -y', & -r, \end{array}$$

so that

$$\left.\begin{array}{llll} R & = \sqrt{x'^2 + y'^2 + r^2}, & r'.s^{(t)} & = -x' \sin\theta_t, \\ \frac{\partial R}{\partial v'} & = \frac{r}{R}, & \frac{\partial r'}{\partial v'}.s^{(t)} = s_z^{(t)} & = -\cos\theta_t. \end{array}\right\} \tag{38}$$

(37) becomes on substitution

$$J = ik_0 Q_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{R}\left(\frac{r}{R} + n\cos\theta_t\right) e^{ik_0(R - nx'\sin\theta_t)} dx' dy' \tag{39}$$

It will be convenient to introduce an angle $\phi$ and an associated unit vector $s$, defined by

$$n\sin\theta_t = \sin\phi, \tag{40}$$

$$s_x = -\sin\phi, \quad s_y = 0, \quad s_z = -\cos\phi. \tag{41}$$

Since $k_0 r$ was assumed to be large compared with unity, the exponent in (39) will also be large compared with unity and the exponential term will oscillate rapidly and change sign many times as the point $(x', y')$ explores the domain of integration. Under these conditions, a good approximation to the value of $\boldsymbol{J}$ is obtained by the application of the following formula, which is a consequence of the principle of stationary phase (see Appendix III, eq. (20)):

$$\iint_D g\left(x', y'\right) e^{ik_0 f(x', y')} dx' dy' \sim \frac{2\pi i}{k_0} \sum_j \frac{\sigma_j}{\sqrt{\left|\alpha_j \beta_j - \gamma_j^2\right|}} g\left(x'_j, y'_j\right) e^{ik_0 f\left(x'_j, y'_j\right)}. \tag{42}$$

Here $(x', y',)$ are the points in the domain $D$ of integration at which $f$ is stationary and

$$\alpha_j = \left(\frac{\partial^2 f}{\partial x'^2}\right)_{x'_j, y'_j}, \quad \beta_j = \left(\frac{\partial^2 f}{\partial y'^2}\right)_{x'_j, y'_j}, \quad \gamma_j = \left(\frac{\partial^2 f}{\partial x'_j \partial y'_j}\right)_{x', y'_j}. \tag{43a}$$

and

$$\left.\begin{array}{rlll} & +1 & \text{when} & \alpha_j\beta_j > \gamma_j^2, \\ \sigma_j = & -1 & \text{when} & \alpha_j > 0, \quad , \quad \alpha_j < 0, \gamma_j^2. \\ & -i & \text{when} & \alpha_j\beta_j < \gamma_j^2, \end{array}\right\} \tag{43b}$$

In the present case, we have, if we also use (40) and (41),

$$g = \frac{1}{R}\left(\frac{r}{R} + \frac{\sin\phi}{\sin\theta_t}\cos\theta_t\right), \quad f = R - x'\sin\phi, \tag{44}$$

and consequently

$$\frac{\partial f}{\partial x'} = \frac{x'}{\sqrt{x'^2 + y'^2 + r^2}} - \sin\phi, \quad \frac{\partial f}{\partial y'} = \frac{y'}{\sqrt{x'^2 + y'^2 + r^2}}.$$

For $f$ to be stationary,

$$\frac{\partial f}{\partial x'} = \frac{\partial f}{\partial y'} = 0.$$

Hence (noting that the positive square root is to be taken in evaluating the partial derivatives), $f$ is seen to be stationary only when

$$x' = x'_1, \quad y' = y'_1,$$

where

$$x'_1 = r\tan\phi, \quad y'_1 = 0. \tag{45}$$

The expressions in (43) and (44) take the following values at this point:

$$\alpha_1 = \left(\frac{\partial^2 f}{\partial x'^2}\right)_{x'_1,y'_1} = \frac{\cos^3\phi}{r}, \quad \beta_1 = \left(\frac{\partial^2 f}{\partial y'^2}\right)_{x'_1,y'_1} = \frac{\cos\phi}{r}, \quad \gamma_1 = 0, \tag{46a}$$

$$\sigma_1 = 1, \tag{46b}$$

$$R_1 = (R)_{x'_1,y'_1} = \frac{r}{\cos\phi}, \tag{46c}$$

$$f(x'_1, y'_1) = r\cos\phi = r\cdot s, \tag{46 d}$$

$$g(x'_1, y'_1) = \frac{\cos\phi}{r}\frac{\sin(\phi+\theta_t)}{\sin\theta_t}. \tag{46e}$$

The formula (42) now gives

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{1}{R}\left(\frac{r}{R} + n\cos\theta_t\right)e^{ik_0(R - nx'\sin\theta_t)}dx'dy' = \frac{2\pi i}{k_0}\frac{\sin(\phi+\theta_t)}{\cos\phi\sin\theta_t}e^{ik_0(r.s)}$$

so that (39) becomes

$$J = \int_\Sigma\left\{Q\frac{\partial G}{\partial \nu'} - G\frac{\partial Q}{\partial \nu'}\right\}dS' = -2\pi\frac{\sin(\phi+\theta_t)}{\cos\phi\sin\theta_t}Q_0e^{ik_0(r.s)} \tag{47}$$

We now substitute this expression into the integro-differential equation (23), which expresses the extinction of the incident wave. We have

$$\operatorname{curl} Q_0e^{ik_0(r.s)} = ik_0(s\wedge Q_0)e^{ik_0(r.s)}$$
$$\operatorname{curlcurl} Q_0e^{ik_0(r.s)} = -{k_0}^2 s\wedge(s\wedge Q_0)e^{ik_0(r.s)}$$

Hence (23) reduces to

$$A_0^{(i)} e^{ik_0(r.s^{(i)})} = -2\pi k_0^2 \frac{\sin(\phi+\theta_t)}{\cos\phi \sin\theta_t} s \wedge (s \wedge Q_0)\, e^{ik_0(r.s)} \tag{48}$$

Now this must be an identity for all points $r$ on the boundary. Hence

$$s = s^{(i)}, \tag{49}$$

and

$$A_0^{(i)} = -2\pi k_0^2 \frac{\sin(\phi+\theta_t)}{\cos\phi \sin\theta_t} \left\{ s^{(i)} \wedge \left( s^{(i)} \wedge Q_0 \right) \right\} \tag{50}$$

Eq. (49) expresses the law of refraction § 1.5 (8). For it implies that

$$\phi = \theta_i \tag{51}$$

or, by (40),

$$n \sin\theta_t = \sin\theta_i. \tag{52}$$

Moreover (41) and (32) show that the vector $s^{(t)}$ lies in the plane specified by $s^{(i)}$ and the normal to $\Sigma$.

Eq. (50) connects the amplitudes of the incident wave (31) and the transmitted wave (34). Denoting the (vector) amplitude of the transmitted wave by $T_0$,

$$T_0 = 4\pi k_0^2 Q_0 \tag{53}$$

and expanding the vector triple product, (50) becomes

$$A_0^{(i)} = \frac{1}{2} \frac{\sin(\theta_i+\theta_t)}{\cos\theta_i \sin\theta_t} \left[ T_0 - s^{(i)} \left( s^{(i)}.T_0 \right) \right] \tag{54}$$

Let $A_\perp, A_\parallel$and $T_\perp, T_\mathrm{l}$ be the resolutes of ${A_0}^{(i)}$ and of $T_0$ in directions perpendicular and parallel to the plane of incidence. Then, remembering that ${A_0}^{(i)}$ is orthogonal to $s^{(i)}$ and $T_0$ is orthogonal to $s^{(t)}$, and that $s^{(i)}$ and $s^{(t)}$ are inclined to each other at an angle ( $\theta_i - \theta_t$ ), it follows from (54) that

$$A_\perp = \frac{1}{2} \frac{\sin(\theta_i+\theta_t)}{\cos\theta_i \sin\theta_t} T_\perp \tag{55a}$$

$$A_\mathbb{I} = \frac{1}{2} \frac{\sin(\theta_i+\theta_t)\cos(\theta_i-\theta_t)}{\cos\theta_i \sin\theta_t} T_1 \tag{55b}$$

These relations will be recognized as the Fresnel formulae § 1.5 (20a) for refraction.

Finally we consider the case when the point of observation is outside the medium ${}_1z > 0$ ). The calculations are strictly analogous, only $z' = -r$ must now be replaced by $z' = r$ in the appropriate formulae. This is equivalent to replacing $s_z$ by $-s_z$ in (46d), i.e. to replacing $\phi = \theta_i$ by $\phi = \theta_r$, where

$$\theta_r = \pi - \theta_i \tag{56}$$

In place of the unit vector $\boldsymbol{s} = \boldsymbol{s}^{(i)}$ we now introduce the unit vector $\boldsymbol{s}^{(r)}$ with components

$$s_x^{(r)} = -\sin\theta_r = -\sin\theta_i, \quad s_y^{(r)} = 0, \quad s_z^{(r)} = -\cos\theta_r = \cos\theta_i, \tag{57}$$

and obtain instead of (47),

$$J' = \int_\Sigma \left\{ Q\frac{\partial G}{\partial v'} - G\frac{\partial Q}{\partial v'} \right\} dS' = -2\pi \frac{\sin(\theta_i - \theta_t)}{\cos\theta_i \sin\theta_t} Q_0 e^{ik_0\left(r.s^{(r)}\right)} \tag{58}$$

(30) now gives for the time-free part of the reflected wave

$$A_0^{(r)} e^{ik_0\left(r.s^{(r)}\right)} = 2\pi k_0^2 \frac{\sin(\theta_i - \theta_t)}{\cos\theta_i \sin\theta_t} \left\{ s^{(r)} \wedge \left( s^{(r)} \wedge Q_0 \right) \right\} e^{ik_0\left(r.s^{(r)}\right)} \tag{59}$$

(59) represents a plane wave propagated in the direction specified by the unit vector $s^{(r)}$, this direction being related to the direction $s^{(i)}$ of the incident wave by (56); relation (56) expresses the law of reflection, in agreement with § 1.5 (7).

In terms of the amplitude of the transmitted wave, the amplitude ${A_0}^{(r)}$ of the reflected wave is

$$A_0^{(r)} = -\frac{1}{2} \frac{\sin(\theta_i - \theta_t)}{\cos\theta_i \sin\theta_t} \left[ T_0 - s^{(r)} \left( s^{(r)}.T_0 \right) \right] \tag{60}$$

Denoting by $R_\perp$ and $R_\parallel$the resolutes of ${A_0}^{(r)}$ in directions perpendicular and parallel to the plane of incidence, and using (55), we obtain

$$R_\perp = -\frac{1}{2} \frac{\sin(\theta_i - \theta_t)}{\cos\theta_i \sin\theta_t} T_\perp = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)} A_\perp \tag{61a}$$

and

$$R_\mathrm{l} = -\frac{1}{2} \frac{\sin(\theta_i - \theta_t)\cos(\theta_r - \theta_t)}{\cos\theta_i \sin\theta_t} T_\mathrm{n} = \frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)} A_\mathrm{n}. \tag{61b}$$

Relations (61) are the Fresnel formulae § 1.5 (21a) for reflection.

## 9.4 Foundations of Geometrical Optics

### 9.4.1 3.1 Approximation for Very Short Wavelengths

The electromagnetic field associated with the propagation of visible light is characterized by very rapid oscillations (frequencies of the order of $10^{14}\mathrm{sec}^{-1}$ ) or, what amounts to the same thing, by the smallness of the wavelength (of order $10^{-5}$ cm ). It may therefore be expected that a good first approximation to the propagation laws in such cases may be obtained by a complete neglect of the finiteness of the wavelength. It is found that for many optical problems such a procedure is entirely adequate; in fact, phenomena which can be attributed to departures from this approximate theory (so-called diffraction phenomena, studied in Chapter VIII) can only be demonstrated by means of carefully conducted experiments.

The branch of optics which is characterized by the neglect of the wavelength, i.e. that corresponding to the limiting case $\lambda_0 \to 0$, is known as geometrical optics,* since in this approximation the optical laws may be formulated in the language of geometry. The energy may then be regarded as being transported along certain curves (light rays). A physical model of a pencil of rays may be obtained by allowing the light from a source of negligible extension to pass through a very small opening in an opaque screen. The light which reaches the space behind the screen will fill a region the boundary of which (the edge of the pencil) will, at first sight, appear to be sharp. A more careful examination will reveal, however, that the light intensity near the boundary varies rapidly but continuously from darkness in the shadow to lightness in the illuminated region, and that the variation is not monotonic but is of an oscillatory character, manifested by the appearance of bright and dark bands, called diffraction fringes.

The region in which this rapid variation takes place is only of the order of magnitude of the wavelength. Hence, as long as this magnitude is neglected in comparison with the dimensions of the opening, we may speak of a sharply bounded pencil of rays. † On reducing the size of the opening down to the dimensions of the wavelength phenomena appear which need more refined study. If, however, one considers only the limiting case of negligible wavelengths, no restriction on the size of the opening is imposed, and we may say that an opening of vanishingly small dimensions defines an infinitely thin pencil-the light ray. It will be shown that the variation in the cross-section of a pencil of rays is a measure of the variation of the intensity of the light. Moreover it will be seen that it is possible to associate a state of polarization with each ray, and to study its variation along the ray.

Further it will be seen that for small wavelengths the field has the same general character as that of a plane wave, and, moreover, that within the approximation of geometrical optics the laws of refraction and reflection established for plane waves incident upon a plane boundary remain valid under more general conditions. Hence if a light ray falls on a sharp boundary (e.g. the surface of a lens) it is split into a reflected ray and a transmitted ray, and the changes in the state of polarization as well as the reflectivity and transmissivity may be calculated from the corresponding formulae for plane waves.

The preceding remarks imply that, when the wavelength is small enough, the sum total of optical phenomena may be deduced from geometrical considerations, by determining the paths of the light rays and calculating the associated intensity and polarization. We shall now formulate the appropriate laws by considering the implications of Maxwell's equations when $\lambda_0 \to 0$.*

### 3.1.1 Derivation of the eikonal equation

We consider a general time-harmonic field

$$\left.\begin{aligned} \boldsymbol{E}(\boldsymbol{r},\boldsymbol{t}) &= \boldsymbol{E}_0(\boldsymbol{r})e^{-i\omega t}, \\ \boldsymbol{H}(\boldsymbol{r},t) &= \boldsymbol{H}_0(\boldsymbol{r})e^{-i\omega t}, \end{aligned}\right\} \tag{1}$$

in a non-conducting isotropic medium. $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ denote complex vector functions of positions, and, as explained in § 1.4.3, the real parts of the expressions on the right-hand side of (1) are understood to represent the fields.

The complex vectors $\boldsymbol{E}_0$ and $\boldsymbol{H}_0$ will satisfy Maxwell's equations in their time-free form, obtained on substituting (1) into (1)-(4) of § 1.1. In regions free of currents and charges ($j = \rho = 0$), these equations are

$$\begin{aligned} \operatorname{curl} H_0 + ik_0\varepsilon E_0 &= 0, \\ \operatorname{curl} E_0 - ik_0\mu H_0 &= 0, \\ \operatorname{div} \varepsilon E_0 &= 0, \\ \operatorname{div} \mu H_0 &= 0. \end{aligned} \tag{5}$$

0

- The historical development of geometrical optics is described by M. Herzberger, Strahlenoptik (Berlin, Springer, 1931), 179; Z. Instrumentenkunde, 52 (1932), 429-435, 485-493, 534-542, C. Carathodory, Geometrische Optik (Berlin, Springer, 1937) and E. Mach, The Principles of Physical Optics, A Historical and Philosophical Treatment (First German edition 1913, English translation: London, Methuen, 1926; reprinted by Dover Publications, New York, 1953).
  † That the boundary becomes sharp in the limit as $\lambda_0 \to 0$ was first shown by G. Kirchnoff, Vorlesungen . Math. Phys., 2 (Mathematische Optik), (Leipzig, Teubner, 1891), p. 33. See also B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 2nd edition, 1950), p. 79, and A. Sommerfeld, Optics (New York, Academic Press, 1954), § 35.

Here the material relations $D = \varepsilon E, B = \mu H$ have been used and, as before, $k_0 = \omega/c = 2\pi/\lambda_0, \lambda_0$ being the vacuum wavelength.

We have seen that a homogeneous plane wave in a medium of refractive index $n = \sqrt{\varepsilon\mu}$, propagated in the direction specified by the unit vector $s$, is represented by

$$E_0 = ee^{ik_0 n(s.r)}, \quad H_0 = he^{ik_0 n(s.r)}, \tag{6}$$

where $e$ and $h$ are constant, generally complex vectors. For a (monochromatic) electric dipole field in the vacuum we found (cf. § 2.2) that

$$E_0 = ee^{ik_0 r}, \quad H_0 = he^{ik_0 r}, \tag{7}$$

$r$ being the distance from the dipole. Here $\boldsymbol{e}$ and $\boldsymbol{h}$ are no longer constant vectors, but at distances sufficiently far away from the dipole ( $r \gg \lambda_0$ ) these vectors are, with suitable normalization of the dipole moment, independent of $k_0$.

These examples suggest that in regions which are many wavelengths away from the sources we may represent more general types of fields in the form

$$E_0 = e(r)e^{ik_0 \mathscr{S}(r)}, \quad H_0 = h(r)e^{ik_0 \mathscr{S}(r)}, \tag{8}$$

where $\mathscr{S}(r)$, "the optical path", is a real scalar function of position, and $e(r)$ and $h(r)$ are vector functions of position, which may in general be complex.* With (8) as trial solution, Maxwell's equations lead to a set of relations between $\boldsymbol{e}, \boldsymbol{h}$ and $\mathscr{S}$. It will be shown that for large $k_0$ (small $\lambda_0$ ) these relations demand that $\mathscr{S}$ should satisfy a certain differential equation, which is independent of the amplitude vectors $\boldsymbol{e}$ and $\boldsymbol{h}$.

From (8), using well-known vector identities,

$$\operatorname{curl} \boldsymbol{H}_0 = (\operatorname{curl} \boldsymbol{h} + ik_0 \operatorname{grad} \mathscr{S} \wedge \boldsymbol{h})\, e^{ik_0 \mathscr{S}}, \tag{9}$$

$$\operatorname{div} \mu\boldsymbol{H}_0 = (\mu \operatorname{div} \boldsymbol{h} + \boldsymbol{h} \cdot \operatorname{grad} \mu + ik_0 \mu \boldsymbol{h} \cdot \operatorname{grad} \mathscr{S})\, e^{ik_0 \mathscr{S}}, \tag{10}$$

with similar expressions for curl $\boldsymbol{E}_0$ and div $\varepsilon\boldsymbol{E}_0$. Hence (2)-(5) become

$$\operatorname{grad} \mathscr{S} \wedge h + \varepsilon e = -\frac{1}{ik_0} \operatorname{curl} h \tag{11}$$

$$\operatorname{grad} \mathscr{S} \wedge e - \mu h = -\frac{1}{ik_0} \operatorname{curl} e \tag{12}$$

$$e \cdot \operatorname{grad} \mathscr{S} = -\frac{1}{ik_0}(e \cdot \operatorname{grad} \log \varepsilon + \operatorname{div} e) \tag{13}$$

$$h \cdot \operatorname{grad} \mathscr{S} = -\frac{1}{ik_0}(h \cdot \operatorname{grad} \log \mu + \operatorname{div} h) \tag{14}$$

0

- It was first shown by A. Sommerfeld and J. Runge, Ann.d. Physik, 35 (1911), 289, using a suggestion of P. Debye, that the basic equation of geometrical optics (the eikonal equation (15b)) may be derived from the (scalar) wave equation in the limiting case $\lambda_0 \to 0$. Generalizations which take into account the vectorial character of the electromagnetic field are due to W. Innatowsky, Trans. State Opt. Institute (Petrograd), 1 (1919), III; V. A. Fock, ibid., 3 (1924), 3; S. M. Rytov, Compt. Rend. (Doklady) Acad. Sci. URSS, 18 (1938), 263; N. Arley, Det. Kgl. Danske Videns Selsk., 22 (1945), No. 8; F. G. Friedlander, Proc. Cambr. Phil. Soc., 43 (1947), 284; K. Suchy, Ann. d. Physik., 11 (1952), 113, ibid., 12 (1953), 423, and ibid., 13 (1953), 178; R. S. Ingarden and A. Krzywicki, Acta Phys. Polonica, 14 (1955), 255.

We are interested in the solution for very large values of $k_0$. Hence as long as the multiplicative factors of $1/ik_0$ on the right-hand side are not exceptionally large they may be neglected, and the equations then reduce to

$$\begin{aligned} \operatorname{grad}\mathscr{S} \wedge \boldsymbol{h} + \varepsilon \boldsymbol{e} &= 0 \\ \operatorname{grad}\mathscr{S} \wedge \boldsymbol{e} - \mu \boldsymbol{h} &= 0 \\ \boldsymbol{e} \cdot \operatorname{grad}\mathscr{S} &= 0 \\ \boldsymbol{h} \cdot \operatorname{grad}\mathscr{S} &= 0 \end{aligned} \tag{14a}$$

We can confine our attention to equations (11a) and (12a) alone, since (13a) and (14a) follow from them on scalar multiplication with grad $\mathscr{S}$. Now (11a) and (12a) may be regarded as a set of six simultaneous homogeneous linear scalar equations for the Cartesian components $e_x, h_x, \ldots$ of $e$ and $h$. These simultaneous equations have non-trivial solutions only if a consistency condition (the vanishing of the associated

determinant) is satisfied. This condition may be obtained simply by eliminating $\boldsymbol{e}$ or $\boldsymbol{h}$ between (11a) and (12a). Substituting for $h$ from (12a), (11a) becomes

$$\frac{1}{\mu}\left[(e \cdot \operatorname{grad}\mathscr{S}) \operatorname{grad}\mathscr{S} - e(\operatorname{grad}\mathscr{S})^2\right] + \varepsilon e = 0.$$

The first term vanishes on account of (13a), and the equation then reduces, since $e$ does not vanish everywhere, to

$$(\operatorname{grad}\mathscr{S})^2 = n^2, \tag{15a}$$

or, written explicitly,

$$\left(\frac{\partial\mathscr{S}}{\partial x}\right)^2 + \left(\frac{\partial\mathscr{S}}{\partial y}\right)^2 + \left(\frac{\partial\mathscr{S}}{\partial z}\right)^2 = n^2(x, y, z) \tag{15b}$$

where as before $n = \sqrt{\varepsilon\mu}$ denotes the refractive index. The function $\mathscr{S}$ is often called the eikonal* and (15b) is known as the eikonal equation; it is the basic equation of geometrical optics. † The surfaces

$$\mathscr{S}(\boldsymbol{r}) = \text{ constant}$$

may be called the geometrical wave surfaces or the geometrical wave-fronts. ‡
The eikonal equation was derived here by using the first-order Maxwell's equations, but it may also be derived from the second-order wave equations for the electric or magnetic field vectors. To show this one substitutes from (1) and (8) into the wave equation § 1.2 (5) and obtains, after a straightforward calculation,

$$\boldsymbol{K}(\boldsymbol{e}, \mathscr{S}, n) + \frac{1}{ik_0}\boldsymbol{L}(\boldsymbol{e}, \mathscr{S}, n, \mu) + \frac{1}{(ik_0)^2}\mathbf{M}(\boldsymbol{e}, \varepsilon, \mu) = 0, \tag{16}$$

where

$$\begin{aligned} \boldsymbol{K}(\boldsymbol{e}, \mathscr{S}, n) &= \left\{n^2 - (\operatorname{grad}\mathscr{S})^2\right\}\boldsymbol{e}, \\ \mathbf{L}(\boldsymbol{e}, \mathscr{S}, n, \mu) &= \left\{\operatorname{grad}\mathscr{S} \cdot \operatorname{grad}\log\mu - \nabla^2\mathscr{S}\right\}\boldsymbol{e} - 2\{\boldsymbol{e} \cdot \operatorname{grad}\log n\}\operatorname{grad}\mathscr{S} \\ &\quad - 2\{\operatorname{grad}\mathscr{S} \cdot \operatorname{grad}\}\boldsymbol{e}, \\ \mathbf{M}(\boldsymbol{e}, \varepsilon, \mu) &= \operatorname{curl}\boldsymbol{e} \wedge \operatorname{grad}\log\mu - \nabla^2\boldsymbol{e} - \operatorname{grad}(\boldsymbol{e} \cdot \operatorname{grad}\log\varepsilon). \end{aligned}$$

0

- Complex $e$ and $h$ are necessary, if all possible states of polarization are to be included. According to § 1.4 (75) real $e$ and $h$ correspond to fields which are linearly polarized.

The corresponding equation involving $\boldsymbol{h}$, obtained on substitution into the wave equation (6) in § 1.2 for $\boldsymbol{H}$ (or more simply by using the fact that Maxwell's equations remain unchanged when $\boldsymbol{E}$ and $\boldsymbol{H}$ and simultaneously $\varepsilon$ and $-\mu$ are interchanged), is

$$\mathbf{K}(\boldsymbol{h}, \mathscr{S}, n) + \frac{1}{(ik_0)}\mathbf{L}(\boldsymbol{h}, \mathscr{S}, n, \varepsilon) + \frac{1}{(ik_0)^2}\mathbf{M}(\boldsymbol{h}, \mu, \varepsilon) = 0. \tag{17}$$

For sufficiently large $k_0$ the second and third terms may in general be neglected; then $K = 0$, giving again the eikonal equation. It will be seen later that the terms in the first power of $1/(ik_0)$ in (16) and (17) also possess a physical interpretation.

It may be shown that in many cases of importance the spatial parts $\boldsymbol{E_0}$ and $\boldsymbol{H_0}$ of the field vectors may be developed into asymptotic series of the form*

$$E_0 = e^{ik_0\mathscr{S}} \sum_{m\geqslant 0} \frac{e^{(m)}}{(ik_0)^m}, \quad H_0 = e^{ik_0\mathscr{S}} \sum_{m\geqslant 0} \frac{h^{(m)}}{(ik_0)^m}, \tag{18}$$

where $\boldsymbol{e}^{(m)}$ and $\boldsymbol{h}^{(m)}$ are functions of position, and $\mathscr{S}$ is the same function as before. † Geometrical optics corresponds to the leading terms of these expansions.

### 3.1.2 The light rays and the intensity law of geometrical optics

From (8), and from (54) and (55) in § 1.4, it follows that the time averages of the electric and magnetic energy densities $\langle w_e \rangle$ and $\langle w_m \rangle$ are given by

$$\langle w_e \rangle = \frac{\varepsilon}{16\pi} e \cdot e^\star, \quad \langle w_m \rangle = \frac{\mu}{16\pi} h \cdot h^\star \tag{19}$$

Substitution for $e^\star$ from (11a) and for $h$ from (12a) gives

$$\langle w_e \rangle = \langle w_m \rangle = \frac{1}{16\pi} \left[ e, h^\star, \operatorname{grad} \mathscr{S} \right], \tag{20}$$

the square bracket denoting the scalar triple product. Hence, within the accuracy of geometrical optics, the time-averaged electric and magnetic energy densities are equal.

The time average of the Poynting vector is obtained from (8) and § 1.4 (52):

$$\langle S \rangle = \frac{c}{8\pi} \mathscr{R} \left( e \wedge h^\star \right)$$

Using (12a), we obtain

$$\langle \boldsymbol{S} \rangle = \frac{c}{8\pi\mu} \left\{ (\boldsymbol{e} \cdot \boldsymbol{e}^\star) \operatorname{grad} \mathscr{S} - (\boldsymbol{e} \cdot \operatorname{grad} \mathscr{S}) \boldsymbol{e}^\star \right\}.$$

The last term vanishes on account of (13a) so that we have, if use is made of the expression for $\langle w_e \rangle$ and of Maxwell's relation $\varepsilon\mu = n^2$,

---

0

- The term eikonal (from Greek $\varepsilon\iota\kappa\tilde{\omega}\nu$ = image) was introduced in 1895 by H. Bruns to describe certain related functions (cf. p. 133), but has come to be used in a wider sense.
  † The eikonal equation may also be regarded as the equation of the characteristics of the wave equations (5) and (6), in § 1.2, for $\boldsymbol{E}$ and $\boldsymbol{H}$, and describes rigorously the propagation of discontinuities of the solutions of these equations. In geometrical optics we are, however, not concerned with the propagation of discontinuities but with time-harmonic (or nearly time-harmonic) solutions. The formal equivalence of the two interpretations is demonstrated in Appendix VI.

The eikonal equation will also be recognized as the Hamilton-Jacobi equation of the variational problem $\delta \int n ds = 0$, the optical counterpart of which goes back to Fermat (cf. §3.3.2 and Appendix I).
‡ In future we shall drop the adjective "geometrical" when there is no risk of confusion.

$$\langle \boldsymbol{S} \rangle = \frac{2c}{n^2} \langle w_e \rangle \operatorname{grad} \mathscr{S}. \tag{21}$$

Since $\langle w_e \rangle = \langle w_m \rangle$, the term $2\langle w_e \rangle$ represents the time average $\langle w \rangle$ of the total energy density (i.e. $\langle w \rangle = \langle w_e \rangle + \langle w_m \rangle$ ). Also, on account of the eikonal equation, $(\operatorname{grad} \mathscr{S})/n$ is a unit vector ( $s$ say),

$$s = \frac{\operatorname{grad} \mathscr{S}}{n} = \frac{\operatorname{grad} \mathscr{S}}{|\operatorname{grad} \mathscr{S}|}, \tag{22}$$

and (21) shows that $s$ is in the direction of the average Poynting vector. If, as before, we set $c/n = v$, (21) becomes

$$\langle \boldsymbol{S} \rangle = v \langle w \rangle \boldsymbol{s}. \tag{23}$$

Hence the average Poynting vector is in the direction of the normal to the geometrical wave-front, and its magnitude is equal to the product of the average energy density and the velocity $v = c/n$. This result is analogous to the relation (9) in § 1.4 for plane waves,

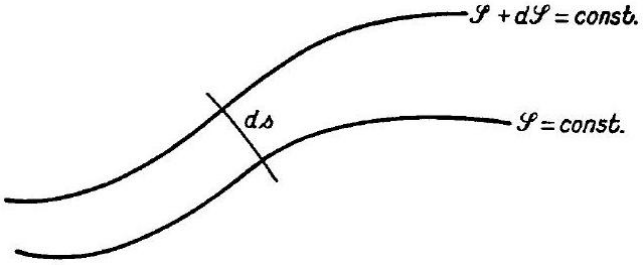



Fig. 3.1. Illustrating the meaning of the relation $n s=\operatorname{grad} \mathscr{S}$.

and shows that within the accuracy of geometrical optics the average energy density is propagated with the velocity $v = c/n$.

The geometrical light rays may now be defined as the orthogonal trajectories to the geometrical wave-fronts $\mathscr{S} =$ constants. We shall regard them as oriented curves whose direction coincides everywhere with the direction of the average Poynting vector.* If $r(s)$ denotes the position vector of a point $P$ on a ray, considered as a function of the length of arc $s$ of the ray, then $dr/ds = s$, and the equation of the ray may be written as

$$n \frac{dr}{ds} = \operatorname{grad} \mathscr{S}. \tag{24}$$

From (13a) and (14a) it is seen that the electric and magnetic vectors are at every point orthogonal to the ray.

0

- We assume here that only one geometrical wave-front passes through each point. In some cases, for example when reflection takes place at obstacles present in the medium, several wave-fronts may pass through each point. The resulting field is then represented by the addition of series of the above type.
† The theory of such asymptotic expansions has its origin chiefly in the work of R . K. Luneburg, Propagation of Electromagnetic Waves (mimeographed lecture notes, New York University, 1947-1948). See also M. Kline, Comm. Pure and Appl. Math., 4 (1951), 225; ibid., 8 (1955), 595 and W. Braunbeck, Z. Naturforsch., 6 (1951), 672. A comprehensive account of the theory is given in M. Kline and I. W. Kay, Electromagnetic Theory and Geometrical Optics [New York, Interscience Publishers, (1965)].

The meaning of (24) may be made clearer from the following remarks. Consider two neighbouring wave-fronts $\mathscr{S} = \text{constant}$ and $\mathscr{S} + d\mathscr{S} = \text{constant}$ (Fig. 3.1). Then

$$\frac{d\mathscr{S}}{ds} = \frac{dr}{ds} \cdot \operatorname{grad} \mathscr{S} = n. \tag{25}$$

Hence the distance $ds$ between points on the opposite ends of a normal cutting the two wave-fronts is inversely proportional to the refractive index, i.e. directly proportional to $v$.

The integral $\int_C n ds$ taken along a curve $C$ is known as the optical length of the curve. Denoting by square brackets the optical length of the ray which joins points $P_1$ and $P_2$, we have

$$[P_1 P_2] = \int_{P_1}^{P_2} n ds = \mathscr{S}(P_2) - \mathscr{S}(P_1) \tag{26}$$

Since, as we have seen, the average energy density is propagated with the velocity $v = c/n$ along the ray,

$$nds = \frac{c}{v} ds = cdt,$$

where $dt$ is the time needed for the energy to travel the distance $ds$ along the ray; hence

$$[P_1 P_2] = c \int_{P_1}^{P_2} dt \tag{27}$$

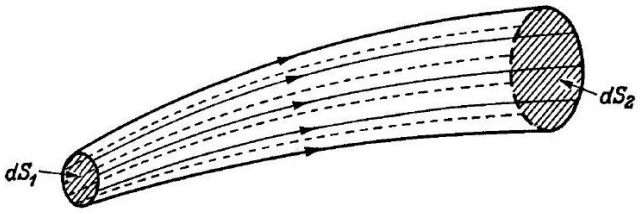



Fig. 3.2. Illustrating the intensity law of geometrical optics.

i.e. the optical length $[P_1 P_2]$ is equal to the product of the vacuum velocity of light and the time needed for light to travel from $P_1$ to $P_2$.

The intensity of light $I$ was defined as the absolute value of the time average of the Poynting vector. We therefore have from (23),

$$I = |\langle S \rangle| = v \langle w \rangle, \tag{28}$$

and the conservation law § 1.4 (57) gives

$$\operatorname{div}(Is) = 0. \tag{29}$$

To see the implications of this relation we take a narrow tube formed by all the rays proceeding from an element $dS_1$ of a wave-front $\mathscr{S}(r) = a_1$ ( $a_1$ being a constant), and denote by $dS_2$ the corresponding element in which these rays intersect another wavefront $\mathscr{S}(\boldsymbol{r}) = a_2$ (Fig. 3.2). Integrating (29) throughout the tube and applying Gauss' theorem we obtain

0

- This definition of light rays is appropriate for isotropic media only. We shall see later (Chapter XIV) that in an anisotropic medium the direction of the wave-front normal does not, in general, coincide with the direction of the Poynting vector.

$$\int Is.vdS = 0 \tag{30}$$

$\nu$ denoting the outward normal to the tube. Now

$$\begin{aligned} s \cdot \nu &= 1 \text{ on } dS_2 \\ &= -1 \text{ on } dS_1 \\ &= 0 \text{ elsewhere,} \end{aligned}$$

so that (30) reduces to

$$I_1 dS_1 = I_2 dS_2, \tag{31}$$

$I_1$ and $I_2$ denoting the intensity on $dS_1$ and on $dS_2$ respectively. Hence $IdS$ remains constant along a tube of rays. This result expresses the intensity law of geometrical optics.

We shall see later that in a homogeneous medium the rays are straight lines. The intensity law may then be expressed in a somewhat different form. Assume first that $dS_1$, and consequently also $dS_2$, are bounded by segments of lines of curvature (see Fig. 3.3). If $R_1$ and $R_1'$ are the principal radii of curvature (cf. §4.6.1) of the segments $A_1B_1$ and $B_1C_1$, then

$$A_1B_1 = R_1 d\theta, \quad B_1C_1 = R_1' d\phi,$$

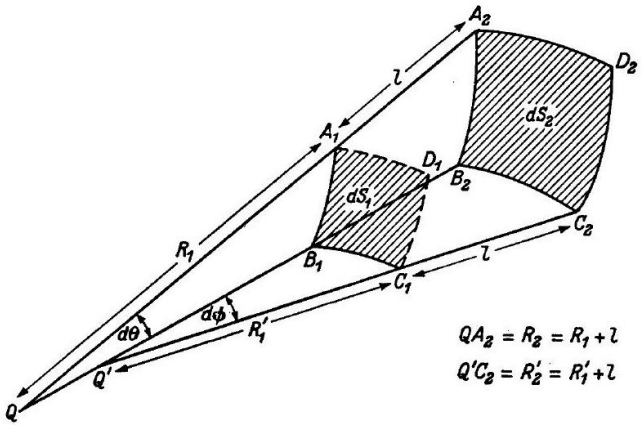



Fig. 3.3. Illustrating the intensity law of geometrical optics for rectilinear rays.

where $d\theta$ and $d\phi$ are the angles which $A_1B_1$ and $B_1C_1$ subtend at the respective centres of curvature $Q$ and $Q'$. Hence

$$dS_1 = A_1B_1 \cdot B_1C_1 = R_1R_1' d\theta d\phi \tag{32}$$

similarly for an element $dS_2$ in which the bundle of rays through $dS_1$ meets another wavefront of the family,

$$dS_2 = A_2B_2.B_2C_2 = R_2R_2' d\theta d\phi \tag{33}$$

If $l$ is the distance between $dS_1$ and $dS_2$ measured along the rays, then

$$R_2 = R_1 + l, \quad R_2' = R_1' + l,$$

and it follows that

$$\frac{I_2}{I_1} = \frac{dS_1}{dS_2} = \frac{R_1 R_1'}{R_2 R_2'} = \frac{R_1 R_1'}{(R_1 + l)(R_1' + l)}. \tag{34}$$

If the areas $dS_1$ and $dS_2$ are bounded by arbitrary curves, (34) still holds. This can immediately be seen if we regard them as made up of a number of elements bounded by lines of curvature, and sum their contributions.

If $R_1 \ll l, R_1' \ll l$, (34) reduces to

$$\frac{I_2}{I_1} = \frac{R_1 R_1'}{l^2}. \tag{35}$$

This formula is sometimes used in connection with the scattering of radiation.
The reciprocal $1/RR'$ of the product of the two principal radii of curvature is called the Gaussian curvature (or second curvature) of the surface. (34) shows that the
intensity at any point of a rectilinear ray is proportional to the Gaussian curvature of the wave-front which passes through that point. In particular if all the (rectilinear) rays have a point in common, the wave-fronts are spheres centred at that point; then $R_1 = R_1'$, $R_2 = R_2'$ and we obtain (dropping the suffixes) the inverse square law

$$I = \frac{\text{constant}}{R^2} \tag{36}$$

Returning to the general case of an arbitrary pencil of rays (curved or straight), we can write down an explicit expression in terms of the $\mathscr{S}$ function for the variation of the intensity along each ray. Substituting for $s$ from (22) into (29), and using the identities $\operatorname{div} uv = u \operatorname{div} v + v \cdot \operatorname{grad} u$, and $\operatorname{div} \operatorname{grad} = \nabla^2$, we obtain

$$\frac{I}{n}\nabla^2 \mathscr{S} + \operatorname{grad} \mathscr{S} . \operatorname{grad} \frac{I}{n} = 0.$$

This may also be written as

$$\nabla^2 \mathscr{S} + \operatorname{grad} \mathscr{S} \cdot \operatorname{grad} \log \frac{I}{n} = 0. \tag{37}$$

Let us now introduce the operator

$$\frac{\partial}{\partial \tau} = \operatorname{grad} \mathscr{S} . \operatorname{grad}, \tag{38}$$

where $\tau$ is a parameter which specifies position along the ray. Then (37) may be written as

$$\frac{\partial}{\partial \tau} \log \frac{I}{n} = -\nabla^2 \mathscr{S}$$

whence, on integration,

$$I = n e^{-\int^\tau \nabla^2 \mathscr{S} d\tau}$$

But by (38), (15), and (25),

$$d\tau = \frac{d\mathscr{S}}{(\operatorname{grad} \mathscr{S})^2} = \frac{1}{n^2} d\mathscr{S} = \frac{1}{n} ds, \tag{39}$$

so that we finally obtain the following expressions for the ratio of the intensities at any two points of a ray:

$$\frac{I_2}{I_1} = \frac{n_2}{n_1} e^{-\int_{\mathscr{S}_1}^{\mathscr{S}_2} \frac{\nabla^* \mathscr{S}}{n^2} d\mathscr{S}} = \frac{n_2}{n_1} e^{-\int_{s_1}^{s_2} \frac{\nabla^* \mathscr{S}}{n} ds}, \tag{40}$$

the integrals being taken along the ray*.

### 3.1.3 Propagation of the amplitude vectors

We have seen that, when the wavelength is sufficiently small, the transport of energy may be represented by means of a simple hydrodynamical model which may be completely described in terms of the real scalar function $\mathscr{S}$, this function being a solution of the eikonal equation (15). According to traditional terminology, one understands by geometrical optics this approximate picture of energy propagation, using the concept of rays and wave-fronts. In other words polarization properties

are excluded. The reason for this restriction is undoubtedly due to the fact that the simple laws of geometrical optics concerning rays and wave-fronts were known from experiments long before the electromagnetic theory of light was established. It is however possible, and from our point of view quite natural, to extend the meaning of geometrical optics to embrace also certain geometrical laws relating to the propagation of the "amplitude vectors" $\boldsymbol{e}$ and $\boldsymbol{h}$. These laws may be easily deduced from the wave equations (16)-(17).

Since $\mathscr{S}$ satisfies the eikonal equation, it follows that $K = 0$, and we see that when $k_0$ is sufficiently large ( $\lambda_0$ small enough), only the $L$-terms need to be retained in (16) and (17). Hence, in the present approximation, the amplitude vectors and the eikonal are connected by the relations $L = 0$. If we use again the operator $\partial/\partial\tau$ introduced by (38), the equations $L = 0$ become

$$\frac{\partial e}{\partial \tau} + \frac{1}{2}\left(\nabla^2 \mathscr{S} - \frac{\partial \log \mu}{\partial \tau}\right) e + (e \cdot \operatorname{grad} \log n) \operatorname{grad} \mathscr{S} = 0 \tag{41}$$

$$\frac{\partial h}{\partial \tau} + \frac{1}{2}\left(\nabla^2 \mathscr{S} - \frac{\partial \log \varepsilon}{\partial \tau}\right) h + (h \cdot \operatorname{grad} \log n) \operatorname{grad} \mathscr{S} = 0 \tag{42}$$

These are the required transport equations for the variation of $\boldsymbol{e}$ and $\boldsymbol{h}$ along each ray. The implications of these equations can best be understood by examining separately the variation of the magnitude and of the direction of these vectors.

We multiply (41) scalarly by $e^\star$ and add to the resulting equation the corresponding equation obtained by taking the complex conjugate. This gives

$$\frac{\partial}{\partial \tau}(\boldsymbol{e} \cdot \boldsymbol{e}^\star) + \left(\nabla^2 \mathscr{S} - \frac{\partial \log \mu}{\partial \tau}\right) \boldsymbol{e} \cdot \boldsymbol{e}^\star = 0. \tag{43}$$

On account of the identity $\operatorname{div} uv = u \operatorname{div} v + v \cdot \operatorname{grad} u$, the second and third term may be combined as follows:

$$\nabla^2 \mathscr{S} - \frac{\partial \log \mu}{\partial \tau} = \nabla^2 \mathscr{S} - \operatorname{grad} \mathscr{S} \cdot \operatorname{grad} \log \mu = \mu \operatorname{div}\left(\frac{1}{\mu} \operatorname{grad} \mathscr{S}\right). \tag{44}$$

Integrating (43) along a ray, the following expression for the ratio of $e \cdot e^\star$ at any two points of the ray is obtained:*

$$\frac{(e \cdot e^\star)_2}{(e \cdot e^\star)_1} = e^{-\int_{\tau_1}^{\tau_2} \mu \operatorname{div}\left(\frac{1}{\mu} \operatorname{grad} \mathscr{S}\right) d\tau} = e^{-\int_{s_1}^{s_2} \sqrt{\frac{\mu}{\varepsilon}} \operatorname{div}\left(\frac{1}{\mu} \operatorname{grad} \mathscr{S}\right) ds} \tag{45}$$

0

- It has been shown by M. Kline, Comm. Pure and Appl. Maths., 14 (1961), 473 that the intensity ratio (40) may be expressed in terms of an integral which involves the principal radii of curvature of the associated wavefronts. Kline's formula is a natural generalization, to inhomogeneous media, of the formula (34). See also M. Kline and I. W. Kay, Electromagnetic Theory and Geometrical Optics (New York, Interscience Publishers, 1965), p. 184.

- This relation may also be written in the alternative form

$$\left(\frac{e\cdot e^\star}{\mu}\right)_2 = \left(\frac{e\cdot e^\star}{\mu}\right)_1 e^{-\int_{s_1}^{s_2}\frac{\nabla^2\mathscr{S}}{n}ds} \qquad (45a)$$

which follows when (43) is re-written in the form

$$\frac{\partial}{\partial\tau}\left[\log\left(\frac{e\cdot e^\star}{\mu}\right)\right] = -\nabla^2\mathscr{S},$$

and the integral is taken along a ray. (45a) is in fact only another way of expressing the relation (40) for the variation of intensity, and follows from it when the relation

$$I = \frac{2c}{n}\langle w_e\rangle = \frac{c\varepsilon}{8\pi n}(e\cdot e^\star)$$

and the Maxwell formula $\varepsilon\mu = n^2$ are used.
Similarly

$$\frac{(\boldsymbol{h}\cdot\boldsymbol{h}^\star)_2}{(\boldsymbol{h}\cdot\boldsymbol{h}^\star)_1} = e^{-\int_{s_1}^{s_2}\sqrt{\frac{\varepsilon}{\mu}}\,\mathrm{div}\left(\frac{1}{\varepsilon}\,\mathrm{grad}\,\mathscr{S}\right)ds} \qquad (46)$$

Next consider the variation of the complex unit vectors

$$u = \frac{e}{\sqrt{e\cdot e^\star}}, \quad v = \frac{h}{\sqrt{h\cdot h^\star}} \qquad (47)$$

along each ray. Substitution into (41) gives

$$\frac{\partial u}{\partial\tau} + \frac{1}{2}\left[\frac{\partial\log(e\cdot e^\star)}{\partial\tau} + \nabla^2\mathscr{S} - \frac{\partial\log\mu}{\partial\tau}\right]u + (u\cdot\mathrm{grad}\log n)\,\mathrm{grad}\,\mathscr{S} = 0$$

The second, third and fourth terms vanish on account of (43), and it follows that

$$\frac{du}{d\tau} \equiv n\frac{du}{ds} = -(u\cdot\mathrm{grad}\log n)\,\mathrm{grad}\,\mathscr{S} \qquad (48)$$

and similarly

$$\frac{d\mathbf{v}}{d\tau} \equiv n\frac{d\mathbf{v}}{ds} = -(\mathbf{v}\cdot\mathrm{grad}\log n)\,\mathrm{grad}\,\mathscr{S}. \qquad (49)$$

This is the required law for the variation of $u$ and $v$ along each ray.* In particular, for a homogeneous medium ( $n$ = constant) (48) and (49) reduce to $du/ds = dv/ds = 0$ so that $u$ and $v$ then remain constant along each ray.

Finally we note that for a time-harmonic homogeneous plane wave in a homogeneous medium, $\mathscr{S} = ns.r$ and $e, h, \varepsilon$ and $\mu$ are all constants, and consequently $K = L = M \equiv 0$ in (16). Such a wave (whatever its frequency) therefore obeys rigorously the laws of geometrical optics.

### 3.1.4 Generalizations and the limits of validity of geometrical optics

The considerations of the preceding sections apply to a strictly monochromatic field. Such a field, which may be regarded as a typical Fourier component of an arbitrary field, is produced by a harmonic oscillator, or by a set of such oscillators of the same frequency.

In optics one usually deals with a source which emits light within a narrow, but nevertheless finite, frequency range. The source may then be regarded as arising from a large number of

harmonic oscillators whose frequencies fall within this range. To obtain the intensity at a typical field point $P$ one has to sum the individual fields produced by each oscillator (element of the source):

$$E = \sum_n E_n, \quad H = \sum_n H_n \tag{50}$$

The intensity is then given by (using real representation)

$$\begin{aligned} I(P) = |\langle S \rangle| &= \frac{c}{4\pi} |\langle E \wedge H \rangle| = \frac{c}{4\pi} \left| \sum_{n,m} \langle E_n \wedge H_m \rangle \right| \\ &= \frac{c}{4\pi} \left| \sum_n \langle E_n \wedge H_n \rangle + \sum_{n \neq m} \langle E_n \wedge H_m \rangle \right|. \end{aligned} \tag{51}$$

In many optical problems it is usually permissible to assume that the second sum in (51) vanishes (the fields are then said to be incoherent), so that

$$I(P) = \frac{c}{4\pi} \left| \sum_n \langle E_n \wedge H_n \rangle \right| = \left| \sum_n \langle S_n \rangle \right|, \tag{52}$$

$\boldsymbol{S}_n$ denoting the Poynting vector due to the $n$th element of the source. It is not possible to discuss at this stage the conditions under which the neglect of the second

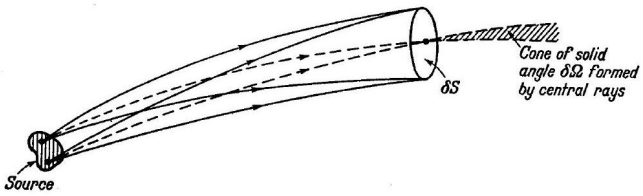



Fig. 3.4. Illustrating the intensity law of geometrical optics for an extended incoherent source.

term in (51) is justified, but this point will be considered fully later, in connection with partial coherence (Chapter X).

Let $\delta S$ be a small portion of a wave-front associated with one particular element of the source. Every element of the source sends through $\delta S$ a tube of rays, and the central rays of these tubes fill a cone of solid angle $\delta\Omega$ (Fig. 3.4). If the semi-vertical angle of this cone is small enough, we may neglect the variation of $S_n$ with direction, and (52) may then be replaced by

---

0

- The relations (48) and (49) have an interesting interpretation in terms of non-Euclidean geometry. If we consider the associated non-Euclidean space whose line element is given by

$$ds' = nds = n\sqrt{dx^2 + dy^2 + dz^2}$$

then the geometrical light rays correspond to geodesics in this space, and (48) and (49) may be shown to imply that each of the two vectors $u$ and $v$ is transferred parallel to itself (in the sense of Levi-Civita parallelism) along each ray. Cf. F. Bortolotti, Rend. R. Acc. Naz. Linc., 6a, 4 (1926), 552 ; R. K. Luneburg, Mathematical Theory of Optics (mimeographed lecture notes, Brown University, Providence, R.I., 1944, p. 55-59; printed version published by University of California Press, Berkeley and Los Angeles, 1964, p. 51-55); M. Kline and I. W. Kay, Electromagnetic Theory and Geometrical Optics (New York, Interscience Publishers, 1965), p. 180-183.

$$I(P) = \sum_n |\langle S_n \rangle| = \sum_n I_n. \tag{53}$$

Now the number of elements (oscillators) may be regarded as being so large that no appreciable error is introduced by treating the distribution as continuous. The contribution due to each element is then infinitesimal, but the total effect is finite. The sum (integral) is then proportional to $\delta\Omega$ :

$$I(P) = B\delta\Omega,$$

and the total (time-averaged) energy flux $\delta F$ which crosses the element $\delta S$ per unit time is given by

$$\delta F = B\delta\Omega\delta S. \tag{54}$$

This formula is of importance in photometry, and will be used later.
We must now briefly consider the limits of validity of geometrical optics. The eikonal equation was derived on the assumption that the terms on the right-hand sides of (11) and (12) may be negleeted. If the dimensionless quantities $\varepsilon, \mu$ and
$|\operatorname{grad} \mathscr{S}|$ are assumed to be of order unity, we see that this neglect will be justified provided that the magnitudes of the changes in $\boldsymbol{e}$ and $\boldsymbol{h}$ are small compared with the magnitudes of $\boldsymbol{e}$ and $\boldsymbol{h}$ over domains whose linear dimensions are of the order of a wavelength. This condition is violated, for example, at boundaries of shadows, for across such boundaries the intensity (and therefore also $\boldsymbol{e}$ and $\boldsymbol{h}$ ) changes rapidly. In the neighbourhood of points where the intensity distribution has a very sharp maximum (e.g. at a focus, see § 8.8), geometrical optics likewise cannot be expected to describe correctly the behaviour of the field.

The transport equations (41) and (42) for the complex amplitude vector $e$ and $h$ were obtained on the assumption that $\mathscr{S}$ satisfies the eikonal equation, and that the terms $\lambda_0|M(e, \varepsilon, \mu)|$ and $\lambda_0|M(h, \mu, \varepsilon)|$ are small compared with $|L(e, \mathscr{S}, n, \mu)|$ and $|\boldsymbol{L}(\boldsymbol{h}, \mathscr{S}, n, \varepsilon)|$ respectively. This imposes certain additional restrictions on, not only the first, but also the second derivatives of $\boldsymbol{e}$ and $\boldsymbol{h}$. These conditions are rather complicated and will not be studied here.

It is, of course, possible to obtain improved approximations by retaining some of the higher-order terms in the expansions (18) for the field vectors.* In problems of instrumental optics, the practical advantage of such a procedure is, however, doubtful, since the closer the special regions are approached the more terms have to be included, and the expansions usually break down completely at points of particular interest (e.g. at a focus or at a caustic surface). A more powerful approach to the study of the intensity distribution in such regions is offered by methods which will be discussed in the chapters on diffraction.

Finally we stress that the simplicity of the geometrical optics model arises essentially from the fact that, in general, the field behaves locally as a plane wave. At optical wavelengths, the regions for which this simple geometrical model is inadequate are an exception rather than a rule; in fact for most optical problems geometrical optics furnishes at least a good starting point for more refined investigations.

## 9.4.2 3.2 General Properties of Rays

### 3.2.1 The differential equation of light rays

The light rays have been defined as the orthogonal trajectories to the geometrical wavefronts $\mathscr{S}(x, y, z) =$ constant and we have seen that, if $r$ is a position vector of a typical point on a ray and $s$ the length of the ray measured from a fixed point on it, then

$$n\frac{dr}{ds} = \operatorname{grad}\mathscr{S}. \tag{1}$$

This equation specifies the rays by means of the function $\mathscr{S}$, but one can easily derive from it a differential equation which specifies the rays directly in terms of the refractive index function $n(r)$.

Differentiating (1) with respect to $s$ we obtain

$$\begin{aligned}\frac{a}{ds}\left(n\frac{dr}{ds}\right) &= \frac{d}{ds}(\operatorname{grad}\mathscr{S}) \\ &= \frac{dr}{ds}\cdot\operatorname{grad}(\operatorname{grad}\mathscr{S}) \\ &= \frac{1}{n}\operatorname{grad}\mathscr{S}\cdot\operatorname{grad}(\operatorname{grad}\mathscr{S}) \quad (\text{ by (1)}) \\ &= \frac{1}{2n}\operatorname{grad}\left[(\operatorname{grad}\mathscr{S})^2\right] \\ &= \frac{1}{2n}\operatorname{grad}n^2 \quad (\text{ by §3.1(15)})\end{aligned}$$

i.e.

$$\frac{d}{ds}\left(n\frac{dr}{ds}\right) = \operatorname{grad}n. \tag{2}$$

This is the vector form of the differential equations of the light rays. In particular, in a homogeneous medium $n =$ constant and (2) then reduces to

$$\frac{d^2r}{ds^2} = 0,$$

whence

$$\boldsymbol{r} = s\boldsymbol{a} + \boldsymbol{b}, \tag{3}$$

$a$ and $b$ being constant vectors. (3) is a vector equation of a straight line in the direction of the vector $\boldsymbol{a}$, passing through the point $\boldsymbol{r} = \boldsymbol{b}$. Hence in a homogeneous medium the light rays have the form of straight lines.

As an example of some interest, let us consider rays in a medium which has spherical symmetry, i.e. where the refractive index depends only on the distance $r$ from a fixed point $O$ :

$$n = n(r). \tag{4}$$

This case is approximately realized by the earth's atmosphere, when the curvature of the earth is taken into account.

Consider the variation of the vector $r \wedge [n(r)s]$ along the ray. We have

---

0

- It has been suggested by J. B. Keller [J. Appl. Phys., 28 (1957), 426; also Calculus of Variations and its Application, ed. L. M. Graves (New York, McGraw-Hill, 1958), 27] that the behaviour of the contributions represented by the higher-order terms may be studied by means of a model which is an extension of ordinary geometrical optics. In this theory the concept of a diffracted ray is introduced, which obeys a generalized Fermat's principle. With each such ray an appropriate field is associated and is assumed to satisfy the same propagation laws as the geometrical optics field. Some applications of the theory were described by J. B. Keller, Trans. Inst. Radio Eng., A.P.-4 (1956), 312 and J. B. Keller, R. M. Lewis and B. D. Seckler, J. Appl. Phus., 28 (1957), 570. See also M. Kline and I. W. Kay, loc. cit.

$$\frac{d}{ds}(r \wedge ns) = \frac{dr}{ds} \wedge ns + r \wedge \frac{d}{ds}(ns). \tag{5}$$

Since $dr/ds = s$, the first term on the right vanishes. The second term may, on account of (2), be written as $r \wedge \operatorname{grad} n$. Now from (4)

$$\operatorname{grad} n = \frac{r}{r}\frac{dn}{dr},$$

so that the second term on the right-hand side of (5) also vanishes. Hence

$$r \wedge ns = \text{ constant}. \tag{6}$$

This relation implies that all the rays are plane curves, situated in a plane through the origin, and that along each ray

$$nr \sin \phi = \text{constant}, \tag{7}$$

where $\phi$ is the angle between the position vector $r$ and the tangent at the point $r$ on the ray (see Fig. 3.5). Since $r \sin \phi$ represents the perpendicular distance $d$ from the origin to the tangent, (7) may also be written as

$$nd = \text{ constant}. \tag{8}$$

This relation is sometimes called the formula of Bouguer and is the analogue of a well-known formula in dynamics, which expresses the conservation of angular momentum of a particle moving under the action of a central force.

To obtain an explicit expression for the rays in a spherically symmetrical medium, we recall from elementary geometry that, if $(r, \theta)$ are the polar coordinates of a

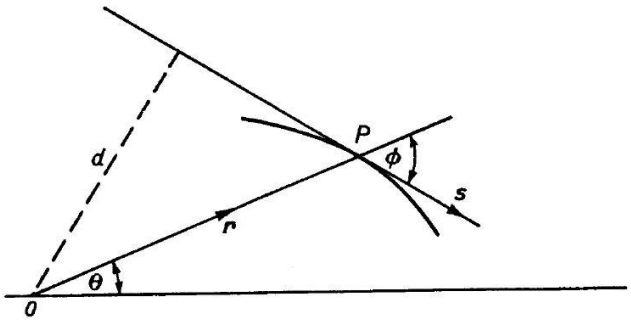



Fig. 3.5. Illustrating Bouguer's formula $nd$ = constant, for rays in a medium with spherical symmetry.

plane curve, then the angle $\phi$ between the radius vector to a point $P$ on the curve and the tangent at $P$ is given by*

$$\sin \phi = \frac{r(\theta)}{\sqrt{r^2(\theta) + \left(\frac{dr}{d\theta}\right)^2}} \tag{9}$$

From (7) and (9)

$$\frac{dr}{d\theta} = \frac{r}{c}\sqrt{n^2 r^2 - c^2} \tag{10}$$

$c$ being a constant. The equation of rays in a medium with spherical symmetry may therefore be written in the form

$$\theta = c \int^r \frac{dr}{r\sqrt{n^2 r^2 - c^2}} \tag{11}$$

Let us now return to the general case and consider the curvature vector of a ray, i.e. the vector

$$K = \frac{ds}{ds} = \frac{1}{\rho}\nu \tag{12}$$

whose magnitude $1/\rho$ is the reciprocal of the radius of curvature; $\nu$ is the unit principal normal at a typical point of the ray.

From (2) and (12) it follows that

$$nK = \operatorname{grad} n - \frac{dn}{ds}s. \tag{13}$$

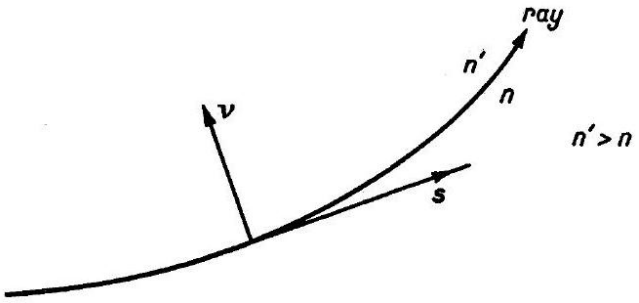



Fig. 3.6. Bending of a ray in a heterogeneous medium.

This relation shows that the gradient of the refractive index lies in the osculating plane of the ray.

If we multiply (13) scalarly by $K$ and use (12) we find that

$$|K| = \frac{1}{\rho} = \nu \cdot \operatorname{grad} \log n. \tag{14}$$

Since $\rho$ is always positive, this implies that as we proceed along the principal normal the refractive index increases i.e. the ray bends towards the region of higher refractive index (Fig. 3.6).

### 3.2.2 The laws of refraction and reflection

So far it has been assumed that the refractive index function $n$ is continuous. We must now discuss the behaviour of rays when they cross a surface separating two homogeneous media of different refractive indices. It has been shown by Sommerfeld

0

- See, for example, R. Courant, Differential and Integral Calculus, Vol. I (Glasgow, Blackie, 2nd edition, 1942), p. 265.

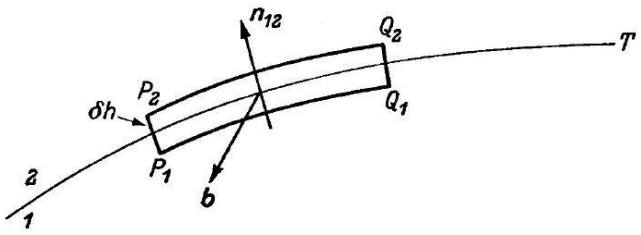



Fig. 3.7. Derivation of the laws of refraction and reflection.

and Runge (loc. cit.) that the behaviour can easily be determined by an argument similar to that used in deriving the conditions relating to the changes in the field vectors across a surface discontinuity (cf. § 1.1.3).

It follows from (1), on account of the identity curl grad $\equiv 0$, that the vector $ns = ndr/ds$, called sometimes the ray vector, satisfies the relation

$$\operatorname{curl} ns = 0. \tag{15}$$

As in § 1.1.3 we replace the discontinuity surface $T$ by a transition layer throughout which $\varepsilon, \mu$ and $n$ change rapidly but continuously from their values near $T$ on one side to their values near $T$ on the other. Next we take a plane element of area with its sides $P_1Q_1$ and $P_2Q_2$ parallel and with $P_1P_2$ and $Q_1Q_2$ perpendicular to $T$ (Fig. 3.7).

If $\boldsymbol{b}$ denotes the unit normal to this area, then we have from (15), on integrating throughout the area and applying Stokes' theorem,

$$\int (\operatorname{curl} ns) \cdot b dS = \int ns \cdot dr = 0 \tag{16}$$

the second integral being taken along the boundary curve $P_1Q_1Q_2P_2$. Proceeding to the limit as the height $\delta h \to 0$, in a strictly similar manner as in the derivation of § 1.1 (23), we obtain

$$n_{12} \wedge (n_2 s_2 - n_1 s_1) = 0 \tag{17}$$

where $n_{12}$ is the unit normal to the boundary surface pointing from the first into the second medium. (17) implies that the tangential component of the ray vector $ns$ is

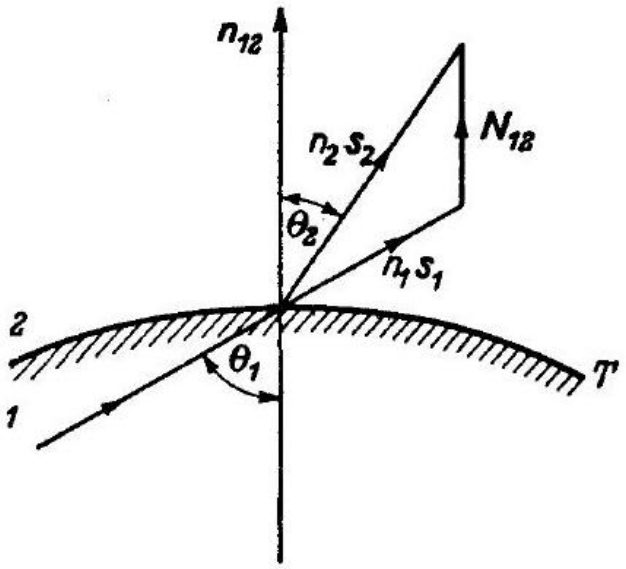



Fig. 3.8 (a). Illustrating the law of refraction.

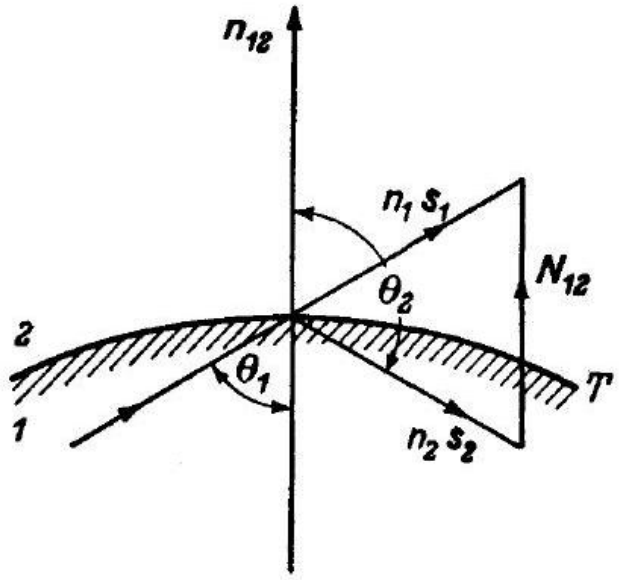



Fig. 3.8 (b). Illustrating the law of reflection.

continuous across the surface or, what amounts to the same thing, the vector $N = n_2 s_2 - n_1 s_1$ is normal to the surface.

Let $\theta_1$ and $\theta_2$ be the angles which the incident ray and the refracted ray make with the normal $\boldsymbol{n}_{12}$ to the surface (see Fig. 3.8a). Then it follows from (17) that
so that

$$n_2 (n_{12} \wedge s_2) = n_1 (n_{12} \wedge s_1) \tag{18}$$

$$n_2 \sin\theta_2 = n_1 \sin\theta_1 \tag{19}$$

(18) implies that the refracted ray lies in the same plane as the incident ray and the normal to the surface (the plane of incidence) and (19) shows that the ratio of the sine of the angle

of refraction to the sine of the angle of incidence is equal to the ratio $n_1/n_2$ of the refractive indices. These two results express the law of refraction (Snell's law). This law has already been derived in § 1.5 for the special case of plane waves. But whilst the earlier discussion concerned a plane wave of arbitrary wavelength falling upon a plane refracting surface, the present analysis applies to waves and refracting surfaces of more general form, provided that the wavelength is sufficiently small ( $\lambda_0 \to 0$ ). This condition means, in practice, that the radii of curvature of the incident wave and of the boundary surface must be large compared to the wavelength of the incident light.

As in the case treated in § 1.5 we must expect that there will be another wave, the reflected wave, propagated back into the first medium. Setting $n_2 = n_1$ in (18) and (19) (see Fig. 3.8b) it follows that the reflected ray lies in the plane of incidence and that $\sin\theta_2 = \sin\theta_1$; hence

$$\theta_2 = \pi - \theta_1. \tag{20}$$

The last two results express the law of reflection.

### 3.2.3 Ray congruences and their focal properties

The relation (15), namely

$$\operatorname{curl} ns = 0, \tag{21}$$

characterizes all the ray systems which can be realized in an isotropic medium and distinguishes them from more general families of curves. In a homogeneous isotropic medium $n$ is constant, and (21) then reduces to

$$\operatorname{curl} s = 0. \tag{22}$$

Rays in a heterogeneous isotropic medium can also be characterized by a relation independent of $n$. It may be obtained by applying to (21) the identity $\operatorname{curl} ns = n \operatorname{curl} s + (\operatorname{grad} n) \wedge s$ and taking the scalar product with $s$. It then follows that a system of rays in any isotropic medium must satisfy the relation

$$\boldsymbol{s} \cdot \operatorname{curl} \boldsymbol{s} = \boldsymbol{0}. \tag{23}$$

A system of curves which fills a portion of space in such a way that in general a single curve passes through each point of the region is called a congruence. If there exists a family of surfaces which cut each of the curves orthogonally the congruence is said to be normal; if there is no such family, it is said to be skew. For ordinary geometrical optics (light propagation) only normal congruences are of interest, but in electron optics (see Appendix II) skew congruences also play an important part.

If each curve of the congruence is a straight line the congruence is said to be rectilinear; (23) and (22) are the necessary and sufficient conditions that the curves should represent a normal and a normal rectilinear congruence respectively.*

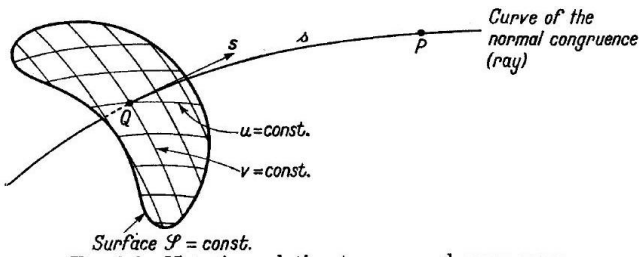



Fig 3.9. Notation relating to a normal congruence.

Let us choose a set of curvilinear coordinate lines $u, v$ on one of the orthogonal surfaces $\mathscr{S}(x, y, z) =$ constant. To every point $Q(u, v)$ of this surface there will then correspond one curve of the congruence, namely that curve which meets $\mathscr{S}$ in $Q$. Let $r$ denote the position vector of a point $P$ on the curve. $r$ may then be regarded as a function of the coordinates $(u, v)$ and of the length of arc $s$ between $Q$ and $P$, measured along the curve (Fig. 3.9).

Consider two neighbouring curves of the congruence passing through the points $(u, v)$ and $(u + du, v + dv)$ on $\mathscr{S}$, and let us examine whether there are points on these curves such that the distance between them is of the second or higher order (one says that the curves cut to first order at such points). Points with this property are called foci and must satisfy the equation

$$r(u, v, s) = r(u + du, v + dv, s + ds) \tag{24}$$

to the first order.

Expanding (24) we obtain

$$\boldsymbol{r}_u du + \boldsymbol{r}_v dv + \boldsymbol{s} ds = 0 \tag{25}$$

where $r_u, r_v$ are the partial derivatives with respect to $u$ and $v$. Condition (25) implies that $r_u, r_v$ and $s$ are coplanar. This is equivalent to saying that the scalar triple product of the three vectors vanishes, i.e.

$$[\boldsymbol{r}_u, \boldsymbol{r}_v, \boldsymbol{s}] = 0 \tag{26}$$

The number of foci on a given curve $(u, v)$ depends on the number of values of $s$ which satisfy (26). If $r$ is a polynomial in $s$ of degree $m$, then since $s = dr/ds$, it is seen that (26) is an equation of degree $3m - 1$ in $s$. In particular, if the congruence is rectilinear, $r$ is a linear function of $s (m = 1)$, showing that there are two foci on each ray of a rectilinear congruence.

If $u$ and $v$ take on all possible values, the foci will describe a surface, represented by (26), known as the focal surface; in optics it is called the caustic surface. Any curve of the congruence is tangent to the focal surface at each focus of the curve. The tangent plane at any point of the focal surface is known as the focal plane.

We shall mainly be concerned with rays in a homogeneous medium, i.e. with rectilinear congruences. Some further properties of such congruences will be discussed in § 4.6, in connection with astigmatic pencils of rays.

## 9.4.3 3.3 Other Basic Theorems of Geometrical Optics

With the help of the relations established in the preceding sections, we shall now derive a number of theorems concerning rays and wave-fronts.

### 3.3.1 Lagrange's integral invariant

Assume first that the refractive index $n$ is a continuous function of position. Then as in § 3.2 (16) it follows on applying Stokes' theorem to the integral, taken over any open surface, of the normal component of curl $ns$, that

$$\oint ns.dr = 0. \tag{1}$$

0

- For a more detailed discussion of congruences of curves see, for example, C. E. Weatherburn, Differential Geometry of Three Dimensions (Cambridge University Press), Vol. I (1927), Chapter X; Vol. II (1930), Chapter XIII.

The integral extends over the closed boundary curve $C$ of the surface. (1) is known as Lagrange's integral invariant* and implies that the integral

$$\int_{P_1}^{P_2} ns \cdot dr \tag{2}$$

taken between any two points $P_1$ and $P_2$ in the field, is independent of the path of integration.

With the help of the law of refraction it is easily seen that (1) also holds when the curve $C$ intersects a surface which separates two homogeneous media of different refractive indices. To show this, let $C_1$ and $C_2$ be the portions of $C$ on each side of the refracting surface $T$ (Fig. 3.10), and let the points of intersection of $C$ with the surface $T$ be joined by another curve $K$ in the surface. On taking (1) along each of the loops $C_1K$ and $C_2K$ and on adding the equations, we obtain

$$\int_{C_1} n_1 s_1 \cdot dr + \int_{C_2} n_2 s_2 \cdot dr + \int_K (n_2 s_2 - n_1 s_1) \cdot dr = 0 \tag{3}$$

The integral over $K$ vanishes, since according to the law of refraction the vector $N = n_1 s_1 - n_2 s_2$ is at each point of $K$ perpendicular to the surface, and consequently (3) reduces to (1).

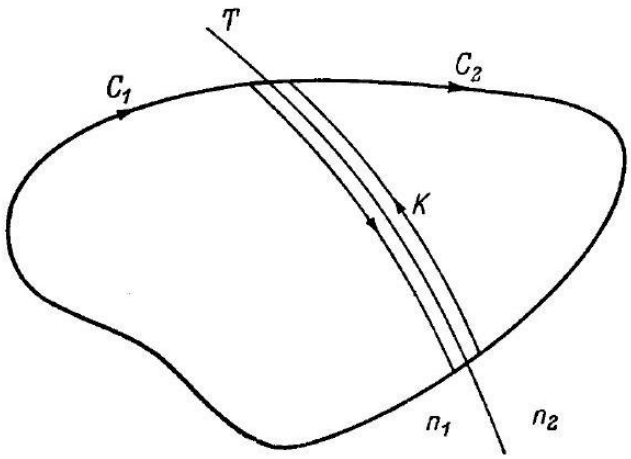



Fig. 3.10. Derivation of the Lagrange's integral invariant in the presence of a surface of discontinuity of the refractive index.

### 3.3.2 The principle of Fermat

The principle of Fermat, known also as the principle of the shortest optical path* asserts that the optical length

$$\int_{P}^{P_2} nds \tag{4}$$

of an actual ray between any two points $P_1$ and $P_2$ is shorter than the optical length of any other curve which joins these points and which lies in a certain regular neighbourhood of it. By

0

• Sometimes called Poincare's invariant. In fact it is only a special one-dimensional case of much more general integral invariants discussed by J. H. Poincar in his Les Mthodes Nouvelles de la Mcanique Cleste, 3 (Paris, Gauthier-Villars, 1899). Cf. E. Cartan, Leons sur les Invariants Integraux (Paris, Hermann, 1922). See also our Appendix I, eq. (85).

a regular neighbourhood we mean one that may be covered by rays in such a way that one (and only one) ray passes through each point of it. Such a covering is exhibited, for example, by rays from a point source $P_1$ in that domain around $P_1$ where the rays on account of refraction or reflection or on account of their curvature do not intersect each other.

Before proving this theorem it may be mentioned that it is possible to formulate Fermat's principle in a form which is weaker but which has a wider range of validity. According to this formulation the actual ray is distinguished from other curves (no

- Since by § 3.1 (27)

$$\int_{P_1}^{P_2} nds = c \int_{P_1}^{P_2} dt$$

it is also known as the principle of least time.

longer restricted to lie in a regular neighbourhood) by a stationary value of the integral.*

To prove Fermat's principle, we take a pencil of rays and compare a segment $P_1P_2$ of a ray $\bar{C}$ with an arbitrary curve $C$ joining $P_1$ and $P_2$ (Fig. 3.11). Let two neighbouring orthogonal trajectories (wave-fronts) of the pencil intersect $C$ in $Q_1$ and $Q_2$ and $\bar{C}$ in $\bar{Q}_1$ and $\bar{Q}_2$. Further let $Q_2'$ be the point of intersection of the trajectory $Q_2\bar{Q}_2$ with the ray $\bar{C}''$ which passes through $Q_1$.

Applying Lagrange's integral relation to the small triangle $Q_1Q_2Q_2'$, we have

$$(ns.dr)_{Q_1Q_2} + (ns.dr)_{Q_2Q_2'} - (nds)_{Q_1Q_2} = 0. \tag{5}$$

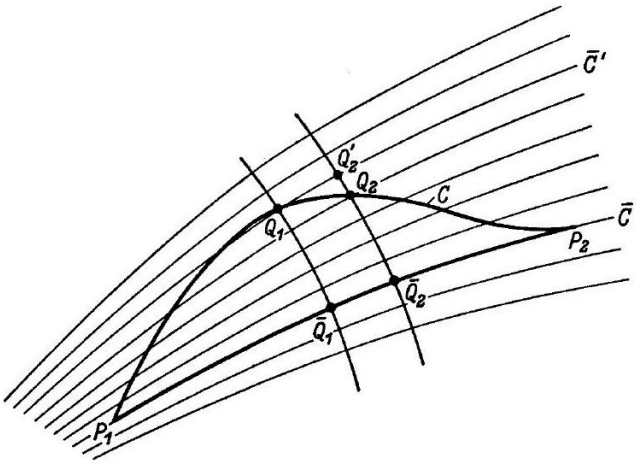



Fig. 3.11. Illustrating Fermat's principle.

Now from the definition of the scalar product

$$(ns.dr)_{Q_1Q_2} \leqslant (nds)_{Q_1Q_2}.$$

Further, $s$ is orthogonal to $dr$ on the wave-front, so that

$$(ns \cdot dr)_{Q_2Q_2'} = 0.$$

Also from § 3.1 (25), since $Q_1, Q_2'$ and $\bar{Q}_1, \bar{Q}_2$ are corresponding points on the two wave-fronts,

$$(nds)_{Q_1Q_2'} = (nds)_{\bar{Q}_1\bar{Q}_2}.$$

On substituting from the last three relations into (5) we find that

$$(nds)_{\bar{Q}_1\bar{Q}_2} \leqslant (nds)_{Q_1Q_2}, \tag{6}$$

whence, on integration,

$$\int_{\bar{C}} nds \leqslant \int_C nds. \tag{7}$$

Moreover, the equality sign could only hold if the directions of $s$ and $dr$ were coincident at every point of $C$, i.e. if the comparison curve was an actual ray. This case is excluded by our assumption that not more than one ray passes through any point of the neighbourhood. Hence the optical length of the ray is smaller than the optical length of the comparison curve, which is Fermat's principle.

It can easily be seen that, when the regularity condition is not fulfilled, the optical length of the ray may no longer be a minimum. Consider for example a field of rays

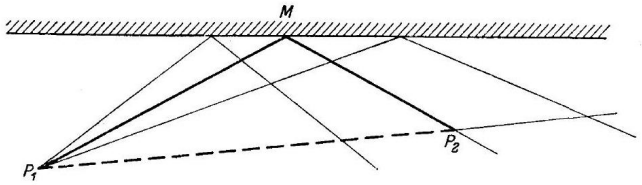



Fig. 3.12. Field of rays obtained by reflection of light from a point source on a plane mirror.

from a point source $P_1$ in a homogeneous medium, reflected by a plane mirror (Fig. 3.12). Two rays then pass through each point $P_2$; the optical length of the direct ray $P_1P_2$ is an absolute minimum but the reflected ray $P_1MP_2$ gives a minimum only relative to curves in a certain restricted neighbourhood of it. In general when rays from a point source $P_1$ are refracted or reflected at boundaries between homogeneous media, the regular neighbourhood will terminate on the envelope (caustic) formed by the rays. The point $P_1'$ at which a ray from a point source at $P_1$ touches the envelope is called the conjugate of $P_1$ on the particular ray. For the optical length of a ray $P_1P_2$

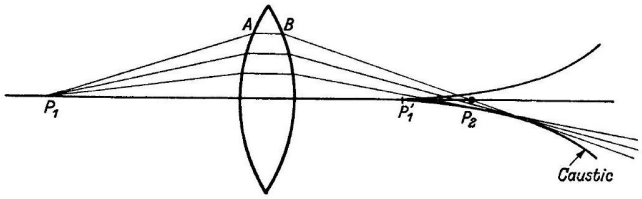



Fig. 3.13. Caustic formed by rays from an axial point source after passing through a lens.

0

- To find the curves for which the integral has a stationary value we must apply in general the methods of the variational calculus, described in Appendix I. It is shown there that such curves satisfy the Euler Differential Equations AI (7). In the present case these are nothing but the equations § 3.2. (2) of the rays as shown in section 11 of Appendix I.

It has been stressed by C. Carathodory (loc. cit.) that the stationary value is never a true maximum. In the weaker formulation of Fermat's principle it is therefore appropriate to speak of a stationary value rather than of an extremal value. The minimal formulation on the other hand corresponds to a "strong minimum" in the sense of Jacobi (Appendix I, section 10).

to be a minimum, $P_2$ must lie between $P_1$ and $P_1'$, i.e. $P_1$ and $P_2$ must lie on the same side of the caustic. For example, in the case of an uncorrected lens (Fig. 3.13) the central ray from $P_1$ has a minimal optical length only up to the tip ( $P_1'$ ) of the caustic (the Gaussian image of $P_1$ ). For any point $P_2$ which lies behind the envelope the optical length of the direct path $P_1P_1'P_2$ exceeds that of the broken path $P_1ABP_2$.

### 3.3.3 The theorem of Malus and Dupin and some related theorems

The light rays have been defined as the orthogonal trajectories of the wave surfaces $\mathscr{S}(x, y, z) =$ constant, $\mathscr{S}$ being a solution of the eikonal equation (15) in § 3.1. This is a natural way of introducing the light rays when the laws of geometrical optics are to be deduced from Maxwell's equations. Historically, however, geometrical optics developed as the theory of light rays which were defined differently, namely as curves
for which the line integral $\int nds$ has a stationary value. Formulated this way geometrical optics may then be developed purely along the lines of calculus of variations.*

Variational considerations are of considerable importance as they often reveal analogies between different branches of physics. In particular there is a close analogy between geometrical optics and the mechanics of a moving particle; this was brought out very clearly by the celebrated investigations of Sir W. R. Hamilon, whose approach became of great value in modern physics, especially in applications to DE Broglie's wave mechanics. In order not to interrupt the optical considerations, an account of the relevant parts of the calculus of variation and of the Hamiltonian analogy are given in separate sections (Appendix I and II). Here we shall only show

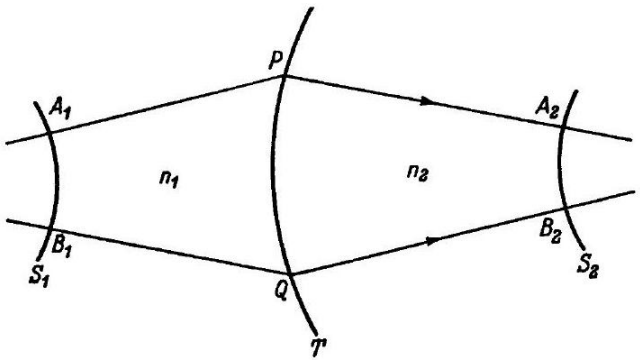



Fig. 3.14. Illustrating the theorem of Malus and Dupin.

how several theorems, which played an important part in the development of geometrical optics, may be derived from Lagrange's integral invariant.

Consider rays in a homogeneous medium: if they all have a point in common, e.g. when they then proceed from a point source, they are said to form a homocentric pencil. Such a pencil forms a normal congruence, since every ray of the pencil is cut orthogonally by spheres centred on the mutual point of intersection of the rays. In 1808 Malus † showed that, if a homocentric pencil of rectilinear rays is refracted or reflected at a surface, the resulting pencil (in general no longer homocentric) will again form a normal congruence. Later Dupin (1816), Quetelet (1825), and Gergonne (1825) generalized Malus's result. These investigations lead to the following theorem, known sometimes as the theorem of Malus and Dupin: A normal rectilinear congruence remains normal after any number of refractions or reflections. ‡

It will be sufficient to establish the theorem for a single refraction. Consider a normal rectilinear congruence of rays in a homogeneous medium of refractive index $n_1$ and assume

that the rays undergo a refraction at a surface $T$ which separates this medium from another homogeneous medium of refractive index $n_2$ (Fig. 3.14).

Let $S_1$ be one of the orthogonal trajectories (wave-fronts) in the first region, and let $A_1$ and $P$ be the points of intersections of a typical ray in the first medium with

$S_1$ and with $T$ respectively, and let $A_2$ be any point on the refracted ray. If the point $A_1$ is displaced to another point $B_1$ on the wave-front, the point $P$ will be displaced to another point $Q$ on the refracting surface. Now take a point $B_2$, on the ray which is refracted at $Q$, such that the optical path from $B_1$ to $B_2$ is equal to the optical path from $A_1$ to $A_2$ :

$$[A_1PA_2] = [B_1QB_2]\,. \tag{8}$$

As $B_1$ takes on all possible positions on $S_1$ the point $B_2$ describes a surface $S_2$. It will now be shown that the refracted ray $QB_2$ is perpendicular to this surface.

Applying Lagrange's integral invariant to the closed path $A_1PA_2B_2QB_1A_1$, it follows that

$$\int_{A_1PA_2} nds + \int_{A_2B_2} ns\cdot dr + \int_{B_2QB_1} nds + \int_{B_1A_1} ns\cdot dr = 0 \tag{9}$$

Now by (8).

$$\int_{A_1PA_2} nds + \int_{B_2QB_1} nds = 0 \tag{10}$$

Moreover, since on $S_1$ the unit vector $s$ is everywhere orthogonal to $S_1$,

$$\int_{B_1A_1} ns\cdot dr = 0 \tag{11}$$

so that (9) reduces to

$$\int_{A_2B_2} ns\cdot dr = 0 \tag{12}$$

This relation must hold for every curve on $S_2$. This is only possible if $s.dr = 0$ for every linear element $dr$ of $S_2$, i.e. if the refracted rays are orthogonal to the surface; in other words if the refracted rays form a normal congruence. The proof for reflection is strictly analogous.

Since $[A_1PA_2] = [B_1QB_2]$ it follows that the optical path length between any two orthogonal surfaces (wave-fronts) is the same for all rays. This result clearly remains valid when several successive refractions or reflections takes place and, as is immediately obvious from eq. (26) in § 3.1 it also applies to rays in a medium with continuously varying refractive index. This theorem is known as the principle of equal optical path; it implies that the orthogonal trajectories (geometrical wavefronts) of a normal congruence of rays, or of a set of normal congruences generated by successive refractions or reflections, are "optically parallel" to each other (cf. Appendix I).

---

0

- A systematic treatment of this kind is given for example in C. Carathodory (loc. cit.).
† E. Malus, Optique Dioptrique, J. cole polytechn., 7 (1808), 1-44, 84-129. Also his "Trait d'optique", Mm. prsent. l'Institut par divers savants, 2 (1811), 214-302. References and an account of the interesting history of the Malus-Dupin theorem can be found in the Mathematical Papers of Sir William Rowan Hamilton, 1 (Geometrical Optics), edited by A. W. Conway and J. L. Synge (Cambridge University Press, 1931), p. 463.
‡ T. Levi-Civita, Rend. R. Acc. Naz. Linc., 9 (1900) 237 established the converse theorem, namely that in general two normal rectilinear congruences may be transformed into each other by a single refraction or reflection.

A related theorem, first put forward by Huygens* asserts that each element of a wave-front may be regarded as the centre of a secondary disturbance which gives rise to spherical wavelets; and moreover that the position of the wave-front at any later time is the envelope of all such wavelets. This result, sometimes called Huygens' construction, is essentially a rule for the construction of a set of surfaces which are "optically parallel" to each other. If the medium is homogeneous, one can use in the construction wavelets of finite radius, in other cases one has to proceed in infinitesimal steps.

Huygens' theorem was later extended by Fresnel and led to the formulation of the so-called Huygens-Fresnel principle, which is of great importance in the theory of diffraction (see §8.2), and which may be regarded as the basic postulate of the wave theory of light.

## 9.5 Geometrical Theory of Optical Imaging

### 9.5.1 4.1 The Characteristic Functions of Hamilton

In § 3.1 it was shown that, within the approximations of geometrical optics, the field may be characterized by a single scalar function $\mathscr{S}(\boldsymbol{r})$. Since $\mathscr{S}(\boldsymbol{r})$ satisfies the eikonal equation (15) in § 3.1, this function is fully specified by the refractive index function $n(r)$ alone, together with the appropriate boundary conditions.

Instead of the function $\mathscr{S}(\boldsymbol{r})$ closely related functions known as characteristic functions of the medium are often used. They were introduced into optics by W. R. Hamilton, in a series of classical papers.* Although on account of algebraic complexity it is impossible to determine the characteristic functions explicitly for all but the simplest media, Hamilton's methods nevertheless form a very powerful tool for systematic analytical investigations of the general properties of optical systems.

In discussing the properties of these functions and their applications, an isotropic but generally heterogeneous medium will be assumed.

#### 4.1.1 The point characteristic

Let $(\, x_0, y_0, z_0 \,)$ and $(\, x_1, y_1, z_1 \,)$ be respectively the coordinates of two points $P_0$ and $P_1$ each referred to a different set of mutually parallel, rectangular axes † (Fig. 4.1). If the two points are imagined to be joined by all possible curves, there will, in general, be some amongst them, the optical rays, which satisfy Fermat's principle. Assume for the present that not more than one ray joins any two arbitrary points. The characteristic function $V$, or the point characteristic, is then defined as the optical length $[P_0P_1]$ of the ray between the two points, considered as a function of their coordinates,

$$V\,(x_0, y_0, z_0; x_1, y_1, z_1) = \int_{P_0}^{P_1} n ds \tag{1}$$

It is important to note that this function is defined by the medium.

0

- Trait de la Lumire (Leyden, 1690); English translation (Treatise on Light) by S. P. Thompson (London, Macmillan & Co., 1912).

0

- Sir W. R. Hamilton, Trans. Roy. Irish Acad., 15 (1828), 69; ibid., 16 (1830), 1; ibid., 16 (1831), 93; ibid., 17 (1837), 1. Reprinted in The Mathematical Papers of Sir W. R. Hamilton, Vol. I (Geometrical Optics), edited by A. W. Conway and J. L. Synge (Cambridge University Press, 1931).

Many years later Bruns independently considered similar functions which he called eikonals. (H. Bruns, Abh. Kgl. Schs. Ges. Wiss., math-phys. Kl., 21 (1895), 323.) As already mentioned on p. 112, this term has come to

From (1) and § 3.1 (26) it follows that

$$V(x_0, y_0, z_0; x_1, y_1, z_1) = \mathscr{S}(x_1, y_1, z_1) - \mathscr{S}(x_0, y_{0s} z_0) \tag{2}$$

where the function $\mathscr{S}$ is now associated with any pencil of rays to which the natural ray through $P_0$ and $P_1$ belongs [for example, a pencil produced by a point source at $P_0$].* Then by §3.1(24), the unit vectors $s_0$ and $s_1$ at $P_0$ and $P_1$, in the direction of the ray, are given by

$$\left.\begin{aligned} \operatorname{grad}^0 V &= -n_0 s_0, \\ \operatorname{grad}^1 V &= n_1 s_1, \end{aligned}\right\} \tag{3}$$

the superscripts 0 and 1 implying that the operator grad is taken with respect to the coordinates ( $x_0, y_0, z_0$ ) and ( $x_1, y_1, z_1$ ) respectively.

The vector

$$g = ns \tag{4}$$

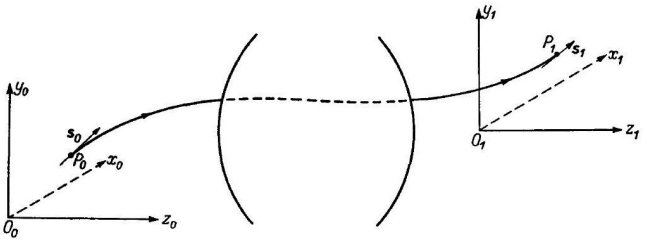



Fig. 4.1. Illustrating the definition of the point characteristic function.

is sometimes called the ray vector. If $\alpha, \beta$ and $\gamma$ are the angles which the ray vector makes with the coordinate axes, its projections †

$$p = n\cos\alpha, \quad q = n\cos\beta, \quad m = n\cos\gamma, \tag{5}$$

on to the axes are called the ray components. On account of the identity

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1,$$

they satisfy the relation

$$p^2 + q^2 + m^2 = n^2. \tag{6}$$

From (3) it follows that the ray components at $P_0$ and $P_1$ are given by

$$p_0 = -\frac{\partial V}{\partial x_0}, \quad p_1 = \frac{\partial V}{\partial x_1}, \tag{7}$$

with similar expressions for $q_0, q_1$, and $m_0, m_1$. These relations show that, from the knowledge of the point characteristic, the components of the ray which joins any two points in the medium can immediately be determined. Further it follows from (6)

be used in a wider sense. The characteristic functions of Hamilton are themselves often referred to as eikonals.

A useful introduction to Hamiton's methods is a monograph by J. L. Synge, Geometrical Optics (Cambridge University Press, 1937). The relationship between the work of Hamilton and Bruns was discussed by F. Klein in Z. Math. Phys., 46 (1901), 376, and Ges. Math. Abh., 2 (1922), 603, C. Carathodory, Geometrische Optik (Berlin, Springer, 1937), p. 4, and in a polemic between M. Herzberger and J. L. Synge, J. Opt. Soc. Amer., 27 (1937), 75, 133, 138.

† The use of two reference systems has some advantages, since $P_0$ and $P_1$ are often situated in different regions, namely, the object- and image-spaces of an optical system.

0

and (7) that the point characteristic satisfies the eikonal equation in both sets of variables:

$$\left(\frac{\partial V}{\partial x_0}\right)^2 + \left(\frac{\partial V}{\partial y_0}\right)^2 + \left(\frac{\partial V}{\partial z_0}\right)^2 = n_0^2 \tag{8}$$

$$\left(\frac{\partial V}{\partial x_1}\right)^2 + \left(\frac{\partial V}{\partial y_1}\right)^2 + \left(\frac{\partial V}{\partial z_1}\right)^2 = n_1^2 \tag{9}$$

Instead of the point characteristic it is often convenient to use certain related functions (also introduced by Hamilon), known as the mixed characteristic and the angle characteristic. They may be derived from the point characteristic by means of LEGENDRE transformations,* and are particularly useful when either $P_0$ or $P_1$ or both these points are at infinity.

### 4.1.2 The mixed characteristic

The mixed characteristic function $W$ is defined by the equation †

$$W = V - \Sigma p_1 x_1, \tag{10}$$

where $\Sigma$ denotes summation over the three similar terms with suffix one. (10) expresses $W$ as a function of nine variables, but in general only six (and in a homogeneous medium only five) are independent. To show this, consider the effect of small displacements of the points $P_0$ and $P_1$. The corresponding change in $W$ is then given by

$$\delta W = \delta V - \Sigma p_1 \delta x_1 - \Sigma x_1 \delta p_1. \tag{11}$$

Now by (7),

$$\delta V = \Sigma p_1 \delta x_1 - \Sigma p_0 \delta x_0 \tag{12}$$

From (11) and (12),

$$\delta W = -\Sigma p_0 \delta x_0 - \Sigma x_1 \delta p_1. \tag{13}$$

This relation shows that in general $W$ can be expressed as a function of the six variables $x_0, y_0, z_0, p_1, q_1$ and $m_1$, and that, when it is so expressed,

$$p_0 = -\frac{\partial W}{\partial x_0}, \quad x_1 = -\frac{\partial W}{\partial p_1} \tag{14}$$

with similar expressions for $q_0, y_1, m_0$ and $z_1$. On account of (6), $W\,(x_0, y_0, z_0; p_1, q_1, m_1)$ satisfies the eikonal equation

$$\left(\frac{\partial W}{\partial x_0}\right)^2 + \left(\frac{\partial W}{\partial y_0}\right)^2 + \left(\frac{\partial W}{\partial z_0}\right)^2 = n_0^2 \tag{15}$$

It is to be observed (see Fig. 4.2) that the sum $\Sigma p_1 x_1$ has a simple geometrical interpretation:

$$\Sigma p_1 x_1 = n_1 d_1 \tag{16}$$

---

* In the language of the calculus of variations, $\mathscr{S}$ represents the solution which involves a two-parameter family ( $\infty^2$ ) of extremals, of the Hamilton-Jacobi equation, associated with Fermat's variational problem. The point characteristic function $V$ represents the general solution, which involves all ( $\infty^4$ ) extremals (cf. Appendix I).

† This "asymmetrical" notation is purposely chosen here to remind us that on account of the identity (6), only two of the ray components are independent.

where $d_1 = Q_1P_1$ is the projection of $O_1P_1$ on to the tangent to the ray at $P_1$. If $P_1$ is situated in a homogeneous region, the portion of the ray near $P_1$ coincides with the line segment $Q_1P_1$; according to (10) and (16) $W$ then represents the optical length of

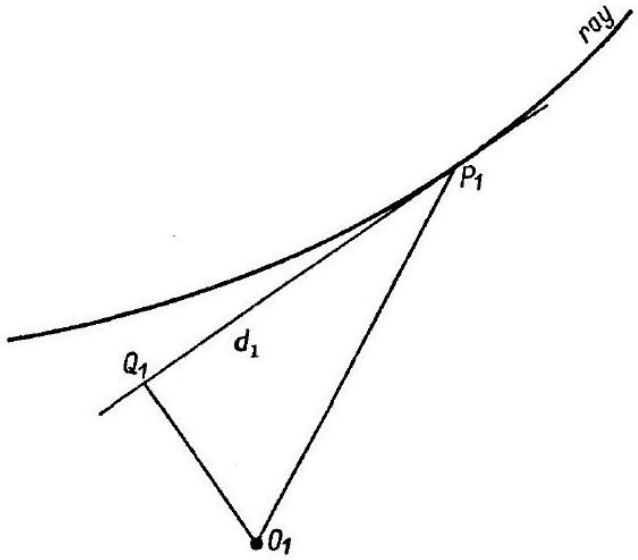



Fig. 4.2. Illustrating the meaning of the mixed characteristic function.

the ray from $P_0$ to the foot $Q_1$ of the perpendicular drawn from the origin $O_1$ on to the final portion of the ray (Fig. 4.3):

$$W = [P_0Q_1]\,. \tag{17}$$

Since in this case the refractive index of the medium around $P_1$ has a constant value it follows from (6) that

$$\delta m_1 = -\frac{p_1\delta p_1 + q_1\delta q_1}{m_1}, \tag{18}$$

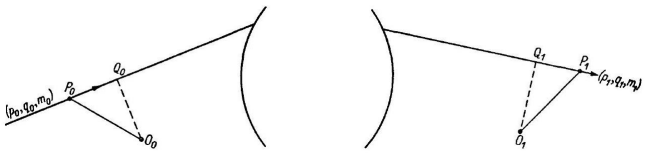


Fig. 4.3. Interpretation of Hamilton's characteristic functions when the initial and final media are homogeneous:

---

• A LEGENDRE transformation transforms in general a function $f(x, y)$ into a function $g(x, z)$ where $z = \partial f/\partial y$, in such a way that the derivative of $g$ with respect to the new variable $z$ is equal to the old variable $y$.

† To bring out more clearly its physical meaning, we follow Synge in defining the mixed characteristic with opposite sign to that used by Hamilton.

One can also define a mixed characteristic

$$W' = V + \Sigma p_0 x_0$$

the summation being taken over similar terms with suffix zero. The properties of $W$ and $W'$ are, of course, strictly analogous.

$$V(x_0, y_0, z_0; x_1, y_1, z_1) = [P_0 P_1]$$
$$W(x_0, y_0, z_0; p_1, q_1) = [P_0 Q_1]$$
$$T(p_0, q_0; p_1, q_1) = [Q_0 Q_1]$$

and (13) becomes on substitution from (18)*

$$\delta W = -\Sigma p_0 \delta x_0 - \left(x_1 - \frac{p_1}{m_1} z_1\right) \delta p_1 - \left(y_1 - \frac{q_1}{m_1} z_1\right) \delta q_1. \tag{19}$$

Hence when the final medium is homogeneous, the mixed characteristic is expressible as a function of five variables:

$$W = W(x_0, y_0, z_0; p_1, q_1), \tag{20}$$

and its derivatives then satisfy the relations

$$p_0 = -\frac{\partial W}{\partial x_0}, \quad q_0 = -\frac{\partial W}{\partial y_0}, \quad m_0 = -\frac{\partial W}{\partial z_0} \tag{21}$$

$$x_1 - \frac{p_1}{m_1} z_1 = -\frac{\partial W}{\partial p_1}, \quad y_1 - \frac{q_1}{m_1} z_1 = -\frac{\partial W}{\partial q_1} \tag{22}$$

(21) and (22) show that if a point on the ray in the initial medium and the components of the ray in the final medium are given, the ray components in the initial medium and points on the ray in the final medium may immediately be determined from the knowledge of the mixed characteristic.

### 4.1.3 The angle characteristic

The angle characteristic $T$ may be defined by means of the relation

$$T = V + \Sigma p_0 x_0 - \Sigma p_1 x_1. \tag{23}$$

If $P_0$ and $P_1$ are slightly displaced, the corresponding change in $T$ is given by

$$\delta T = \Sigma x_0 \delta p_0 - \Sigma x_1 \delta p_1, \tag{24}$$

where (12) was used. Hence $T$ is expressible as a function of the six ray components, and when expressed in this way,

$$x_0 = \frac{\partial T}{\partial p_0}, \quad x_1 = -\frac{\partial T}{\partial p_1}, \tag{25}$$

with similar relations for the other coordinates.
It is seen from (23) that if the regions in which $P_0$ and $P_1$ are situated are both homogeneous, $T$ represents the optical length of the ray between the feet $Q_0$ and $Q_1$ of perpendiculars drawn from $O_0$ and $O_1$ on to the initial and final portions of the ray (see Fig. 4.3),

$$T = [Q_0 Q_1]. \tag{26}$$

0

- If a function depends on variables which are connected by subsidiary relations, such as (6), some of the variables may be eliminated, or alternatively the relations may be used to express it as a homogeneous function in all the variables. The alternative procedure, which is somewhat more difficult to handle, was frequently employed by Hamiton.

In this case the angle characteristic may be expressed as a function of four variables only. For if we substitute for $\delta m_1$ from (18) and for $\delta m_0$ from a similar relation, (24) becomes

$$\begin{aligned}\delta T = &\left(x_0 - \frac{p_0}{m_0} z_0\right) \delta p_0 + \left(y_0 - \frac{q_0}{m_0} z_0\right) \delta q_0 \\ &- \left(x_1 - \frac{p_1}{m_1} z_1\right) \delta p_1 - \left(y_1 - \frac{q_1}{m_1} z_1\right) \delta q_1\end{aligned} \tag{27}$$

This relation shows that when the initial and final media are homogeneous, the angle characteristic is expressible as a function of the four variables $p_0, q_0, p_1$ and $q_1$ :

$$T = T\left(p_0, q_0; p_1, q_1\right) \tag{28}$$

and its derivatives then satisfy the relations

$$\left.\begin{array}{ll} x_0 - \frac{p_0}{m_0} z_0 = \frac{\partial T}{\partial p_0}, & y_0 - \frac{q_0}{m_0} z_0 = \frac{\partial T}{\partial q_0} \\ x_1 - \frac{p_1}{m_1} z_1 = -\frac{\partial T}{\partial p_1}, & y_1 - \frac{q_1}{m_1} z_1 = -\frac{\partial T}{\partial q_1}. \end{array}\right\} \tag{29}$$

If the components of the initial and final portion of a ray are given, the coordinates of points on these portions may, according to (29), be determined immediately from the knowledge of the angle characteristic.

### 4.1.4 Approximate form of the angle characteristic of a refracting surface of revolution

Let

$$z = c_2\left(x^2 + y^2\right) + c_4\left(x^2 + y^2\right)^2 + \ldots \tag{30}$$

where $c_2, c_4, \ldots$ are constants, be the equation of a refracting surface of revolution, referred to Cartesian axes, whose origin $O$ coincides with the axial point (called pole) of the surface, and whose $z$-direction is along the axis of symmetry. If $r$ denotes the radius of curvature at the pole of the surface (measured as positive when the surface is convex towards light incident from the negative $z$-direction), then

$$c_2 = \frac{1}{2r} \tag{31}$$

For a spherical surface of radius $r, c_4 = 1/8r^3$. For a general surface of revolution we may write

$$c_4 = \frac{1}{8r^3}(1 + b) \tag{32}$$

the constant $b$ (sometimes called the deformation coefficient) is a rough measure of the departure of the surface from spherical form. In terms of $r$ and $b$,

$$z = \frac{x^2 + y^2}{2r} + \frac{\left(x^2 + y^2\right)^2}{8r^3}(1 + b) + \ldots \tag{33}$$

It will be assumed that the regions on either side of the surface are homogeneous and of refractive indices $n_0$ and $n_1$ respectively. The angle characteristic will be referred to systems of axes parallel to those at $O$ and with origins at axial points $O_0\left(0, 0, a_0\right)$ and $O_1\left(0, 0, a_1\right)$ $[a_0 < 0, a_1 > 0, r > 0$ in Fig. 4.4].

If $P$ is the point of intersection of the incident ray with the refracting surface, and if $Q_0$ and $Q_1$ are the feet of the perpendiculars drawn from $O_0$ and $O_1$ on to the incident and the refracted ray, the angle characteristic $T$ is then, according to (26),

$$\begin{aligned} T &= [Q_0P] + [PQ_1] \\ &= \{xp_0 + yq_0 + (z - a_0)\, m_0\} - \{xp_1 + yq_1 + (z - a_1)\, m_1\} \end{aligned} \tag{34}$$

where ( $x, y, z$ ) are the coordinates of $P$ with respect to the axes at $O$, and $(p_0, q_0, m_0)$, ( $p_1, q_1, m_1$ ) are the components of the ray incident and refracted at $P$.

The coordinates $(x, y, z)$ may be eliminated from (34) with the help of the law of refraction. According to § 3.2.2, the law of refraction is equivalent to the assertion
that the vector $N\,(p_0 - p_1, q_0 - q_1, m_0 - m_1)$ is normal to the surface at $P$. Hence if (33) is written in the form

$$F(x, y, z) \equiv z - \frac{x^2 + y^2}{2r} - \frac{(x^2 + y^2)^2}{8r^3}(1 + b) - \ldots = 0, \tag{35}$$

then

$$\left.\begin{aligned} \lambda \tfrac{\partial F}{\partial x} = -\lambda \left[\tfrac{x}{r} + \ldots\right] &= p_0 - p_1, \\ \lambda \tfrac{\partial F}{\partial y} = -\lambda \left[\tfrac{y}{r} + \ldots\right] &= q_0 - q_1, \\ \lambda \tfrac{\partial F}{\partial z} = \lambda &= m_0 - m_1. \end{aligned}\right\} \tag{36}$$

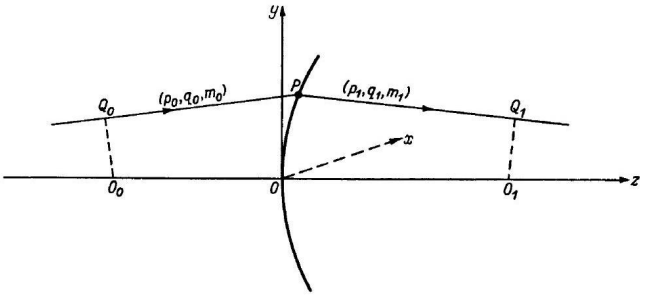



Fig. 4.4. The angle characteristic of a refracting surface of revolution. (The points $O_{0}, O, O_{1}, Q_{1}, P, Q_{0}$ are not necessarily coplanar.)

These equations imply that

$$\left.\begin{aligned} x &= -r\tfrac{p_0 - p_1}{m_0 - m_1} + \Delta x \\ y &= -r\tfrac{q_0 - q_1}{m_0 - m_1} + \Delta y \end{aligned}\right\} \tag{37}$$

where $\Delta x$ and $\Delta y$ are quantities of the third order in $p, q, x/r, y/r$. To express $z$ in terms of the ray components, we substitute from (37) into (35), and obtain

$$\begin{aligned} z = \frac{r}{2\,(m_1 - m_0)^2}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right] + \frac{1}{m_1 - m_0}\left[\Delta x\,(p_0 - p_1) + \Delta y\,(q_0 - q_1)\right] \\ + \frac{1}{8}\frac{r(1 + b)}{(m_1 - m_0)^4}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]^2 + \ldots \end{aligned} \tag{38}$$

To find the expansion of $T$ up to and including the fourth-order terms it is not necessary to evaluate $\Delta x$ and $\Delta y$; for when we substitute from (37) and (38) into (34), the contributions

involving $\Delta x$ and $\Delta y$ are seen to be of order higher than the fourth and may therefore be neglected. (34) then becomes

$$T(p_0, q_0, m_0; p_1, q_1, m_1) = -m_0 a_0 + m_1 a_1 + \frac{r}{2(m_1 - m_0)}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right] - \frac{1}{8}\frac{r(1+b)}{(m_1 - m_0)^3}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]^2 \tag{39}$$

(39) is the expansion of the angle characteristic up to the fourth order, the angle characteristic being considered as a function of all the six ray components. Two of the components may be eliminated by using the identity (6). We have from (6)

$$\left.\begin{aligned} m_0 &= n_0 - \tfrac{1}{2n_0}(p_0^2 + q_0^2) - \tfrac{1}{8n_0^3}(p_0^2 + q_0^2)^2 + \dots \\ m_1 &= n_1 - \tfrac{1}{2n_1}(p_1^2 + q_1^2) - \tfrac{1}{8n_1^3}(p_1^2 + q_1^2)^2 + \dots \end{aligned}\right\} \tag{40}$$

so that

$$\frac{1}{m_1 - m_0} = \frac{1}{n_1 - n_0}\left[1 - \frac{1}{2n_0(n_1 - n_0)}(p_0^2 + q_0^2) + \frac{1}{2n_1(n_1 - n_0)}(p_1^2 + q_1^2) + \dots\right] \tag{41}$$

(39) becomes, on substitution from (41):

$$\begin{aligned} T(p_0, q_0; p_1, q_1) &= n_1 a_1 - n_0 a_0 \\ &+ \frac{r}{2(n_1 - n_0)}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right] + \frac{a_0}{2n_0}(p_0^2 + q_0^2) - \frac{a_1}{2n_1}(p_1^2 + q_1^2) \\ &- \frac{r}{4(n_1 - n_0)^2}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]\left[\frac{p_0^2 + q_0^2}{n_0} - \frac{p_1^2 + q_1^2}{n_1}\right] \\ &- \frac{(1+b)r}{8(n_1 - n_0)^3}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]^2 \\ &+ \frac{a_0}{8n_0^3}(p_0^2 + q_0^2)^2 - \frac{a_1}{8n_1^3}(p_1^2 + q_1^2)^2 + \dots \end{aligned} \tag{42}$$

The four variables $p_0, q_0, p_1$ and $q_1$ are seen to enter this expression only in the three combinations*

$$p_0^2 + q_0^2 = u^2, \quad p_1^2 + q_1^2 = v^2, \quad \text{and} \quad p_0 p_1 + q_0 q_1 = w^2. \tag{43}$$

With this substitution (42) becomes, on separating into orders,

$$T(p_0, q_0; p_1, q_1) = T^{(0)} + T^{(2)} + T^{(4)} + \dots$$

0

- It can be shown more generally that the angle characteristic of any medium which is rotationally symmetrical about the $z$-axis depends on the four variables only through the three combinations (43). To see this, we use a result proved in §5.1, according to which any function $F(x_0, y_0; x_1, y_1)$ which is invariant with respect to rotation of axes about the origin in the $xy$-plane depends only on the three scalar products

$$r_0^2 = x_0^2 + y_0^2, \quad r_1^2 = x_1^2 + y_1^2, \quad r_0 \cdot r_1 = x_0 x_1 + y_0 y_1,$$

of the two vectors $r_0(x_0, y_0), r_1(x_1, y_1)$. Identifying $r_0$ and $r_1$ with the projections of the propagation vectors $g_0(p_0, q_0, m_0)$ and $g_1(p_1, q_1, m_1)$ on to the $xy$-planes, the result follows.

where

$$\left.\begin{aligned} T^{(0)} &= n_1 a_1 - n_0 a_0, \\ T^{(2)} &= \mathscr{A}u^2 + \mathscr{B}v^2 + \mathscr{C}w^2, \\ T^{(4)} &= \mathscr{D}u^4 + \mathscr{E}v^4 + \mathscr{F}w^4 + \mathscr{G}u^2v^2 + \mathscr{H}u^2w^2 + \mathscr{K}v^2w, \end{aligned}\right\} \tag{44}$$

and

$$\left.\begin{aligned} \mathscr{A} &= \tfrac{1}{2}\left[\tfrac{r}{n_1-n_0} + \tfrac{a_0}{n_0}\right], \\ \mathscr{B} &= \tfrac{1}{2}\left[\tfrac{r}{n_1-n_0} - \tfrac{a_1}{n_1}\right], \\ \mathscr{C} &= -\tfrac{r}{n_1-n_0}, \\ \mathscr{D} &= -\tfrac{r}{4(n_1-n_0)^2}\left[\tfrac{1+b}{2(n_1-n_0)} + \tfrac{1}{n_0}\right] + \tfrac{a_0}{8{n_0}^3}, \\ \mathscr{E} &= -\tfrac{r}{4(n_1-n_0)^2}\left[\tfrac{1+b}{2(n_1-n_0)} - \tfrac{1}{n_1}\right] - \tfrac{a_1}{8{n_1}^3}, \\ \mathscr{F} &= -(1+b)r, \\ \mathscr{G} &= -\tfrac{r}{4(n_1-n_0)^3}, \\ \mathscr{H} &= \tfrac{r}{2(n_1-n_0)^2}\left[\tfrac{1+b}{n_1-n_0} + \tfrac{1}{n_0} - \tfrac{1}{n_1}\right], \\ \mathscr{K} &= \tfrac{r}{n_1-n_0} + \tfrac{1}{n_0}\Big], \\ 2\,(n_1-n_0)^2 & \left[\tfrac{1+b}{n_1-n_0} - \tfrac{1}{n_1}\right]. \end{aligned}\right\} \tag{45}$$

### 4.1.5 Approximate form of the angle characteristic of a reflecting surface of revolution

The expansion up to fourth degree for the angle characteristic associated with a reflecting surface of revolution can be derived in a similar manner. It is however not necessary to carry out the calculations in full. Using the same notation as in the preceding section (cf. Fig. 4.4 and 4.5) all the equations of § 4.1 .4 up to and including (39) apply without change in the present case; hence (39) is also the angle characteristic of a reflecting surface of revolution, when regarded as a function of all the six ray components. However, when $m_0$ and $m_1$ are eliminated from (39) with the help of the two identities connecting the ray components, different expressions for $T$ (as a function of four ray components) are obtained in the two cases. Denoting by $n$ the refractive index of the medium in which the rays are situated, we have in place of (40),

$$\left.\begin{aligned} m_0 = \quad & n - \tfrac{1}{2n}\left({p_0}^2 + q_0^2\right) - \tfrac{1}{8n^3}\left({p_0}^2 + {q_0}^2\right)^2 + \ldots. \\ m_1 = -\Big[ & n - \tfrac{1}{2n}\left({p_1}^2 + {q_1}^2\right) - \tfrac{1}{8n^3}\left({p_1}^2 + {q_1}^2\right)^2 + \ldots\Big]. \end{aligned}\right\} \tag{46}$$

In the second relation the negative square root $-\sqrt{n^2 - (p^2+q^2)}$ has been taken, as we assume that the reflected ray returns into the region from which the light is propagated ($z < 0$); the direction cosine of the reflected ray with respect to the positive $z$-direction, and consequently $m$, is therefore negative. Since (40) reduces to (46) on setting $n_0 = -n_1 = n$, it follows that the angle characteristic, considered as a function of the four ray components $p_0, q_0, p_1$ and $q_1$, of a reflecting surface of revolution, can be obtained

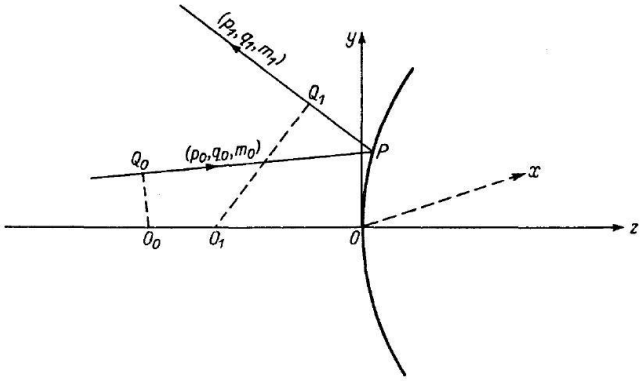



Fig. 4.5. The angle characteristic of a reflecting surface of revolution.
(The points $O_{0}, O_{1}, O, Q_{0}, P, Q_{1}$ are not necessarily coplanar.)

from the angle characteristic of a refracting surface of revolution by setting $n_0 = -n_1 = n$. Hence, for the case of reflection, we have

$$\left.\begin{aligned} T^{(0)} &= -n\,(a_0 + a_1) \\ T^{(2)} &= \mathscr{A}'u^2 + \mathscr{B}'v^2 + \mathscr{C}'w^2 \\ T^{(4)} &= \mathscr{D}'u^4 + \mathscr{E}'v^4 + \mathscr{F}'w^4 + \mathscr{G}'u^2v^2 + \mathscr{H}'u^2w^2 + \mathscr{K}'v^2w^2 \end{aligned}\right\} \tag{47}$$

where

$$\begin{aligned} \mathscr{A}' &= \frac{1}{2n}\left[-\frac{1}{2}r + a_0\right], \\ \mathscr{B}' &= \frac{1}{2n}\left[-\frac{1}{2}r + a_1\right], \\ \mathscr{C}' &= \frac{r}{2n}, \\ \mathscr{D}' &= -\frac{r}{16n^3}\left[-\frac{1+b}{4} + 1\right] + \frac{a_0}{8n^3}, \\ \mathscr{E}' &= -\frac{r}{16n^3}\left[-\frac{1+b}{4} + 1\right] + \frac{a_1}{8n^3}, \\ \mathscr{F}' &= \frac{(1+b)r}{16n^3}, \\ \mathscr{G}' &= \frac{r}{16n^3}\left[\frac{1+b}{2} - 2\right], \\ \mathscr{H}' &= \frac{r}{8n^3}\left[-\frac{1+b}{2} + 1\right], \\ \mathscr{K}' &= \frac{r}{8n^3}\left[-\frac{1+b}{2} + 1\right]. \end{aligned} \tag{48}$$

## 9.5.2 4.2 Perfect Imaging

Consider the propagation of light from a point source situated at a point $P_0$ in a medium specified by a refractive index function $n(x, y, z)$. An infinite number of rays will then proceed from $P_0$, but in general only a finite number will pass through any other point of the medium. In special cases, however, a point $P_1$ may be found through which an infinity of rays pass. Such a point $P_1$ is said to be a stigmatic (or a sharp) image of $P_0$.

In an ideal optical instrument every point $P_0$ of a three-dimensional region, called the object space, will give rise to a stigmatic image $P_1$. The totality of the image points defines the image space. The corresponding points in the two spaces are said to be conjugate points. In general not all the rays which proceed from $P_0$ will reach the image space; some, for example, will be excluded by the diaphragms of the instrument. Those rays which reach the image space will be said to lie in the field of the instrument. When $P_0$ describes a curve $C_0$ in the object space, $P_1$ will describe a conjugate curve $C_1$. The two curves will not necessarily be geometrically similar to each other. If every curve $C_0$ of the object space is geometrically similar to its image, we may say that the imaging between the two spaces is perfect. In a similar way we may define perfect imaging between two surfaces.

Optical instruments which are perfect in the sense just defined are of considerable interest, and accordingly we shall formulate some general theorems relating to perfect, or at least sharp, imaging of three-dimensional domains. Some results relating to sharp imaging of two-dimensional domains (surfaces) will be briefly discussed in § 4.2.3.

### 4.2.1 General theorems

An optical system $\mathscr{I}$ which images stigmatically a three-dimensional domain is often called an absolute instrument. It will be shown that in an absolute instrument the optical length of any curve in the object space is equal to the optical length of its image. This theorem was first put forward by Maxwell* in 1858 for the special case when both the object and the image space are homogeneous. More rigorous proofs were later given by H. Bruns (1895), F. Klein (1901) and H. Liebmann (1916). †

The theorem was later shown by Carathodory ‡ not to be restricted to homogeneous media, but to be valid also when the media are heterogeneous and anisotropic. In proving this theorem the method of Carathodory will be used, but our discussion will be restricted to absolute instruments with isotropic (but generally heterogeneous) object and image spaces. §

Let $A_0B_0$ and $A_1B_1$ be ray-segments in the object and image spaces (Fig. 4.6) of a ray which lies in the field of an absolute instrument $\mathscr{I}$. Any other ray with a lineelement which neither in position nor in direction departs appreciably from an element of $A_0B_0A_1B_1$, will also lie in the field of the instrument.

If each element of a curve (which is assumed to have a continuously turning tangent) coincides with an element of some ray which lies completely in the field of $\mathscr{I}$, the curve will be said to lie tangentially in the field of $\mathscr{I}$. If a "polygon" with a sufficient number of sides is inscribed into such a curve, each side of the polygon will coincide with an element of a ray which lies completely in the field of the instrument.

According to the principle of equal optical path (see § 3.3.3), all the rays which join $A_0$ to

0

- J. C. Maxwell, Quart. J. Pure Appl. Maths., 2 (1858), p. 233. Also his Scientific Papers, I (Cambridge University Press, 1890), 271.
  † H. Bruns, "Das Eikonal", Abh. Kgl. schs. Ges. Wiss., math-phys. Kl., 21 (1895), 370; F. Klein, Z. Math. Phys., 46 (1901), 376; Ges. Math. Abh., 2 (1922), 607. (See also E. T. Whittaker, The Theory of Optical Instruments (Cambridge University Press, 1907), 47. H. Liebmann, Sitzgsber. bayer. Akad. Wiss., Math-naturw. Abt. (1916), 183. An account of these researches will also be found in an article by H. Boegehold in S. Czapski and O. Eppenstein, Grundzge der Optischen Instrumente nach Abbe (Leipzig, Barth, 3rd edition (1924), p. 213).
  ‡ C. Carathodory, Sitzgsber. bayer. Akad. Wiss. Math-naturw. Abt., 56 (1926), 1.
  § In light optics heterogeneous substances are not used, but they are employed in microwave optics [see, for example, J. Brown, Microwave Lenses (London, Methuen), 1953]. In electron optics the "refractive index" (cf. Appendix II) is also a function of position. Our discussion is not, therefore, of purely academic interest.

its image $A_1$ have the same optical length. We shall denote this optical path length by $V(A_0)$ and will show that it is in fact independent of $A_0$.

Let $B_0$ and $B_1$ be another pair of conjugate points. Then (see Fig. 4.6)

$$[A_1B_1] = [A_0B_0] + V(B_0) - V(A_0). \tag{1}$$

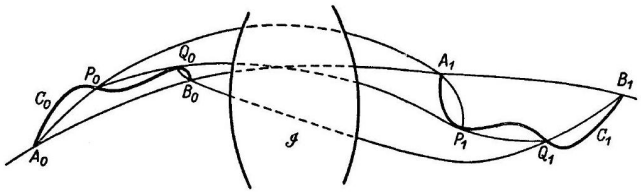



Fig. 4.6. An absolute optical instrument.

Let $C_0$ be a curve which joins $A_0$ and $B_0$ and which lies tangentially in the field of the instrument, and let $C_1$ be its image. We inscribe into $C_0$ a polygon $A_0P_0Q_0B_0$ and denote the image points of $P_0$ and $Q_0$ by $P_1$ and $Q_1$. Then, on applying (1) to the side $A_1P_1$, we have

$$[A_1P_1] = [A_0P_0] + V(P_0) - V(A_0),$$

and similarly for the other sides:

$$[P_1Q_1] = [P_0Q_0] + V(Q_0) - V(P_0),$$

and

$$[Q_1B_1] = [Q_0B_0] + V(B_0) - V(Q_0).$$

Hence

$$[A_1P_1] + [P_1Q_1] + [Q_1B_1] = [A_0P_0] + [P_0Q_0] + [Q_0B_0] + V(B_0) - V(A_0).$$

Obviously this result can be extended to a polygon of any number of sides $N$. Proceeding to the limit, as $N \to \infty$ in such a way that the greatest side tends to zero, we obtain the relation

$$L_1 = L_0 + V(B_0) - V(A_0), \tag{2}$$

where

$$L_0 = \int_{C_0} n_0 ds_0, \quad L_1 = \int_{C_1} n_1 ds_1 \tag{3}$$

are the optical lengths of the curves $C_0$ and $C_1$. Next, it will be shown that $V(B_0) = V(A_0)$. The points on the two curves are in a one-one correspondence, which may be expressed by relations of the form

$$x_1 = f(x_0, y_0, z_0), \quad y_1 = g(x_0, y_0, z_0), \quad z_1 = h(x_0, y_0, z_0). \tag{4}$$

An element $ds_1$ of $C_1$ is a function of the corresponding element $ds_0$,

$$ds_1 = \sqrt{\left(\frac{dx_1}{ds_0}\right)^2 + \left(\frac{dy_1}{ds_0}\right)^2 + \left(\frac{dz_1}{ds_0}\right)^2}\, ds_0. \tag{5}$$

Hence

$$L_1 = \int_{C_1} F\left(x_1, y_1, z_1, \frac{dx_1}{ds_0}, \frac{dy_1}{ds_0}, \frac{dz_1}{ds_0}\right) ds_0 \tag{6}$$

where

$$F\left(x_1, y_1, z_1, \frac{dx_1}{ds_0}, \frac{dy_1}{ds_0}, \frac{dz_1}{ds_0}\right) = n_1\left(x_1, y_1, z_1\right) \sqrt{\left(\frac{dx_1}{ds_0}\right)^2 + \left(\frac{dy_1}{ds_0}\right)^2 + \left(\frac{dz_1}{ds_0}\right)^2}$$

is a homogeneous function of the first degree in the derivatives $dx_1/ds_0, dy_1/ds_0$ and $dz_1/ds_0$; moreover $F$ remains unchanged when $dx_1/ds_0, \ldots$, is replaced by $-dx_1/ds_0, \ldots$. Now from (4),

$$\frac{dx_1}{ds_0} = \frac{\partial f}{\partial x_0}\frac{dx_0}{ds_0} + \frac{\partial f}{\partial y_0}\frac{dy_0}{ds_0} + \frac{\partial f}{\partial z_0}\frac{dz_0}{ds_0} \tag{7}$$

with similar expressions for $dy_1/ds_0$ and $dz_1/ds_0$. Hence using (7) and (4), $F$ can be expressed in the form

$$F\left(x_1, y_1, z_1, \frac{dx_1}{ds_0}, \frac{dy_1}{ds_0}, \frac{dz_1}{ds_0}\right) = \Phi\left(x_0, y_0, z_0, \frac{dx_0}{ds_0}, \frac{dy_0}{ds_0}, \frac{dz_0}{ds_0}\right) \tag{8}$$

$\Phi$ being also a homogeneous function of the first degree in $dx_0/ds_0, \ldots$; moreover, $\Phi$ remains unchanged when $dx_0/ds_0, \ldots$, is replaced by $-dx_0/ds_0, \ldots$,

$$\Phi\left(x_0, y_0, z_0, -\frac{dx_0}{ds_0}, -\frac{dy_0}{ds_0}, -\frac{dz_0}{ds_0}\right) = \Phi\left(x_0, y_0, z_0, \frac{dx_0}{ds_0}, \frac{dy_0}{ds_0}, \frac{dz_0}{ds_0}\right). \tag{9}$$

From (2), (6), and (8) it follows that

$$\int_{C_0} \left(n_0 - \Phi\right) ds_0 = V\left(A_0\right) - V\left(B_0\right) \tag{10}$$

showing that the value of the curvilinear integral in (10) depends only on the end points $A_0, B_0$ and not on the choice of $C_0$. The curve $C_0$ is, however, not quite arbitrary, for it must lie tangentially in the field of the instrument. Nevertheless, it may be concluded that the expression $\left(n_0 - \Phi\right) ds_0$ must be a complete differential of some function $\Psi$,

$$n_0 - \Phi = \frac{\partial \Psi}{\partial x_0}\frac{dx_0}{ds_0} + \frac{\partial \Psi}{\partial y_0}\frac{dy_0}{ds_0} + \frac{\partial \Psi}{\partial z_0}\frac{dz_0}{ds_0}$$

If now the derivatives $dx_0/ds_0, \ldots$, are replaced by $-dx_0/ds_0$, the right-hand side will change sign, but the left-hand side will, on account of (9), remain unchanged. This is only possible if each side vanishes ( $\Psi$ = constant); hence

$$\Phi = n_0. \tag{11}$$

Equation (10) shows that $V\left(A_0\right) = V\left(B_0\right)$ and consequently (2) reduces to the relation $L_1 = L_0$. Hence for any curve, whether or not it lies tangentially in the field, provided only it has an image,

$$\int_{C_0} n_0 ds_0 = \int_{C_1} n_1 ds_1 \tag{12}$$

This is Maxwell's theorem for an absolute instrument.*

From Maxwell's theorem a number of interesting conclusions can immediately be drawn. Consider a small triangle whose sides are of lengths ${ds_0}^{(1)}, {ds_0}^{(2)}, {ds_0}^{(3)}$ and let ${ds_1}^{(1)}, {ds_1}^{(2)}, {ds_1}^{(3)}$ be the sides of its image formed by an absolute instrument. Further let $n_0$

and $n_1$ be the refractive indices of the regions where the triangles are situated. By Maxwell's theorem,

$$n_0 ds_0^{(1)} = n_1 ds_1^{(1)}, \quad n_0 ds_0^{(2)} = n_1 ds_1^{(2)}, \quad n_0 ds_0^{(3)} = n_1 ds_1^{(3)}. \tag{13}$$

Hence the two triangles are similar to each other and corresponding sides are in the inverse ratio of the refractive indices. The angle between any two curves will therefore be preserved in the imaging, i.e. the imaging must be a conformal transformation. Now there is a general theorem due to Liouville † according to which a conformal transformation of a three-dimensional domain on to a three-dimensional domain can only be a projective transformation (collineation), an inversion ‡ or the combination of these two transformations. We have thus established the following theorem due to

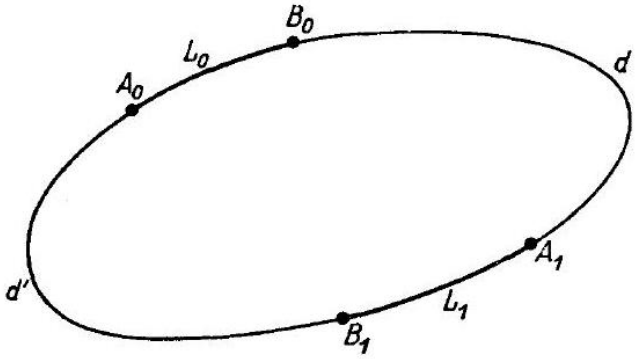



Fig. 4.7. Illustrating Lenz's proof of Maxwell's theorem for an absolute instrument.

must pass through these points. Hence each ray must be a closed curve and by the principle of equal optical path it follows that (see Fig. 4.7)

$$[A_0A_1]_{\text{clockwise}} = [A_0A_1]_{\text{anticlockwise}}$$

and

$$[B_0B_1]_{\text{clockwise}} = [B_0B_1]_{\text{anticlockwise}}$$

Let

$$[A_0B_0] = L_0, \quad [A_1B_1] = L_1, \quad [B_0A_1] = d, \quad [B_1A_0] = d'.$$

The two equations then become

$$L_0 + d = d' + L_1$$
$$d + L_1 = L_0 + d'$$

and on subtraction we obtain

$$L_0 = L_1.$$

0

- The following less general, but very simple and elegant proof of Maxwell's theorem was given by W. Lenz (contribution in Probleme der Modernen Physik, edited by P. Debye (Leipzig, Hirzel, 1928), p. 198.

Assume that all the rays from each object point reach the image and let ( $A_0, A_1$ ) and ( $B_0, B_1$ ) be two conjugate pairs. By hypothesis, the ray $A_0B_0$ must pass through $A_1$ and $B_1$. Likewise $B_0A_0$

This proves Maxwell's theorem for the special case when the curve is a portion of a ray. Generalization to an arbitrary curve may be obtained as in our main proof, by regarding the curve as a limiting form of a polygon formed by a large number of ray segments.
† See, for example, W. Blaschke, Vorlesungen ber Differential-Geometrie I (Berlin, Springer, 2nd ed. (1924), p. 68; 4th ed. (1945), p. 101)).
‡ An inversion transforms each point $P_0$ into a point $P_1$ on the line joining $P_0$ with a fixed origin $O$, and the product $OP_0.OP_1$ is constant.

Carathodory: The imaging by an absolute instrument is either a projective transformation, an inversion, or a combination of the two.

Let us now briefly consider the case when the imaging between the two spaces is not only stigmatic but is perfect, i.e. it is such that any figure is transformed into one which is geometrically similar to it. Clearly the imaging must be a projective transformation, since it transforms lines into lines.* It then follows from (13) that the magnification $ds_1/ds_0$ between any two conjugate linear elements is equal to the ratio $n_0/n_1$ of the refractive indices. In particular, if $n_0 = n_1 =$ constant, then $ds_1/ds_0 = 1$, so that a perfect imaging between two homogeneous spaces of equal refractive indices is always trivial in the sense that it produces an image which is congruent with the object. A plane mirror (or a combination of plane mirrors) is the only known instrument which produces such imaging.

These general considerations imply that, in order to obtain non-trivial imaging between homogeneous spaces of equal refractive indices, the requirement of exact stigmatism or of strict similarity between the object and the image must be dropped.

### 4.2.2 Maxwell's "fish-eye"

A simple and interesting example of an absolute instrument is presented by the medium which is characterized by the refractive index function

$$n(r) = \frac{1}{1 + (r/a)^2} n_0, \tag{14}$$

where $r$ denotes the distance from a fixed point $O$, and $n_0$ and $a$ are constants. It is known as the "fish-eye" and was first investigated by Maxwell. †

It was shown in § 3.2 that in a medium with spherical symmetry the rays are plane curves which lie in planes through the origin, and that the equation of the rays may be written in the form (cf. § 3.2 (11))

$$\theta = c \int^r \frac{dr}{r\sqrt{n^2(r)r^2 - c^2}},$$

$c$ being a constant. On substituting from (14) and setting

$$\rho = \frac{r}{a}, \quad K = \frac{c}{an_0}, \tag{15}$$

0

- Cf. F. Klein, Elementary Mathematics from an Advanced Standpoint, Vol. II (translated from third German ed., London, Macmillan, 1939, 89; reprinted by Dover Publications, New York).
  † J. C. Maxwell, Cambridge and Dublin Math.J., 8 (1854), 188; also Scientific Papers, I (Cambridge University Press), p. 76.

Interesting generalizations of Maxwell's fish-eye were found by W. Lenz, contribution in Probleme der Modernen Physik, edited by P. Debye (Leipzig, Hirzel, 1928), 198 and R. Stettler, Optik, 12 (1955), 529. The latter paper also includes a generalization of the so-called Luneburg lens which, because of its wide angle scanning capabilities, has useful applications in microwave antenna design. This lens, first considered by R. K. Lunuburg in his Mathematical theory of Optics (mimeographed lecture notes, Brown University, Providence, R.I., 1944;

we obtain

$$\theta = \int^{\rho} \frac{K\left(1+\rho^2\right) d\rho}{\rho\sqrt{\rho^2 - K^2\left(1+\rho^2\right)^2}}. \tag{16}$$

It may be verified that

$$\frac{K\left(1+\rho^2\right)}{\rho\sqrt{\rho^2 - K^2\left(1+\rho^2\right)^2}} = \frac{d}{d\rho}\left[\arcsin\left(\frac{K}{\sqrt{1-4K^2}}\frac{\rho^2-1}{\rho}\right)\right],$$

so that (16) becomes

$$\sin(\theta - \alpha) = \frac{c}{\sqrt{a^2 n_0^2 - 4c^2}}\frac{r^2 - a^2}{ar}, \tag{17}$$

where $\alpha$ is a constant of integration.

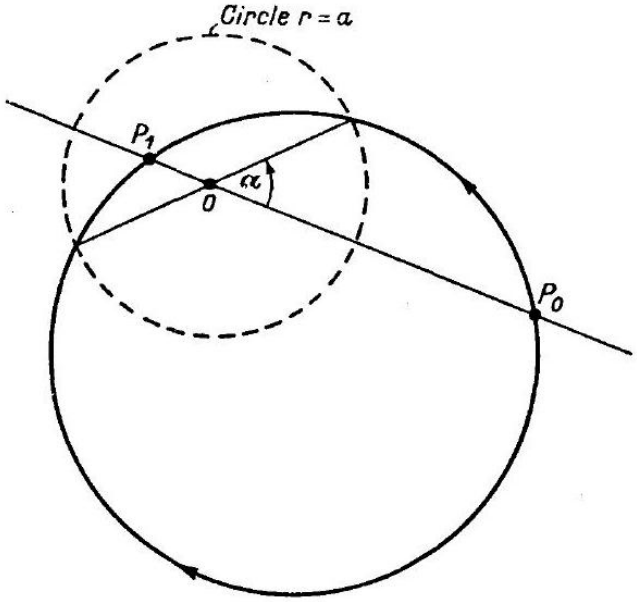



Fig. 4.8. Rays in Maxwell's "fish-eye".

(17) is the polar equation of the rays. The one-parameter family of rays through a fixed point $P_0\left(r_0, \theta_0\right)$ is, therefore, given by

$$\frac{r^2 - a^2}{r\sin(\theta - \alpha)} = \frac{r_0^2 - a^2}{r_0 \sin\left(\theta_0 - \alpha\right)}. \tag{18}$$

It is seen that whatever the value of $\alpha$, this equation is satisfied by $r = r_1, \theta = \theta_1$, where

$$r_1 = \frac{a^2}{r_0}, \quad \theta_1 = \pi + \theta_0, \tag{19}$$

printed version published by University of California Press, Berkeley and Los Angeles, 1964, § 29), is an inhomogeneous sphere with the refractive index function $n(r) = \sqrt{2 - r^2}(0 \leqslant r \leqslant 1)$. When placed in a homogeneous medium of unit refractive index, it brings to a sharp focus every incident pencil of parallel rays. See also R. F. Rinehart, J. Appl. Phys., 19 (1348), 860 ; A. Fletcher, T. Murphy and A. Young, Proc. Roy. Soc., A 223 (1954), 216; and G. Toraldo di Francla, Optica Acta, 1 (1954-1955), 157.

showing that all the rays from an arbitrary point $P_0$ meet in a point $P_1$ on the line joining $P_0$ to $O$; $P_0$ and $P_1$ are on opposite sides of $O$ and $OP_0.OP_1 = a^2$. Hence the fish-eye is an absolute instrument in which the imaging is an inversion.

We note that (17) is satisfied by $r = a, \theta = \alpha$ and $r = a, \theta = \pi + \alpha$; each ray therefore intersects the fixed circle $r = a$ in diametrically opposite points (see Fig. 4.8).

To obtain the equation of the rays in Cartesian coordinates, we put $x = r\cos\theta$, $y = r\sin\theta$ in (17), and find

$$y\cos\alpha - x\sin\alpha = \frac{c}{a\sqrt{a^2n_0^2 - 4c^2}}\left(x^2 + y^2 - a^2\right),$$

or

$$(x + b\sin\alpha)^2 + (y - b\cos\alpha)^2 = a^2 + b^2 \tag{20}$$

where

$$b = \frac{a}{2c}\sqrt{a^2n_0^2 - 4c^2}$$

(20) shows that each ray is a circle.

### 4.2.3 Stigmatic imaging of surfaces

So far we have been concerned only with perfect or sharp imaging of three-dimensional domains. We saw that when the object space and image space are homogeneous and of equal refractive indices, perfect imaging can only be of a trivial kind, producing a mirror image of the object. It is natural to inquire whether non-trivial imaging may be obtained, when it is required that only certain surfaces should be imaged perfectly (or at least sharply) by the instrument. This question has been investigated by a number of authors,* who found that when the object space and image space are homogeneous, not more than two surfaces may in general † be sharply imaged by a rotationally symmetrical system. For the proof of this theorem the papers by Boegehold and Herzberger and by Smith should be consulted. Here we shall only consider in detail a simple case of sharp imaging of a spherical surface which is of particular interest in practice.

Consider the refraction at a solid homogeneous sphere $S$ embedded in a homogeneous medium. Let $O$ be the centre of the sphere, $r$ its radius, and $n$ and $n'$ the refractive indices of the sphere and of the surrounding medium respectively. Further, let $AQ$ be a ray incident upon the sphere. The refracted ray $QB$ can easily be found by means of the following construction:

Let $S_0$ and $S_1$ be two spheres whose centres are at $O$ and whose radii are

$$r_0 = \frac{n}{n'}r, \quad r_1 = \frac{n'}{n}r. \tag{21}$$

If $P_0$ is the point of intersection of $AQ$ with $S_0$, and $P_1$ is the point at which $OP_0$ meets $S_1$, then $QP_1$ is the refracted ray. For one has, by construction (see Fig. 4.9)

$$\frac{OQ}{OP_0} = \frac{OP_1}{OQ} = \frac{n'}{n} \tag{22}$$

Moreover

$$\widehat{QOP_0} = \widehat{QOP_1} \tag{23}$$

Hence the triangles $OQP_0$ and $OP_1Q$ are similar, and consequently

$$\frac{\sin\phi_0}{\sin\phi_1} = \frac{OP_0}{OQ} = \frac{n}{n'}, \tag{24}$$

where $\phi_0 = \widehat{OQP_0}$ and $\phi_1 = \widehat{OQP_1}$ are the angles of incidence and refraction respectively. $\phi_0$ and $\phi_1$ satisfy the law of refraction, and consequently $QP_1$ is the refracted ray.

The construction implies that all rays which diverge from a point $P_0$ on $S_0$ will form a (virtual) stigmatic image at the point $P_1$ in which the diameter $OP_0$ intersects $S_1$. Hence the sphere $S_1$ is a stigmatic image of $S_0$, and vice versa.

It will be useful to express (24) in a somewhat different form. If we denote by $\theta_0$ and $\theta_1$ the angles which the two conjugate rays make with the line $P_0P_1$, i.e. $\theta_0 = O\widehat{P_0Q}$,

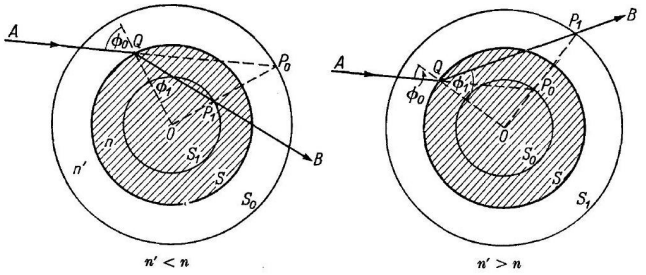



Fig. 4.9. Refraction at a spherical surface. Aplanatic points.

$\theta_1 = \widehat{OP_1Q}$, then, since the two triangles are similar, it follows that $\theta_0 = \phi_1$ and $\theta_1 = \phi_0$; hence

$$\frac{\sin\theta_1}{\sin\theta_0} = \frac{n}{n'} = \text{ constant.} \tag{25}$$

Equation (25) is a special case of the so-called sine condition whose significance will be explained in § 4.5. In accordance with the terminology of § 4.5, $P_0$ and $P_1$ are called aplanatic points of the spherical surface $S$.

The existence of the aplanatic points for refraction at a spherical surface is made use of, as will be shown in §6.6, in the construction of certain microscope objectives.

## 9.5.3 4.3 Projective Transformation (Collineation) with Axial Symmetry

It has been shown in the previous section that perfect imaging between threedimensional domains must necessarily be a projective transformation, since it transforms lines into lines. But even when the requirements for perfect imaging are not strictly fulfilled, the properties of projective transformations are of great importance. For, as will be seen later, the relationship between the object and the image in any optical system is, at least to a first approximation, a transformation of this kind. It will therefore be convenient to study the general properties of a projective transformation before deriving the laws of image formation in actual instruments.

0

- H. Boegehold and M. Herzberger, Compositio Mathematica, 1 (1935), 448; M. Herzberger, Ann. New York Acad. Sci., 48 (1946), Art. 1, p. 1; T. Smith, Proc. Phys. Soc., 60 (1948), 293. See also C. G. Wynne, Proc. Phys. Soc., 65B (1952), 436.
  † "In general" implies here that certain degenerate cases in which the whole object space is imaged sharply (e.g. by reflection on a plane mirror) are excluded.

Though this preliminary discussion is essentially of a geometrical nature, it will be convenient to retain, where possible, the terminology of optics.

### 4.3.1 General formulae

Let $(\ x, y, z\ )$ be the coordinates of a point $P$ of the object space and $(\ x', y', z'\ )$ the coordinates of a point $P'$ in the image space, both referred at present to the same set of Cartesian rectangular axes, chosen arbitrarily. A projective relationship between the two spaces is mathematically expressed by relations of the form

$$x' = \frac{F_1}{F_0}, \quad y' = \frac{F_2}{F_0}, \quad z' = \frac{F_3}{F_0} \tag{1}$$

where

$$F_i = a_i x + b_i y + c_i z + d_i \quad (i = 0, 1, 2, 3) \tag{2}$$

Pairs of points related by (1) will be said to form a conjugate pair.
Solving (1) for $x, y, z$, we obtain relations of the same form

$$x = \frac{F_1'}{F_0''}, \quad y = \frac{F_2'}{F_0''}, \quad z = \frac{F_3'}{F_0'}, \tag{3}$$

where

$$F_i' = a_i' x' + b_i' y' + c_i' z' + d_i'$$

From (1) it follows that the image of any point situated in the plane $F_0 = 0$ will lie at infinity. Similarly from (3) it is seen that all object points whose images lie in the plane $F_0' = 0$ are at infinity. The plane $F_0 = 0$ is called the focal plane of the object space and the plane $F_0' = 0$ is known as the focal plane of the image space.* Rays which are parallel in the object space will be transformed into rays which intersect in a point on the focal plane $F_0' = 0$. Similarly rays from a point in the focal plane $F_0 = 0$ will transform into a pencil of parallel rays. In special cases both the focal planes may lie at infinity. The transformation is then said to be affine or telescopic. Then to finite values $(\ x, y, z\ )$ there correspond finite values of $(\ x', y', z'\ )$, so that in a telescopic transformation one always has $F_0 \neq 0$ and $F_0' \neq 0$. This clearly is only possible when $a_0 = b_0 = c_0 = 0$ and $a_0' = b_0' = c_0' = 0$.

Of special importance for optics is the case of axial symmetry, since the majority of optical systems consist of surfaces of revolution with a common axis (called usually centred systems). It then follows from symmetry that the image of each point $P_0$ lies in the plane which contains $P_0$ and the axis; in considering the properties of the associated projective transformation we may therefore restrict our discussion to points situated in such a meridional plane. Let the meridional plane be the $yz$-plane and take the $z$-axis along the axis of symmetry. A point $(0, y, z)$ of the object space will then be transformed into a point $(\ 0, y', z'\ )$ of the image space, where

$$y' = \frac{b_2 y + c_2 z + d_2}{b_0 y + c_0 z + d_0}, \quad z' = \frac{b_3 y + c_3 z + d_3}{b_0 y + c_0 z + d_0}. \tag{4}$$

Now it follows from symmetry that $z'$ remains unchanged when $y$ is changed into $-y$. This, in general, is only possible if $b_0 = b_3 = 0$. Further it follows that, if $y \to -y$, then $y' \to -y'$, which implies that $c_2 = d_2 = 0$. Hence (4) reduces to

$$y' = \frac{b_2 y}{c_0 z + d_0}, \quad z' = \frac{c_3 z + d_3}{c_0 z + d_0} \tag{5}$$

These equations contain five constants but only their ratios are significant. Hence a projective transformation with axial symmetry is characterized by four parameters.

Solving (5) for $y$ and $z$, we obtain

$$y = \frac{c_0 d_3 - c_3 d_0}{b_2} \frac{y'}{c_0 z' - c_3}, \quad z = \frac{-d_0 z' + d_3}{c_0 z' - c_3}. \tag{6}$$

From (5) and (6) it is seen that the focal planes are given by

$$F_0 \equiv c_0 z + d_0 = 0, \quad F_0' \equiv c_0 z' - c_3 = 0;$$

hence the focal planes intersect the axis at right angles in the points whose abscissae are

$$z = -\frac{d_0}{c_0}, \quad z' = \frac{c_3}{c_0}. \tag{7}$$

These points are called the principal foci and are denoted by $F$ and $F'$ in Fig. 4.10.

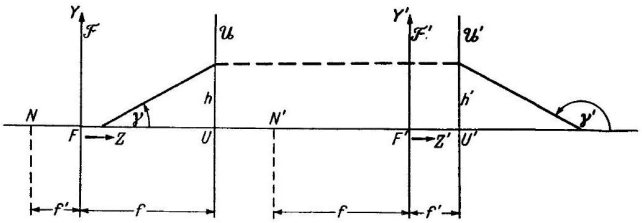



Fig. 4.10. The cardinal points and planes of an optical system: $F\hat{}\{\}\{\prime\}$, $F\hat{}\{\}\{\prime\}=$ foci ; $U, U' =$ unit (or principal) points; $N, N' =$ nodal points; $\mathscr{F}, \mathscr{F}' =$ focal planes; $\mathscr{U}, \mathscr{U}' =$ unit (principal) planes.

It will now be convenient to introduce a separate coordinate system for each of the two spaces, measuring the $z$ coordinates from the principal foci; i.e. we set

$$\left.\begin{aligned} y = Y, \quad c_0 z + d_0 = c_0 Z, \\ y' = Y', \quad c_0 z' - c_3 = c_0 Z'. \end{aligned}\right\} \tag{8}$$

(5) then become

$$Y' = \frac{b_2}{c_0} \frac{Y}{Z}, \quad Z' = \frac{c_0 d_3 - c_3 d_0}{c_0^2 Z}.$$

Further we set

$$f = \frac{b_2}{c_0}, \quad f' = \frac{c_0 d_3 - c_3 d_0}{b_2 c_0}. \tag{9}$$

With this substitution, the equations of the transformation take the simple form

$$\frac{Y'}{Y} = \frac{f}{Z} = \frac{Z'}{f'}. \tag{10}$$

The second relation, $ZZ' = ff'$, is usually called Newton's equation. The constant $f$ is known as the focal length of the object space and $f'$ as the focal length of the image space.

0

- The terms "focal plane" and "focal points" have here a somewhat different meaning than in connection with normal congruences (§ 3.2.3) and astigmatic pencils (§ 4.6).

From (10) it follows that, for fixed object and image planes,

$$\left(\frac{dY'}{dY}\right)_{Z-\text{ const.}} = \frac{Y'}{Y} = \frac{f}{Z} = \frac{Z'}{f'}. \tag{11}$$

This quantity is known as the lateral magnification. Further, we have independently of $Y$ and $Y'$, the longitudinal magnification

$$\frac{dZ'}{dZ} = -\frac{Z'}{Z} = -\frac{ff'}{Z^2} = -\frac{Z'^2}{ff'}. \tag{12}$$

From (11) and (12) it is seen that the two magnifications are related by the formula

$$\frac{dZ'}{dZ} = -\frac{f'}{f}\left(\frac{dY'}{dY}\right)^2_{Z=\text{ const.}} \tag{13}$$

Since the lateral magnification depends on $Z$ but not on $Y$ it follows that an object which is situated in a plane perpendicular to the axis will be transformed into one which is geometrically similar to it.

The lateral magnification is equal to unity when $Z = f$ and $Z' = f'$. These planes are called the unit or principal planes and are denoted by the letters $\mathscr{U}$ and $\mathscr{U}'$ in Figs. 4.10 and 4.11. The points $U$ and $U'$ in which these planes intersect the axis are known as the unit or principal points.

Let $h$ be the distance from the axis at which a ray from an axial point $(0, 0, Z)$ meets the unit plane of the object space. The conjugate ray meets the other unit plane at the same distance ( $h$ ) from the axis; and the angles $\gamma$ and $\gamma'$ which the two rays make with the axis are given by (see Fig. 4.10)

$$\tan\gamma = \frac{h}{f - Z}, \quad \tan\gamma' = \frac{h}{f' - Z'}.$$

The ratio

$$\frac{\tan\gamma'}{\tan\gamma} = \frac{f - Z}{f' - Z'} = -\frac{Z}{f'} = -\frac{f}{Z'} \tag{14}$$

is known as the angular magnification or convergence ratio. It is independent of $h$ and $h'$ and is equal to unity when $Z = -f'$ and $Z' = -f$. The two axial points $N$ and $N'$ specified by these expressions are known as nodal points. They are the two conjugate points which are characterized by the property that conjugate rays passing through them are parallel to each other. The distance between the nodal points is equal to the distance between the unit points. When $f = -f'$ the nodal points and the unit points coincide. The foci, the unit points, and the nodal points are called cardinal points of the transformation.

It is sometimes convenient to measure distances from the unit planes rather than from the focal planes. This means that in place of $Z$ and $Z'$ we use the variables

$$\zeta = Z - f, \quad \zeta' = Z' - f'. \tag{15}$$

Newton's equation $ZZ' = ff'$ then becomes

$$\frac{f}{\zeta} + \frac{f'}{\zeta'} = -1. \tag{16}$$

Using the properties of the focal points and the principal planes, we can find the point $P'$ conjugate to a given point $P$ by a simple geometrical construction. Two lines are drawn through $P$, one parallel to the axis and the other through the focus $F$ (Fig. 4.11). Let $A$ and $B$ be the

points in which these lines intersect the unit plane $\mathscr{U}$. From the property of the unit planes it then follows that the points $A'$ and $B'$, which are conjugate to $A$ and $B$, are the points of intersection with $\mathscr{U}'$ of the
two lines through $A$ and $B$ drawn parallel to the axis. Moreover, since $PA$ passes through the focus $F$, $P'A'$ must be parallel to the axis, and since $PB$ is parallel to the axis, $P'B'$ must pass through the other focus $F'$. Hence the image point $P'$ is the point of intersection of $AA'$ with $B'F'$.

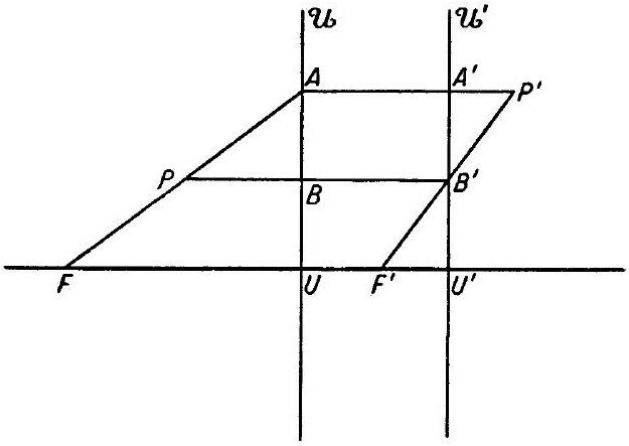



Fig. 4.11. Graphical determination of the image point.

### 4.3.2 The telescopic case

We shall now consider the special case of telescopic (affine) collineation. As already explained it is characterized by having both focal planes at infinity. The coefficient $c_0$ in (5) and (6) then vanishes, and (5) reduces to

$$y' = \frac{b_2 y}{d_0}, \quad z' = \frac{c_3 z + d_3}{d_0} \tag{17}$$

We shall again choose separate coordinate systems for the object space and for the image space. As origins we take any conjugate axial pair of points. Referred to the new set of axes, the equations of the transformation take the simple form

$$Y' = \alpha Y, \quad Z' = \beta Z \tag{18}$$

where $\alpha = b_2/d_0$ and $\beta = c_3/d_0$.
The lateral as well as the longitudinal magnification is now constant. The angular magnification is also constant; to show this, consider two conjugate rays and take the origins of the two coordinate systems at the points in which these rays intersect the axis. The equations of the two rays then are

$$Y = Z \tan\gamma, \quad Y' = Z' \tan\gamma' \tag{19}$$

Hence

$$\frac{\tan\gamma'}{\tan\gamma} = \frac{Y'}{Z'}\frac{Z}{Y} = \frac{\alpha}{\beta} = \frac{b_2}{c_3}. \tag{20}$$

Although in a telescopic transformation both $f$ and $f'$ are infinite, the ratio $f' \mid f$ must be regarded as finite. For by (9), as $c_0 \to 0$,

$$\frac{f'}{f} \to -\frac{c_3 d_0}{{b_2}^2} = -\frac{\beta}{\alpha^2} \tag{21}$$

### 4.3.3 Classification of projective transformations

Projective transformations may be classified according to the signs of the focal lengths.

When the focal lengths are of opposite signs, $ff' < 0$, and according to (12) $dZ'/dZ > 0$. This implies that if the object is displaced in a direction parallel to the axis, the image will be displaced in the same direction. It will be seen later that such imaging occurs whenever the image is produced either by refraction alone, or by an even number of reflections, or by a combination of the two. Such imaging is called concurrent or dioptric.

When the focal lengths have equal signs, $ff' > 0, dZ'/dZ < 0$, so that to a displacement of the object in a direction parallel to the axis, there corresponds a displacement of the image in the opposite direction. This type of imaging is produced by means of an odd number of reflections, or by a combination of an odd number of reflections with any number of refractions. Imaging of this type is called contracurrent or katoptric.

In each group two types of transformations are distinguished, according to the sign of the focal length $f$. It is seen from (11) that for $Z > 0$ the lateral magnification is positive or negative according as $f > 0$ or $f < 0$. Hence an object situated in the right-hand half of the object space will have an image which is similarly oriented or inverted, according as $f$ is positive or negative. In the former case the transformation is said to be convergent; in the latter, it is said to be divergent. This terminology is derived from the fact that an incident bundle of parallel rays, after passing the unit plane of the image space, is rendered convergent in the former case and divergent in the latter. The four cases are summarized in Table XII.

Table XII
Classification of projective transformations (Cartesian sign-convention)

| | Convergent | Divergent |
|---|---|---|
| Concurrent (Dioptric) | $f > 0; f' < 0$ | $f < 0; f' > 0$ |
| Contracurrent (Katoptric) | $f > 0; f' > 0$ | $< 0; f' < 0$ |

In the special case when the transformation is telescopic, the four types are distinguished by the signs of $\alpha$ and $\beta$. It follows from (21) that the transformation will be concurrent when $\beta$ is positive and contracurrent when it is negative. Further it is seen from (18) that it will be convergent or divergent according as to $\alpha$ is positive or negative.

### 4.3.4 Combination of projective transformations

We now consider the combination of two successive projective transformations, which are rotationally symmetrical about the same axis.

Let the subscript 0 refer to the first transformation, and subscript 1 to the second. Then the equations specifying the two transformations are:

$$\left.\begin{array}{l} \frac{Y_0'}{Y_0} = \frac{f_0}{Z_0} = \frac{Z_0'}{f_0'}, \\ \frac{Y_1'}{Y_1} = \frac{f_1}{Z_1} = \frac{Z_1'}{f_1'}, \end{array}\right\} \tag{22}$$

Let $c$ be the distance between the focal points $F_0'$ and $F_1$. Since the image space of the first transformation coincides with the object space of the second,

$$Z_1 = Z_0' - c, \quad Y_1 = Y_0'. \tag{23}$$

Elimination of the coordinates of the intermediate space from (22) by means of (23) gives

$$\left.\begin{aligned} Y_1' &= \frac{Z_1' Y_1}{f_1'} = \frac{Z_1' Y_0'}{f_1'} = \frac{Z_1' f_0 Y_0}{f_1' Z_0} = \frac{f_0 f_1 Y_0}{f_0 f_0' - c Z_0}, \\ Z_1' &= \frac{f_1 f_1'}{Z_1} = \frac{f_1 f_1'}{Z_0' - c} = \frac{f_1 f_1'}{\frac{f_0 f_0'}{Z_0} - c} = \frac{f_1 f_1' Z_0}{f_0 f_0' - c Z_0} \end{aligned}\right\} \tag{24}$$

Let

$$\left.\begin{aligned} Y &= Y_0, \quad Z = Z_0 - \frac{f_0 f_0'}{c}, \\ Y' &= Y_1', \quad Z' = Z_1' + \frac{f_1 f_1'}{c}. \end{aligned}\right\} \tag{25}$$

Equations (25) express a change of coordinates, the origins of the two systems being shifted by distances $f_0' f_0/c$ and $-f_1 f_1'/c$ respectively in the $Z$-direction. In terms of these variables, the equations of the combined transformation become

$$\frac{Y'}{Y} = \frac{f}{Z} = \frac{Z'}{f'}, \tag{26}$$

where

$$f = -\frac{f_0 f_1}{c}, \quad f' = \frac{f_0' f_1'}{c} \tag{27}$$

The distance between the origins of the new and the old systems of coordinates, i.e. the distances $\delta = F_0 F$ and $\delta' = F_1' F'$ of the foci of the equivalent transformation from the foci of the individual transformations are seen from (25) to be

$$\delta = \frac{f_0 f_0'}{c}, \quad \delta' = -\frac{f_1 f_1'}{c}. \tag{28}$$

If $c = 0$, then $f = f' = \infty$ so that the equivalent collineation is telescopic. The equations (24) then reduce to

$$\left.\begin{aligned} Y_1' &= \frac{f_1}{f_0'} Y_0 \\ Z_1' &= \frac{f_1 f_1'}{f_0 f_0'} Z_0; \end{aligned}\right\} \tag{29}$$

the constants $\alpha$ and $\beta$ in (18) of the equivalent transformation, are therefore

$$\alpha = \frac{f_1}{f_0'}, \quad \beta = \frac{f_1 f_1'}{f_0 f_0'} \tag{30}$$

The angular magnification is now

$$\frac{\tan \gamma'}{\tan \gamma} = \frac{\alpha}{\beta} = \frac{f_0}{f_1'}. \tag{31}$$

If one or both of the transformations are telescopic, the above considerations must be somewhat modified.

## 9.5.4 4.4 Gaussian Optics

We shall now study the elementary properties of lenses, mirrors, and their combinations. In this elementary theory only those points and rays will be considered which lie in the immediate neighbourhood of the axis; terms involving squares and higher powers of off-axis distances, or of the angles which the rays make with the axis, will be neglected. The resulting theory is known as Gaussian optics.*

### 4.4.1 Refracting surface of revolution

Consider a pencil of rays incident on a refracting surface of revolution which separates two homogeneous media of refractive indices $n_0$ and $n_1$. To begin with, points and rays in both media will be referred to the same Cartesian reference system, whose origin will be taken at the pole $O$ of the surface, with the $z$-direction along the axis of symmetry.

Let $P_0\,(x_0, y_0, z_0)$ and $P_1\,(x_1, y_1, z_1)$ be points on the incident and on the refracted ray respectively. Neglecting terms of degree higher than first, it follows from §4.1 (29), §4.1 (40), and §4.1 (44) that the coordinates of these points and the components of the two rays are connected by the relations

$$\left.\begin{aligned} x_0 - \tfrac{p_0}{n_0} z_0 &= \tfrac{\partial T^{(2)}}{\partial p_0} = 2\mathscr{A} p_0 + \mathscr{C} p_1, \\ x_1 - \tfrac{p_1}{n_1} z_1 &= -\tfrac{\partial T^{(2)}}{\partial p_1} = -2\mathscr{B} p_1 - \mathscr{C} p_0, \\ y_0 - \tfrac{q_0}{n_0} z_0 &= \tfrac{\partial T^{(2)}}{\partial q_0} = 2\mathscr{A} q_0 + \mathscr{C} q_1, \\ y_1 - \tfrac{q_1}{n_1} z_1 &= -\tfrac{\partial T^{(2)}}{\partial q_1} = -2\mathscr{B} q_1 - \mathscr{C} q_0, \end{aligned}\right\} \tag{1b}$$

where, according to §4.1 (45),

$$\mathscr{A} = \mathscr{B} = \frac{1}{2}\frac{r}{n_1 - n_0}, \quad \mathscr{C} = -\frac{r}{n_1 - n_0} \tag{2}$$

$r$ being the paraxial radius of curvature of the surface.

Let us examine under what conditions all the rays from $P_0$ (which may be assumed to lie in the plane $x = 0$ ) will, after refraction, pass through $P_1$. The coordinates of $P_1$ will then depend only on the coordinates of $P_0$ and not on the components of the rays, so that when $q_1$ is eliminated from (1b), $q_0$ must also disappear.

Now from the first equation (1b)

$$q_1 = \frac{1}{\mathscr{C}}\left\{ y_0 - q_0 \left( 2\mathscr{A} + \frac{1}{n_0} z_0 \right) \right\}, \tag{3}$$

and substituting this into the second equation, we obtain

$$y_1 = -\left( 2\mathscr{B} - \frac{1}{n_1} z_1 \right) \frac{1}{\mathscr{C}} y_0 + \left[ \frac{1}{\mathscr{C}} \left( 2\mathscr{B} - \frac{1}{n_1} z_1 \right) \left( 2\mathscr{A} + \frac{1}{n_0} z_0 \right) - \mathscr{C} \right] q_0. \tag{4}$$

Hence $P_1$ will be a stigmatic image of $P_0$ if

$$\left( 2\mathscr{A} + \frac{1}{n_0} z_0 \right) \left( 2\mathscr{B} - \frac{1}{n_1} z_1 \right) = \mathscr{C}^2 \tag{5}$$

or, on substituting from (2), if

$$\left[ \frac{r}{n_1 - n_0} + \frac{z_0}{n_0} \right] \left[ \frac{r}{n_1 - n_0} - \frac{z_1}{n_1} \right] = \frac{r^2}{(n_1 - n_0)^2} \tag{6}$$

(6) may be written in the form

$$n_0 \left[ \frac{1}{r} - \frac{1}{z_0} \right] = n_1 \left[ \frac{1}{r} - \frac{1}{z_1} \right]. \tag{7}$$

0

- As before, the usual sign convention of analytical geometry (Cartesian sign convention) is used. The various sign conventions employed in practice are very fully discussed in a Report on the Teaching of Geometrical Optics published by the Physical Society (London) in 1934.

It is seen that, within the present degree of approximation, every point gives rise to a stigmatic image; and the distances of the conjugate planes from the pole $O$ of the surface are related by (7). Moreover, equation (4), subject to (5), implies that the imaging is a projective transformation.

The expressions on either side of (7) are known as Abbe's (refraction) invariant and play an important part in the theory of optical imaging. (7) may also be written in the form

$$\frac{n_1}{z_1} - \frac{n_0}{z_0} = \frac{n_1 - n_0}{r}. \tag{8}$$

The quantity $(n_1 - n_0)/r$ is known as the power of the refracting surface and will be denoted by $\mathscr{P}$,

$$\mathscr{P} = \frac{n_1 - n_0}{r} \tag{9}$$

According to (4) and (5) the lateral magnification $y_1/y_0$ is equal to unity when $z_1/n_1 = 2\mathscr{B} + \mathscr{C}$. But by (2), $2\mathscr{B} + \mathscr{C} = 0$. Hence the unit points $U_0$ and $U_1$ are given by $z_0 = z_1 = 0$, i.e. the unit points coincide with the pole of the surface. Further, from (8),

$$z_0 \to -\frac{n_0 r}{n_1 - n_0} \quad \text{as } z_1 \to \infty$$

and

$$z_1 \to \quad \frac{n_1 r}{n_1 - n_0} \quad \text{as } z_0 \to -\infty$$

so that the abscissae of the foci $F_0$ and $F_1$ are $-n_0 r/(n_1 - n_0)$ and $n_1 r/(n_1 - n_0)$ respectively. The focal lengths $f_0 = F_0U_0, f_1 = F_1U_1$ are therefore given by
and

$$\left.\begin{aligned} f_0 &= \tfrac{n_0 r}{n_1 - n_0} \\ f_1 &= -\tfrac{n_1 r}{n_1 - n_0} \end{aligned}\right\} \tag{10}$$

or, in terms of the power $\mathscr{P}$ of the surface,*

$$\frac{n_0}{f_0} = -\frac{n_1}{f_1} = \mathscr{P}. \tag{II}$$

Since $f_0$ and $f_1$ have different signs, the imaging is concurrent (cf. §4.3.3). If the surface is convex towards the incident light ( $r > 0$ ) and if $n_0 < n_1$ then $f_0 > 0, f_1 < 0$, and the imaging is convergent. If $r > 0$ and $n_0 > n_1$ it is divergent (Fig. 4.12). When the surface is concave towards the incident light the situation is reversed.

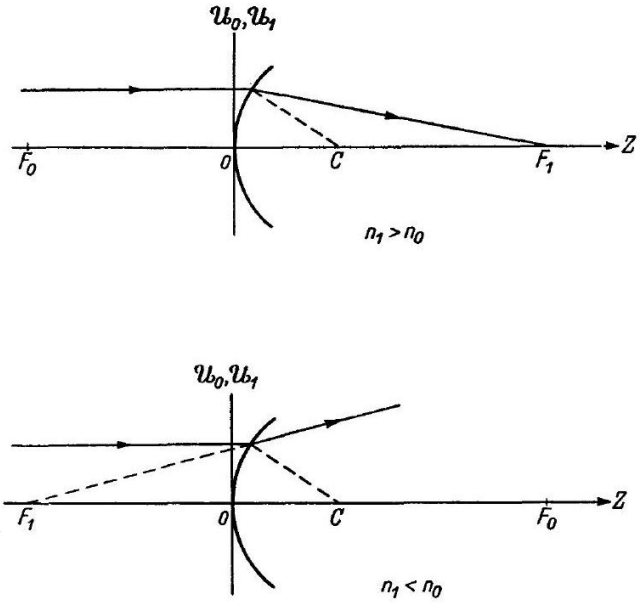



Fig. 4.12. Position of the cardinal points for refraction at a surface of revolution.

Using the expressions (10) for the focal lengths, (8) becomes

$$\frac{f_0}{z_0} + \frac{f_1}{z_1} = -1; \tag{12}$$

and the coefficients (2) may be written in the form

$$\mathscr{A} = \frac{f_0}{2n_0}, \quad \mathscr{B} = -\frac{f_1}{2n_1}, \quad \mathscr{C} = -\frac{f_0}{n_0} = \frac{f_1}{n_1}. \tag{13}$$

Let us introduce separate coordinate systems in the two spaces, with the origins at the foci, and with the axes parallel to those at $O$ :

$$\begin{array}{ll} X_0 = x_0, & X_1 = x_1 \\ Y_0 = y_0, & Y_1 = y_1 \\ Z_0 = z_0 + f_0, & Z_1 = z_1 + f_1. \end{array}$$

Equations (12) and (4), subject to (5), then reduce to the standard form §4.3 (10):

$$\frac{Y_1}{Y_0} = \frac{f_0}{Z_0} = \frac{Z_1}{f_1}. \tag{14}$$

### 4.4.2 Reflecting surface of revolution

It is seen that with a notation strictly analogous to that used in the previous section, equations of the form (la) and (lb) also hold when reflection takes place on a surface of revolution, the coefficients $\mathscr{A}, \mathscr{B}$ and $\mathscr{C}$ now being replaced by the corresponding coefficients $\mathscr{A}', \mathscr{B}'$ and

0

- It will be seen later that the relation $n_0/f_0 = -n_1/f_1$ is not restricted to a single refracting surface, but holds in general for any centred system, the quantities with suffix zero referring to the object space, those with the suffix 1 to the image space. (11) may therefore be regarded as defining the power of a general centred system. The practical unit of power is a dioptre; it is the reciprocal of the focal length, when the focal length is expressed in meters. The power is positive when the system is convergent ( $f_0 > 0$ ) and negative when it is divergent ( $f_0 < 0$ ).

$\mathscr{C}'$ of §4.1.5. It has been shown in §4.1.5 that $\mathscr{A}', \mathscr{B}'$ and $\mathscr{C}'$ can be obtained from $\mathscr{A}, \mathscr{B}$ and $\mathscr{C}$ on setting $n_0 = -n_1 = n, n$ denoting the refractive index of the medium in which the rays are situated. Hence the appropriate

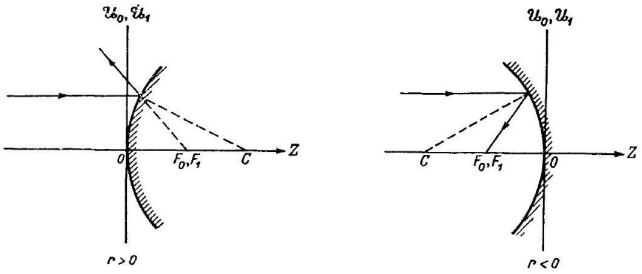



Fig. 4.13. Position of the cardinal points for reflection on a mirror of revolution.

formulae for reflection may be immediately written down by making this substitution in the preceding formulae. In particular, (7) gives

$$\frac{1}{z_0} - \frac{1}{r} = -\frac{1}{z_1} + \frac{1}{r}, \tag{15}$$

the expression on either side of (15) being *Abbe* 's (reflection) invariant. (15) may also be written as

$$\frac{1}{z_0} + \frac{1}{z_1} = \frac{2}{r} \tag{16}$$

The focal lengths $f_0$ and $f_1$ are now given by

$$f_0 = f_1 = -\frac{r}{2}, \tag{17}$$

and the power $\mathscr{P}$ is

$$\mathscr{P} = -\frac{2n}{r}. \tag{18}$$

Since $f_0 f_1 > 0$, the imaging is contracurrent (katoptric). When the surface is convex towards the incident light $(r > 0) f_0 < 0$, and the imaging is then divergent; when it is concave towards the incident light, $(r < 0), f_0 > 0$ and the imaging is convergent (Fig. 4.13).

### 4.4.3 The thick lens

Next we derive the Gaussian formulae relating to imaging by two surfaces which are rotationally symmetrical about the same axis.

Let $n_0, n_1$ and $n_2$ be the refractive indices of the three regions, taken in the order in which light passes through them, and let $r_1$ and $r_2$ be the radii of curvature of the surfaces at their axial points, measured positive when the surface is convex towards the incident light.

By (10), the focal lengths of the first surface are given by

$$f_0 = \frac{n_0 r_1}{n_1 - n_0}, \quad f_0' = -\frac{n_1 r_1}{n_1 - n_0}, \tag{19}$$

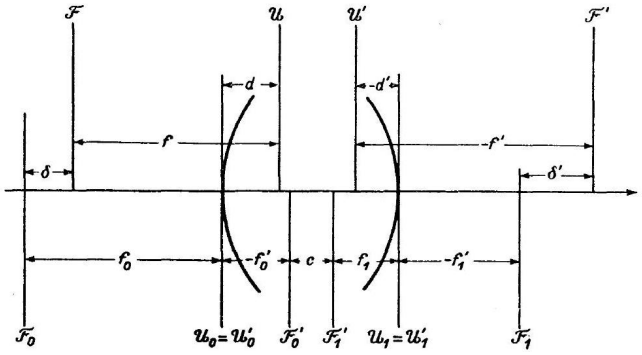



Fig. 4.14. The cardinal points of a combined system (thick lens).

and of the second surface by

$$f_1 = \frac{n_1 r_2}{n_2 - n_1'}, \quad f_1' = -\frac{n_2 r_2}{n_2 - n_1}. \tag{20}$$

According to § 4.3 (27) the focal lengths of the combination are

$$f = -\frac{f_0 f_1}{c}, \quad f' = \frac{f_0' f_1'}{c}, \tag{21}$$

where $c$ is the distance between the foci $F_0'$ and $F_1$. Let $t$ be the axial thickness of the lens, i.e. the distance between the poles of the two surfaces; then (see Fig. 4.14)

$$c = t + f_0' - f_1. \tag{22}$$

On substituting into (22) for $f_0'$ and for $f_1$ we obtain

$$c = \frac{D}{(n_1 - n_0)(n_2 - n_1)}, \tag{23}$$

where

$$D = (n_1 - n_0)(n_2 - n_1)\, t - n_1 \left[(n_2 - n_1)\, r_1 + (n_1 - n_0)\, r_2\right]. \tag{24}$$

The required expressions for the focal lengths of the combination are now obtained on substituting for $f_0, f_1, f_0', f_1'$ and $c$ into (21). This gives

$$f = -n_0 n_1 \frac{r_1 r_2}{D}; \quad f' = n_1 n_2 \frac{r_1 r_2}{D}. \tag{25}$$

Since $ff'$ is negative, the imaging is concurrent. The power $\mathscr{P}$ of the lens is

$$\begin{aligned} \mathscr{P} &= \frac{n_0}{f} = -\frac{n_2}{f'} = -\frac{1}{n_1}\frac{D}{r_1 r_2} \\ &= \mathscr{P}_1 + \mathscr{P}_2 - \frac{t}{n_3}\mathscr{P}_1 \mathscr{P}_2 \end{aligned} \tag{26}$$

where $\mathscr{P}_1$ and $\mathscr{P}_2$ are the powers of the two surfaces.
By §4.3 (28), the distances $\delta = F_0 F$ and $\delta' = F_1' F'$ are seen to be

$$\delta = -n_0 n_1 \frac{n_2 - n_1}{n_1 - n_0}\frac{r_1^2}{D}, \quad \delta' = n_1 n_2 \frac{n_1 - n_0}{n_2 - n_1}\frac{r_2^2}{D}. \tag{27}$$

The distances $d$ and $d'$ of the principal planes $\mathscr{U}$ and $\mathscr{U}'$ from the poles of the surfaces are (see Fig. 4.14)

$$\left.\begin{aligned} d &= \delta + f - f_0 = -n_0\,(n_2 - n_1)\,\tfrac{r_1 t}{D}, \\ d' &= \delta' + f' - f_1' = n_2\,(n_1 - n_0)\,\tfrac{r_2 t}{D}. \end{aligned}\right\} \tag{28}$$

Of particular importance is the case where the media on both sides of the lens are of the same refractive index, i.e. when $n_2 = n_0$. If we set $n_1/n_0 = n_1/n_2 = n$, the formulae then reduce to

$$\left.\begin{aligned} f &= -f' = -\tfrac{n r_1 r_2}{\Delta} \\ \delta &= n\tfrac{r_1^2}{\Delta}, \quad \delta' = -n\tfrac{r_2^2}{\Delta} \\ d &= (n-1)\tfrac{r_1 t}{\Delta}, \quad d' = (n-1)\tfrac{r_2 t}{\Delta}, \end{aligned}\right\} \tag{29}$$

where

$$\Delta = (n-1)\,[n\,(r_1 - r_2) - (n-1)t] \tag{30}$$

Referred to axes at the foci $F$ and $F'$, the abscissae of the unit points are $Z = f$ and $Z' = f'$, and the nodal points are given by $Z = -f', Z = -f$. Since $f = -f'$ the unit points and the nodal points now coincide. The formula § 4.3 (16) which relates the distances $\zeta$ and $\zeta'$ of conjugate planes from the unit planes becomes

$$\frac{1}{\zeta} - \frac{1}{\zeta'} = -\frac{1}{f} \tag{31}$$

The lens is convergent ( $f > 0$ ) or divergent ( $f < 0$ ) according as

$$f = -n\frac{r_1 r_2}{\Delta} \gtrless 0 \tag{32}$$

i.e. according as

$$\frac{1}{r_2} - \frac{1}{r_1} \lessgtr \frac{n-1}{n}\frac{t}{r_1 r_2} \tag{33}$$

When $f = \infty$, we have the intermediate case of telescopic imaging. Then $\Delta = 0$, i.e.

$$r_1 - r_2 = \frac{n-1}{n}t \tag{34}$$

The three cases may be illustrated by considering a double convex lens, $r_1 > 0$, $r_2 < 0$ (cf. Fig. 4.15). If, for simplicity, we assume that both radii are numerically equal, i.e. $r_1 = -r_2 = r$, the imaging will be convergent or divergent according as $t \lessgtr 2nr/(n-1)$ and telescopic when $t = 2nr/(n-1)$.

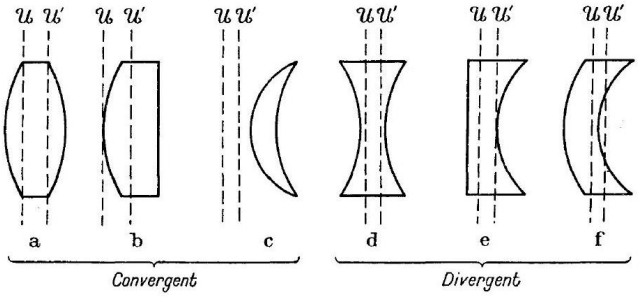



Fig. 4.15. Common types of lenses:

(a) Double-convex; (b) Plano-convex; (c) Convergent meniscus;
(d) Double-concave; (e) Plano-concave; (f) Divergent meniscus.
$\mathscr{U}$ and $\mathscr{U}'$ are the unit planes, light being assumed to be incident from the left.

### 4.4.4 The thin lens

The preceding formulae take a particularly simple form when the lens is so thin that the axial thickness $t$ may be neglected. Then, according to (28), $d = d' = 0$, so that the unit planes pass through the axial point of the (infinitely thin) lens. Consequently, the rays which go through the centre of the lens will not suffer any deviation; this implies that imaging by a thin lens is a central projection from the centre of the lens.

From (26) it follows, on setting $t = 0$, that

$$\mathscr{P} = \mathscr{P}_1 + \mathscr{P}_2 = \frac{n_1 - n_0}{r_1} + \frac{n_2 - n_1}{r_2}, \tag{35}$$

i.e. the power of a thin lens is equal to the sum of the powers of the surfaces forming it. If the media on the two sides of the lens are of equal refractive indices ( $n_0 = n_2$ ), we have from (25)

$$\frac{1}{f} = -\frac{1}{f'} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right), \tag{36}$$

where as before, $n = n_1/n_0 = n_1/n_2$. Assuming that $n > 1$, as is usually the case, $f$ is seen to be positive or negative according as the curvature $1/r_1$ of the first surface is greater or smaller than the curvature $1/r_2$ of the second surface (an appropriate sign being associated with the curvature). This implies that thin lenses whose thickness decreases from centre to the edge are convergent, and those whose thickness increases to the edge are divergent.

For later purposes we shall write down an expression for the focal lengths $f$ and $f'$ of a system consisting of two centred thin lenses, situated in air. According to § 4.3 (27) we have, since $f_0 = -f_0', f_1 = -f_1'$,

$$\frac{1}{f} = -\frac{1}{f'} = -\frac{c}{f_0 f_1}, \tag{37}$$

$c$ being the distance between the foci $F_0'$ and $F_1$ (see Fig. 4.16). If $l$ is the distance between the two lenses, then

$$l = f_0 + c + f_1. \tag{38}$$

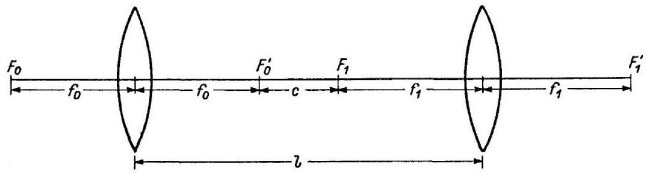



Fig. 4.16. System formed by two centred thin lenses.

Hence

$$\frac{1}{f} = -\frac{1}{f'} = \frac{1}{f_0} + \frac{1}{f_1} - \frac{l}{f_0 f_1}. \tag{39}$$

When the lenses are in contact ( $l = 0$ ), ( 39 ) may also be written in the form, $\mathscr{P} = \mathscr{P}_1 + \mathscr{P}_2$, so that the power of the combination is then simply equal to sum of the powers of the two lenses.

### 4.4.5 The general centred system

Within the approximations of Gaussian theory, a refraction or a reflection at a surface of revolution was seen to give rise to a projective relationship between the object and the image spaces.* Since according to § 4.3 successive applications of projective transformations are equivalent to a single projective transformation, it follows that imaging by a centred system is, to the present degree of approximation, also a transformation of this type. The cardinal points of the equivalent transformation may be found by the application of the formulae of §4.4.1, §4.4.2 and §4.3.4. We shall mainly confine our discussion to the derivation of an important invariant relation valid (within the present degree of accuracy) for any centred system.

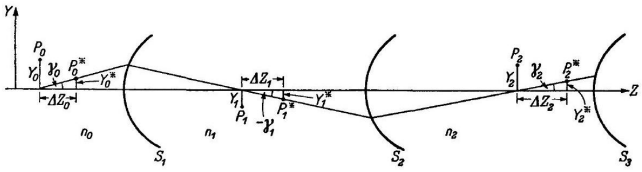


Fig. 4.17. Illustrating the Smith-Helmholtz formula.

Let $S_1, S_2, \ldots S_m$ be the successive surfaces of the system, $f_0, f_1, \ldots f_{m-1}$ the corresponding focal lengths, and $n_0, n_1, \ldots n_m$ the refractive indices of the successive spaces (Fig. 4.17). Further, let $P_0, P_0^\star$ be two points in the object space, situated in a

meridional plane, and let $P_1, P_1^\star, P_2, P_2^\star \ldots$, be their images in the successive surfaces. Referred to axes at the foci of the first surface, the coordinates of $P_0, P_0^\star$ and $P_1, P_1^\star$ are related by eq. (14), viz.

$$\frac{Y_1}{Y_0} = \frac{f_0}{Z_0} = \frac{Z_1}{f_1}, \tag{40}$$

$$\frac{Y_1^\star}{Y_0^\star} = \frac{f_0}{Z_0^\star} = \frac{Z_1^\star}{f_1}. \tag{41}$$

Hence

$$\frac{Z_1^\star - Z_1}{f_1} = \frac{Y_1^\star}{Y_0^\star} - \frac{Y_1}{Y_0} = \frac{Y_0 Y_1^\star - Y_1 Y_0^\star}{Y_0 Y_0^\star}, \tag{42}$$

and

$$\frac{Z_0^\star - Z_0}{f_0} = \frac{Y_0^\star}{Y_1^\star} - \frac{Y_0}{Y_1} = \frac{Y_1 Y_0^\star - Y_0 Y_1^\star}{Y_1 Y_1^\star}. \tag{43}$$

0

- As in the case of a single surface, the object and the image spaces are regarded as superposed on to each other and extending indefinitely in all directions. The part of the object space which lies in front of the first surface (counted in the order in which the light traverses the system) is said to form the real portion of the object space and the portion of the image space which follows the last surface is called the real portion of the image space. The remaining portions of the two spaces are said to be virtual. In a similar manner we may define the real and virtual parts of any of the intermediate spaces of the system.

Let

$$\left.\begin{aligned} Z_0^\star - Z_0 &= \Delta Z_0, \\ Z_1^\star - Z_1 &= \Delta Z_1. \end{aligned}\right\} \tag{44}$$

Then from (42) and (43),

$$\frac{f_0 Y_0 Y_0^\star}{\Delta Z_0} = -\frac{f_1 Y_1 Y_1^\star}{\Delta Z_1}. \tag{45}$$

Now by (10), $f_0/f_1 = -n_0/n_1$, so that this equation may be written as

$$\frac{n_0 Y_0 Y_0^\star}{\Delta Z_0} = \frac{n_1 Y_1 Y_1^\star}{\Delta Z_1}. \tag{46}$$

Similarly, for refraction at the second surface,

$$\frac{n_1 Y_1 Y_1^\star}{\Delta Z_1} = \frac{n_2 Y_2 Y_2^\star}{\Delta Z_2}, \tag{47}$$

and generally

$$\frac{n_{i-1} Y_{i-1} Y_{i-1}^\star}{\Delta Z_{i-1}} = \frac{n_i Y_i Y_i^\star}{\Delta Z_i} \quad (1 \leqslant i \leqslant m). \tag{48}$$

Hence $n_i Y_i Y_i^\star / \Delta Z_i$ is an invariant in the successive transformations. This result plays an important part in the geometrical theory of image formation. If we set $Y_i^\star / \Delta Z_i = \tan \gamma_i$ (see Fig. 4.17), (48) becomes

$$n_{i-1} Y_{i-1} \tan \gamma_{i-1} = n_i Y_i \tan \gamma_i.$$

Since to the present degree of accuracy, $\tan \gamma$ and $\tan \gamma'$ may be replaced by $\gamma$ and $\gamma'$ respectively, we obtain the Smith-Helmholtz formula*

$$n_{i-1} Y_{i-1} \gamma_{i-1} = n_i Y_i \gamma_i. \tag{49}$$

The quantity $n_i Y_i \gamma_i$ is known as the Smith-Helmholtz invariant.
From (48) and (49) a number of important conclusions may be drawn. As one is usually interested only in relations between quantities pertaining to the first and the last medium (object and image space), we shall drop the suffixes and denote quantities which refer to these two spaces by unprimed and primed symbols respectively.

Let $(Y, Z)$ and $(Y + \delta Y, Z + \delta Z)$ be two neighbouring points in the object space and ( $Y', Z'$ ) and ( $Y' + \delta Y', Z' + \delta Z'$ ) the conjugate points. The Smith-Helmholtz formula gives, by successive application, the following relation:

$$\frac{nY(Y + \delta Y)}{\delta Z} = \frac{n'Y'(Y' + \delta Y')}{\delta Z'} \tag{50}$$

In the limit as $\delta Y \to 0$ and $\delta Z \to 0$, this reduces to

$$\frac{dZ'}{dZ} = \frac{n'}{n} \frac{Y'^2}{Y^2} \tag{51}$$

0

- This formula is also associated with the names of Lagrange and Clausius. It was actually known in more restricted forms to earlier writers, e.g. Huygens and Cotes. (Cf. Lord Rayleigh, Phil. Mag., (5) 21 (1886), 466).

According to §4.3 (11), we may write $Y'/Y = (dY'/dY)_{Z=\text{ const.}}$ and (51) becomes

$$\frac{dZ'}{dZ} = \frac{n'}{n}\left(\frac{dY'}{dY}\right)^2_{Z=\text{ const.}} \tag{52}$$

known as Maxwell's elongation formula. It implies that the longitudinal magnification is equal to the square of the lateral magnification multiplied by the ratio $n'/n$ of the refractive indices. Now in § 4.3 we derived an analogous formula [(13)] connecting the magnifications and the ratio of the focal lengths. On comparing these two formulae it is seen that

$$\frac{f'}{f} = -\frac{n'}{n} \tag{53}$$

i.e. the ratio of the focal lengths of the instrument is equal to the ratio $n'/n$ of the refractive indices, taken with a negative sign.

From the Smith-Helmholtz formula it also follows, that

$$\frac{dY'}{dY}\frac{\gamma'}{\gamma} = \frac{n}{n'} \tag{54}$$

showing that the product of the lateral and the angular magnification is independent of the choice of the conjugate planes.

It has been assumed so far that the system consists of refracting surfaces only. If one of the surfaces (say the $i$ th one) is a mirror, we obtain in place of (48),

$$\frac{Y_{i-1}Y^\star_{i-1}}{\Delta Z_{i-1}} = -\frac{Y_iY^\star_i}{\Delta Z_i}$$

the negative sign arising from the fact that for reflection $f_{i-1}/f_i = 1$, whereas for refraction $f_{i-1}/f_i = -n_{i-1}/n_i$. In consequence $n'$ must be replaced by $-n'$ in the final formulae. More generally $n'$ must be replaced by $-n'$ when the system contains an odd number of mirrors; when it contains an even number of mirrors, the final formulae remain unchanged.

## 9.5.5 4.5 Stigmatic Imaging with Wide-Angle Pencils

The laws of Gaussian Optics were derived under the assumption that the size of the object and the angles which the rays make with the axis are sufficiently small. Often one also has to consider systems where the object is of small linear dimensions, but where the inclination of the rays is appreciable. There are two simple criteria, known
as the sine condition* and the Herschel condition † relating to the stigmatic imaging in such instruments.

Let $O_0$ be an axial object point and $P_0$ any point in its neighbourhood, not necessarily on the axis. Assume that the system images these two points stigmatically, and let $O_1$ and $P_1$ be the stigmatic images.

Let ( $x_0, y_0, z_0$ ) and ( $x_1, y_1, z_1$ ) be the coordinates of $P_0$ and $P_1$ respectively, $P_0$ being referred to rectangular axes at $O_0$ and $P_1$ to parallel axes at $O_1$, the $z$-directions being taken along the axis of the system (Fig. 4.18). By the principle of equal optical path, the path lengths of all the rays joining $P_0$ and $P_1$ are the same. Hence, if $V$ denotes the point characteristic of the medium,

$$V\,(x_0, y_0, z_0; x_1, y_1, z_1) - V(0, 0, 0; 0, 0, 0) = F\,(x_0, y_0, z_0; x_1, y_1, z_1)\,, \tag{1}$$

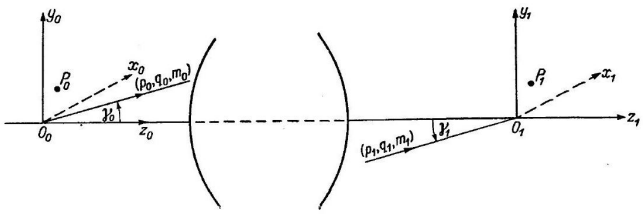



Fig. 4.18. Illustrating the sine condition and the Herschel condition.

$F$ being some function which is independent of the ray components. Using the basic relations §4.1 (7) which express the ray component in terms of the point characteristic, we have from (1), if terms above the first power in distances are neglected,

$$\left(p_1{}^{(0)}x_1 + q_1{}^{(0)}y_1 + m_1{}^{(0)}z_1\right) - \left(p_0{}^{(0)}x_0 + q_0{}^{(0)}y_0 + m_0{}^{(0)}z_0\right) = F\left(x_0, y_0, z_0; x_1, y_1, z_1\right), \qquad (2)$$

$( p_0{}^{(0)}, q_0{}^{(0)}, m_0{}^{(0)} )$ and $( p_1{}^{(0)}, q_1{}^{(0)}, m_1{}^{(0)} )$ being the ray components of any pair of corresponding rays through $O_0$ and $O_1$. It is to be noted that, although the points $P_0$ and $P_1$ are assumed to be in the neighbourhood of $O_0$ and $O_1$, no restriction as to the magnitude of the ray components is imposed.
Two cases are of particular interest, namely when (i) $P_0$ and $P_1$ lie in the planes $z_0 = 0$ and $z_1 = 0$ respectively, and (ii) when $P_0$ and $P_1$ are on the axis of symmetry. The two cases will be considered separately.

### 4.5.1 The sine condition

Without loss of generality we may again consider only points in a meridional plane ($x_0 = x_1 = 0$). If $P_0$ lies in the plane $z_0 = 0$ and $P_1$ in the plane $z_1 = 0$, (2) becomes

$$q_1^{(0)}y_1 - q_0^{(0)}y_0 = F\left(0, y_0, 0; 0, y_1, 0\right). \qquad (3)$$

This relation holds for each pair of conjugate rays. In particular it must, therefore, hold for the axial pair $p_0{}^{(0)} = q_0{}^{(0)} = 0, p_1{}^{(0)} = q_1{}^{(0)} = 0$. Hence

$$F\left(0, y_0, 0; 0, y_1, 0\right) = 0. \qquad (4)$$

Relation (3) becomes

$$q_1{}^{(0)}y_1 = q_0{}^{(0)}y_0, \qquad (5)$$

or, more explicitly,

$$n_1 y_1 \sin\gamma_1 = n_0 y_0 \sin\gamma_0, \qquad (6)$$

$\gamma_0$ and $\gamma_1$ being the angles which the corresponding rays through $O_0$ and $O_1$ make with the $z$-axis, and $n_0$ and $n_1$ being the refractive indices of the object and image spaces. (6) is known

0

- The sine condition was first derived by R. Clausius (Pogg. Ann., 121 (1864), 1) and by H. Helmholtz (Pogg. Ann. Jubelband (1874), 557) from thermodynamical considerations. Its importance was, however, not recognized until it was rediscovered by E. Abbe (Jenaisch. Ges. Med. Naturw. 1879), 129, also Carl. Repert. Phys., 16 (1880), 303).

The derivations given here are essentially due to C. Hockin, J. Roy. Micro. Soc. (2) 4 (1884), 337. † J. F. W. Herschel, Phil. Trans. Roy. Soc., 111 (1821), 226.

as the sine condition, and is the required condition under which a small region of the object plane in the neighbourhood of the axis is imaged sharply by a pencil of any angular divergence. If the angular divergence is sufficiently small, $\sin\gamma_0$ and $\sin\gamma_1$ may be replaced by $\gamma_0$ and $\gamma_1$ respectively, and the sine condition reduces to the Smith-Helmholtz formula, §4.4 (49).

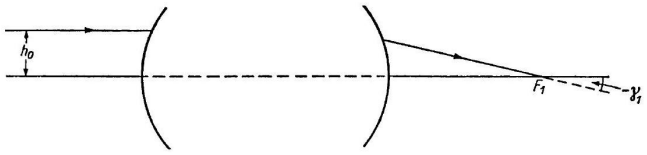



Fig. 4.19. The sine condition, when the object is at infinity.

If the object lies at infinity, the sine condition takes a different form. Assume first that the axial object point is at a great distance from the first surface. If $Z_0$ is the abscissa of this point referred now to axes at the first focus, and $h_0$ is the height above the axis at which a ray from the axial point meets the first surface, then $\sin\gamma_0 \sim -h_0/Z_0$; more precisely $Z_0 \sin\gamma_0/h_0 \to -1$ as $Z_0 \to -\infty$ whilst $h_0$ is kept fixed. Hence, if $Z_0$ is large enough, (6) may be written as

$$\frac{n_0}{n_1}h_0 = -\frac{y_1}{y_0}Z_0 \sin\gamma_1. \tag{7}$$

But by §4.3(10), $y_1 Z_0/y_0 = f_0$, and by §4.4(53) $n_0/n_1 = -f_0/f_1$, so that in the limit (6) reduces to (see Fig. 4.19)

$$\frac{h_0}{\sin\gamma_1} = f_1. \tag{8}$$

This implies that each ray which is incident in the direction parallel to the axis intersects its conjugate ray on a sphere of radius $f_1$, which is centred at the focus $F_1$.

Axial points which are stigmatic images of each other, and which, in addition, have the property that conjugate rays which pass through them satisfy the sine condition, are said to form an aplanatic pair. We have already encountered such point pairs when studying the refraction at a spherical surface (§ 4.2.3).

In the terminology of the theory of aberrations (cf. Chapter V), axial stigmatism implies the absence of all those terms in the expansion of the characteristic function which do not depend on the off-axis distance of the object, i.e. it implies the absence of spherical aberration of all orders. If, in addition, the sine condition is satisfied, then all terms in the characteristic function which depend on the first power of the
off-axis distance must also vanish; these terms represent aberrations known as circular coma.

Since the sine condition gives information about the quality of the off-axis image in terms of the properties of axial pencils it is of great importance for optical design.

### 4.5.2 The Herschel condition

Next consider the case when $P_0$ and $P_1$ lie on the axis of the system ( $x_0 = y_0 = 0$, $x_1 = y_1 = 0$ ). The condition (2) for sharp imaging reduces to

$$m_1{}^{(0)} z_1 - m_0{}^{(0)} z_0 = F\,(0, 0, z_0; 0, 0, z_1)\,, \tag{9}$$

or, in terms of $\gamma_0$ and $\gamma_1$,

$$n_1 z_1 \cos\gamma_1 - n_0 z_0 \cos\gamma_0 = F\,(0, 0, z_0; 0, 0, z_1)\,. \tag{10}$$

In particular for the axial ray this gives

$$F(0,0,z_0;0,0,z_1) = n_1 z_1 - n_0 z_0. \tag{11}$$

Hence (10) may be written as

$$n_1 z_1 \sin^2(\gamma_1/2) = n_0 z_0 \sin^2(\gamma_0/2) \tag{12}$$

This is one form of the Herschel condition. Since the distances from the origin are assumed to be small, we have, by §4.4(52),

$$\frac{z_1}{z_0} = \frac{n_1}{n_0}\left(\frac{y_1}{y_0}\right)^2 \tag{13}$$

so that the Herschel condition may also be written in the form

$$n_1 y_1 \sin(\gamma_1/2) = n_0 y_0 \sin(\gamma_0/2) \tag{14}$$

When the Hersohel condition is satisfied, an element of the axis in the immediate neighbourhood of $O_0$ will be imaged sharply by a pencil of rays, irrespective of the angular divergence of the pencil.

It is to be noted that the sine condition and the Herschel condition cannot hold simultaneously unless $\gamma_1 = \gamma_0$. Then $y_1/y_0 = z_1/z_0 = n_0/n_1$, i.e. the longitudinal and lateral magnifications must then be equal to the ratio of the refractive indices of the object and image space.

## 9.5.6 4.6 Astigmatic Pencils of Rays

Rectilinear rays which have a point in common are said to form a homocentric pencil. The associated wave-fronts are then spherical, centred on their common point of intersection. It was with such pencils that we were concerned in the preceding sections.

In general, the homocentricity of a pencil is destroyed on refraction or reflection. It will therefore be useful to study the properties of more general pencils of rectilinear rays.

### 4.6.1 Focal properties of a thin pencil

Let $S$ denote one of the orthogonal trajectories (wave-fronts) of a pencil of rectilinear rays, and let $P$ be any point on it (Fig. 4.20). We take a plane through the ray at $P$ and denote by $C$ the curve in which this plane intersects $S$. Since the rays of the
pencil are all normal to $S$, the centre of the circle of curvature at $P$ will be on the ray through $P$.

If now the plane is gradually rotated around the ray, the curve $C$ and consequently the radius of curvature will change continuously. When the plane has undergone a rotation of 180°, the radius of curvature will have passed through its maximum and minimum values. It can be shown by elementary geometry* that the two planes which contain the shortest and the longest radius of curvature are perpendicular to each other. These two planes are known as the principal planes † at $P$ and the corresponding radii are called principal radii of curvature. The curves on $S$ which have the property that at each point they are tangential to the principal planes form two mutually orthogonal families of curves, called the lines of curvature.

In general two normals at adjacent points of a surface do not intersect, to first order. But if they are drawn from adjacent points on a line of curvature they will intersect to this order, and their point of intersection is a focus of the congruence formed by

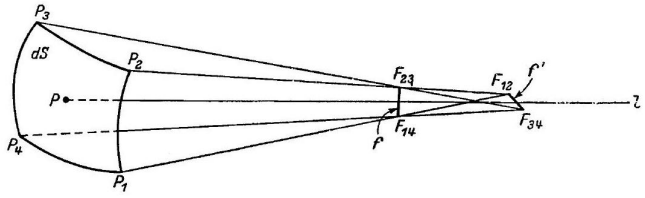



Fig. 4.20. A thin pencil of rays.

the normals (rays). Hence, in agreement with the general conclusions of §3.2.3 there are two foci on each normal, these being the two principal centres of curvature. The caustic surface of a pencil of rectilinear rays has therefore, in general, two branches and is the evolute of the wave-fronts; conversely the wave-fronts are the involutes of the caustic surface. If the wave-fronts are surfaces of revolution, one branch of the caustic surface degenerates into a segment of the axis of revolution; the other is a surface of revolution whose meridional section is the evolute of the meridional section of the wave-front.

Let us consider a thin pencil consisting of all rays which intersect an element $dS$ of the wave-front. It will be convenient to take as boundary of $dS$ two pairs of lines of curvature, which may be assumed to be arcs of circles. Two of them ( $P_1P_2$ and $P_4P_3$ ) may be taken to be vertical, and the other two ( $P_1P_4$ and $P_2P_3$ ) to be horizontal (Fig. 4.20).

To the first order in small quantities, all the rays which pass through the arc $P_1P_2$ will intersect in a focus $F_{12}$, and those through $P_3P_4$ will intersect at $F_{34}$. The line $f'$ joining $F_{12}$ and $F_{34}$ is known as a focal line of the pencil and is seen to be horizontal. Similarly the rays through $P_1P_4$ and $P_2P_3$ will give rise to a vertical focal line $f$.

If the lines of curvature are drawn through any point on $dS$, the corresponding two foci will lie on the two focal lines, and conversely. Hence an approximate model of a thin pencil of rays is obtained by joining all pairs of points on two mutually orthogonal elements of lines.

The ray $l$ through the centre point $P$ is called the central (or principal) ray of the pencil, and the distance between the focal lines, measured along this ray, is called the astigmatic focal distance of the pencil. The two planes specified by $f$ and $l$, and by $f'$ and $l$, are known as the focal planes of the pencil, and are mutually perpendicular. It is not, however, necessarily true (as is often incorrectly asserted in the literature) that the focal lines are perpendicular to the central ray. Consider for example a family of wave-fronts which have cylindrical symmetry about a common axis (Fig. 4.21). Let $dS$ be a surface element (assumed not to contain an axial point) of one of the wave-fronts, and let $ds$ be the curve of intersection of $dS$ with a plane which contains the axis. Then, clearly, the focal line $f$ at the centre of curvature $K$ of $ds$ is perpendicular to this plane. The other focal line, $f'$, coincides with a portion of the axis, namely with that portion which is bounded by the normals at the end points of ds. In general this focal line is not perpendicular to $l$.

0

- See, for example, C. E. Weatherburn, Differential Geometry of Three Dimensions (Cambridge University Press, 1927), p. 185.
  † The terms principal planes and focal planes have now a different meaning than in connection with projective transformations discussed in §4.3.

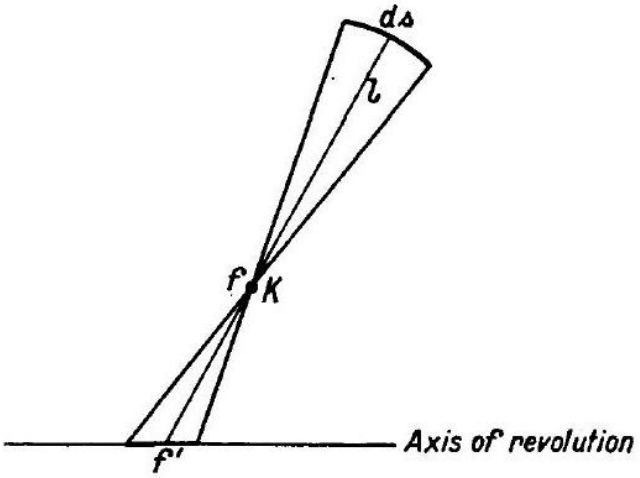



Fig. 4.21. Focal lines of a wave-front with cylindrical symmetry.

### 4.6.2 Refraction of a thin pencil

It was seen that a thin pencil of rays is completely specified by its central ray and its two focal lines. Suppose that a pencil specified in this way is incident on a refracting surface. It is of importance to determine the central ray and the focal lines of the refracted pencil. We shall consider the case of particular importance in practice, namely, when one of the principal planes of the incident pencil coincides with a

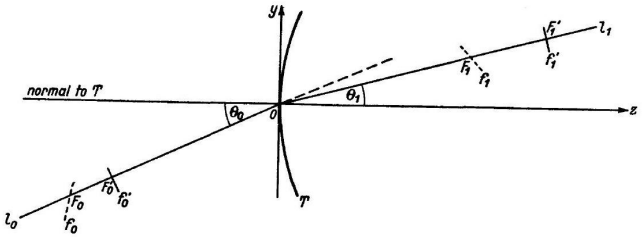



Fig. 4.22. Refraction of a thin astigmatic pencil.

principal plane of curvature of the surface at the point $O$ at which the central ray meets it (Fig. 4.22).

Take Cartesian axes at $O$, with $Oz$ along the normal to the surface $T$ and with $Ox$ and $Oy$ in the direction of the principal lines of curvature of $T$.

Further, let $\theta_0$ and $\theta_1$ be the angles which the central rays $l_0$ and $l_1$ of the two pencils make with $Oz$. Let $F_0$ and $F_0'$ be the foci of the incident pencil, situated at $z = \zeta_0$ and $z = \zeta_0'$ respectively. The focal line at $F_0$ will be assumed to be perpendicular to the plane of incidence; $F_0$ is then said to be a primary focus, and the corresponding focal line $f_0$ the primary focal line. The focus $F_0'$ is called the secondary focus ; the corresponding focal line $f_0'$ (which lies in the plane of incidence) is known as the secondary focal line. In the case of a centred system, the primary and secondary
focus of a pencil whose central ray lies in the meridional plane are known as the tangential and the sagittal focus respectively.

To find the focal lines of the refracted pencil it is necessary to write down first an expression for the angle characteristic of the refracting surface. If the radii of curvature of the surface in the principal directions $Ox$ and $Oy$ are $r_x$ and $r_y$ respectively, the equation of the surface is

$$z = \frac{x^2}{2r_x} + \frac{y^2}{2r_y} + \dots \tag{1}$$

Now according to §4.1 (34), the angle characteristic referred to a set of axes at $O$ is (taking $a_0 = a_1 = 0$ )

$$T = (p_0 - p_1)\, x + (q_0 - q_1)\, y + (m_0 - m_1)\, z. \tag{2}$$

From the law of refraction it follows, in a way similar to that described in §4.1 (which corresponds to the case $r_x = r_y$ ), when terms of the lowest degree only are retained, that

$$\left.\begin{aligned} x &= -r_x \tfrac{p_0 - p_1}{m_0 - m_1} \\ y &= -r_y \tfrac{q_0 - q_1}{m_0 - m_1} \end{aligned}\right\} \tag{3}$$

Substitution into (1) gives

$$z = \frac{1}{2\mu^2}\left[r_x\,(p_0 - p_1)^2 + r_y\,(q_0 - q_1)^2\right] \tag{4}$$

with

$$\mu = m_1 - m_0 = n_1 \cos\theta_1 - n_0 \cos\theta_0. \tag{5}$$

Substitution from (3) and (4) then leads to the required expression for the angle characteristic:

$$T\,(p_0, q_0; p_1, q_1) = \frac{1}{2\mu}\left[r_x\,(p_0 - p_1)^2 + r_y\,(q_0 - q_1)^2\right] + \dots \tag{6}$$

On using this expression in §4.1 (29), the equations of the incident and refracted ray are obtained:

$$x_0 - \frac{p_0}{m_0} z_0 = \frac{1}{\mu} r_x\,(p_0 - p_1)\,, \tag{7}$$

$$x_1 - \frac{p_1}{m_1} z_1 = \frac{1}{\mu} r_x\,(p_0 - p_1)\,, \tag{8}$$

the corresponding equations involving the $y$ coordinates being strictly analogous.
Consider now the changes in the various quantities as we pass from the central ray to a neighbouring ray. From (7) and (8)

$$\delta x_0 - \frac{p_0}{m_0}\delta z_0 = z_0 \delta\left(\frac{p_0}{m_0}\right) + r_x\left\{\delta\left(\frac{p_0}{\mu}\right) - \delta\left(\frac{p_1}{\mu}\right)\right\}, \tag{9}$$

$$\delta x_1 - \frac{p_1}{m_1}\delta z_1 = z_1 \delta\left(\frac{p_1}{m_1}\right) + r_x\left\{\delta\left(\frac{p_0}{\mu}\right) - \delta\left(\frac{p_1}{\mu}\right)\right\} \tag{10}$$

Now the components of the central ray of the incident pencil are

$$p_0 = 0, \quad q_0 = n_0 \sin\theta_0, \quad m_0 = n_0 \cos\theta_0 \tag{11}$$

so that

$$\delta\left(\frac{p_0}{m_0}\right) = \frac{1}{m_0^2}\left(m_0\delta p_0 - p_0\delta m_0\right)$$
$$= \frac{1}{n_0}\sec\theta_0\delta p_0 \tag{12}$$

and

$$\delta\left(\frac{q_0}{m_0}\right) = \frac{1}{m_0^2}\left(m_0\delta q_0 - q_0\delta m_0\right)$$
$$= \frac{1}{n_0}\sec^3\theta_0\delta q_0 \tag{13}$$

Here use was made of the identity $m_0\delta m_0 + p_0\delta p_0 + q_0\delta q_0 = 0$.
Equations (9) and (10) become

$$\delta x_0 = \frac{z_0}{n_0}\sec\theta_0\delta p_0 - \frac{1}{\mu}r_x\left(\delta p_1 - \delta p_0\right) \tag{14}$$
$$\delta x_1 = \frac{z_1}{n_1}\sec\theta_1\delta p_1 - \frac{1}{\mu}r_x\left(\delta p_1 - \delta p_0\right) \tag{15}$$

In deriving (15) use was also made of the fact that $p_1 = 0$; this result follows from the law of refraction and the assumption that $p_0 = 0$. In a similar way we find

$$\delta y_0 - \left(\tan\theta_0\right)\delta z_0 = \frac{z_0}{n_0}\left(\sec^3\theta_0\right)\delta q_0 - r_y\left\{\delta\left(\frac{q_1}{\mu}\right) - \delta\left(\frac{q_0}{\mu}\right)\right\}, \tag{16}$$
$$\delta y_1 - \left(\tan\theta_1\right)\delta z_1 = \frac{z_1}{n_1}\left(\sec^3\theta_1\right)\delta q_1 - r_y\left\{\delta\left(\frac{q_1}{\mu}\right) - \delta\left(\frac{q_0}{\mu}\right)\right\}. \tag{17}$$

Consider now those rays of the pencil which pass through the focus $F_0$. Then $z_0 = \zeta_0$, $\delta x_0 = \delta y_0 = \delta z_0 = 0$. Since all these rays also intersect the focal line $f_0', \delta p_0 = 0$. With this substitution (14) and (16) give

$$\delta p_1 = 0 \tag{14a}$$

and

$$\frac{\zeta_0}{n_0}\sec^3\theta_0\delta q_0 - \frac{1}{\mu}r_y\left(\delta q_1 - \delta q_0\right) = 0 \tag{16a}$$

(14a) shows that the corresponding refracted rays lie in the $yz$-plane. Now since all the rays from $F_0$ pass through the focus $F_1\,(z_1 = \zeta_1)$, (17) must hold with $z_1 = \zeta_1$, $\delta x_1 = \delta y_1 = \delta z_1 = 0$, whatever the value of $\delta q_0$, so that

$$\frac{\zeta_1}{n_1}\sec^3\theta_1\delta q_1 - \frac{1}{\mu}r_y\left(\delta q_1 - \delta q_0\right) = 0 \tag{17a}$$

Equations (16a) and (17a) can be satisfied simultaneously for an arbitrary value of $\delta q_0$ only if

$$\frac{n_0\cos^3\theta_0}{\zeta_0} - \frac{n_1\cos^3\theta_1}{\zeta_1} = \frac{n_0\cos\theta_0 - n_1\cos\theta_1}{r_y} \tag{18}$$

This relation gives the position of the focus $F_1$ of the refracted rays. From (14a) it is seen that the focal line through $F_1$ is perpendicular to the $yz$-plane so that $F_1$ is a primary focus.

To find the position of the other focus, consider the rays which proceed from $F_0'$. Then $z_0 = \zeta_0', \delta x_0 = \delta y_0 = \delta z_0 = 0$. Since all these rays intersect the focal line $f_0$, $\delta q_0 = \delta m_0 = 0$. Equations (14) and (16) now give

$$\frac{\zeta_0'}{n_0} \sec\theta_0 \delta p_0 - \frac{1}{\mu} r_\alpha (\delta p_1 - \delta p_0) = 0, \tag{14b}$$

and

$$\delta q_1 = 0. \tag{16b}$$

(16b) shows that the refracted rays now lie in the $xz$-plane. All these rays will pass through the other focus $F_1'$ $(z_1 = \zeta_1')$, so that (15) must be satisfied with $z_1 = \zeta_1'$, $\delta x_1 = \delta y_1 = \delta z_1 = 0$, whatever the value of $\delta p_0$. Hence,

$$\frac{\zeta_1'}{n_1} \sec\theta_1 \delta p_1 - \frac{1}{\mu} r_x (\delta p_1 - \delta p_0) = 0. \tag{15b}$$

Since (15b) and (14b) hold simultaneously for any arbitrary value of $\delta p_0$, it follows that

$$\frac{n_0 \cos\theta_0}{\zeta_0'} - \frac{n_1 \cos\theta_1}{\zeta_1'} = \frac{n_0 \cos\theta_0 - n_1 \cos\theta_1}{r_x}. \tag{19}$$

This relation gives the position of the secondary focus $F_1'$.
It is often convenient to specify the position of the foci by means of their distances from $O$ rather than by means of their $z$ coordinates. If $OF_0 = {d_0}^{(t)}, OF_0' = {d_0}^{(s)}$, $OF_1 = {d_1}^{(t)}, OF_1' = {d_1}^{(s)}$ (in Fig. 4.22${d_0}^{(t)} < 0, {d_0}^{(s)} < 0, {d_1}^{(t)} > 0, {d_1}^{(s)} > 0$ ), then

$$\left.\begin{array}{ll} \zeta_0 = {d_0}^{(t)} \cos\theta_0, & \zeta_1 = {d_1}^{(t)} \cos\theta_1, \\ \zeta_0' = {d_0}^{(s)} \cos\theta_0, & \zeta_1' = {d_1}^{(s)} \cos\theta_1, \end{array}\right\} \tag{20}$$

and the two relations (18) and (19) become

$$\frac{n_0 \cos^2\theta_0}{d_0^{(t)}} - \frac{n_1 \cos^2\theta_1}{d_1^{(t)}} = \frac{n_0 \cos\theta_0 - n_1 \cos\theta_1}{r_y}, \tag{21}$$

and

$$\frac{n_0}{d_0^{(s)}} - \frac{n_1}{d_1^{(s)}} = \frac{n_0 \cos\theta_0 - n_1 \cos\theta_1}{r_x}. \tag{22}$$

The corresponding relations for reflection may be obtained by setting $n_1 = -n_0$.

## 9.5.7 4.7 Chromatic Aberration. Dispersion by a Prism

In Chapter II it was shown that the refractive index is not a material constant but depends on colour, i.e. on the wavelength of light. We shall now discuss some elementary consequences of this result in relation to the performance of lenses and prisms.

### 4.7.1 Chromatic aberration

If a ray of polychromatic light is incident upon a refracting surface, it is split into a set of rays, each of which is associated with a different wavelength. In traversing an optical system, light of different wavelengths will therefore, after the first refraction, follow slightly different paths. In consequence, the image will not be sharp and the system is said to suffer from chromatic aberration.

We shall again confine our attention to points and rays in the immediate neighbourhood of the axis, i.e. it will be assumed that the imaging in each wavelength obeys the laws of Gaussian optics. The chromatic aberration is then said to be of the first order, or primary. If $Q_\alpha$ and $Q_\beta$ are the images of a point $P$ in two different wavelengths (Fig. 4.23), the projections of $Q_\alpha Q_\beta$ in the directions parallel and perpendicular to the axis are known as longitudinal and lateral chromatic aberration respectively.

Consider the change $\delta f$ in the focal length of a thin lens, due to a change $\delta n$ in the refractive index. According to §4.4(36) the quantity $(\ n-1\ )\ f$ will, for a given lens, be independent of the wavelength. Hence

$$\frac{\delta f}{f} + \frac{\delta n}{n-1} = 0. \tag{1}$$

The quantity

$$\Delta = \frac{n_F - n_C}{n_D - 1} \tag{2}$$

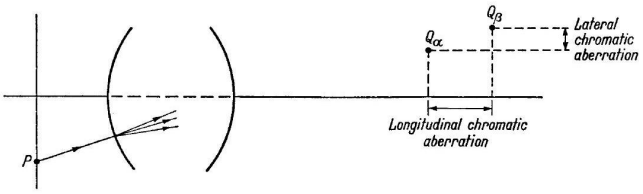



Fig. 4.23. The longitudinal and lateral chromatic aberration.

where $n_F, n_D$ and $n_C$ are the refractive indices for the Fraunhofer $F, D$ and $C$ lines ( $\lambda =$ 4861Å, 5893Å, 6563Å respectively) is a rough measure of the dispersive properties of the glass, and is called the dispersive power. From (1) it is seen that it is approximately equal to the distance between the red and blue image divided by the focal length of the lens, when the object is at infinity. The variation with wavelength, of the refractive index of the usual types of glass employed in optical systems is shown in Fig. 4.24. The corresponding values of $\Delta$ lie between about 1/60 and 1/30.

To obtain an image of good quality, the monochromatic as well as the chromatic aberrations must be small. Usually a compromise has to be made, since in general it is impossible to eliminate all the aberrations simultaneously. Often it is sufficient to eliminate the chromatic aberration for two selected wavelengths only. The choice of these wavelengths will naturally depend on the purpose for which the system is designed; for example, since the ordinary photographic plate is more sensitive to the blue region than is the human eye, photographic objectives are usually "achromatized" for colours nearer to the blue end of the spectrum than is the case in visual instruments. Achromatization with respect to two wavelengths does, of course, not secure a complete removal of the colour error. The remaining chromatic aberration is known as the secondary spectrum.

Let us now examine under what conditions two thin lenses will form an achromatic combination with respect to their focal lengths. According to §4.4 (39) the reciprocal of the focal length of a combination of two thin lenses separated by a distance $l$ is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{l}{f_1 f_2}. \tag{3}$$

It is seen that $\delta f = 0$ when

$$\frac{\delta f_1}{f_1^2} + \frac{\delta f_2}{f_2^2} - \frac{l}{f_1 f_2}\left[\frac{\delta f_1}{f_1} + \frac{\delta f_2}{f_2}\right] = 0. \tag{4}$$

If the achromatization is made for the $C$ and $F$ lines, we have, using (1) and (2),

$$l = \frac{\Delta_1 f_2 + \Delta_2 f_1}{\Delta_1 + \Delta_2}, \tag{5}$$

where $\Delta_1$ and $\Delta_2$ are the dispersive powers of the two lenses.

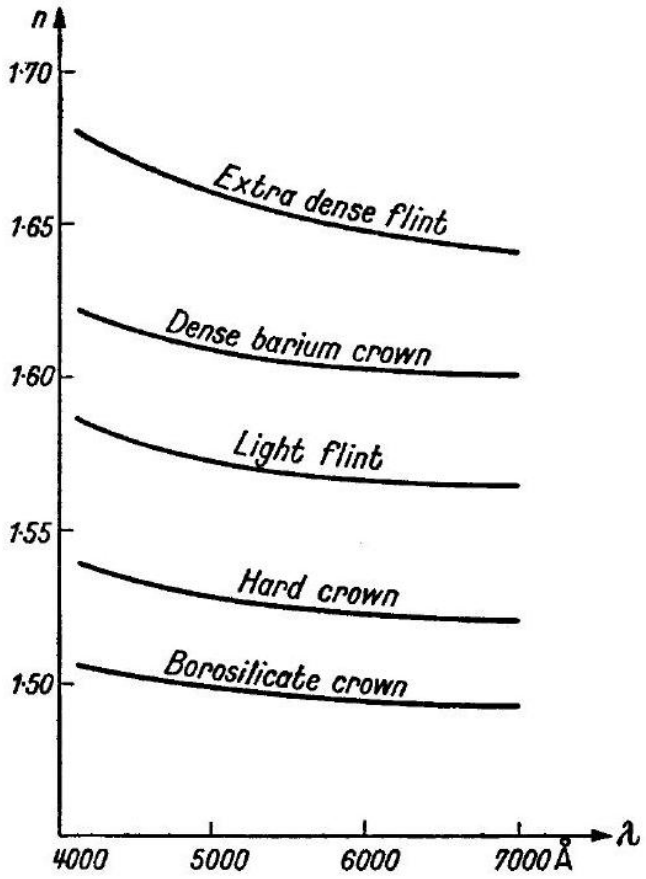



Fig. 4.24. Typical dispersion curves of various types of glass.

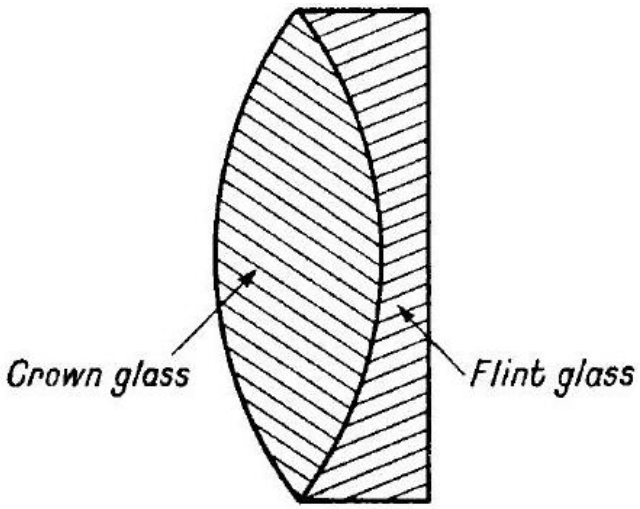



Fig. 4.25. An achromatic doublet.

One method of reducing the chromatic aberration is to employ two thin lenses in contact (Fig. 4.25), one made of crown glass, and the other of flint glass. In this case, since $l = 0$, we have, according to (5),

$$\frac{\Delta_1}{f_1} + \frac{\Delta_2}{f_2} = 0, \tag{6}$$

or, using (3),

$$\frac{1}{f_1} = \frac{1}{f}\frac{\Delta_2}{\Delta_2 - \Delta_1}, \quad \frac{1}{f_2} = -\frac{1}{f}\frac{\Delta_1}{\Delta_2 - \Delta_1}. \tag{7}$$

Now for given glass, and with a fixed value of the focal length $f$, (7) specifies $f_1$ and $f_2$ uniquely. But $f_1$ and $f_2$ depend on three radii of curvature; hence one of the radii may be chosen arbitrarily. This degree of freedom is sometimes used to make the spherical aberration as small as possible.

Another method of obtaining an achromatic system is to employ two thin lenses made of the same glass ( $\Delta_1 = \Delta_2$ ) and separated by a distance equal to half of the sum of their focal lengths

$$l = \frac{1}{2}(f_1 + f_2). \tag{8}$$

That such a combination is achromatic follows immediately from (5).

An instrument consisting of several components cannot, in general, be made achromatic with respect to both position and magnification, unless each component is itself achromatic in this sense. We shall prove this for the case of two centred thin lenses separated by a distance $l$. Since according to § 4.4.4 the imaging by a thin lens is a central projection from the centre of the lens, we have (see Fig. 4.26)

$$\frac{Y_1'}{Y_1} = -\frac{\zeta_1'}{\zeta_1}, \quad \frac{Y_2'}{Y_2} = -\frac{\zeta_2'}{\zeta_2}. \tag{9}$$

Since $Y_2 = Y_1'$, the magnification is

$$\frac{Y_2'}{Y_1} = \frac{\zeta_1'}{\zeta_1} \cdot \frac{\zeta_2'}{\zeta_2}. \tag{10}$$

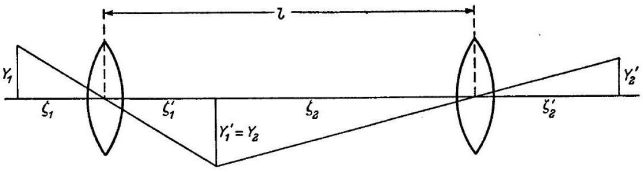



Fig. 4.26. Achromatization of two thin lenses.

If now the wavelength is altered, $\zeta_1$ will remain unchanged, and if the position of the image is assumed to be achromatized, $\zeta_2'$ will also remain unchanged. Hence the condition for achromatization of the magnification may be expressed by the formula

$$\delta\left(\frac{\zeta_1'}{\zeta_2}\right) = \frac{1}{\zeta_2^2}\left(\zeta_2\delta\zeta_1' - \zeta_1'\delta\zeta_2\right) = 0. \tag{11}$$

Since $\zeta_1' + \zeta_2 = l, \delta\zeta_1' = -\delta\zeta_2$, and it is seen that (11) can only be satisfied if $\delta\zeta_1' = \delta\zeta_2 = 0$, i.e. if each of the lenses is achromatized.

So far we have only considered the primary chromatic aberration of a thin lens and of a combination of two such lenses. Expressions for the primary chromatic aberration of a general centred system will be derived in Chapter V.

### 4.7.2 Dispersion by a prism

We shall now briefly discuss the passage of light through a prism.
Let $\alpha$ be the angle between the two faces of the prism. It is assumed that the edge $A$ in which the two faces meet is perpendicular to the plane which contains the incident, transmitted, and emergent rays (Fig. 4.27). To begin with the light will be assumed to be strictly monochromatic.

Let $B_1$ and $B_2$ be the points of intersection of the incident and the emergent ray with the two faces, $\phi_1$ and $\psi_1$ the angles of incidence and refraction at $B_1$, and $\psi_2$ and $\phi_2$ the inner and outer angles at $B_2$ (i.e. the angles which the ray $B_1B_2$ and the emergent ray make with the normal at $B_2$ ). Further let $C$ be the point of intersection of the normals to the prism at $B_1$ and $B_2$, and $D$ the point of intersection of the incident and the emergent rays, when these are prolonged sufficiently far.

If $\varepsilon$ is the angle of deviation, i.e. the angle which the emergent ray makes with the incident ray, then

$$\phi_1 + \phi_2 = \varepsilon + \alpha \tag{12}$$
$$\psi_1 + \psi_2 = \alpha \tag{13}$$

Further, by the law of refraction,

$$\left.\begin{aligned}\sin\phi_1 &= n\sin\psi_1,\\ \sin\phi_2 &= n\sin\psi_2,\end{aligned}\right\} \tag{14}$$

where $n$ is the refractive index of the glass with respect to the surrounding air. The deviation $\varepsilon$ will have an extremum when

$$\frac{d\varepsilon}{d\phi_1} = 0. \tag{15}$$

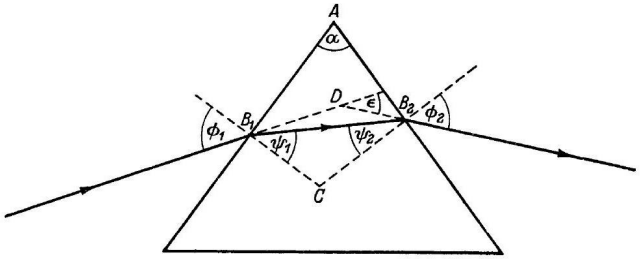



Fig. 4.27. Passage of a ray through a prism.

Using (12) this implies that

$$\left(\frac{d\phi_2}{d\phi_1}\right)_{\text{extr.}} = -1. \tag{16}$$

Now we have from (13) and (14),

$$\left.\begin{aligned} \tfrac{d\psi_1}{d\phi_1} &= -\tfrac{d\psi_2}{d\phi_1} \\ \cos\phi_1 &= n\cos\psi_1 \tfrac{d\psi_1}{d\phi_1} \\ \cos\phi_2 \tfrac{d\phi_2}{d\phi_1} &= n\cos\psi_2 \tfrac{d\psi_2}{d\phi_1} \end{aligned}\right\} \tag{17}$$

and hence, on elimination

$$\frac{d\phi_2}{d\phi_1} = -\frac{\cos\phi_1\cos\psi_2}{\cos\psi_1\cos\phi_2}. \tag{18}$$

From (16) and (18) it follows, that for an extremum,

$$\frac{\cos\phi_1\cos\psi_2}{\cos\psi_1\cos\phi_2} = 1, \tag{19}$$

whence, on squaring and using (14)

$$\frac{1-\sin^2\phi_1}{n^2-\sin^2\phi_1} = \frac{1-\sin^2\phi_2}{n^2-\sin^2\phi_2}. \tag{20}$$

This equation is satisfied by
then

$$\left.\begin{aligned} \phi_1 &= \phi_2; \\ \psi_1 &= \psi_2. \end{aligned}\right\} \tag{21}$$

To determine the nature of the extremum we must evaluate $d^2\varepsilon/d{\phi_1}^2$. From (12) and (18),

$$\begin{aligned} \frac{d^2\varepsilon}{d\phi_1^2} &= \frac{d^2\phi_2}{d\phi_1^2} = \frac{d\phi_2}{d\phi_1}\frac{d}{d\phi_1}\left[\log\left(-\frac{d\phi_2}{d\phi_1}\right)\right] \\ &= \frac{d\phi_2}{d\phi_1}\left[-\tan\phi_1 - \tan\psi_2\frac{d\psi_2}{d\phi_1} + \tan\psi_1\frac{d\psi_1}{d\phi_1} + \tan\phi_2\frac{d\phi_2}{d\phi_1}\right] \end{aligned} \tag{22}$$

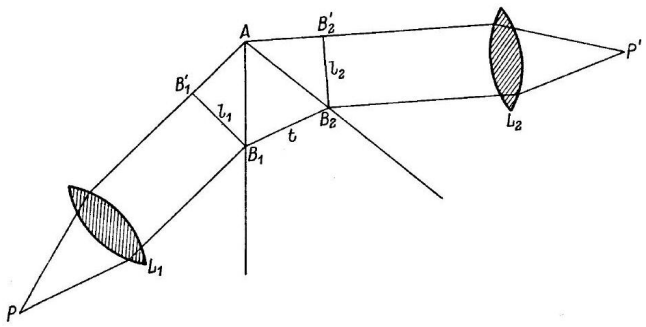



Fig. 4.28. Dispersion by a prism.

When $\phi_1 = \phi_2, \psi_1 = \psi_2$, this becomes with the help of (16), (17) and (14),

$$\left(\frac{d^2\varepsilon}{d\phi_1^2}\right)_{\text{extr.}} = 2\tan\phi_1 - 2\tan\psi_1 \frac{\cos\phi_1}{n\cos\psi_1} = 2\tan\phi_1\left(1 - \frac{\tan^2\psi_1}{\tan^2\phi_1}\right). \tag{23}$$

Since $n > 1, \phi_1 > \psi_1$; also since $0 < \phi_1 < \pi/2, \tan\phi_1 > 0$. Hence $(d^2\varepsilon/d{\phi_1}^2) > 0$, so that the deviation is a minimum. According to (21) it takes place when the passage of the rays through the prism is symmetrical. The minimal value of the deviation then is

$$\varepsilon_{\min} = 2\phi_1 - \alpha. \tag{24}$$

In terms of $\varepsilon_{\min}$ and $\alpha$, the angle of incidence and the angle of refraction at the first face of the prism are

$$\phi_1 = \frac{1}{2}\left(\varepsilon_{\min} + \alpha\right), \quad \psi_1 = \frac{1}{2}\alpha, \tag{25}$$

so that

$$n = \frac{\sin\phi_1}{\sin\psi_1} = \frac{\sin\left[\frac{1}{2}\left(\varepsilon_{\min} + \alpha\right)\right]}{\sin\left(\frac{1}{2}\alpha\right)}. \tag{26}$$

This formula is often used in determinations of the refractive index of glass. One measures $\varepsilon_{\min}$ and $\alpha$ with the help of a spectrometer and evaluates $n$ from (26).

Instead of a single ray, let us now consider the passage of a pencil of parallel rays through the prism, for example from a point source $P$ placed in the focal plane of a lens $L_1$ (Fig. 4.28), the light still being assumed to be monochromatic. Let $B_1'$ and $B_2'$ be the feet of the perpendiculars dropped from $B_1$ and $B_2$ on to the rays which pass through the edge $A$. Then $B_1B_1'$ and $B_2B_2'$ are the lines of intersection of two
wave-fronts with the plane of incidence (plane of the figure). These two lines are inclined to each other at the angle of deviation $\varepsilon$. We set

$$B_1B_1' = l_1, \quad B_2B_2' = l_2', \quad B_1B_2 = t \tag{27}$$

Consider now a parallel beam of polychromatic instead of monochromatic light. If the lens $L_1$ is corrected for chromatic aberration, $B_1B_1'$ will still be in the wave front of the incident pencil. On the other hand the line $B_2B_2'$ will no longer be unique, but will depend on the wavelength $\lambda$. For the refractive index of the prism is a function of the wavelength,

$$n = n(\lambda) \tag{28}$$

and consequently the deviation $\varepsilon$ also depends on $\lambda$ :

$$\varepsilon = \varepsilon(\lambda). \tag{29}$$

The quantity

$$\frac{d\varepsilon}{d\lambda} = \frac{d\varepsilon}{dn}\frac{dn}{d\lambda} \tag{30}$$

formed for a constant value of the angle of incidence $\phi_1$ is often called the angular dispersion of the prism. In (30) the first factor on the right depends entirely on the geometry of the arrangement, whilst the second factor characterizes the dispersive power of the glass of which the prism is made. Since $\phi_1 =$ constant, we have from (12) and (13)

$$\frac{d\varepsilon}{dn} = \frac{d\phi_2}{dn}, \quad \frac{d\psi_1}{dn} = -\frac{d\psi_2}{dn} \tag{31}$$

and from (14)

$$\begin{aligned} \sin\psi_1 + n\cos\psi_1\frac{d\psi_1}{dn} &= 0 \\ \cos\phi_2\frac{d\phi_2}{dn} &= \sin\psi_2 + n\cos\psi_2\frac{d\psi_2}{dn} \end{aligned} \tag{32}$$

whence on elimination

$$\frac{d\varepsilon}{dn} = \frac{\sin(\psi_1+\psi_2)}{\cos\phi_2\cos\psi_1} = \frac{\sin\alpha}{\cos\phi_2\cos\psi_1} \tag{33}$$

From the triangle $AB_1B_2$, we have, by the sine theorem,

$$AB_2 = \frac{\cos\psi_1}{\sin\alpha}t \tag{34}$$

and from the triangle $AB_2B_2'$,

$$AB_2 = l_2\sec\phi_2. \tag{35}$$

Using (34) and (35), (33) gives

$$\frac{d\varepsilon}{d\lambda} = \frac{t}{l_2}\frac{dn}{d\lambda} \tag{36}$$

In the position of minimum deviation, we have by symmetry $l_1 = l_2$. If, moreover, the lenses are so large that the pencil completely fills the prism, then $t$ will be equal to the length $b$ of the base of the prism. Equation (36) then gives for the angular dispersion $\delta\varepsilon$, i.e. for the angle by which the emergent wave-front is rotated when the wavelength is changed from $\lambda$ to $\lambda+\delta\lambda$, the following expression:

$$\delta\varepsilon = \frac{b}{l_1}\frac{dn}{d\lambda}\delta\lambda \tag{37}$$

## 9.5.8 4.8 Photometry and Apertures

The branch of optics concerned with the measurement of light is called photometry. Strictly, photometry is not part of geometrical optics, but it seems appropriate to include a short account of it in the present chapter, since in many practical applications the approximate geometrical picture of an optical field forms an adequate basis for photometric investigations. We shall,

therefore, take over the geometrical model according to which light is regarded as the flow of luminous energy along the geometrical rays, subject to the geometrical law of conservation of energy. This states (cf. § 3.1, eq. (31)), that the energy which is transmitted in unit time through any cross-section of a tube of rays is constant.

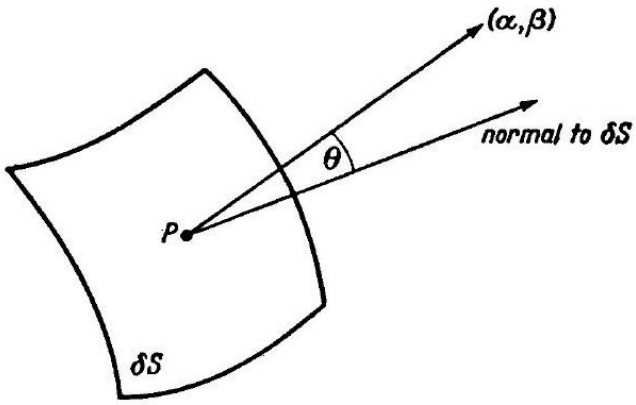



Fig. 4.29. Energy transmission from a surface element.

### 4.8.1 Basic concepts of photometry*

In photometry we are concerned essentially with the light energy emerging from a portion of a surface $S$. This surface may be fictitious, or it may coincide with the actual radiating surface of a source, or with an illuminated surface of a solid. If the latter is opaque, it is the reflected light which is considered; if it is transparent or semi-transparent (in which case the light is partly absorbed or scattered), it is the transmitted light which is usually measured.

Let $P(\xi, \eta)$ be a typical point on $S$ referred to any convenient set of curvilinear coordinates on the surface. The (time averaged) amount of energy which emerges per unit time from the element $\delta S$ of the surface at $P$ and which falls within an element $\delta\Omega$ of the solid angle around a direction specified by the polar angles ( $\alpha, \beta$ ), may be expressed in the form

$$\delta F = B \cos\theta \delta S \delta\Omega. \tag{1}$$

Here $\theta$ is the angle which the direction ( $\alpha, \beta$ ) makes with the normal to the surface element (see Fig. 4.29), and $B$ is a factor which in general depends on ( $\xi, \eta$ ) and ( $\alpha, \beta$ ),

$$B = B(\xi, \eta; \alpha, \beta). \tag{2}$$

The factor $\cos\theta$ is introduced in (1) since it is the projection of $\delta S$ on to a plane normal to the direction ( $\alpha, \beta$ ), rather than $\delta S$ itself, which is the physically significant quantity. $B$ is called the photometric brightness at the point $(\xi, \eta)$ in the direction $(\alpha, \beta)$. It must be distinguished from the visual sensation of brightness, from which *

it will in general differ, because the eye is not equally sensitive to all colours;* this point will be discussed more fully later.
$\delta F$ is usually decomposed in two different ways, to show the explicit dependence on $\delta\Omega$ and $\delta S$ :

0

- Only the fundamental notions of photometry will be discussed. For further information and for the description of instruments used for measuring light see, for example, J. W. T. Walsh, Photometry (London, Constable, 2nd ed., 1953) or Measurement of Radiant Energy (McGraw-Hill, New York and London, 1937), edited by W. E. Forsythe.

$$\delta F = \delta I \delta\Omega = \delta E \delta S \tag{3}$$

Comparison of (1) and (3) gives

$$\delta I = \frac{\delta F}{\delta\Omega} = B\cos\theta\delta S, \tag{4}$$

$$\delta E = \frac{\delta F}{\delta S} = B\cos\theta\delta\Omega. \tag{5}$$

The integral

$$I(\alpha,\beta) = \int B\cos\theta dS \tag{6}$$

taken over a piece of surface is called the photometric intensity † in the direction ( $\alpha,\beta$ ), and the integral

$$E(\xi,\eta) = \int B\cos\theta d\Omega \tag{7}$$

taken throughout a solid angle is called the photometric illumination at the point ( $\xi,\eta$ ). The variation of $B$ with direction will depend on the nature of the surface, especially on whether it is rough or smooth, whether it is self-luminous, or whether it transmits or reflects other light. Often it is permissible to assume that, to a good approximation, $B$ is independent of the direction. The radiation is then said to be isotropic. If the radiation is isotropic and if the radiating surface is plane, (6) reduces to

$$I(\alpha,\beta) = I_0\cos\theta \tag{8}$$

where

$$I_0 = \int B dS$$

The photometric intensity in any direction then varies as the cosine of the angle between that direction and the normal to the surface. (8) is known as Lambert's (cosine) law, and when satisfied one speaks of diffuse emission or diffuse reflection, according as to whether the surface is emitting or reflecting.

The measurement of the quantities $F, B, I$ and $E$ involves the determination of a time interval, an area, a direction, a solid angle and an energy. The averages involved are often small and consequently sensitive instruments have to be used. They are essentially of two kinds. First, those which react to the heat developed in an absorbing medium (e.g. bolometer, thermo-couple, etc.) and are mainly used in studying heat

---

0

- We use the adjective "photometric" when we wish to stress that a particular quantity is evaluated with regard to its true physical, rather than its visual, effect.
  † In Chapter I the light intensity was defined as the time average of the amount of energy which crosses per second a unit area perpendicular to the direction of the flow. This quantity must not be confused with the photometric intensity as defined by (6). It is unfortunate that the same word is used to denote two different quantities. Except in the present section we shall always understand by "intensity" the quantity introduced in Chapter I. If the surface element $\delta S$ at $P$ is orthogonal to the Poynting vector, the intensity (in the sense of Chapter I) is equal to the photometric illumination $\delta E$ at $P$.

radiation (infra-red); secondly, those which are based on the (surface) photo-electric effect, i.e. on the phenomenon of emission of electrons from a metal, caused by the incidence of light on the surface of the metal (in monochromatic light the number of emitted electrons, i.e. the current produced, is proportional to the energy of the incident light). Instruments of this kind are used, for instance, as exposure meters in photography.

In technical photometry, however, indirect methods are used. One defines and constructs a standard source of light, and expresses its photometric data in absolute energy units. Measurements are then made relative to this source, often using the eye as a null indicator, namely as an indicator of equal brightness. The comparison is based on a simple law which holds for the illumination due to a very small source* $Q$ :

Let $\delta S$ be a surface element at $P$ and let $QP = r$. If $\theta$ is the angle which $QP$ makes with the normal to $\delta S$ (see Fig. 4.30), then the energy which the source sends through $\delta S$ per unit time is $I\delta\Omega$, where $I$ is the photometric intensity of the source in the

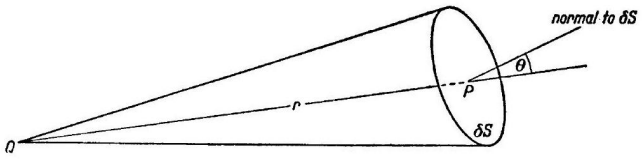



Fig. 4.30. Illumination from a point source.

direction $QP$, and $\delta\Omega$ is the solid angle which $\delta S$ subtends at $Q$. Now by elementary geometry,

$$\cos\theta\delta S = r^2\delta\Omega \tag{9}$$

Hence, using (3), we have

$$E = \frac{I\cos\theta}{r^2} \tag{10}$$

(10) is the basic equation of practical photometry. It expresses the so-called cosine law of illumination ( $E$ proportional to $\cos\theta$ ) and the inverse square law ( $E$ inversely proportional to $r^2$ ), and enables a comparison of the intensities of sources with the help of simple geometry. If a surface element $\delta S$ is illuminated by two point sources $Q_1$ and $Q_2$ of photometric intensities $I_1$ and $I_2$, and if the lines connecting $\delta S$ with the sources make angles $\theta_1$ and $\theta_2$ with the normal to $\delta S$ (Fig. 4.31) then, when the photometric illuminations are equal, we have

$$I_1 : I_2 = \frac{r_1^2}{\cos\theta_1} : \frac{r_2^2}{\cos\theta_2} \tag{11}$$

With the help of (11), the strength of a given source may be determined by a comparison with a standard source.

The equality of the illumination produced by two sources can be detected either by physical means, or directly, by the eye. The comparison by the eye is relatively easy when the light

0

- The photometric intensity $I$ of a point source is defined by a procedure often used in connection with limiting concepts. Let us suppose that the area of the surface decreases towards zero, while at the same time $B$ increases towards infinity, in such a way that the integral (6) remains finite. Then $I$ is a function of $\alpha$ and $\beta$ and of the position of the point source.

In calculating the photometric illumination, the finite extension of the source is usually neglected when its linear dimensions are less than about 1/20 of its distance from the illuminated surface.

from the two sources is monochromatic and of the same frequency, but in general one has to compare sources which emit light of different spectral compositions. This simple procedure can then no longer be

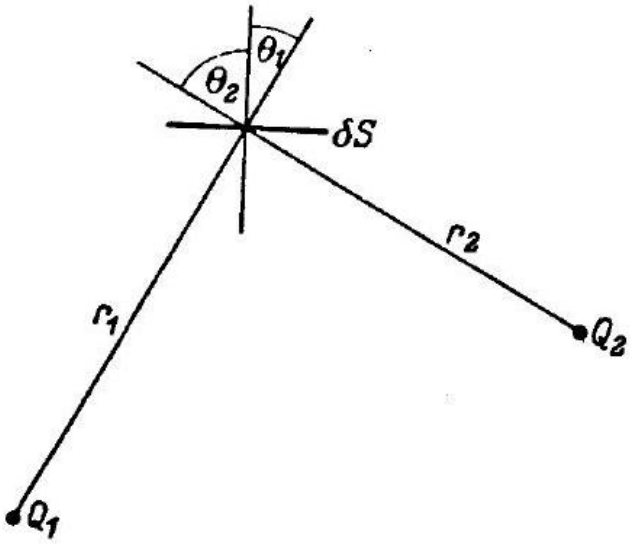



Fig. 4.31. Comparison of the intensities of two point sources.

used, since the eye is not equally sensitive to light of all wavelengths. Filters, each of which transmits light in a narrow range around a known wavelength of the spectrum, must be used instead; the determination of equality of illumination for different colours is then reduced to the determination of the relative energy values, taking into account the relative visibility curve of the eye. This is the curve obtained by plotting the reciprocal $K_\lambda$ of that amount of flux which produces the same sensation of brightness, against the wavelength. The curve depends to some extent on the strength of the illumination. For bright light it has a maximum near 550 m$\mu$. (See Fig. 4.32 and Table XIII.) With decreasing strength of illumination the curve retains its shape but the maximum shifts towards the blue end of the spectrum, being at about 507 m$\mu$ for very faint light. This phenomenon is known as the Purkinje effect.
If the flux of energy is evaluated with respect to the visual sensation which it

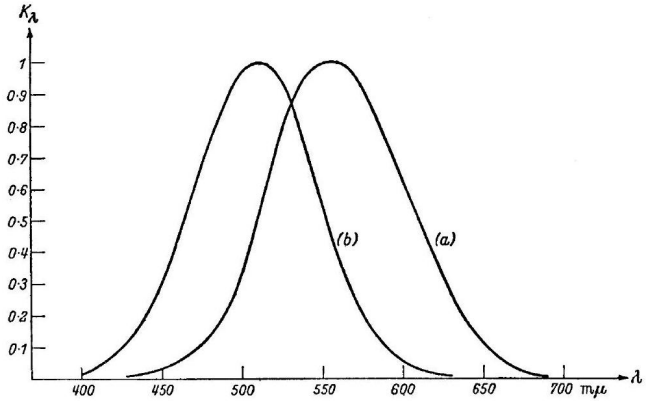



Fig. 4.32. The relative visibility curve $K_{\lambda}$ of the average human eye:
(a) for bright light and (b) for feeble light.

produces, rather than with regard to its true physical magnitude, one speaks of the luminous energy $F'$ :

$$F' = \frac{\int K_\lambda F_\lambda d\lambda}{\int K_\lambda d\lambda} \tag{12}$$

$F_\lambda d\lambda$ being the amount of energy in the range ( $\lambda, \lambda + d\lambda$ ), $K_\lambda$ the relative visibility and the integration being taken throughout the spectral range. The quantities $B', I'$ and $E'$, which bear the same relations to $B, I$ and $E$ as $F'$ bears to $F$, are usually called the luminance, the luminous intensity (or candle power) and the illumination respectively. These concepts are extensively used in visual photometry.

There are practical units for each of the four quantities $F', B', I'$ and $E'$. Since it is easier to maintain a standard of luminous intensity rather than of luminous flux, the unit of luminous intensity is usually considered to be the basic photometric unit, and those for $F', B'$ and $E'$ are expressed in terms of it. The adopted standard of luminous intensity was at one time the international candle, a standard preserved by

Table XIII
The Relative Visibility Factor $K_{\lambda}$ of an average human eye (bright light)

| $\lambda$( in m$\mu$) | $K_\lambda$ | $\lambda$( in $\mathbf{m}\mu$) | $\boldsymbol{K_\lambda}$ |
|---|---|---|---|
| 400 | 0.0004 | 600 | 0.631 |
| 410 | 0.0012 | 610 | 0.503 |
| 420 | 0.0040 | 620 | 0.381 |
| 430 | 0 · 0116 | 630 | 0.265 |
| 440 | 0 · 023 | 640 | 0.175 |
| 450 | 0.038 | 650 | 0 · 107 |
| 460 | 0 · 060 | 660 | 0.061 |
| 470 | 0.091 | 670 | 0 · 032 |
| 480 | 0 · 139 | 680 | 0.017 |
| 490 | 0 · 208 | 690 | 0.0082 |
| 500 | 0.323 | 700 | 0.0041 |
| 510 | 0.503 | 710 | 0.0021 |
| 520 | 0 · 710 | 720 | 0.00105 |
| 530 | 0.862 | 730 | 0.00052 |
| 540 | 0.954 | 740 | 0.00025 |
| 550 | 0.995 | 750 | 0 · 00012 |
| 560 | 0.995 | 760 | 0.00006 |
| 570 | 0.952 | | |
| 580 | 0.870 | | |
| 590 | 0.757 | | |

a number of carbon lamps kept at various national laboratories. In more recent times it has been replaced by a new standard called the candela (cd); this is defined as one-sixtieth of the luminous intensity per square centimetre of a black body radiator at the temperature of solidification of platinum ( 2042°K approx.). The value of the luminous intensity of light of a different spectral composition must be evaluated by the procedure already explained, taking into account the relative visibility curve.

The unit of luminous flux is called the lumen. It is the luminous flux emitted within a unit solid angle by a uniform point source of luminous intensity 1 candela.

The unit of illumination depends on the unit of length employed. The metric unit of illumination is the lux (lx), sometimes also called the metre-candle; it is the illumination of a

surface area of one square metre receiving a luminous flux of one lumen. The British unit is the lumen per square foot, formerly called the foot-candle (f.c.).

The unit of luminance is the candela per square centimetre, termed the stilb (sb), and the candela per square metre termed the nit. In the British system, the units used are the candela per square inch or candela per square foot.

Other units are also used, but for their definitions and their relation to the units here discussed we must refer to books on photometry.

### 4.8.2 Stops and pupils*

The amount of light which reaches the image space of an optical system depends not only on the brightness of the object but also on the dimensions of the optical elements (lenses, mirrors) and of the stops. A stop (or diaphragm) is an opening, usually circular, in an opaque screen. The opaque parts of the screen prevent some of the severely aberrated rays from reaching the image. For the purposes of the present

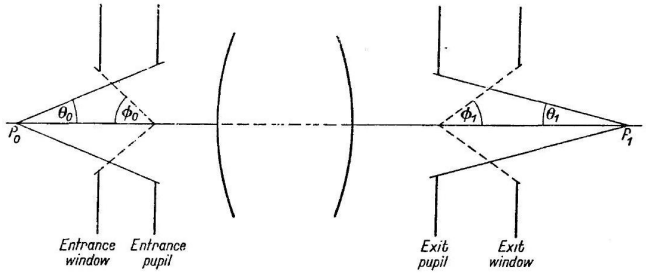



Fig. 4.33. Stops and pupils.

discussion it will be convenient to include the edges of the lenses and mirrors as well as diaphragms in the term "stop".

Consider all the rays from the axial object point $P_0$. The stop which determines the cross-section of the image-forming pencil is called the aperture stop. To determine its position, the Gaussian image of each stop must be found in the part of the system which precedes it; the image which subtends the smallest angle at $P_0$ is called the entrance pupil. The physical stop which gives rise to the entrance pupil is the aperture stop. (If it lies in front of the first surface it is identical with the entrance pupil.) The angle $2\theta_0$ which the diameter of the entrance pupil subtends at $P_0$ is called the angular aperture on the object side, or simply angular aperture (Fig. 4.33).

The image of the aperture stop formed by the part of the system which follows it (also the image of the entry pupil by the whole system) is known as the exit pupil; the angle $2\theta_1$ which its diameter subtends at the image $P_1$ may be called the angular aperture on the image side (sometimes also the projection angle).

In the pencil of rays which proceeds from each object point, there will be a ray which passes through the centre of the entrance pupil. This special ray is known as the principal ray (also the chief or the reference ray) of the pencil, and is of particular importance in the theory of aberrations. In the absence of aberrations, the principal ray will also pass through the centre of the aperture stop and through the centre of the exit pupil.

If the aperture stop is situated in the back focal plane of that part of the system which precedes it, then the entrance pupil will be at infinity and all the principal rays

0

• The theory of stops was formulated by E. Abbe, Jena Z. Naturwiss., 6 (1871), 263, and was extended by

in the object space will be parallel to the axis. Such a system is said to be telecentric on the object side. If the aperture stop is in the front focal plane of the part of the system which follows it, then the exit pupil will be at infinity and the principal rays in the image space will be parallel to the axis; the system is then said to be telecentric on the image side. Telecentric arrangements are useful in measurements of the size of the object.

If other parameters are kept fixed, the angular aperture is a measure of the amount of light which traverses the system. There are other quantities which are frequently used to specify the "light-gathering power" of an optical system, for example the numerical aperture (N.A.) of a microscope objective. This is defined as the sine of the angular semi-aperture in the object space multiplied by the refractive index of the object space,

$$N.A. = n \sin \theta_0. \tag{13}$$

When a system is designed to work with objects at a great distance, as in the case of telescopes or certain photographic lenses, a convenient measure of its light-gathering

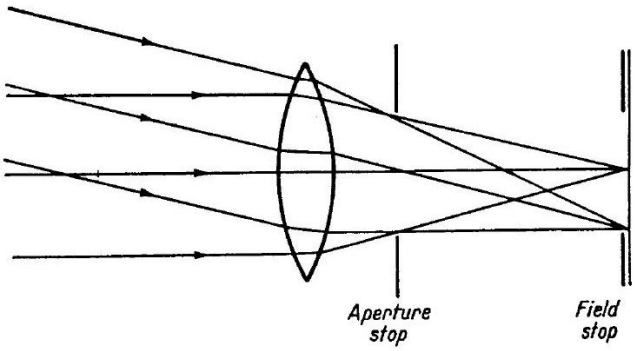



Fig. 4.34. Illustrating the distinction between the aperture stop and a field stop.

power is the so-called " $F$ number" or "nominal focal ratio". It is the ratio of the focal length $f$ of the system to the diameter $d$ of the entrance pupil:

$$F = f/d. \tag{14}$$

Thus for a lens of 10 cm focal length which works with an aperture of 2 cm, $F = 5$. It is called an $f/5$ lens and is said to work at a "speed" of $f/5$.

Quantities such as the angular aperture, the numerical aperture, and the $F$ number, which may be taken as the measure of the light-gathering power of the instrument, are often called relative apertures.

In addition to aperture stops, optical systems also possess field stops; they determine what proportion of the surface of an extended object is imaged by the instrument. The distinction between the two types of stops is illustrated in Fig. 4.34.

To determine the field stop, we again find first the image of each stop in the part of the system which precedes it. That image which subtends the smallest angle ( $2\phi_0$ in Fig. 4.33) at the centre of the entrance pupil is called the entrance window, and the angle $2\phi_0$ is called the field angle or the angular field of view. The physical stop which corresponds to the entrance window is then the required field stop.

The image of the entrance window by the instrument (that is also the image of the field stop by the part of the system which follows it), is called the exit window. The angle ( $2\phi_1$ in

M. von Rohr, Zentr. Ztg. Opt. Mech., 41 (1920), 145, 159, 171.

Fig. 4.33) which the diameter of the exit window subtends at the centre of the exit pupil is sometimes called the image field angle.

It may happen that, although the aperture stop is smaller than the lenses, some of the rays miss part of a lens entirely and parts of a lens may receive no light at all from certain regions of the object. This effect is known as vignetting, and is illustrated in Fig. 4.35. It is seldom encountered in systems like telescopes which have a relatively

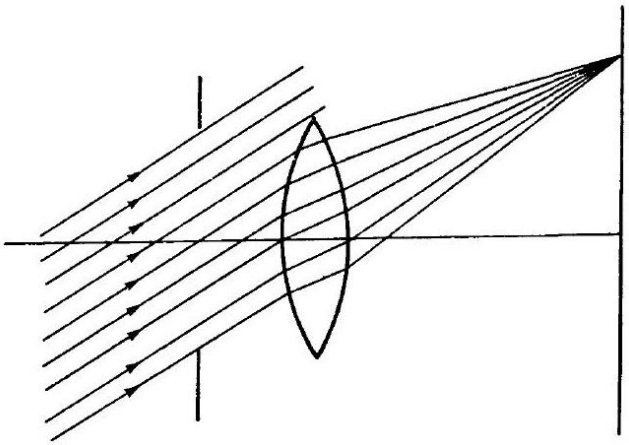


Fig. 4.35. Vignetting.

small field of view, but is of importance in other instruments, such as photographic objectives. Designers sometimes rely on vignetting to obliterate undesirable off-axis aberrations.

### 4.8.3 Brightness and illumination of images

We shall now briefly consider the relations between the basis photometric quantities which characterize the radiation in the image and object space.

Assume that the object is a small plane element of area $\delta S_0$, perpendicular to the axis and radiating in accordance with Lambert's law. The photometric brightness $B_0$ is then independent of direction. The amount of energy $\delta F_0$ which falls per unit time on to an annular element of the entry pupil centred on the axis is

$$\delta F_0 = B_0 \cos\gamma_0 \delta S_0 \delta\Omega_0 \tag{15}$$

where

$$\delta\Omega_0 = 2\pi \sin\gamma_0 \delta\gamma_0, \tag{16}$$

$\gamma_0$ being the angle which a typical ray passing through the annulus makes with the axis. Hence if $\theta_0$ denotes, as before, the angular semi-aperture on the object side, the total flux of energy which falls on to the entrance pupil per unit time is

$$\begin{aligned} F_0 &= 2\pi B_0 \delta S_0 \int_0^{\theta_0} \sin\gamma_0 \cos\gamma_0 d\gamma_0 \\ &= \pi B_0 \delta S_0 \sin^2\theta_0 \end{aligned} \tag{17}$$

The energy flux $F_1$ emerging from the exit pupil may be expressed in a similar form:

$$F_1 = \pi B_1 \delta S_1 \sin^2\theta_1 \tag{18}$$

$F_1$ cannot exceed $F_0$ and can only be equal to it if there are no losses due to reflection, absorption or scattering within the system; hence

$$B_1 \sin^2 \theta_1 \delta S_1 \leqslant B_0 \sin^2 \theta_0 \delta S_0. \tag{19}$$

Now the ratio $\delta S_1/\delta S_0$ is equal to the square of the lateral magnification $M$ :

$$\frac{\delta S_1}{\delta S_0} = M^2 \tag{20}$$

If further it is assumed that the system obeys the sine condition (§ 4.5, eq. (6)),

$$\frac{n_0 \sin \theta_0}{n_1 \sin \theta_1} = M \tag{21}$$

On substituting from (20) and (21) into (19) it follows that

$$B_1 \leqslant \left(\frac{n_1}{n_0}\right)^2 B_0 \tag{22}$$

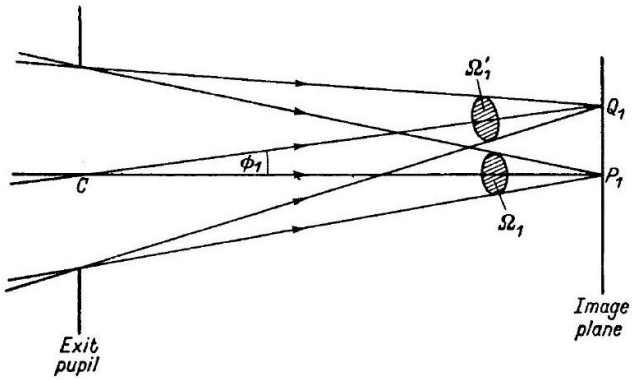



Fig. 4.36. Illumination at an off-axis image point.

In particular, if the refractive indices of the object and image spaces are equal, then according to (22), the photometric brightness of the image cannot exceed that of the object, and can only be equal to it if the losses of light within the system are negligible.

From (18) and (22) it follows, assuming the losses to be negligible, that

$$F_1 = \pi \left(\frac{n_1}{n_0}\right)^2 B_0 \delta S_1 \sin^2 \theta_1 \tag{23}$$

so that the photometric illumination $E_1 = F_1/\delta S_1$ at the axial point $P_1$ of the image is

$$E_1 = \pi \left(\frac{n_1}{n_0}\right)^2 B_0 \sin^2 \theta_1 \tag{24}$$

If $\theta_1$ is small, the solid angle $\Omega_1$, which the exit pupil subtends at $P_1$, is approximately equal to $\pi \sin^2 \theta_1$, so that (24) may then be written as

$$E_1 = \left(\frac{n_1}{n_0}\right)^2 B_0 \Omega_1 \tag{25}$$

This relation applies to the axial image, but the off-axis image may be treated in a similar way. If $\phi_1$ is the angle which the principal ray $CQ_1$ makes with the axis (see Fig. 4.36), then we have in place of (25),

$$E_1 = \left(\frac{n_1}{n_0}\right)^2 B_0 \Omega_1' \cos\phi_1 \tag{26}$$

$\Omega_1'$ being the solid angle which the exit pupil subtends at $Q_1$. With the help of (9) it follows that

$$\frac{\Omega_1'}{\Omega_1} = \cos\phi_1 \left(\frac{CP_1}{CQ_1}\right)^2 = \cos^3\phi_1, \tag{27}$$

so that

$$E_1 = \left(\frac{n_1}{n_0}\right)^2 B_0 \Omega_1 \cos^4\phi_1 \tag{28}$$

This formula shows that the illumination in the image decreases as the fourth power of the cosine of the angle which the principal ray through the image point makes with the axis, it being assumed that the object radiates according to Lambert's law, that there are no losses of light within the system, and that the angular semi-aperture $\theta_1$ is small.

In applying the preceding formulae, it should be remembered that they have been derived with the help of the laws of geometrical optics. For a very small source they may no longer give a good approximation. For example, the image of a point source is not a point but a bright disc surrounded by rings (the Airx pattern, see § 8.5.2.); the light is then distributed over the whole diffraction pattern, and consequently the illumination at the geometrical focus is smaller than that given by (24).

## 9.5.9 4.9 Ray Tracing* (!!!????)

(??? is exactly this theory is used in 3d graphics now???)

In designing optical instruments it is necessary to determine the path of the light with a greater accuracy than that given by Gaussian optics. This may be done by algebraic analysis, taking into account the higher-order terms in the expansion of the characteristic function, a procedure discussed in the next chapter. Alternatively one may determine the path of the light rays accurately with the help of elementary geometry, by successive application of the law of refraction (or reflection); this method, which will now be briefly described, is known as ray tracing and is extensively employed in practice.

### 4.9.1 Oblique meridional rays †

We consider first the tracing of an oblique meridional ray, i.e. a meridional ray from an extra-axial object point. Let $A$ be the pole of the first surface of the system. The surface will be assumed to be a spherical refracting surface of radius $r$ centred at a point $C$, and separating media of refractive indices $n$ and $n'$. An incident ray $OP$ (see Fig. 4.37) in the meridional plane is specified by the angle $U$ which it makes with the axis, and by the distance $L = AB$ between the pole of $A$ and the point $B$ at which it meets the axis. Let $I$ be the angle between the incident ray and the normal

0

- For a fuller discussion of ray tracing, see for example A. E. Conrady, Applied Optics and Optical Design, Part I (Oxford University Press, 1929; reprinted by Dover Publications, Inc., New York, 1957); M. von Rohr, The Geometrical Investigation of Formation of Images in Optical Systems, translated from German by R. Kanthack (London, H.M. Stationery Office, 1920), and M. Herzberger, Modern Geometrical Optics (New York, Interscience Publishers, 1958). A method for the tracing of rays with the help of electronic computing machines was described by G. Black, Proc. Phys. Soc. B, 68 (1954), 569. A method for the

$PC$. The corresponding quantities relating to the refracted ray are denoted by primed symbols.

The following sign convention is used: The quantities $r, L$ and $L'$ are taken to be positive when $C, B$ and $B'$ are to the right of $A$, the light being assumed to be incident from the left. The angles $U$ and $U'$ are considered to be positive if the axis can be brought into coincidence with the rays $PB$ and $PB'$ by a clockwise rotation of less than 90° about $B$ or $B'$ respectively. The angles $I$ and $I'$ are taken to be positive if the incident and the refracted rays may be made to coincide with the normal $PC$ by a clockwise rotation of less than 90° about the point $P$ of incidence.

The quantities $L$ and $U$ which specify the incident ray may be assumed to be known.

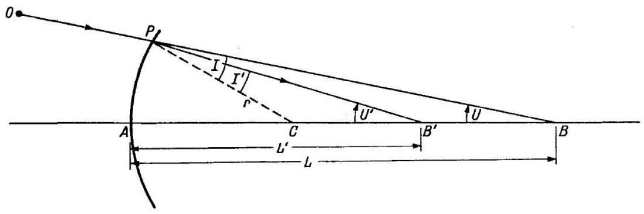



Fig. 4.37. Notation used in an oblique meridional ray trace.

It is then necessary to calculate $L'$ and $U'$. Assuming also for the moment that both $L$ and $r$ are finite, we have, from the triangle $PCB$ :

$$\sin I = \frac{L - r}{r} \sin U. \tag{1}$$

By the law of refraction

$$\sin I' = \frac{n}{n'} \sin I. \tag{2}$$

Also from the figure

$$U' = U + I - I'. \tag{3}$$

Finally, from the triangle $PCB'$,

$$L' = \frac{\sin I'}{\sin U'} r + r. \tag{4}$$

By successive application of the refraction equations (1)-(4), the quantities $L'$ and $U'$, which specify the refracted ray $PB'$, are obtained.

The refracted ray $PB'$ now becomes the incident ray for the second surface. Writing $L'_1$ in place of $L'$ and $U'_1$ in place of $U'$, and denoting by $L_2, U_2$ the corresponding values, with $L_2$ referred to the pole of the second surface, we have the transfer equations

$$L_2 = L'_1 - d, \tag{5}$$

$$U_2 = U'_1, \tag{6}$$

tracing of rays through non-spherical surfaces was proposed by T. Smith, Proc. Phys. Soc., 57 (1945), 286; see also W. Weinstein, Proc. Phys. Soc. B, 65 (1952), 731.

† In §4.9.1, §4.9.2, and §4.10 the notation and the sign convention usually employed in a meridional ray trace is used. The sign convention for angles differs from the Cartesian sign convention used throughout the rest of the book; to revert to the Cartesian sign convention set $U = -\gamma$, $U' = -\gamma'$.

where $d > 0$ is the distance between the poles of the two surfaces.
Next the "incident values" given by (5) and (6) are substituted into the refraction equations (1)-(4). Solving for the primed quantities, the ray is then traced through the
second surface. In this way, by the repeated application of the refraction and the transfer equations, the values $L'$ and $U'$, relating to the ray in the image space, are obtained. The point of intersection of this ray with the image plane may then be determined. In practice, one would naturally trace not a single ray, but a number of suitably selected rays through the system; their intersection points with the image plane then give a rough estimate of the performance of the system.

If one of the surfaces (say the $k$ th) is a mirror, the appropriate formulae to be used may be formally deduced from the preceding ones by setting $n'_k = -n_k$. Then $d_k$ must be considered to be negative. Moreover all the remaining refractive indices and

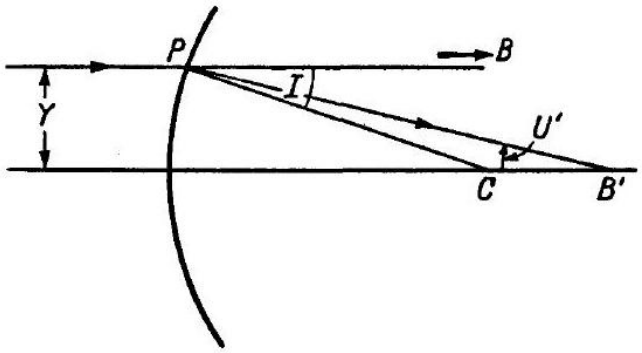



Fig. 4.38 (a). The special case $L = \infty$.

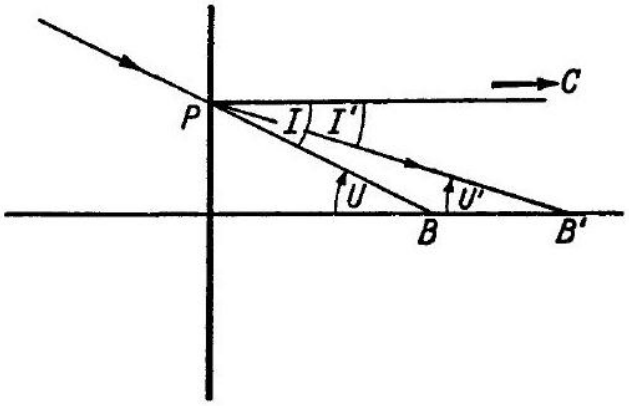



Fig. 4.38 (b). The special case $r = \infty$.

the subsequent $d$ values must also be considered to be negative, unless a second reflection takes place, when they revert to positive signs.

Next consider the two special cases which were so far excluded. If the incident ray is parallel to the axis ( $L = \infty$ ) the equation

$$\sin I = \frac{Y}{r}, \tag{7}$$

is used in place of (1), where $Y$ is the distance of the ray from the axis (see Fig. 4.38(a)).

If the surface is plane ( $r = \infty$ ) we have, in place of (1)-(4), the following set of equations (cf. Fig. 4.38(b)):

$$I = -U \tag{8}$$

$$\sin U' = \frac{n}{n'} \sin U \tag{9}$$

$$I' = -U' \tag{10}$$

$$L' = \frac{\tan U}{\tan U'} L \tag{11}$$

On account of (9), equation (11) may also be written in the form

$$L' = \frac{n'}{n} \frac{\cos U'}{\cos U} L, \tag{11a}$$

which is moro convenient for computation than (11) if the angles are small.

It is useful to determine also the coordinates ( $Y_k, Z_k$ ) of the point $P_k$ of incidence at the $k$ th surface, and the distance $D_k = P_k P_{k+1}$ between two successive points of incidence (see Fig. 4.39).

From the figure,

$$Y_k = r_k \sin (U_k + I_k) \tag{12}$$

$$Z_k = r_k - r_k \cos (U_k + I_k) = \frac{Y_k^2}{[1 + \cos (U_k + I_k)]\, r_k} \tag{13}$$

$$D_k = \{d_k + Z_{k+1} - Z_k\} \sec U'_k \tag{14}$$

In terms of $Y_k, Z_k$ and $U'_k$,

$$L'_k = Z_k + \frac{Y_k}{\tan U'_k}. \tag{15}$$

This relation may be used as a check on the value computed from (11a) or (11).

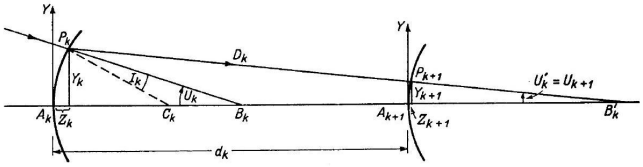



Fig. 4.39. Oblique meridional ray trace through two successive refracting surfaces.

In the special case when $L_k$ is infinite, $U_k = 0$, and (12)-(15) still holds. If $r_k$ is infinite, $Y_k$ may be computed from the relation

$$Y_k = L_k \tan U_k, \tag{16}$$

$Z_k$ then being zero.

### 4.9.2 Paraxial rays

If the inclination of a ray to the axis is sufficiently small, the sines of the various angles may be replaced, in the preceding formulae, by the angles themselves. The formulae then reduce to the Gaussian approximation for the path of the light. Such "paraxial ray-tracing formulae" are

used in practice for computing the Gaussian magnification and the focal length of the system. A brief summary of these formulae will therefore be given here.

It is customary to denote quantities which refer to the paraxial region by small letters. The refraction equations (1)-(4) become

$$i = \frac{l - r}{r} u, \tag{17}$$

$$i' = \frac{n}{n'} i, \tag{18}$$

$$u' = u + i - i', \tag{19}$$

$$l' = \frac{i'}{u'} r + r. \tag{20}$$

The transfer equations (5) and (6) take the form

$$l_2 = l_1' - d, \tag{21}$$

$$u_2 = u_1'. \tag{22}$$

In a similar way the paraxial equations for the cases $L = \infty$ and $r = \infty$ are obtained from (7)-(1la).

The paraxial equation for the incidence height, needed later, follows from (12):

$$y_k = r_k \left( u_k + i_k \right). \tag{23}$$

Although the relations (17)-(20) involve the angles which the incident ray and the refracted ray make with the axis, $l'$ is independent of these quantities. This result, established in a different manner in § 4.4.1 follows when $i, i'$ and $u'$ are eliminated from the above relations. It is then found that $u$ also disappears, and we obtain

$$n' \left( \frac{1}{r} - \frac{1}{l'} \right) = n \left( \frac{1}{r} - \frac{1}{l} \right). \tag{24}$$

This will be recognized as Abbe's relation § 4.4 (7).

To determine the lateral Gaussian magnification $M$, it is only necessary to trace a paraxial ray from the axial object point. Then according to §4.4 (54),

$$M = \frac{n_1}{n_l} \frac{u_1}{u_l} \tag{25}$$

where the subscripts 1 and $l$ refer to the first and last medium.

The focal length $f'$ of the system may be obtained by tracing a paraxial ray at any desired height $y_1$ from an infinitely distant object. The equation of the conjugate ray in the image space, referred to axes at the second focal point, is then $y_l / z_l = -u_l'$, and it follows from § 4.3 (10) that

$$f' = -\frac{y_1}{u_l'}. \tag{26}$$

### 4.9.3 Skew rays

So far only rays which lie in a meridional plane were considered. We shall now briefly discuss the tracing of skew rays, i.e. rays which are not coplanar with the axis. The tracing of such

rays is much more laborious and is usually carried out only in the design of systems with very high apertures.

A ray will now be specified by its direction cosines and by the coordinates of the point at which it meets a particular surface of the system. We take Cartesian rectangular axes at the pole $A_1$ of the first surface, with the $Z$-axis along the axis of the system. Let $L_1, M_1, N_1, (L_1{}^2 + M_1{}^2 + N_1{}^2 = 1)$ be the direction cosines of a ray incident at the point $P_1\,(X_1, Y_1, Z_1)$ (see Fig. 4.40).

The first step is to calculate the cosine of the angle $I_1$ of incidence. If $r_1$ is the radius of the first surface, the direction cosines of the normal at $P_1$ are

$$\bar{L}_1 = -\frac{X_1}{r_1}, \quad \bar{M}_1 = -\frac{Y_1}{r_1}, \quad \bar{N}_1 = \frac{r_1 - Z_1}{r_1},$$

so that

$$\begin{aligned} \cos I_1 &= L_1\bar{L}_1 + M_1\bar{M}_1 + N_1\bar{N}_1 \\ &= N_1 - \frac{1}{r_1}\left(L_1X_1 + M_1Y_1 + N_1Z_1\right). \end{aligned} \tag{27}$$

The next step is to determine the direction cosines $L_1', M_1', N_1'$ of the refracted ray. This is done in two stages: One calculates first the cosine of the angle of refraction, using the law of refraction in the form

$$n' \cos I_1' = \sqrt{n'^2 - n^2 + n^2 \cos^2 I_1}. \tag{28}$$

One then uses the fact that the refracted ray lies in the plane specified by the incident ray and the surface normal. Denoting by $s_1, s_1'$ and $\bar{s}_1$ the unit vectors along the incident ray, the refracted ray, and the normal, i.e. the vectors with components $(L_1, M_1, N_1)\,, (L_1', M_1', N_1')$ and $\left(\bar{L}_1, \bar{M}_1, \bar{N}_1\right)$, the coplanarity condition gives

$$s_1' = \lambda s_1 + \mu \bar{s}_1, \tag{29}$$

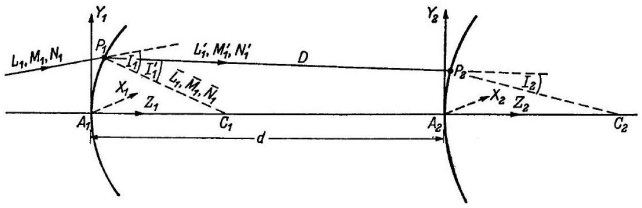



Fig. 4.40. Tracing of a skew ray.

where $\lambda$ and $\mu$ are certain scalar functions. To determine $\lambda$ and $\mu$ we first multiply (29) scalarly by $s_1$, and use the fact that (see Fig. 4.40) $s_1 \cdot s_1' = \cos\left(I_1 - I_1'\right)$, $s_1 \cdot \bar{s}_1 = \cos I_1$. This gives

$$\cos\left(I_1 - I_1'\right) = \lambda + \mu \cos I_1.$$

Next we multiply (29) scalarly by $\bar{s}_1$, and use the relations $\bar{s}_1 \cdot s_1' = \cos I_1'$, $\bar{s}_1 \cdot s_1 = \cos I_1$. We then obtain

$$\cos I_1' = \lambda \cos I_1 + \mu.$$

From the last two relations it follows that

$$\lambda = \frac{\sin I_1'}{\sin I_1} = \frac{n}{n'}, \quad \mu = \frac{1}{n'}\left(n' \cos I_1' - n \cos I_1\right),$$

and (29) then gives the following three equations for the direction cosines of the refracted ray:

$$\left.\begin{array}{ll} n'L_1' & = nL_1 - \sigma X_1 \\ n'M_1' & = nM_1 - \sigma Y_1 \\ n'N_1' & = nN_1 - \sigma\left(Z_1 - r_1\right) \end{array}\right\} \tag{30}$$

where

$$\sigma = \frac{1}{r_1}\left(n' \cos I_1' - n \cos I_1\right). \tag{31}$$

This completes the ray trace through the first surface, by means of the refraction equations (27), (28), (30) and (31).

The refracted ray now becomes the incident ray for the second surface. Taking parallel axes at the pole $A_2$ of the second surface, we have the transfer equations for the direction cosines

$$L_2 = L_1', \quad M_2 = M_1', \quad N_2 = N_1'. \tag{32}$$

Let $d$ be the distance $A_1A_2$ between the poles of the two surfaces. Denoting by $X_1^+, Y_1^+, Z_1^+$the coordinates of the point $P_1$ referred to the axes at $A_2$, we have the transfer equations for the coordinates

$$X_1^+ = X_1, \quad Y_1^+ = Y_1, \quad Z_1^+ = Z_1 - d. \tag{33}$$

Next the coordinates $(X_2, Y_2, Z_2)$ of the point $P_2$ in which the refracted ray meets the second surface must be determined. If $D$ denotes the distance from $P_1$ and $P_2$, then

$$\left.\begin{array}{l} X_2 = X_1^+ + L_2 D, \\ Y_2 = Y_1^+ + M_2 D, \\ Z_2 = Z_1^+ + N_2 D. \end{array}\right\} \tag{34}$$

To determine $D$ we use the fact that $P_2$ lies on the second surface. If $r_2$ is the radius of this surface, then

$${X_2}^2 + {Y_2}^2 + {Z_2}^2 - 2Z_2 r_2 = 0. \tag{35}$$

On substituting into this equation from (34) the following equation for $D$ is obtained:

$$D^2 - 2Fr_2 D + Gr_2 = 0, \tag{36}$$

where

$$\left.\begin{array}{l} F = N_2 - \frac{1}{r_2}\left(L_2 X_1^+ + M_2 Y_1^+ + N_2 Z_1^+\right), \\ G = \frac{1}{r_2}\left[\left(X_1^+\right)^2 + \left(Y_1^+\right)^2 + \left(Z_1^+\right)^2\right] - 2Z_1^+. \end{array}\right\} \tag{37}$$

Equation (36) gives the following expression for $D$, if we also make use of the fact that $D = d$ for the axial ray:

$$D = r_2\left(F - \sqrt{F^2 - \frac{1}{r_2}G}\right). \tag{38}$$

The coordinates of $P_2$ are determined on substituting for $D$ into (34).
Finally, to complete the cycle, the cosine of the angle $I_2$ of incidence at the second surface must be calculated. It is given by a relation strictly analogous to (27):

$$\cos I_2 = N_2 - \frac{1}{r_2}\left(L_2X_2 + M_2Y_2 + N_2Z_2\right), \tag{39}$$

or, using (34) and (38)

$$\begin{aligned}\cos I_2 &= F - \frac{1}{r_2}D \\ &= \sqrt{F^2 - \frac{1}{r_2}G}.\end{aligned} \tag{40}$$

Hence the last stage consists of the evaluation of the formulae (40), (38) and (34).
Because of the labour involved in the tracing of a general skew ray, the calculations are sometimes restricted to skew rays which lie in the immediate neighbourhood of a selected principal ray. Such skew rays may be traced through the system with the help of simplified schemes,* which are similar to those used for paraxial ray trace and are adequate for determining the position of the sagittal focal surface.

## 9.5.10 4.10 Design of Aspheric Surfaces

In the great majority of optical systems lenses and mirrors are employed, the surfaces of which have plane, spherical or paraboloidal form. The restriction to surfaces of such simple form is mainly due to the practical difficulties encountered in the production of surfaces of more complicated shapes with the high degree of precision required in optics. The restriction to surfaces of simple form naturally imposes limitations on the ultimate performance which systems of conventional design can attain. For this reason, in spite of the difficulties of manufacture, surfaces of more complicated form, called aspheric surfaces, are employed in certain systems. As early as 1905, K. Schwarzschild* considered a class of telescope objectives consisting of two aspheric mirrors, and showed that such systems can be made aplanatic.

In 1930, Bernhard Schmidt, an optician of Hamburg, constructed a telescope of a new type, which consisted of a spherical mirror and a suitably designed aspheric lens placed at its centre of curvature. The performance of this system (considered more fully in §6.4) was found to be quite outstanding. By means of such a telescope it is possible to photograph on one plate a very large region of the sky, many hundred times larger than can be obtained with telescopes of conventional design. The Schmidt camera has since become an important tool in astronomical research. Aspheric systems which use the principle of the Schimidt camera are also employed in certain projectiontype television receivers, † in X-ray fluorescent screen photography, and certain fast low-dispersion spectrographs. Aspheric surfaces find also useful application in microscopy (cf. § 6.6).

By making one surface of any centred system aspherical, it is possible, in general, to ensure exact axial stigmatism; by means of two aspheric surfaces any centred system may in general be made aplanatic. In this section formulae will be derived for the design of such aspheric surfaces.

---

0

- Cf. H. H. Hopkins, Proc. Phys. Soc., 58 (1946), 663; also his Wave Theory of Aberrations (Oxford, Clarendon Press, 1950), pp. 59, 65.

### 4.10.1 Attainment of axial stigmatism ‡

Consider the rays from an axial object point $P$. In the image space of the system the rays from different zones of the exit pupil will in general intersect the axis at different points. Let $S^{(0)}$ be the last surface of the system and $O$ its axial point (Fig. 4.41). It will be shown that it is possible to modify the profile of $S^{(0)}$ in such a way as to compensate exactly for the departure from the homocentricity of the image-forming pencil; more precisely we shall show that it is possible in general to replace $S^{(0)}$ by a new surface $S$ which will ensure that all the rays in the image space meet the axis at any prescribed point $Q$.

On account of symmetry only meridional rays need be considered. Each ray in the space which precedes the last surface will be specified by the following parameters: the angle $U$ which it makes with the axis and the distance $H = OM$ at which it intersects the $Y$-axis (see Fig. 4.41). It will be convenient to label the rays: Let $t$ be

any convenient parameter, for example the angle $U_0$ which the corresponding ray in the object space makes with the axis, or the height at which it meets the first surface. The pencil may then be completely specified by the two functional relations

$$U = U(t), \quad H = H(t). \tag{1}$$

It may be assumed that $t = 0$ for the axial ray. In general the relations (1) will not be known in an explicit form; a table of values of $U$ and $H$ for any prescribed set of $t$ values may, however, be obtained from a ray trace.

Let $W$ be the orthogonal surface (wave-front) through $O$ of the pencil (Fig. 4.41), and let $N$ be the point in which the ray through $M$ meets it. Assuming that a correcting surface $S$ with the required property can be found, it follows that in the corrected

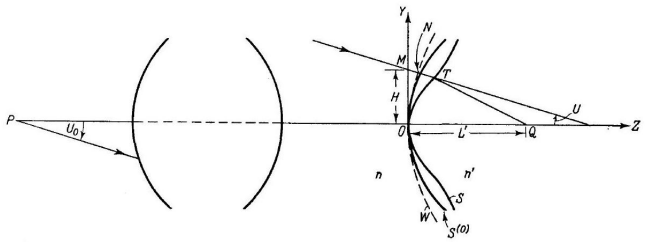



Fig. 4.41. Design of an aspheric surface to attain axial stigmatism.

system the optical path length from $N$ to $Q$ must be equal to the optical path length from $O$ to $Q$. Hence, if $T(Y, Z)$ is the point in which the ray meets $S$,

$$[NT] + [TQ] = [OQ]. \tag{2}$$

If $n$ is the refractive index of the space which precedes $S$, and $n'$ the refractive index of the image space, we have*

0

- K. Schwarzschild, Astr. Mitt. Königl. Sternwarte Göttingen (1905). Reprinted from Abh. Königl. Ges. Wiss. Gottingen, Math. Phys. Klasse, 4 (1905-1906), No. 2.

Two telescopes of this type were later constructed, one with aperture of 24 in . at the University of Indiana and another, with a $12 -$ in. aperture, at Brown University.

† See, for example, I. G. Malloff and D. W. Epstein, Electronics, 17 (1944), 98.

‡ The methods here described are due to E. Wolf and W. S. Preddy, Proc. Phys. Soc., 59 (1947), 704; also E. Wolf, Proc. Phys. Soc., 61 (1948), 494. Similar formulae were also given by R. K. Luneburg in his Mathematical Theory of Optics (mimeographed notes of lectures delivered at Brown University, R.I. (1944); printed version published by University of California Press, Berkeley and Los Angeles (1964), § 24.

$$\begin{aligned}
[NT] &= [MT] - [MN] = nZ\sec U - [MN] \\
[TQ] &= n'\sqrt{(L'-Z)^2 + (H - Z\tan U)^2} \\
[OQ] &= n'L'
\end{aligned} \tag{3}$$

$L'$ denoting the distance from $O$ to $Q$.
The optical path [ $MN$ ] which occurs in the expression for [ $NT$ ] may also be evaluated in terms of the given quantities. By applying Lagrange's invariant relation §3.3(1) to the curve formed by the segment $OM$ of the $Y$-axis, the segment $MN$ of the ray and the curve $NO$ on the wave-front $W$, we have

$$\int_{OM} ns\cdot dr + \int_{MN} ns\cdot dr + \int_{NO} ns\cdot dr = 0 \tag{4}$$

where $s$ is the unit vector along the ray and $dr$ an element of the path of integration. Now from the figure,

$$\left.\begin{array}{l}
\int_{OM} ns\cdot dr = -n\int_0^H \sin U dH = -n\int_0^t \sin U \frac{dH}{dt}dt \\
\int_{MN} ns\cdot dr = [MN] \\
\int_{NO} ns\cdot dr = 0
\end{array}\right\} \tag{5}$$

and hence (4) gives

$$[MN] = n\int_0^t \sin U\frac{dH}{dt}dt \tag{6}$$

(6) expresses $[MN]$ as an integral which may be evaluated numerically from a table of $U$ and $H$ values.*

Substitution from (3) and (6) into (2) leads to the following equation for $Z$ :

$$\mathscr{A}Z^2 + 2\mathscr{B}Z\cos U + \mathscr{C}\cos^2 U = 0 \tag{7}$$

with

$$\left.\begin{array}{ll}
\mathscr{A} & = n'^2 - n^2 \\
\mathscr{B} & = n^2\int_0^t \sin U\frac{dH}{dt}dt - n'^2\left(L'\cos U + H\sin U\right) + nn'L' \\
\mathscr{C} & = n'^2H^2 - \left(n\int_0^t \sin U\frac{dH}{dt}dt\right)\left(n\int_0^t \sin U\frac{dH}{dt}dt + 2n'L'\right)
\end{array}\right\} \tag{8}$$

Hence

$$Z = \frac{\cos U}{\mathscr{A}}\left[-\mathscr{B} \pm \sqrt{\mathscr{B}^2 - \mathscr{A}\mathscr{C}}\right] \tag{9}$$

Also, from the figure,

$$Y = H - Z\tan U. \tag{10}$$

Since $Z = U = H = 0$ when $t = 0$, and since we assume $L'$ to be positive (see Fig. 4.41), the positive sign must be taken in front of the square root in (9). Finally, combining (9) and (10), we obtain

0

- As in §4.9.1, the angle $U$ is considered to be positive if the axis can be brought into coincidence with the ray by a clockwise rotation of less than 90° about the axial point.

$$Z + iY = \frac{-\mathscr{B} + \sqrt{\mathscr{B}^2 - \mathscr{A}\mathscr{C}}}{\mathscr{A}} e^{-iU} + iH \tag{11}$$

(11) is an exact parametric equation of the aspheric surface $S$ in terms of the free parameter $t$.

The special case when the focus $Q$ is at infinity ( $L' = \infty$ ) is also of interest. To derive the appropriate formula we note that both $\mathscr{B}$ and $\mathscr{C}$ contain $L'$ in the first power only. Hence for sufficiently large $L'$,

$$\begin{aligned} -\mathscr{B} + \sqrt{\mathscr{B}^2 - \mathscr{A}\mathscr{C}} &= \mathscr{B}\left[-1 + \sqrt{1 - \frac{\mathscr{A}\mathscr{C}}{\mathscr{D}^2}}\right] \\ &= -\frac{1}{2}\frac{\mathscr{A}\mathscr{C}}{\mathscr{B}} - \frac{1}{8}\frac{\mathscr{A}^2\mathscr{C}^2}{\mathscr{B}^3} - \cdots \\ &= -\frac{1}{2}\frac{\mathscr{A}\mathscr{C}}{\mathscr{B}} + O\left(\frac{1}{L'}\right) \end{aligned}$$

In the limit, as $L' \to \infty$, (11) therefore reduces to

$$Z + iY = \frac{ne^{-iU}}{n - n'\cos U} \int_0^t \sin U \frac{dH}{dt} dt + iH. \tag{12}$$

We have only considered the case when the aspheric surface is the last surface of the system but the method may be extended to the design of an aspheric surface situated in the interior of an optical system. The computation is, however, much more laborious in such cases and will not be considered here.*

### 4.10.2 Attainment of aplanatism †

We have seen that by making one surface of a system aspherical, it is possible to attain exact axial stigmatism. We shall now consider the design of two aspheric

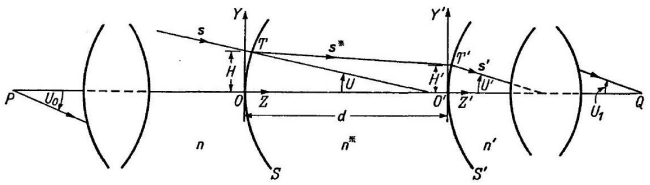



Fig. 4.42. Design of two aspheric surfaces to attain aplanatism.

surfaces which will ensure not only axial stigmatism but also the satisfaction of the sine condition.

Let $S$ and $S'$ be the two aspheric surfaces, the profiles of which are to be determined. It will be assumed that $S$ and $S'$ are neighbouring surfaces in the system. ‡ They may, however, be separated from the object or image points by any number of refracting or reflecting surfaces.

---

0

- It is, of course, possible to evaluate [ $MN$ ] directly from a ray trace by making use of the property that the optical path from $P$ to $N$ is equal to the optical path from $P$ to $O$. This gives

$$[MN] = [PO] - [PM]$$

Again, we shall be concerned only with the final correction of the system, and assume that all the design data, save the profiles of $S$ and $S'$, are known.

We introduce two sets of Cartesian rectangular axes, with origins at the poles $O$ and $O'$ of $S$ and $S'$, and with the $Z$-axes along the axis of the system. The surface $S$ will be referred to the axes at $O$, and $S'$ to the axes at $O'$.

In the space which precedes $S$, the pencil of rays from the axial object point $P$ will again be specified by a relation of the form (see Fig. 4.42)

$$U = U(t), \quad H = H(t). \tag{13}$$

The pencil of rays in the space which follows $S'$ (in the corrected system) will be specified by a similar relation:

$$U' = U'(t'), \quad H' = H'(t'). \tag{14}$$

(14) may be obtained in a tabulated form by tracing rays backwards from the selected axial image point $P'$.

If the object and image are at finite distances we choose as the parameters $t$ and $t'$ the sines of the angles which the corresponding rays in the object and image spaces make with the axis of the system:*

$$t = \sin U_0, \quad t' = \sin U_1. \tag{15}$$

If the object is at infinity, we choose as the $t$ parameter the distance $H_0$ of the corresponding ray in the object space from the axis; if the image is at infinity we take $t' = H_1, H_1$ being the distance from the axis of the corresponding ray in the image space. In either of these cases, the sine condition demands that

$$\frac{t}{t'} = \text{ constant}. \tag{16}$$

Our problem may now be formulated as follows: Given (13) and (14), to find two surfaces $S$ and $S'$ which ensure that the pencil $(U, H)$ goes over into the pencil $(U', H')$ by successive refractions at the two surfaces; and, moreover, the corresponding rays in the two pencils must satisfy the relation (16).

Let $n$ be the refractive index of the space which precedes $S, n'$ the refractive index of the space which follows $S'$, and $n^\star$ that of the space between them. Further let $s$ be the unit vector along the ray incident at the point $T(Z, Y)$ and $s^\star$ the unit vector along the refracted ray (see Fig. 4.42).

According to the law of refraction (cf. §3.2.2) the vector $N = ns - n^\star s^\star$ must be in the direction of the surface normal at $T$. Hence if $\tau$ is the unit tangent at $T$ of the meridional section of the surface,

$$(ns - n^\star s^\star) \cdot \tau = 0. \tag{17}$$

0

- For details see E. Wolf, Proc. Phys. Soc., 61 (1948), 494. Other methods were described by M. Herzberger and H. O. Hoadley, J. Opt. Soc. Amer., 36 (1946), 334; and D. S. Volosov J. Opt. Soc. Amer., 37 (1947), 342.

† The methods described in this section are due to G. D. Wassermann and E. Wolf, Proc. Phys. Soc. B, 62 (1949), 2.

‡ A generalization to systems where $S$ and $S'$ are not optical neighbours was described by E. M. Vaskas, J. Opt. Soc. Amer., 47 (1957), 669.

Now the components of the vectors $s, s^\star$ and $\tau$ are:

$$\left.\begin{array}{lccc} s: & 0, & -\sin U, & \cos U, \\ s^\star: & 0, & -\sin U^\star, & \cos U^\star, \\ \tau: & 0, & \frac{\dot{Y}}{\sqrt{\dot{Y}^2+\dot{Z}^2}}, & \frac{\dot{Z}}{\sqrt{\dot{Y}^2+\dot{Z}^2}} \end{array}\right\} \tag{18}$$

here $U^\star$ is the angle which the refracted ray $TT'$ makes with the axis of the system, and the dot denotes differentiation with respect to the parameter $t$. Equation (17) becomes

$$n(\dot{Z}\cos U - \dot{Y}\sin U) = n^\star\left(\dot{Z}\cos U^\star - \dot{Y}\sin U^\star\right). \tag{19}$$

Now if $D$ is the distance from $T$ to $T', D_y$ and $D_z$ the projections of $D$ on to the $Y$ and $Z$-axes, and $d$ the axial distance from $O$ to $O'$,

$$\cos U^\star = \frac{D_z}{D}, \quad \sin U^\star = \frac{D_y}{D}, \tag{20}$$

with

$$D_y = Y - Y', \quad D_z = d + Z' - Z, \quad D = \sqrt{D_y^2 + D_z^2}. \tag{21}$$

Also, from the figure,

$$Y = H - Z\tan U, \tag{22}$$

$$Y' = H' - Z'\tan U'. \tag{23}$$

On substituting in (19) for $\cos U^\star$ and $\sin U^\star$ from (20), and for $\dot{Y}$ from (22), we find that

$$\frac{dZ}{dt} = \left(\frac{nD\cos U - n^\star D_z}{nD\sin U - n^\star D_y} + \tan U\right)^{-1}\left(\frac{dH}{dt} - Z\frac{d}{dt}(\tan U)\right) \tag{24}$$

Similarly

$$\frac{dZ'}{dt'} = \left(\frac{n'D\cos U' - n^\star D_z}{n'D\sin U' - n^\star D_y} + \tan U'\right)^{-1}\left(\frac{dH'}{dt'} - Z'\frac{d}{dt'}(\tan U')\right) \tag{25}$$

Equations (21)-(25), subject to the relation (16) and the boundary conditions $Z = Z' = 0$ when $t = t' = 0$, enable a complete computation of the two correcting surfaces to be carried out. For, using (21), (22) and (23) we may eliminate $Y$ and $Y'$ from (24) and (25); and, using (16), we then obtain two first-order simultaneous differential equations for $Z$ and $Z'$ of the type

$$\frac{dZ}{dt} = f(Z, Z', t), \quad \frac{dZ'}{dt} = g(Z, Z', t) \tag{26}$$

These may be integrated by standard methods.* Since, however, it is necessary to determine not only $Z$ and $Z'$ but also $Y$ and $Y'$ for a selected range of the parameter $t$, it is preferable not to eliminate $Y$ and $Y'$ but rather to solve for the unknown quantities step by step.

0

- If instead of the sine condition we wished (in addition to axial stigmatism) to satisfy the Herschel condition, we would choose

$$t = \sin U_0/2, \quad t' = \sin U_1/2.$$

0

## 9.6 Geometrical Theory of Aberrations

In § 4.9 it was mentioned that within the domain of geometrical optics the departure of the path of light from the predictions of the Gaussian theory may be studied either with the help of ray tracing or by means of algebraic analysis. In the latter treatment, which forms the subject matter of this chapter, terms which involve off-axis distances in powers higher than the second in the expansion of the characteristic functions are retained. These terms represent geometrical aberrations.

The discovery of photography in 1839 by Daguerre (1789-1851) was chiefly responsible for early attempts to extend the Gaussian theory. Practical optics, which until then was mainly concerned with the construction of telescope objectives, was confronted with the new task of producing objectives with large apertures and large fields. J. Petzval, a Hungarian mathematician, attacked with considerable success the related problem of supplementing the Gaussian formulae by terms involving higher powers of the angles of inclination of rays with the axis. Unfortunately, Petzval's extensive manuscript on the subject was destroyed by thieves; what is known about this work comes chiefly from semi-popular reports.* Petzval demonstrated the practical value of his calculations by constructing in about 1840 his wellknown portrait lens (shown in Fig. 6.3(b)) which proved greatly superior to any then in existence. The earliest systematic treatment of geometrical aberrations which was published in full is due to Seidel, † who took into account all the terms of the third order in a general centred system of spherical surfaces. Since then, his analysis has been extended and simplified by many writers.

As the wave-fronts are the orthogonal trajectories of the pencils of rays it follows that the lack of homocentricity of the image-forming pencil is accompanied by departures of the associated wave-fronts from the spherical form. The knowledge of the shape of the wave-fronts is of particular importance in more refined treatments of aberrations, based on diffraction theory (see Chapter IX). For this reason, and also because of the intimate connection between the non-homocentricities of the ray pencils and the asphericities of the associated wave-fronts, we shall consider these two defects side by side. Our analysis will in part be based on important investigations of Schwarzschild ‡ adapted somewhat to this purpose.

### 9.6.1 5.1 Wave and Ray Aberrations; The Aberration Function

Consider a rotationally symmetrical optical system. Let $P_0', P_1'$ and $P_1$ be the points in which a ray from an object point $P_0$ intersects the plane of the entrance pupil, the

exit pupil, and the Gaussian image plane respectively. If $P_1^*$ is the Gaussian image of $P_0$, the vector $\delta_1 = \vec{P_1^\star}\vec{P_1}$ will be called the aberration of the ray, or simply the ray aberration (see Fig. 5.1).

Let $W$ be the wave-front through the centre $O_1'$ of the exit pupil, associated with the image-forming pencil which reaches the image space from $P_0$. In the absence of aberrations, $W$

- For example, by Adams' method (cf. E. T. Whittaker and G. Robinson, The Calculus of Observations, Glasgow, Blackie & Son, 4th ed., 1946, 363), or by the method of Runge and Kutta (cf. C. Runge and H. König, Numerisches Rechnen, Berlin, Springer, 1924).

0

- J. Petzval, Bericht ber die Ergebnisse einiger dioptrischer Untersuchungen (Pesth, 1843); also "Bericht ber optische Untersuchungen", Ber. Kais. Akad. Wien, Math. naturwiss. Kl., 24 (1857), 50, 92, 129.
† L. Seidel, Astr. Nachr., 43 (1856), No. 1027, 289, No. 1028, 305, No. 1029, 321.
‡ K. Schwarzschild, Abh. Königl. Ges. Wis. Göttingen. Math-phys. Kl., 4 (1905-1906), Nos. 1,2,3. Reprinted in Astr. Mitt. Königl. Sternwarte Göttingen (1905). An extension of Schwarzschild's analysis to systems without rotational symmetry was made by G. D. Rabinovich, Akad. Nauk. SSSR, Zh. Eksp. Teor. Fiz., 16 (1946), 161.

coincides with a sphere $S$ which is centred on the Gaussian image point $P^\star$

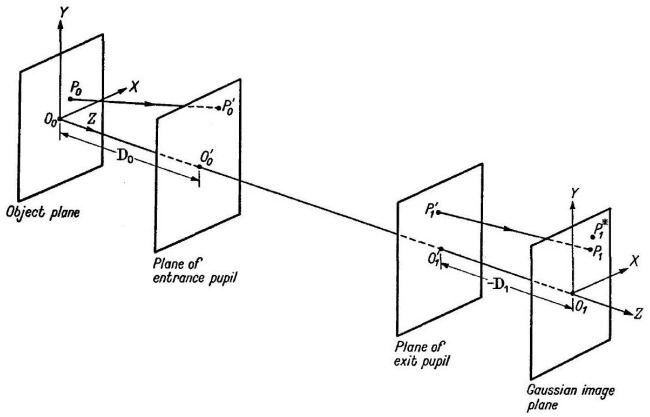



Fig. 5.1. The object plane, the image plane and the pupil planes.

and which passes through $O_1'$. $S$ will be called the Gaussian reference sphere (see Fig. 5.2).

Let $Q$ and $\bar{Q}$ be the points of intersection of the ray $P_1'P_1$ with the Gaussian reference sphere and with the wave-front $W$ respectively. The optical path length $\Phi = [\bar{Q}Q]$ may be called the aberration of the wave element at $Q$, or simply the wave aberration, and will be regarded as positive if $\bar{Q}$ and $P_1$ are on the opposite sides of $Q$. In ordinary instruments, the wave aberrations may be as much as 40 or 50 wavelengths, but in instruments used for more precise work (such as astronomical telescopes or microscopes) they must be reduced to a much smaller value, only a fraction of a wavelength.

One can easily derive expressions for the wave aberration in terms of Hamilton's point characteristic function of the system. If brackets [ . . . ] denote as before the optical path length, we have

$$\begin{aligned}\Phi &= [\bar{Q}Q] \\ &= [P_0Q] - \left[P_0\bar{Q}\right] \\ &= [P_0Q] - [P_0O_1']\,. \end{aligned} \tag{1}$$

Here use was made of the fact that $\bar{Q}$ and $O_1'$ are on the same wave-front, so that $\left[P_0\bar{Q}\right] = [P_0O_1']$.

Let us introduce two sets of mutually parallel Cartesian rectangular axes, with origins at the axial points $O_0$ and $O_1$ of the object and image planes, and with the $Z$ directions along the axis of the system. Points in the object space will be referred to the axes at $O_0$, and points in the image space to the axes at $O_1$. The $Z$ coordinates of the pupil planes will be denoted by $\mathrm{D}_0$ and $\mathrm{D}_1$ ( $\mathrm{D}_1$ is negative in Fig. 5.1).

In terms of the point characteristic $V$, the wave aberration is, according to (1), given by

$$\Phi = V\left(X_0, Y_0, 0; X, Y, Z\right) - V\left(X_0, Y_0, 0; 0, 0, \mathrm{D}_1\right), \tag{2}$$

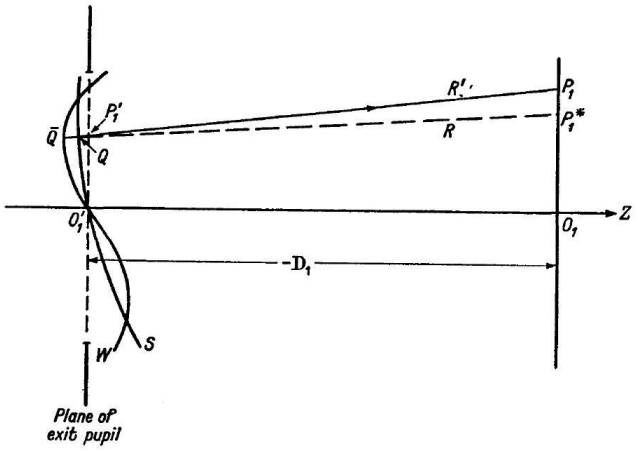



Fig. 5.2. Wave aberration and ray aberration.
(The ray $\bar{Q}$ Q $P_{1}$ does not necessarily lie in the meridional plane $O_1' O_1 P_1^\star$.)

( $X_0, Y_0$ ) being the coordinates of $P_0$ and $(X, Y, Z)$ those of $Q$. Now the coordinates ( $X, Y, Z$ ) are not independent; they are connected by a relation which expresses the fact that $Q$ lies on the Gaussian reference sphere:

$$(X - X_1^\star)^2 + (Y - Y_1^\star)^2 + Z^2 = R^2. \tag{3}$$

Here

$$X_1^\star = MX_0, \quad Y_1^\star = MY_0, \tag{4}$$

are the coordinates of the Gaussian image point $P_1^\star$, $M$ is the Gaussian lateral magnification between the object and image planes, and $R$ is the radius of the Gaussian reference sphere

$$R = \left(X_1^{*2} + Y_1^{*2} + \mathbf{D}_1^2\right)^{1/2}. \tag{5}$$

Using (3), $Z$ may be eliminated from (2), so that $\Phi$ may be regarded as a function of $X_0, Y_0, X$ and $Y$ only, i.e.*

$$\Phi = \Phi\left(X_0, Y_0; X, Y\right).$$

In terms of the aberration function $\Phi\left(X_0, Y_0; X, Y\right)$, simple expressions for the ray aberrations may be obtained. We have, from (2),

$$\frac{\partial \Phi}{\partial X} = \frac{\partial V}{\partial X} + \frac{\partial V}{\partial Z}\frac{\partial Z}{\partial X}. \tag{6}$$

If $\alpha_1, \beta_1$ and $\gamma_1$ are the angles which the ray $QP_1$ makes with the axes, and $(X, Y, Z)$ and ( $X_1, Y_1, 0$ ) are the coordinates of the points $Q$ and $P_1$ respectively, we have, according to §4.1 (7), and from Fig. 5.2,

$$\frac{\partial V}{\partial X} = n_1 \cos\alpha_1 = n_1 \frac{X_1 - X}{R'}, \quad \frac{\partial V}{\partial Z} = n_1 \cos\gamma_1 = -n_1 \frac{Z}{R'}, \tag{7}$$

where

0

- A more general aberration function, suitable for discussing the imaging of extended objects' was introduced by E. Wolf, J. Opt. Soc. Amer., 42 (1952), 547.

$$R' = \left\{(X_1 - X)^2 + (Y_1 - Y)^2 + Z^2\right\}^{1/2} \tag{8}$$

is the distance from $Q$ to $P_1$ and $n_1$ is the refractive index of the image space. Further, from (3),

$$\frac{\partial Z}{\partial X} = -\frac{X - X_1^*}{Z}. \tag{9}$$

Substitution from (7) and (9) into (6) gives the following expressions for the ray aberration components:
and similarly

$$\left.\begin{aligned} X_1 - X_1^\star &= \tfrac{R'}{n_1}\tfrac{\partial \Phi}{\partial X'} \\ Y_1 - Y_1^\star &= \tfrac{R'}{n_1}\tfrac{\partial \Phi}{\partial Y}. \end{aligned}\right\} \tag{10}$$

The relations (10) are exact, but they involve, on the right-hand side, the distance $R'$, which itself depends on the coordinates of $P_1$, i.e. on the ray aberrations*. However, for most practical purposes $R'$ may be replaced by the radius $R$ of the Gaussian reference sphere or by another approximation (cf. eq. (15) below).

It is easily seen that, on account of symmetry, $\Phi$ depends on the four variables only through the three combinations ${X_0}^2 + {Y_0}^2, X^2 + Y^2$ and $X_0X + Y_0Y$. For if polar coordinates are introduced in the $XY$-planes, i.e. if we set

$$\left.\begin{aligned} X_0 &= r_0 \cos\theta_0, \\ Y_0 = r_0 \sin\theta_0, \quad X &= r\cos\theta, \\ Y &= r\sin\theta, \end{aligned}\right\} \tag{11}$$

$\Phi$ becomes a function of $r_0, r, \theta_0$ and $\theta$; or, what amounts to the same thing, a function of $r_0, r, \theta_0 - \theta$ and $\theta$. Suppose now that the $X$ - and $Y$-axes at $O_0$ and $O_1$ are rotated through the same angle and in the same sense about the axis of the system. This leaves $r_0, r$ and $\theta_0 - \theta$ unchanged, whilst $\theta$ increases by the angle of rotation. Since $\Phi$ is invariant with respect to such rotations, it must be independent of the last variable, i.e. it depends on $r_0, r$ and $\theta_0 - \theta$ only. Hence the aberration function $\Phi$ may be expressed as a function of the three scalar products

$$r_0^2 = X_0^2 + Y_0^2, \quad r^2 = X^2 + Y^2 \quad \text{and} \quad r_0 \cdot r = X_0X + Y_0Y, \tag{12}$$

of the two vectors $r_0\,(X_0, Y_0)$ and $r(X, Y)$.
From this result it follows that, if $\Phi$ is expanded into a power series with respect to the four coordinates, the expansion will contain only terms of even degree. There will be no terms of zero degree, since $\Phi(0, 0; 0, 0) = 0$. Moreover, there will be no terms of the second degree, except possibly for a term proportional to $( {X_0}^2 + {Y_0}^2 )$; for according to (10) such terms would give rise to ray aberrations depending linearly on the coordinates, and this contradicts the fact that $P_1^*$ is the Gaussian image of $P_0$. Hence the expansion is of the form

$$\Phi = c\left(X_0^2 + Y_0^2\right) + \Phi^{(4)} + \Phi^{(6)} + \ldots, \tag{13}$$

where $c$ is a constant and $\Phi^{(2k)}$ is a polynomial of degree $2k$ in the coordinates, and contains these coordinates in powers of the three scalar invariants (12) only. A term of a particular degree $2k$ is said to represent a wave aberration of order $2k$. The aberrations of the lowest

0

- A somewhat different pair of exact equations which connect the ray aberrations and the wave aberrations was derived by J. L. Rayces, Optica Acta, 11 (1964), 85.

order ($2k = 4$) are usually called primary or Seidel aberrations* and will be studied in detail in § 5.3.

To show the order of magnitude of certain expressions and the degree of approximation involved in some of our calculations, it is convenient to introduce a parameter $\mu$. This may be any quantity of the first order, e.g. the angular aperture of the system. Then all the rays passing through the system may be assumed to make angles $O(\mu)$ with the optical axis, where the symbol $O(\mu)$ means "not exceeding a moderate multiple" of $\mu$.

Consider the error involved on replacing $R'$ in the basic relations (10) by quantities independent of $X_1$ and $Y_1$. From (3) and (5),

$$Z^2 = \mathrm{D}_1^2 - \left(X^2 + Y^2\right) + 2\left(XX_1^* + YY_1^*\right) \tag{14}$$

and (8) then gives

$$\begin{aligned} R' &= -\mathrm{D}_1 \left[1 + \frac{X_1^2 + Y_1^2 - 2X\left(X_1 - X_1^*\right) - 2Y\left(Y_1 - Y_1^*\right)}{\mathrm{D}_1^2}\right]^{1/2} \\ &= -\mathrm{D}_1 - \frac{X_1^{\star 2} + Y_1^{\star 2}}{2\mathrm{D}_1} + O\left(\mathrm{D}_1\mu^4\right). \end{aligned} \tag{15}$$

The relations (10) for the ray aberration components then become

$$X_1 - X_1^* = -\frac{1}{n_1}\left(\mathrm{D}_1 + \frac{X_1^{*2} + Y_1^{*2}}{2\mathrm{D}_1}\right)\frac{\partial\Phi}{\partial X} + O\left(\mathrm{D}_1\mu^7\right) \tag{16a}$$

$$= -\frac{\mathrm{D}_1}{n_1}\frac{\partial\Phi}{\partial X} + O\left(\mathrm{D}_1\mu^5\right) \tag{16b}$$

$$Y_1 - Y_1^* = -\frac{1}{n_1}\left(\mathrm{D}_1 + \frac{X_1^{*2} + Y_1^{*2}}{2\mathrm{D}_1}\right)\frac{\partial\Phi}{\partial Y} + O\left(\mathrm{D}_1\mu^7\right) \tag{17a}$$

$$= -\frac{\mathrm{D}_1}{n_1}\frac{\partial\Phi}{\partial Y} + O\left(\mathrm{D}_1\mu^5\right) \tag{17b}$$

## 9.6.2 5.2 The Perturbation Eikonal of Schwarzschild

In his researches on geometrical aberrations, Schwarzschild employed a method similar to that used in the calculations of orbital elements in celestial mechanics. In such calculations variables are introduced which remain constant in the unperturbed motion, and the small changes which these quantities undergo in the actual motion are then determined with the help of a perturbation function. By analogy with this procedure, Schwarzschild introduced, in the papers already referred to, certain variables which, within the accuracy of Gaussian optics, have constant values along each ray passing through the optical system. With the help of a certain perturbation function which he introduced and called the Seidel eikonal, he then studied the changes which these variables undergo when the fourth-order terms in the expansion

of the characteristic function are taken into account.* Schwarzschild called these special variables the Seidel variables, as they are related to those employed previously by Seidel.

Within the accuracy of the Seidel theory, the aberration function $\Phi$ defined in the preceding section is very closely related to the perturbation eikonal of Schwarzschild; and by following

0

- Since the ray aberrations associated with wave aberrations of this order are of the third degree in the coordinates, they are also sometimes called the third-order aberrations.

closely Schwarzschild's procedure one can derive the expressions for the fourth coefficients in the expansion of the aberration function for any centred system. This will be done in detail in §5.5. Here the Seidel variables will be defined and the connection between our aberration function and the perturbation function of Schwarzschild will be examined.

Let us introduce new units of length $l_0$ in the object plane and $l_1$ in the Gaussian image plane, such that

$$\frac{l_1}{l_0} = M \tag{1}$$

is the lateral Gaussian magnification between the two planes. Points in the object plane will be specified by the coordinates $x_0, y_0$, and points in the image space by the coordinates $x_1, y_1$, such that

$$\left.\begin{array}{ll} x_0 = \mathrm{C}\frac{X_0}{l_0}, & x_1 = \mathrm{C}\frac{X_1}{l_1} \\ y_0 = \mathrm{C}\frac{Y_0}{l_0}, & y_1 = \mathrm{C}\frac{Y_1}{l_1} \end{array}\right\} \tag{2}$$

where $(X_0, Y_0), (X_1, Y_1)$ are the ordinary coordinates of $P_0$ and $P_1$ (see Fig. 5.1), and $C$ is a constant to be specified later. Within the accuracy of Gaussian optics, $x_1 = x_0$ and $y_1 = y_0$.

The coordinates $( X_0', Y_0' )$ of the points where the ray from $( X_0, Y_0 )$ meets the entrance pupil are connected with the ray components by the relations

$$\frac{X_0' - X_0}{\mathrm{D}_0} = \frac{p_0}{\sqrt{n_0^2 - p_0^2 - q_0^2}}, \quad \frac{Y_0' - Y_0}{\mathrm{D}_0} = \frac{q_0}{\sqrt{n_0^2 - p_0^2 - q_0^2}}, \tag{3}$$

the expressions for the intersection points of the ray with the exit pupil being strictly similar. Within the accuracy of Gaussian optics, the square roots in the denominators may be replaced by $n_0$ and $n_1$ respectively, and the following linear relations between the coordinates are then obtained:

$$\left.\begin{array}{ll} \frac{X_0'-X_0}{\mathrm{D}_0} = \frac{p_0}{n_0}, & \frac{X_1'-X_1}{\mathrm{D}_1} = \frac{p_1}{n_1} \\ \frac{Y_0'-Y_0}{\mathrm{D}_0} = \frac{q_0}{n_0}, & \frac{Y_1'-Y_1}{\mathrm{D}_1} = \frac{q_1}{n_1} \end{array}\right\} \tag{4}$$

Next we introduce new units of length $\lambda_0$ and $\lambda_1$ in the plane of the entrance and the exit pupil such that

$$\frac{\lambda_1}{\lambda_0} = M' \tag{5}$$

is the lateral magnification between the two planes. In place of $X_0', Y_0', X_1', Y_1'$, the following variables will be used:

$$\left.\begin{array}{ll} \xi_0 = \frac{X_0'}{\lambda_0} = \frac{X_0}{\lambda_0} + \frac{\mathrm{D}_0 p_0}{\lambda_0 n_0}, & \xi_1 = \frac{X_1'}{\lambda_1} = \frac{X_1}{\lambda_1} + \frac{\mathrm{D}_1 p_1}{\lambda_1 n_1} \\ \eta_0 = \frac{Y_0'}{\lambda_0} = \frac{Y_0}{\lambda_0} + \frac{\mathrm{D}_0 q_0}{\lambda_0 n_0}, & \eta_1 = \frac{Y_1'}{\lambda_1} = \frac{Y_1}{\lambda_1} + \frac{\mathrm{D}_1 q_1}{\lambda_1 n_1} \end{array}\right\} \tag{6}$$

---

0

- Schwarzschild also considered the general types of fifth-order ray aberrations. Expressions for the associated sixth-order coefficients in the expansion of the characteristic function (Schwarzschild's perturbation eikonal) were first given by his pupil A. Kohlschtter in his Dissertation (Göttingen, 1908).

Fifth-order aberrations were also investigated by M. Herzberger, J. Opt.Soc. Amer., 29 (1939), 395, and also his Modern Geometrical Optics (New York, Interscience Publishers, 1958), H. A. Buchdahl, Optical Aberration Coefficients (Oxford University Press, 1954) and by other workers. A review of higher order aberration theory was published by J. Focke in Progress in Optics, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 1.

Within the accuracy of Gaussian optics, $\xi_1 = \xi_0, \eta_1 = \eta_0$.
In order to simplify later calculations, it will be convenient to choose $C$ as*

$$\mathrm{C} = \frac{n_0 l_0 \lambda_0}{\mathrm{D}_0} = \frac{n_1 l_1 \lambda_1}{\mathrm{D}_1} \tag{7}$$

The equality of the two terms on the right-hand side follows from the SmithHelmholtz formula (§ 4.4 (48)).

The quantities defined by (2) and (6) are the Seidel variables. The inverse relations expressing the old variables in terms of the Seidel variables will also be needed. We have, on solving (2) and (6):

$$\left.\begin{array}{ll} X_0 = \frac{\mathrm{D}_0}{n_0\lambda_0} x_0, & X_1 = \frac{\mathrm{D}_1}{n_1\lambda_1} x_1, \\ Y_0 = \frac{\mathrm{D}_0}{n_0\lambda_0} y_0, & Y_1 = \frac{\mathrm{D}_1}{n_1\lambda_1} y_1, \end{array}\right\} \tag{9}$$

Next the aberration function will be expressed in terms of the Seidel variables. We note first of all that the arguments $X$ and $Y$ may be replaced by $X_1'$ and $Y_1'$ in $\Phi$ without changing the error term in equations (16b) and (17b) of §5.1. Let us denote by $\phi$ the aberration function, when regarded as a function of the Seidel variables,

$$\Phi\left(X_0, Y_0; X_1', Y_1'\right) = \phi\left(x_0, y_0; \xi_1, \eta_1\right). \tag{10}$$

$$\begin{array}{ll} x_0 = kn_0 X_0 \sin\theta_0, & x_1 = -kn_1 X_1 \sin\theta_1. \\ y_0 = kn_0 Y_0 \sin\theta_0, & y_1 = -kn_1 Y_1 \sin\theta_1. \end{array}$$

Then

$$\frac{\partial\Phi}{\partial X_1'} = \frac{\partial\phi}{\partial\xi_1}\frac{\partial\xi_1}{\partial X_1'} = \frac{1}{\lambda_1}\frac{\partial\phi}{\partial\xi_1}, \tag{11a}$$

and from eqs. (2), (1) and (7) and § 5.1 (4)

$$X_1 - X_1^\star = \frac{\mathrm{D}_1}{n_1\lambda_1}\left(x_1 - x_0\right) \tag{llb}$$

With the help of eqs. (11) the formulae (16b) and (17b) of §5.1 give

$$\left.\begin{array}{l} x_1 - x_0 = -\frac{\partial\phi}{\partial\xi_1} + O\left(\mathrm{D}_1\mu^5\right) \\ y_1 - y_0 = -\frac{\partial\phi}{\partial\eta_1} + O\left(\mathrm{D}_1\mu^5\right) \end{array}\right\} \tag{12}$$

It was mentioned earlier that within the accuracy of the Seidel theory, $\phi$ is closely related to a perturbation function introduced by Schwarzschild and called by him the Seidel eikonal. This perturbation function is defined by the expression

$$\psi = T + \frac{\mathrm{D}_0}{2n_0\lambda_0^2}\left(x_0^2 + y_0^2\right) - \frac{\mathrm{D}_1}{2n_1\lambda_1^2}\left(x_1^2 + y_1^2\right) + x_0\left(\xi_1 - \xi_0\right) + y_0\left(\eta_1 - \eta_0\right), \tag{13}$$

0

- Another choice of $C$, which is often useful in the study of images on the basis of diffraction theory, is

$$C = kn_0 l_0 \sin\theta_0 = -kn_1 l_1 \sin\theta_1.$$

Here $k$ is the vacuum wave number of the light and $2\theta_0$ and $2\theta_1$ are the angular apertures of the system, i.e. the angles which the entrance pupil and the exit pupil subtend at the axial object and image points respectively. With this choice of $C$,

where $T = T(p_0, q_0; p_1, q_1)$ is the angle characteristic, referred to origins at $O_0$ and $O_1$. Consider now the effect of small variations in the coordinates. We have, according to eq. (27) of §4.1, with $Z_0 = Z_1 = 0$, that the corresponding change in $T$ is given by

$$\delta T = X_0 dp_0 + Y_0 dq_0 - X_1 dp_1 - Y_1 dq_1 \tag{14}$$

or, in terms of the Seidel variables,

$$\begin{aligned}\delta T = x_0 \left( d\xi_0 - \frac{\mathrm{D}_0}{n_0 \lambda_0^2} dx_0 \right) + y_0 \left( d\eta_0 - \frac{\mathrm{D}_0}{n_0 \lambda_0^2} dy_0 \right) \\ - x_1 \left( d\xi_1 - \frac{\mathrm{D}_1}{n_1 \lambda_1^2} dx_1 \right) - y_1 \left( d\eta_1 - \frac{\mathrm{D}_1}{n_1 \lambda_1^2} dy_1 \right)\end{aligned} \tag{15}$$

Using (15), we deduce from (13) that a small variation in the variables leads to a change in $\psi$ given by

$$d\psi = (\xi_1 - \xi_0)\, dx_0 + (\eta_1 - \eta_0)\, dy_0 + (x_0 - x_1)\, d\xi_1 + (y_0 - y_1)\, d\eta_1 \tag{16}$$

so that $\psi$ may be expressed as function of $x_0, y_0, \xi_1, \eta_1$ and we then have rigorously

$$\left.\begin{array}{ll} \xi_1 - \xi_0 = \frac{\partial \psi}{\partial x_0}, & x_1 - x_0 = -\frac{\partial \psi}{\partial \xi_1} \\ \eta_1 - \eta_0 = \frac{\partial \psi}{\partial y_0}, & y_1 - y_0 = -\frac{\partial \psi}{\partial \eta_1} \end{array}\right\} \tag{17}$$

Hence from the knowledge of $\psi$ the ray aberrations can be determined, both in the image plane and in the plane of the exit pupil by simple differentiation.

Comparison of (17) and (12) shows that within the accuracy of the Seidel theory $\phi - \psi$ must be independent of $\xi_1$ and $\eta_1$, i.e.

$$\phi(x_0, y_0; \xi_1, \eta_1) = \psi(x_0, y_0; \xi_1, \eta_1) + \chi(x_0, y_0) + O\left(\mathrm{D}_1 \mu^6\right) \tag{18}$$

where $\chi(x_0, y_0)$ is some function of $x_0$ and $y_0$. Now from the definition of $\phi$, $\phi(x_0, y_0; 0, 0) = 0$. Hence $\chi(x_0, y_0) = -\psi(x_0, y_0; 0, 0)$ and consequently

$$\phi(x_0, y_0; \xi_1, \eta_1) = \psi(x_0, y_0; \xi_1, \eta_1) - \psi(x_0, y_0; 0, 0) + O\left(\mathrm{D}_1 \mu^6\right) \tag{19}$$

Within the range of validity of the Seidel theory, the error terms in (12) may be neglected. If, however, terms of order higher than fourth are taken into account, the expressions for the ray aberration components in terms of the aberration function $\phi$ are more complicated. On the other hand, the simple relations (17) for the ray aberration components in terms of the perturbation eikonal are exact; the perturbation eikonal appears, however, to have no simple physical meaning.

The determination of terms of order higher than fourth is very laborious in all but the simplest cases. For this reason, algebraic calculations are usually restricted to the domain of the Seidel theory, supplemented where necessary by ray tracing.

## 9.6.3 5.3 The Primary (Seidel) Aberrations

By means of considerations strictly similar to those of § 5.1 relating to the aberration function, it follows on account of symmetry that the power series expansion of the perturbation eikonal of Schwarzschild is of the form

$$\psi = \psi^{(0)} + \psi^{(4)} + \psi^{(6)} + \psi^{(8)} + \ldots \tag{1}$$

where $\psi^{(2k)}$ is a polynomial of degree $2k$ in the four variables; and moreover, that the four variables enter only in the three combinations

$$r^2 = x_0^2 + y_0^2, \quad \rho^2 = \xi_1^2 + \eta_1^2, \quad \kappa^2 = x_0\xi_1 + y_0\eta_1. \tag{2}$$

In (1) the term of the second degree is absent, for the presence of such a term would, according to §5.2 (17), contradict the fact that within the accuracy of Gaussian optics, $x_1 = x_0, y_1 = y_0, \xi_1 = \xi_0$, and $\eta_1 = \eta_0$.

Since the variables enter only in the combinations (2), it follows that $\psi^{(4)}$ must be of the form

$$\psi^{(4)} = -\frac{1}{4}A\mathbf{r}^4 - \frac{1}{4}B\rho^4 - C\kappa^4 - \frac{1}{2}D\mathbf{r}^2\rho^2 + E\mathbf{r}^2\kappa^2 + F\rho^2\kappa^2 \tag{3}$$

where $A, B, \ldots$ are constants. The signs and the numerical factors in (3) have been chosen in agreement with the usual practice; they lead to convenient expressions for the ray aberrations. The evaluation of the aberration constants for any centred system will be discussed in § 5.5.

Clearly the power series expansion of $\phi$ is of the same form as (1), but it contains no term of zero order ( $\phi^{(0)} = 0$ ), and the leading term $\phi^{(4)}$ differs from $\psi^{(4)}$ by the absence of the term $-\frac{1}{4}Ar^4$, as is immediately obvious from eq. (19) of §5.2. Hence the general expression for the wave aberration of the lowest order (fourth) is

$$\phi^{(4)} = -\frac{1}{4}B\rho^4 - C\kappa^4 - \frac{1}{2}Dr^2\rho^2 + Er^2\kappa^2 + F\rho^2\kappa^2 \tag{4}$$

$B, C, \ldots$ being the same coefficients as in (3).

Substitution from (4) into eq. (12) of §5.2 gives the general expression for the ray aberration components of the lowest (third) order:

$$\left.\begin{aligned} \Delta^{(3)}x &= x_1 - x_0 = \tfrac{n_1\lambda_1}{\mathrm{D}_1}\,(X_1 - X_1^\star) \\ &= x_0\,(2C\kappa^2 - E\mathrm{r}^2 - F\rho^2) + \xi_1\,(B\rho^2 + D\mathrm{r}^2 - 2F\kappa^2)\,, \\ \Delta^{(3)}y &= y_1 - y_0 = \tfrac{n_1\lambda_1}{\mathrm{D}_1}\,(Y_1 - Y_1^\star) \\ &= y_0\,(2C\kappa^2 - E\mathrm{r}^2 - F\rho^2) + \eta_1\,(B\rho^2 + D\mathrm{r}^2 - 2F\kappa^2)\,. \end{aligned}\right\} \tag{5}$$

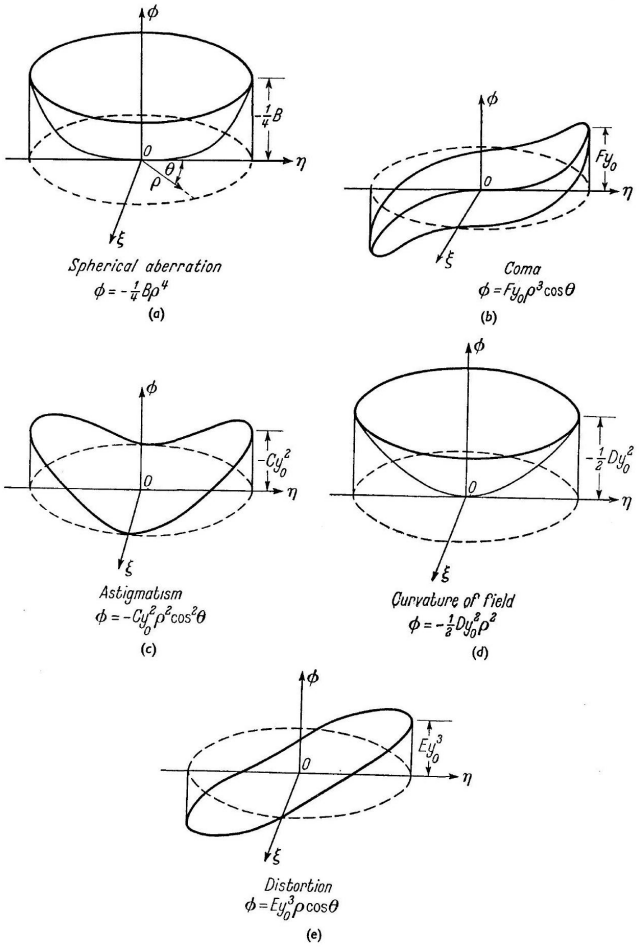



Fig. 5.3. The primary wave aberrations.

The coefficient $A$ does not enter (4) and (5), so that there are altogether five types of aberrations of the lowest order characterized by the five coefficients $B, C, D, E$ and $F$. As already mentioned on p. 207 these aberrations are known as primary or Seidel aberrations.

To discuss the effect of the Seidel aberrations it will be convenient to choose the axes so that the $yz$-plane passes through the object point; then $x_0 = 0$. If further polar coordinates are introduced, so that
(4) becomes

$$\xi_1 = \rho\sin\theta, \quad \eta_1 = \rho\cos\theta \tag{6}$$

$$\phi^{(4)} = -\frac{1}{4}B\rho^4 - Cy_0^2\rho^2\cos^2\theta - \frac{1}{2}Dy_0^2\rho^2 + Ey_0^3\rho\cos\theta + Fy_0\rho^3\cos\theta \tag{7}$$

and (5) takes the form

$$\left.\begin{aligned}\Delta^{(3)}x &= B\rho^3\sin\theta - 2Fy_0\rho^2\sin\theta\cos\theta + Dy_0^2\rho\sin\theta\\ \Delta^{(3)}y &= B\rho^3\cos\theta - Fy_0\rho^2(1+2\cos^2\theta) + (2C+D)y_0^2\rho\cos\theta - Ey_0^3.\end{aligned}\right\} \tag{8}$$

In the special case when all the coefficients in (7) have zero values, the wave-front in the exit pupil coincides (within the present degree of accuracy) with the Gaussian reference sphere (cf. Fig. 5.2). In general, the coefficients will have finite values. Each term then represents a particular type of departure of the wave-front from the ideal spherical form; the five different types are illustrated in Fig. 5.3.
The significance of the ray aberrations associated with a given object point may be illustrated

graphically by means of so-called aberration (or characteristic) curves. These are the curves traced out in the image plane by the intersection points of all the rays emerging from a fixed zone $\rho =$ constant of the exit pupil. The area covered by the aberration curves which correspond to all the possible values of $\rho$ then represent the imperfect image.

We shall consider in turn each of the Seidel aberrations.*

I. Spherical aberration ( $B \neq 0$ )

When all the coefficients except $B$ are zero, (8) reduces to

$$\left.\begin{aligned} \Delta^{(3)}x &= B\rho^3 \sin\theta \\ \Delta^{(3)}y &= B\rho^3 \cos\theta \end{aligned}\right\} \tag{9}$$

The aberration curves are therefore concentric circles whose centres are at the Gaussian image point and whose radii increase with the third power of the zonal radius $\rho$, but are independent of the position ( $y_0$ ) of the object in the field of view. This defect of the image is known as the spherical aberration.

The spherical aberration, being independent of $y_0$, affects the off-axis as well as the axial image. The rays from an axial object point which make an appreciable angle with the axis will intersect the axis in points which lie in front of or behind the Gaussian focus (Fig. 5.4). The point at which the rays from the edge of the aperture intersect the axis is called the marginal focus. If a screen is placed in the image region at right angles to the axis, there is a position for which the circular image spot appearing on the screen is a minimum; this minimal "image" is called the circle of least confusion.

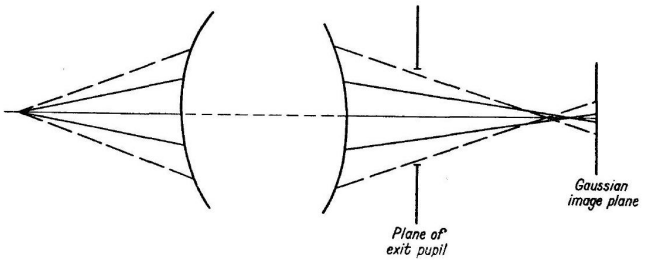



Fig. 5.4. Spherical aberration.

II. Coma($F \neq 0$)

The aberration characterized by the coefficient $F$ is known as coma. According to (8) the ray aberration components are in this case

$$\left.\begin{aligned} \Delta^{(3)}x &= -2F\rho^2 y_0 \sin\theta\cos\theta = -Fy_0\rho^2 \sin 2\theta \\ \Delta^{(3)}y &= -F\rho^2 y_0\,(1 + 2\cos^2\theta) = -Fy_0\rho^2(2 + \cos 2\theta). \end{aligned}\right\} \tag{10}$$

It is seen that, if $y_0$ is fixed and the zonal radius $\rho$ is kept constant, the point $P_1$ in the image plane describes a circle twice over, as $\theta$ runs through the range $0 \leqslant \theta < 2\pi$.

0

- Aberration curves associated with aberrations of higher orders are discussed by G. C. Steward, Trans. Camb. Phil. Soc., 23 (1926), 235 and by N. Chako, Trans. Chalmers University of Technology Gothenburg), Nr. 191 (1957).

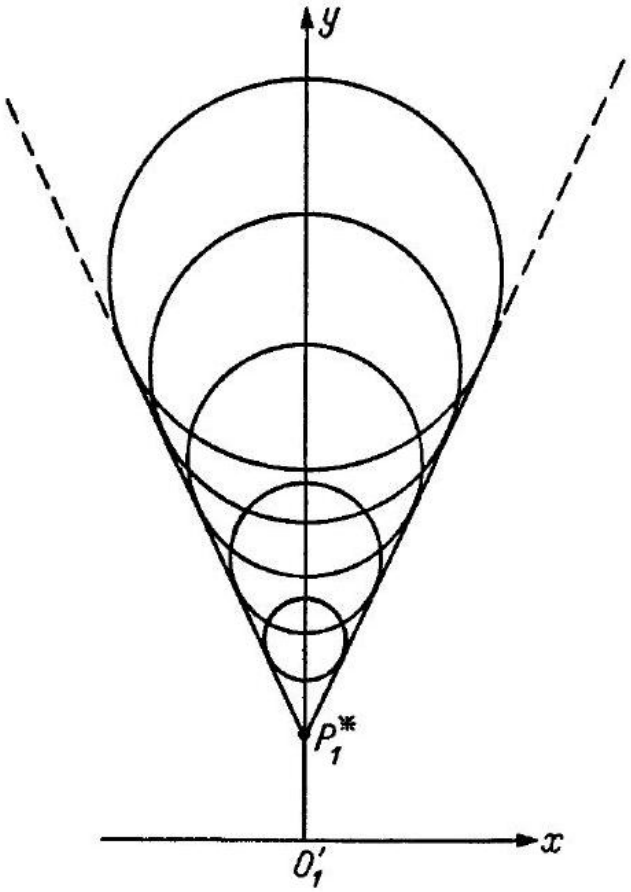



Fig. 5.5. Coma.

III. Astigmatism ( $C \neq 0$ ) and curvature of field ( $D \neq 0$ )

The effects of the aberrations which are characterized by the coefficients $C$ and $D$ can best be studied together. We have from (8), if all the other coefficients have zero values,

$$\left.\begin{aligned} \Delta^{(3)}x &= D\rho {y_0}^2 \sin\theta \\ \Delta^{(3)}y &= (2C + D)\rho {y_0}^2 \cos\theta. \end{aligned}\right\} \tag{11}$$

To see the significance of these aberrations, assume to begin with that the imageforming pencil is very narrow. According to § 4.6, the rays of such a pencil intersect two short lines, of which one (the tangential focal line) is at right angles to the meridional plane, and the other (the sagittal focal line) lies in this plane. Consider now light from all the points in a finite region of the object plane. The focal lines in the image space give rise to two surfaces, the tangential and the sagittal focal surface. To a first approximation these surfaces may be considered to be spheres. Let $R_t$ and $R_s$ be their radii, regarded as positive if the corresponding centre of curvature lies on

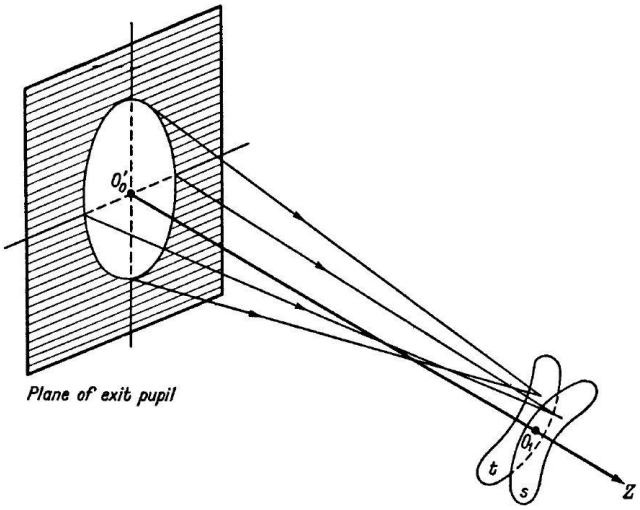



Fig. 5.6. The tangential and sagittal focal surfaces.

the side of the image plane from which the light is propagated ( $R_t > 0, R_s < 0$ in Fig. 5.6).

The radii of curvature may be expressed in terms of the coefficients $C$ and $D$. To show this it will be convenient to calculate the ray aberrations in the presence of curvature, using first the ordinary rather than the Seidel coordinates. We have (see Fig. 5.7)

$$\frac{\Delta^{(3)}Y_1}{Y_1'} = \frac{u}{\mathrm{D}_1 + u} \tag{12}$$

where $u$ is the small distance between the sagittal focal line and the image plane. If $v$ denotes the distance of this focal line from the axis,

$$R_t{}^2 = v^2 + (R_t - u)^2$$

i.e.

$$v^2 = 2R_t u - u^2.$$

If $u$ is regarded as a small quantity of the first order, $v$ may be replaced by $Y_1$, and $u^2$ may be neglected in the last equation, so that

$$u \sim \frac{Y_1{}^2}{2R_t}, \tag{13}$$

and (12) gives, if we also neglect $u$ in comparison with $\mathrm{D}_1$,

$$\Delta^{(3)}Y_1 = \frac{Y_1^2}{2R_t}\frac{Y_1'}{\bar{D}_1}. \tag{14}$$

Similarly

$$\Delta^{(3)}X_1 = \frac{Y_1^2}{2R_s}\frac{X_1'}{\mathrm{D}_1}. \tag{15}$$

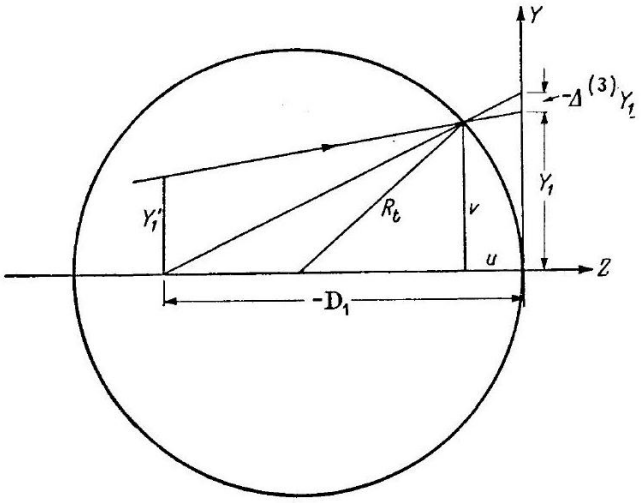



Fig. 5.7. Astigmatism and curvature of field.

Next these relations will be expressed in terms of the Seidel variables. Substitution from eqs. (6) and (8) of §5.2 leads to

$$\Delta^{(3)}y = \frac{n_1\lambda_1}{\mathrm{D}_1}\Delta^{(3)}Y_1 = \frac{n_1\lambda_1}{\mathrm{D}_1}\frac{{Y_1}^2}{2R_t}\frac{Y_1'}{\mathrm{D}_1} = \frac{n_1\lambda_1}{2{\mathrm{D}_1}^2R_t}\frac{{\mathrm{D}_1}^2}{{n_1}^2{\lambda_1}^2}{y_1}^2\lambda_1\eta_1$$

i.e.

$$\Delta^{(3)}y = \frac{y_1^2\eta_1}{2n_1R_t} \tag{16}$$

and similarly

$$\Delta^{(3)}x = \frac{{y_1}^2\xi_1}{2n_1R_3}. \tag{17}$$

In (16) and (17) $y_1$ may be replaced by $y_0$, and we finally obtain, on comparison with (11) and on using (6),

$$\frac{1}{R_t} = 2n_1(2C + D), \quad \frac{1}{R_s} = 2n_1D. \tag{18}$$

The quantity $2C + D$ is usually called the tangential field curvature and $D$ the sagittal field curvature, and the quantity

$$\frac{1}{R} = \frac{1}{2}\left(\frac{1}{R_t} + \frac{1}{R_s}\right) = 2n_1(C + D), \tag{19}$$

which is proportional to their arithmetic mean is called, simply, the field curvature.

From (13) and (18) it is seen that, at height $Y_1$ from the axis, the separation between the two focal surfaces (i.e. the astigmatic distance of the image forming pencil) is

$$\frac{Y_1^2}{2}\left(\frac{1}{R_t} - \frac{1}{R_s}\right) = 2n_1CY_1^2. \tag{20}$$

The semi-difference

$$\frac{1}{2}\left(\frac{1}{R_t} - \frac{1}{R_s}\right) = 2n_1C \tag{21}$$

is accordingly called astigmatism. In the absence of astigmatism ( $C = 0$ ), $R_t = R_s = R$. It will be shown in §5.5.3 that the radius $R$ of the common focal surface may then be calculated from a simple formula which only involves the radii of curvature of the individual surfaces of the system and the refractive indices of all the medii.

IV. Distortion ( $E \neq 0$ )

If the $E$ coefficient alone differs from zero, we have according to (8)

$$\left.\begin{aligned} \Delta^{(3)}x &= 0 \\ \Delta^{(3)}y &= -Ey_0{}^3. \end{aligned}\right\} \tag{22}$$

Since these expressions are independent of $\rho$ and $\theta$, the imaging will be stigmatic and independent of the radius of the exit pupil; the off-axis distance of the image will, however, not be proportional to that of the object. This aberration is therefore called distortion.

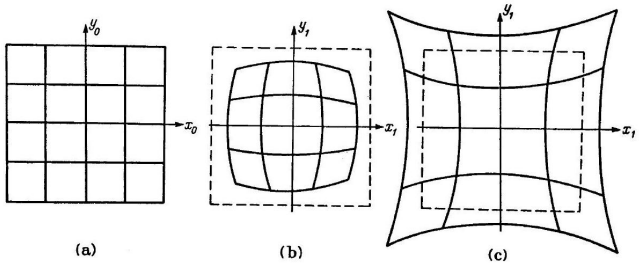



Fig. 5.8.

(a) Object

(b) Image in the presence of barrel distortion ( $E > 0$ )

(c) Image in the presence of pincushion distortion ( $E < 0$ )

If distortion is present, the image of any straight line in the object plane which meets the axis will itself be a straight line, but the image of any other straight line will be curved. This effect is shown in Fig. 5.8 (a), where the object has the form of a mesh formed by equidistant lines parallel to the $x$ - and $y$-axes. Fig. 5.8 (b) illustrates the so-called barrel distortion ( $E > 0$ ), and Fig. 5.8 (c) illustrates pincushion distortion ( $E < 0$ ).

It has been seen that, of the five Seidel aberrations, three (namely spherical aberration, coma and astigmatism) are responsible for lack of sharpness of the image.

The other two (namely curvature of field and distortion) are related to the position and the form of the image. In general it is impossible to design a system which is free from all the primary as well as the higher order aberrations, and a suitable compromise as to their relative magnitudes has, therefore, to be made. In some cases the effects of the Seidel aberrations are reduced by balancing them against aberrations of higher orders. In others one has to eliminate certain aberrations completely even at the price of introducing aberrations of other types. For example, because of the asymmetric appearance of an image in the presence of coma, this aberration must always be suppressed in telescopes, as it would make precise positional measurements impracticable. On the other hand a certain amount of curvature of the field and of distortion is then relatively harmless as these can be eliminated by calculation.

So far we have studied the aberration effects on the basis of geometrical optics only. If, however, the aberrations are very small (wave aberrations of the order of a wavelength or less), diffraction plays an important part. The geometrical theory must then be supplemented by more refined considerations. This will be done in Chapter IX.

## 9.6.4 5.4 Addition Theorem for the Primary Aberrations

Having considered the significance of the primary aberrations, we must turn our attention to the more difficult task of calculating the primary aberration coefficients for a general centred system. As was seen, this is equivalent to determining the fourthorder terms in the power series expansion of the perturbation eikonal of Schwarzschild. In order not to interrupt the main calculations, it will be convenient to consider first the manner in which the perturbation eikonal of an optical system depends on the perturbation eikonals associated with the imaging by each surface of the system.

Consider a centred system which consists of two surfaces of revolution, and let $O_0$ be the axial object point and $O_1$ and $O_2$ its Gaussian images by the first surface and by both the surfaces respectively. Further let

$$T_1 = T_1\left(p_0, q_0; p_1, q_1\right) \tag{1}$$

be the angle characteristic for refraction at the first surface and,

$$T_2 = T_2\left(p_1, q_1; p_2, q_2\right) \tag{2}$$

the angle characteristic for refraction at the second surface, the former being referred to axes at $O_0$ and $O_1$ and the latter to parallel axes at $O_1$ and $O_2$; the $Z$-axes are taken along the axis of the system. Since the media are assumed to be homogeneous, the angle characteristic represents the optical path between the feet of the perpendiculars dropped on to the initial and final portions of the ray from the two origins (see Fig. 4.3). Hence the angle characteristic $T$ of the system referred to axes at $O_0$ and $O_2$ is obtained from $T_1$ and $T_2$ by addition:

$$T = T_1 + T_2 \tag{3}$$

In this expression, the variables $p_1, q_1$ referring to the ray in the intermediate space must be eliminated with the help of formulae describing the imaging by each surface. It will be convenient to carry out the elimination explicitly in the corresponding relation for the perturbation eikonal.

According to § 5.2 (13), the perturbation eikonals $\psi_1$ and $\psi_2$ are given by

$$\psi_1 = T_1 + \frac{\mathrm{D}_0}{2n_0\lambda_0^2}\left(x_0^2 + y_0^2\right) - \frac{\mathrm{D}_1}{2n_1\lambda_1^2}\left(x_1^2 + y_1^2\right) + x_0\left(\xi_1 - \xi_0\right) + y_0\left(\eta_1 - \eta_0\right),$$

and

$$\psi_2 = T_2 + \frac{\mathrm{D}_1}{2n_1\lambda_1^2}\left(x_1^2 + y_1^2\right) - \frac{\mathrm{D}_2}{2n_2\lambda_2^2}\left(x_2^2 + y_2^2\right) + x_1\left(\xi_2 - \xi_1\right) + y_1\left(\eta_2 - \eta_1\right)$$

and the perturbation eikonal $\psi$ of the combination is

$$\psi = T + \frac{\mathrm{D}_0}{2n_0\lambda_0^2}\left(x_0^2 + y_0^2\right) - \frac{\mathrm{D}_2}{2n_2\lambda_2^2}\left(x_2^2 + y_2^2\right) + x_0\left(\xi_2 - \xi_0\right) + y_0\left(\eta_2 - \eta_0\right).$$

Hence, using (3),

$$\psi = \psi_1 + \psi_2 + \left(x_0 - x_1\right)\left(\xi_2 - \xi_1\right) + \left(y_0 - y_1\right)\left(\eta_2 - \eta_1\right),$$

or, using eq. (17) of § 5.2

$$\psi = \psi_1 + \psi_2 + \frac{\partial\psi_1}{\partial\xi_1}\frac{\partial\psi_2}{\partial x_1} + \frac{\partial\psi_1}{\partial\eta_1}\frac{\partial\psi_2}{\partial y_1}. \tag{4}$$

If in (4) the power series expansions for $\psi_1$ and $\psi_2$ are substituted, it follows (since according to §5.3 (1) there are no terms of second order), that

$$\psi = \psi_1^{(0)} + \psi_2^{(0)} + \psi_1^{(4)} + \psi_2^{(4)} + \dots, \tag{5}$$

the terms not shown being of sixth and higher orders. Now in $\boldsymbol{\psi_1}^{(4)}$ and $\boldsymbol{\psi_2}^{(4)}$ we may obviously replace the arguments by their Gaussian values. Hence the variables relating to the intermediate space may be eliminated by means of the following rule: In ${\psi_1}^{(4)}(x_0, y_0; \xi_1, \eta_1)$ replace $\xi_1, \eta_1$ by $\xi_2, \eta_2$ and in ${\psi_2}^{(4)}(x_1, y_1; \xi_2, \eta_2)$ replace $x_1, y_1$ by $x_0, y_0$, and add the resulting expressions. This gives $\psi^{(4)}$ as function of the four variables $x_0, y_0$, and $\xi_2, \eta_2$ as required. This elimination will now be carried out more explicitly.

According to §5.3(3), ${\psi_1}^{(4)}$ and ${\psi_2}^{(4)}$ are of the form

$$\left.\begin{array}{l} {\psi_1}^{(4)} = -\frac{1}{4}A_1 {\mathrm{r}_0}^4 - \frac{1}{4}B_1 {\rho_1}^4 - C_1 {\kappa_{01}}^4 - \frac{1}{2}D_1 {r_0}^2 {\rho_1}^2 + E_1 {r_0}^2 {\kappa_{01}}^2 + F_1 {\rho_1}^2 {\kappa_{01}}^2, \\ \text{with} \quad {r_0}^2 = {x_0}^2 + {y_0}^2, \quad {\rho_1}^2 = {\xi_1}^2 + {\eta_1}^2, \quad {\kappa_{01}}^2 = x_0\xi_1 + y_0\eta_1; \end{array}\right\} \tag{6}$$

$$\left.\begin{array}{l} {\psi_2}^{(4)} = -\frac{1}{4}A_2 {r_1}^4 - \frac{1}{4}B_2 {\rho_2}^4 - C_2 {\kappa_{12}}^4 - \frac{1}{2}D_2 {r_1}^2 {\rho_2}^2 + E {r_1}^2 {\kappa_{12}}^2 + F_2 {\rho_2}^2 {\kappa_{12}}^2, \\ \text{with} \quad {r_1}^2 = {x_1}^2 + {y_1}^2, \quad {\rho_2}^2 = {\xi_2}^2 + {\eta_2}^2, \quad {\kappa_{12}}^2 = x_1\xi_2 + y_1\eta_2; \end{array}\right\} \tag{7}$$

Replacing $\xi_1, \eta_1$ by $\xi_2, \eta_2$ and $x_1, y_1$ by $x_0, y_0$, i.e. $\rho_1$ by $\rho_2, \mathrm{r}_1$ by $\mathrm{r}_0, {\kappa_{12}}^2$ and ${\kappa_{01}}^2$ by

$$\kappa_{02}^2 = x_0\xi_2 + y_0\eta_2, \tag{8}$$

we obtain on addition

$$\begin{aligned} \psi^{(4)} = -\frac{1}{4}(A_1 + A_2) r_0^4 - \frac{1}{4}(B_1 + B_2) \rho_2^4 - (C_1 + C_2) \kappa_{02}^4 - \frac{1}{2}(D_1 + D_2) r_0^2 \rho_2^2 \\ + (E_1 + E_2) r_0^2 \kappa_{02}^2 + (F_1 + F_2) \rho_2^2 \kappa_{02}^2. \end{aligned} \tag{9}$$

A similar result obviously holds for a centred system consisting of any number of surfaces. We have thus proved the following theorem:

Each primary aberration coefficient of a centred system is the sum of the corresponding aberration coefficients associated with the individual surfaces of the system.

It is at this point of our analysis that the advantage of the Seidel variables is particularly evident; for this simple and important result depends basically upon the use of the Seidel variables and has no analogy when ordinary variables are used.

## 9.6.5 5.5 The Primary Aberration Coefficients of a General Centred Lens System

The theorem established in the preceding section reduces the problem of determining the primary aberration coefficients of a general centred system to that of calculating the corresponding coefficients for each of its surfaces. This calculation will now be carried out.

### 5.5.1 The Seidel formulae in terms of two paraxial rays

It may be recalled that the Seidel aberration coefficients may (apart from simple numerical factors) be identified with the coefficients of the fourth-order terms in the power series expansion of the perturbation eikonal $\psi$ of Schwarzschild. According to § 5.2 (13), this function is obtained by adding to the angle characteristic $T$ certain quadratic terms, and by expressing the resulting expression in terms of the Seidel variables. Since the relations between the Seidel variables and the ray components are linear, the order of the terms does not change by transition from the one set of variables to the other. Hence

$$\psi^{(4)}(x_0, y_0; \xi_1, \eta_1) = T^{(4)}(p_0, q_0; p_1, q_1) \tag{1}$$

The expansion of the angle characteristic up to the fourth order for a refracting surface of revolution was derived in §4.1. The fourth-order contribution may, according to §4.1 (42), be written in the form*

$$T^{(4)}(p_0, q_0; p_1, q_1) = -\frac{r}{4(n_1 - n_0)^2}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]\left[\frac{p_0^2 + q_0^2}{n_0} - \frac{p_1^2 + q_1^2}{n_1}\right]$$
$$- \frac{(1+b)r}{8(n_1 - n_0)^3}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]^2 + \frac{a_0}{8n_0^3}\left(p_0^2 + q_0^2\right)^2 - \frac{a_1}{8n_1^3}\left(p_1^2 + q_1^2\right)^2 . (2 \tag{2}$$

It will be convenient to choose the axial points $z = a_0, z = a_1$ as the axial object point and its Gaussian image, and to set (see Fig. 5.9)

$$s = a_0, \quad s' = a_1, \quad t = a_0 + \mathrm{D}_0, \quad t' = a_1 + \mathrm{D}_1. \tag{3}$$

The associated Abbe invariants (§ 4.4 (7)) will be denoted by $K$ and $L$ respectively:

$$n_0\left(\frac{1}{r} - \frac{1}{s}\right) = n_1\left(\frac{1}{r} - \frac{1}{s'}\right) = K, \tag{4}$$

$$n_0\left(\frac{1}{r} - \frac{1}{t}\right) = n_1\left(\frac{1}{r} - \frac{1}{t'}\right) = L. \tag{5}$$

Before substituting into (1) the expressions for the ray components in terms of the Seidel variables, it will be useful to re-write (1) in a slightly different form. Because of (4) we have $\frac{r}{8(n_1-n_0)^3}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]^2 = \frac{r^2}{8(n_1-n_0)^4}\left[\frac{n_1}{s'} - \frac{n_0}{s}\right]\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right]^2$.
Using this relation, (2) may be written as

$$T^{(4)} = \frac{1}{8n_0 s}\left\{\frac{n_0 r}{(n_1 - n_0)^2}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right] - \frac{s}{n_0}\left({p_0}^2 + {q_0}^2\right)\right\}^2$$
$$- \frac{1}{8n_1 s'}\left\{\frac{n_1 r}{(n_1 - n_0)^2}\left[(p_0 - p_1)^2 + (q_0 - q_1)^2\right] - \frac{s'}{n_1}\left({p_1}^2 + {q_1}^2\right)\right\}^2$$
$$- \frac{br}{8(n_1 - n_0)^3}\left\{(p_0 - p_1)^2 + \tfrac{r}{1}(q_0 - q_1)^2\right\}^2 . \tag{6}$$

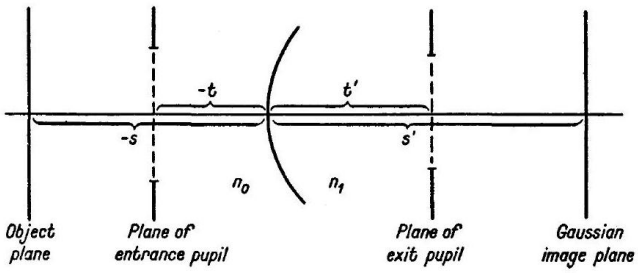



Fig. 5.9. Notation used in the calculation of the primary aberration coefficients.

0

- As in §4.1, $r$ denotes the radius of the surface. This symbol should not be confused with the symbol $r$ which denotes the rotational invariant $\sqrt{{x_0}^2 + {y_0}^2}$ introduced in § 5.3 (2).

In (6), the arguments may be replaced by their Gaussian approximations; in particular, the Seidel variables referring to points on the incident and the refracted ray may be interchanged. In order to obtain $\psi^{(4)}$ as a function of $x_0, y_0, \xi_1$ and $\eta_1$, we may then use in place of §5.2 (9) the relations

$$\left.\begin{array}{ll} p_0 = \frac{n_0\lambda_0}{\mathrm{D}_0}\xi_1 - \frac{1}{\lambda_0}x_0, & p_1 = \frac{n_1\lambda_1}{\mathrm{D}_1}\xi_1 - \frac{1}{\lambda_1}x_0, \\ q_0 = \frac{n_0\lambda_0}{\mathrm{D}_0}\eta_1 - \frac{1}{\lambda_0}y_0, & q_1 = \frac{n_1\lambda_1}{\mathrm{D}_1}\eta_1 - \frac{1}{\lambda_1}y_0. \end{array}\right\} \tag{7}$$

It will be useful to make one further modification. The Gaussian lateral magnification between the object and the image plane ( $l_1/l_0$ ) and between the planes of the entrance and the exit pupil ( $\lambda_1/\lambda_0$ ) may be obtained from § 4.4 (14) and § 4.4 (10), or more simply by noting that imaging by a spherical surface is a projection from the centre of the sphere. Hence, if (4) is also used,

$$\frac{l_1}{l_0} = \frac{r-s'}{r-s} = \frac{n_0 s'}{n_1 s}, \quad \frac{\lambda_1}{\lambda_0} = \frac{r-t'}{r-t} = \frac{n_0 t'}{n_1 t}. \tag{8}$$

Introducing the abbreviations

$$h = \frac{\lambda_0 s}{\mathrm{D}_0} = \frac{\lambda_1 s'}{\mathrm{D}_1}, \quad H = \frac{t}{\lambda_0 n_0} = \frac{t'}{\lambda_1 n_1}, \tag{9}$$

where (8) and §5.2 (7) was used, (7) becomes

$$\left.\begin{array}{ll} p_0 = n_0\left(\frac{h}{s}\xi_1 - \frac{H}{t}x_0\right), & p_1 = n_1\left(\frac{h}{s'}\xi_1 - \frac{H}{t'}x_0\right) \\ q_0 = n_0\left(\frac{h}{s}\eta_1 - \frac{H}{t}y_0\right), & q_1 = n_1\left(\frac{h}{s'}\eta_1 - \frac{H}{t'}y_0\right) \end{array}\right\} \tag{10}$$

If as before, $r^2, \rho^2$ and $\kappa^2$ denote the three rotational invariants

$$r^2 = x_0^2 + y_0^2, \quad \rho^2 = \xi_1^2 + \eta_1^2, \quad \kappa^2 = x_0\xi_1 + y_0\eta_1, \tag{11}$$

the terms in the curly brackets of (6) become

$$\begin{aligned} &(p_0-p_1)^2 + (q_0-q_1)^2 = \left(\frac{n_1-n_0}{\mathrm{r}}\right)^2 \left[H^2\mathrm{r}^2 + h^2\rho^2 - 2hH\kappa^2\right] \\ &\frac{n_0 r}{(n_1-n_0)^2}\left[(p_0-p_1)^2 + (q_0-q_1)^2\right] - \frac{s}{n_0}\left({p_0}^2 + {q_0}^2\right) \\ &= H^2\mathrm{r}^2\left[L - (K-L)\frac{s}{t}\right] + h^2\rho^2 K - 2hH\kappa^2 L \\ &\frac{n_1 r}{(n_1-n_0)^2}\left[(p_0-p_1)^2 + (q_0-q_1)^2\right] - \frac{s'}{n_1}\left({p_1}^2 + {q_1}^2\right) \\ &= H^2\mathrm{r}^2\left[L - (K-L)\frac{s'}{t'}\right] + h^2\rho^2 K - 2hH\kappa^2 L \end{aligned}$$

Substituting from these relations into (6) and recalling (1), we finally obtain the required expression for $\psi^{(4)}$ :

$$\begin{aligned}\psi^{(4)} &= \frac{1}{8}\mathrm{r}^4H^4\left\{\frac{b}{r^3}(n_0-n_1)+L^2\left(\frac{1}{n_0s}-\frac{1}{n_1s'}\right)-2L(K-L)\left(\frac{1}{n_0t}-\frac{1}{n_1t'}\right)\right.\\ &\quad\left.+(K-L)^2\left(\frac{s}{n_0t^2}-\frac{s'}{n_1t^2}\right)\right\}\\ &\quad+\frac{1}{8}\rho^4h^4\left\{\frac{b}{r^3}(n_0-n_1)+K^2\left(\frac{1}{n_0s}-\frac{1}{n_1s'}\right)\right\}\\ &\quad+\frac{1}{2}\kappa^4H^2h^2\left\{\frac{b}{r^3}(n_0-n_1)+L^2\left(\frac{1}{n_0s}-\frac{1}{n_1s'}\right)\right\}\\ &\quad+\frac{1}{4}r^2\rho^2H^2h^2\left\{\frac{b}{r^3}(n_0-n_1)+KL\left(\frac{1}{n_0s}-\frac{1}{n_1s'}\right)-K(K-L)\left(\frac{1}{n_0t}-\frac{1}{n_1t'}\right)\right\}\\ &\quad-\frac{1}{2}r^2\kappa^2H^3h\left\{\frac{b}{r^3}(n_0-n_1)+L^2\left(\frac{1}{n_0s}-\frac{1}{n_1s'}\right)-L(K-L)\left(\frac{1}{n_0t}-\frac{1}{n_1t'}\right)\right\}\\ &\quad-\frac{1}{2}\rho^2\kappa^2Hh^3\left\{\frac{b}{r^3}(n_0-n_1)+KL\left(\frac{1}{n_0s}-\frac{1}{n_1s'}\right)\right\}. \qquad (12)\end{aligned}$$

This formula gives, on comparison with the general expression §5.3 (3), the fourthorder coefficients $A, B, \ldots F$ of the perturbation eikonal of a refracting surface of revolution.

Generalization to a centred system consisting of any number of refracting surfaces is now straightforward. Let us denote by the suffix $i$ quantities referring to the $i$ th surface, and let $n_i$ be the refractive index of the medium which follows the $i$ th surface. Then, using the addition theorem of §5.4 it follows from (12) on comparison with § 5.3 (3) that

$$\begin{aligned}B &= \frac{1}{2}\sum_i h_i^4\left\{\frac{b_i}{r_i^3}(n_i-n_{i-1})+K_i^2\left(\frac{1}{n_is_i'}-\frac{1}{n_{i-1}s_i}\right)\right\}\\ C &= \frac{1}{2}\sum_i H_i^2h_i^2\left\{\frac{b_i}{r_i^3}(n_i-n_{i-1})+L_i^2\left(\frac{1}{n_is_i'}-\frac{1}{n_{i-1}s_i}\right)\right\}\\ D &= \frac{1}{2}\sum_i H_i^2h_i^2\left\{\frac{b_i}{r_i^3}(n_i-n_{i-1})+K_iL_i\left(\frac{1}{n_is_i'}-\frac{1}{n_{i-1}s_i}\right)-K_i(K_i-L_i)\left(\frac{1}{n_it_i'}-\frac{1}{n_{i-1}t_i}\right)\right\}\\ E &= \frac{1}{2}\sum_i H_i^3h_i\left\{\frac{b_i}{r_i^3}(n_i-n_{i-1})+L_i^2\left(\frac{1}{n_is_i'}-\frac{1}{n_{i-1}s_i}\right)-L_i(K_i-L_i)\left(\frac{1}{n_it_i'}-\frac{1}{n_{i-1}t_i}\right)\right\}\\ F &= \frac{1}{2}\sum_i H_ih_i^3\left\{\frac{b_i}{r_i^3}(n_i-n_{i-1})+K_iL_i\left(\frac{1}{n_is_i'}-\frac{1}{n_{i-1}s_i}\right)\right\} \qquad (13)\end{aligned}$$

These are the Seidel formulae for the primary aberration coefficients of a general centred system of refracting surfaces.*

Equations (13) express the primary aberration coefficients in terms of data specifying the passage of two paraxial rays through the system, namely a ray from the axial object point and a ray from the centre of the entrance pupil. It will be useful to summarize the relevant Gaussian formulae. Let $d_i$ be the distance between the poles of the $i$ th and the $(i+1)$ th surface. Since the Gaussian image formed by the first $i$ surfaces of the system is the object for the $(i+1)$ th surface, we have the transfer formulae

$$s_{i+1} = s_i' - d_i, \quad t_{i+1} = t_i' - d_i. \qquad (14)$$

Given the distances $s_1$ and $t_1$ of the object plane and the plane of the entrance pupil from the pole of the first surface, the distances $s_1', t_1', s_2, t_2, s_2', t_2', \ldots$ and the corresponding values $K_1, L_1, K_2, L_2, \ldots$ may then be calculated successively from the Abbe relations

$$\left.\begin{aligned} n_{i-1}\left(\tfrac{1}{r_i}-\tfrac{1}{s_i}\right) &= n_i\left(\tfrac{1}{r_i}-\tfrac{1}{s'_i}\right) = K_i \\ n_{i-1}\left(\tfrac{1}{r_i}-\tfrac{1}{t_i}\right) &= n_i\left(\tfrac{1}{r_i}-\tfrac{1}{t'_i}\right) = L_i \end{aligned}\right\} \tag{15}$$

and from (14). We also have to determine the quantities $h_i$ and $H_i$. If for simplicity the arbitrary length $\lambda_0$ in the plane of the entrance pupil is taken equal to unity, and

the relations $\mathrm{D}_i = t'_i - s'_i = t_{i+1} - s_{i+1}$ are used, it is seen from (9) that $h_i$ and $H_i$ may be calculated in succession from the relations*

$$\left.\begin{aligned} H_1 &= \tfrac{t_1}{n_0} & h_1 &= \tfrac{s_1}{t_1 - s_1} \\ \tfrac{H_{i+1}}{H_i} &= \tfrac{t_{i+1}}{t'_i} & \tfrac{h_{i+1}}{h_i} &= \tfrac{s_{i+1}}{s'_i} \end{aligned}\right\} \tag{16}$$

From (9) and from the Abbe relations (4) and (5) we obtain the following relation, which may be used as check on the calculations and which will be needed later:

$$h_i H_i = \frac{s_i t_i}{n_{i-1}\left(t_i - s_i\right)} = \frac{s'_i t'_i}{n_i\left(t'_i - s'_i\right)} = \frac{1}{L_i - K_i}. \tag{17}$$

### 5.5.2 The Seidel formulae in terms of one paraxial ray

It is often desirable to express the primary aberration coefficients in a form which shows as clearly as possible the dependence of the coefficients on parameters which specify the optical system, and for this purpose (13) are not the most convenient set of formulae. To evaluate them, two rays must be traced through the system in accordance with the laws of Gaussian optics, namely a ray from the axial object point and a ray from the centre of the entrance pupil. It was, however, shown by Seidel that we may eliminate the data relating to the second ray and thus express the primary aberration coefficients by means of quantities which relate to the first ray only. Naturally one quantity depending on the position of the entrance pupil remains in the formulae, since alteration of the position of the stop will obviously influence the aberrations.

To eliminate the data referring to the ray from the centre of the entrance pupil, it is necessary first of all to express the heights $H_i$ in terms of $h_i$. From (16), (14) and (17)

$$\begin{aligned} \frac{H_{i+1}}{h_{r+1}} - \frac{H_i}{h_i} &= \frac{1}{h_i h_{i+1}} H_i h_i \left(\frac{t_{i+1}}{t'_i} - \frac{s_{i+1}}{s'_i}\right) \\ &= \frac{1}{h_i h_{i+1}} H_i h_i d_i \left(\frac{1}{s'_i} - \frac{1}{t'_i}\right) \\ &= \frac{d_i}{n_i h_i h_{i+1}} \end{aligned} \tag{18}$$

If quantities $k_i$ are defined by

$$H_i = k_i h_i, \tag{19}$$

it follows from (18) that

0

- Seidel actually considered centred systems consisting of spherical surfaces only. The effects of asphericities (characterized by the constants $b_i$ ) was taken into account by later writers (cf. M. von Rohr, Geometrical Investigation of Formation of Images in Optical Systems, translated from German by R. Kanthack (London, H.M. Stationery Office, 1920), 344.

$$k_{i+1} = k_1 + \sum_{j=1}^{i} \frac{d_j}{n_j h_j h_{j+1}} \tag{20}$$

where, according to (16),

$$k_1 = \frac{H_1}{h_1} = \frac{t_1\,(t_1 - s_1)}{n_0 s_1} \tag{21}$$

The other quantities in (13) which relate to the paraxial ray from the centre of the entrance pupil may also easily be expressed in terms of quantities referring to the other ray. From (17) and (19)

$$L_i = K_i + \frac{1}{k_i h_i^2} \tag{22}$$

Further from (15) and (22)

$$\begin{aligned}
\frac{1}{n_i t_i'} - \frac{1}{n_{i-1} t_i} =& \frac{1}{n_i r_i} - \frac{1}{n_{i-1} r_i} - \left(\frac{1}{n_i^2} - \frac{1}{n_{i-1}^2}\right) L_i \\
&= \frac{1}{n_i}\left(\frac{K_i}{n_i} + \frac{1}{s_i'}\right) - \frac{1}{n_{i-1}}\left(\frac{K_i}{n_{i-1}} + \frac{1}{s_i}\right) - \left(\frac{1}{n_i^2} - \frac{1}{n_{i-1}^2}\right)\left(K_i + \frac{1}{k_i h_i^2}\right) \\
&= \frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i} - \frac{1}{k_i h_i^2}\left(\frac{1}{n_i^2} - \frac{1}{n_{i-1}^2}\right)
\end{aligned} \tag{23}$$

Substitution from (19), (22), and (23) into (13) finally gives

$$\left.\begin{aligned}
B &= \tfrac{1}{2}\textstyle\sum_i {h_i}^4 \frac{b_i}{{r_i}^3}\,(n_i - n_{i-1}) + {h_i}^4 {K_i}^2 \left(\frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i}\right) \\
C &= \tfrac{1}{2}\textstyle\sum_i {h_i}^4 {k_i}^2 \frac{b_i}{{r_i}^3}\,(n_i - n_{i-1}) + (1 + {h_i}^2 k_i K_i)^2 \left(\frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i}\right) \\
D &= \tfrac{1}{2}\textstyle\sum_i {h_i}^4 {k_i}^2 \frac{b_i}{{r_i}^3}\,(n_i - n_{i-1}) + {h_i}^2 k_i K_i\,(2 + {h_i}^2 k_i K_i)\left(\frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i}\right) \\
&\qquad - K_i \left(\frac{1}{{n_i}^2} - \frac{1}{{n_{i-1}}^2}\right) \\
E &= \tfrac{1}{2}\textstyle\sum_i {h_i}^4 {k_i}^3 \frac{b_i}{{r_i}^3}\,(n_i - n_{i-1}) + k_i\,(1 + {h_i}^2 k_i K_i)\,(2 + {h_i}^2 k_i K_i)\left(\frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i}\right) \\
&\qquad - \frac{1 + {h_i}^2 k_i K_i}{{h_i}^2}\left(\frac{1}{{n_i}^2} - \frac{1}{{n_{i-1}}^2}\right) \\
F &= \tfrac{1}{2}\textstyle\sum_i {h_i}^4 k_i \frac{b_i}{{r_i}^3}\,(n_i - n_{i-1}) + {h_i}^2 K_i\,(1 + k_i {h_i}^2 K_i)\left(\frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i}\right)
\end{aligned}\right\} \tag{24}$$

This is the required form of the Seidel formulae. The position of the entrance pupil enters here only through the factor $k_1$ which is related to the distance $t_1$ of the plane of the entrance pupil from the pole of the first surface by (21). The quantities $s_i, s_i', K_i$ and $h_i$ are again computed from the appropriate relations in (14)-(16), whilst the $k_i$ 's are determined from (20).

0

- The relation $h_{i+1}/h_i = s_{i+1}/s_i'$ implies that $h_i$ is proportional to the height from the axis, at which a paraxial ray from the axial object point intersects the $i$ th surface. The relation $H_{i+1}/H_i = t_{i+1}/t_i{}'$ gives a similar interpretation to $H_i$, in terms of a paraxial ray from the centre of the entrance pupil.

### 5.5.3 Petzval's theorem

From the expressions for the coefficients of astigmatism and curvature we can derive an interesting relation due to Petzval. We have

$$\begin{aligned} C - D =& \frac{1}{2}\sum_i \left\{ \left[ \left(1 + {h_i}^2 k_i K_i\right)^2 - 2{h_i}^2 k_i K_i - {h_i}^4 {k_i}^2 {K_i}^2 \right] \left[ \frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i} \right] \right. \\ & \left. + K_i \left( \frac{1}{{n_i}^2} - \frac{1}{{n_{i-1}}^2} \right) \right\} \\ =& \frac{1}{2}\sum_i \left\{ \frac{1}{n_i s_i'} - \frac{1}{n_{i-1} s_i} + K_i \left( \frac{1}{{n_i}^2} - \frac{1}{{n_{i-1}}^2} \right) \right\} \\ =& \frac{1}{2}\sum_i \frac{1}{r_i} \left( \frac{1}{n_i} - \frac{1}{n_{i-1}} \right) \end{aligned} \tag{25}$$

where (15) was used.
According to § 5.3, $C$ and $D$ determine the sagittal and the tangential field curvature. If $n_\alpha$ denotes the refractive index of the last medium, it follows from eqs. (18) and (21) of § 5.3 that

$$C = \frac{1}{4n_\alpha}\left(\frac{1}{R_t} - \frac{1}{R_s}\right), \quad D = \frac{1}{2n_\alpha}\frac{1}{R_s} \tag{26}$$

Hence (25) may be written as

$$\frac{1}{R_t} - \frac{3}{R_s} = 2n_\alpha \sum \frac{1}{r_i}\left(\frac{1}{n_i} - \frac{1}{n_{i-1}}\right) \tag{27}$$

We thus obtain a relation between the curvatures of the two focal surfaces which contains only the radii of the refracting surfaces of the system and the corresponding refractive indices. If the system is free of spherical aberration, coma, and astigmatism, then a sharp image is formed on a surface of radius $R_s = R_t = R$; and the radius of this surface is according to (27) given by

$$\frac{1}{R} = -n_\alpha \sum_i \frac{1}{r_i}\left(\frac{1}{n_i} - \frac{1}{n_{i-1}}\right) \tag{28}$$

This result is known as Petzval's theorem.
The condition

$$\sum_i \frac{1}{r_i}\left(\frac{1}{n_i} - \frac{1}{n_{i-1}}\right) = 0 \tag{29}$$

is known as Petzval's condition. It is a necessary condition for flatness of the field. It should, however, be remembered that this condition belongs to the domain of the Seidel theory; it loses its significance outside this domain.

Whether or not aberrations are present, the spherical surface which touches the two focal surfaces at their common axial point and whose radius $R$ is given by (28) is called the Petzval surface. According to (27) and (28) the radii of curvature of the sagittal focal surface, the tangential focal surface, and the Petzval surface are related by

$$\frac{3}{R_3} - \frac{1}{R_t} = \frac{2}{R} \tag{30}$$

### 9.6.6 5.6 Example: The Primary Aberrations of a Thin Lens

The Seidel formulae will now be used to find the primary aberration coefficients of a thin lens of refractive index $n$, situated in air (vacuum). In this case

$$n_0 = n_2 = 1, \quad n_1 = n \tag{1}$$

Since the thickness $d$ of the lens is assumed to be negligible, we have according to § 5.5 (14)

$$s_2 = s_1'; \tag{2}$$

and § 5.5 (15) gives

$$\left.\begin{aligned} 1 - \tfrac{1}{r_1} = n\left(\tfrac{1}{r_1} - \tfrac{1}{s_1'}\right) = K_1, \\ n\left(\tfrac{1}{r_2} - \tfrac{1}{s_2}\right) = \left(\tfrac{1}{r_2} - \tfrac{1}{s_2'}\right) = K_2. \end{aligned}\right\} \tag{3}$$

Further from § 5.5 (16) and § 5.5 (20), using (2)
where

$$\left.\begin{aligned} h_1 = h_2 = h, \quad k_1 = k_2 = k, \\ h = \tfrac{s_1}{t_1 - s_1}, \qquad k = \tfrac{t_1(t_1 - s_1)}{s_1}. \end{aligned}\right\} \tag{4}$$

It will be convenient to characterize the lens by a number of simple parameters. Let $\mathscr{P}$ be the power of the lens (cf. eqs. (35) and (36) of §4.4)

$$\mathscr{P} = \frac{1}{f} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right), \tag{5}$$

and let

$$\sigma = (n-1)\left(\frac{1}{r_1} + \frac{1}{r_2}\right). \tag{6}$$

From (3),

$$(n-1)\frac{1}{r_1} = \frac{n}{s_1'} - \frac{1}{s_1}, \quad (n-1)\frac{1}{r_2} = \frac{n}{s_2} - \frac{1}{s_2'}, \tag{7}$$

so that, because of (2),

$$\frac{1}{s_2'} - \frac{1}{s_1} = \mathscr{P} \tag{8}$$

This relation will be written in the form

$$-\frac{1}{s_1} - \frac{\mathscr{P}}{2} = -\frac{1}{s_2'} + \frac{\mathscr{P}}{2} = \mathscr{K}; \tag{9}$$

$\mathscr{K}$ plays a similar role for the lens as $K$ does for a single surface. Accordingly, $\mathscr{K}$ will be referred to as the Abbe invariant of the lens.

The deformation coefficients $b_1$ and $b_2$ of the two surfaces of the lens will be seen later to enter the formulae only through the quantity

$$\beta = (n-1)\left(\frac{b_1}{r_1^3} - \frac{b_2}{r_2^3}\right) \tag{10}$$

this quantity may be called the deformation coefficient of the lens.
We now express the various quantities which enter the Seidel formulae §5.5(24) in terms of the

parameters $\mathscr{P}, \sigma, \beta$ and $\mathscr{K}$. The first three specify the lens and the last specifies the position of the object. First, we obtain from (3)

$$K_1 = \mathscr{K} + \frac{\sigma + n\mathscr{P}}{2(n-1)}, \quad K_2 = \mathscr{K} + \frac{\sigma - n\mathscr{P}}{2(n-1)}. \tag{11}$$

Further

$$\left.\begin{array}{l} \frac{1}{ns_1'} - \frac{1}{s_1} = \frac{n^2-1}{n^2}\mathscr{K} + \frac{\sigma}{2n^2} + \frac{\mathscr{P}}{2}, \\ \frac{1}{s_2'} - \frac{1}{ns_2} = -\frac{n^2-1}{n^2}\mathscr{K} - \frac{\sigma}{2n^2} + \frac{\mathscr{P}}{2}. \end{array}\right\} \tag{12}$$

Substitution into the Seidel formulae § 5.5 (24) leads to the following expressions for the primary aberration coefficients of a thin lens:

$$\left.\begin{array}{l} B = h^4 U \\ F = h^4 k U + h^2 V \\ C = h^4 k^2 U + 2h^2 k V + \frac{1}{2}\mathscr{P} \\ D = h^4 k^2 U + 2h^2 k V + \frac{n+1}{2n}\mathscr{P} \\ E = h^4 k^3 U + 3h^2 k^2 V + k\frac{3n+1}{2n}\mathscr{P} \end{array}\right\} \tag{13}$$

where

$$\left.\begin{array}{lll} U = & \frac{1}{2}\beta + \frac{n^2}{8(n-1)^2}\mathscr{P}^3 & -\frac{n}{2(n+2)}\mathscr{K}^2\mathscr{P} \\ & & +\frac{1}{2n(n+2)}\mathscr{P}\left[\frac{n+2}{2(n-1)}\sigma + 2(n+1)\mathscr{K}\right]^2 \end{array}\right\} \tag{14}$$

Only the case when the entrance pupil coincides with the lens ( $t_1 = 0$ ) will be considered here. According to (4) we then have

$$h = -1, \quad k = 0$$

and the formulae (13) become

$$\left.\begin{array}{rl} B &= U \\ F &= V \\ \frac{1}{2}(C+D) &= \frac{2n+1}{4n}\mathscr{P} \\ \frac{1}{2}(C-D) &= -\frac{1}{4n}\mathscr{P} \\ E &= 0 \end{array}\right\} \tag{15}$$

The quantities $U$ and $V$ defined by (14) are seen to represent the spherical aberration and coma of a lens working with the stop in the plane of the lens. Such a lens is free from distortion ($E = 0$) but astigmatism and field curvature are always present ($C \neq 0, D \neq 0$).

Let us examine whether, for a given lens which is working with the stop in the plane of the lens, there exists a pair of aplanatic points ( $B = F = 0$ ). For coma to be absent we must have, according to (15), $V = 0$,
i.e.

$$\frac{n+1}{2(n-1)}\sigma + (2n+1)\mathscr{K} = 0. \tag{16}$$

Eq. (9) then gives the object distance $s_1$ :

$$\frac{1}{s_1} = -\frac{1}{2}\mathscr{P} + \frac{n+1}{2(n-1)(2n+1)}\sigma. \tag{17}$$

With this choice of $s_1$, the spherical aberration of the lens is fully determined. Hence in general there is no aplanatic point pair for a given lens working with the stop at the lens. However, if only $\mathscr{P}$ and $\sigma$ are given, one may eliminate the spherical aberration by an appropriate choice of the deformation coefficient $\beta$.

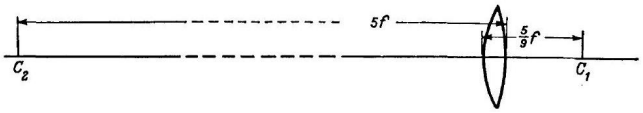



Fig. 5.10. Coma-free thin lens (object at infinity).

Conversely, if the object distance $s_1$ and the focal length $f = 1/\mathscr{P}$ are given one may first choose the radii $r_1$ and $r_2$ in such a way that coma vanishes. According to (17), (5) and (6) we have in this case,

$$\left.\begin{array}{l} \frac{1}{r_1} = \frac{2n+1}{n+1}\frac{1}{s_1} + \frac{n^2}{n^2-1}\mathscr{P} \\ \frac{1}{r_2} = \frac{2n+1}{n+1}\frac{1}{s_1} + \frac{n^2-n-1}{n^2-1}\mathscr{P}. \end{array}\right\} \tag{18}$$

This gives, for example, for an object at infinity ( $s_1 = \infty$ ) and with $n = 1.5$,

$$r_1 = \frac{5}{9}\frac{1}{\mathscr{P}} = \frac{5}{9}f, \quad r_2 = -5\frac{1}{\mathscr{P}} = -5f. \tag{19}$$

Such a coma-free lens is shown in Fig. 5.10, where $C_1$ and $C_2$ are the centres of curvature of the two surfaces of the lens.

Next, the spherical aberration may be eliminated by a suitable choice of $\beta$. When the object is at infinity, we have, according to (9), $\mathscr{K} = -\mathscr{P}/2$, and hence from (16) $\sigma = (2n+1)(n-1)\mathscr{P}/(n+1)$; the condition $U = 0$ for the absence of spherical aberration then gives the following expression for $\beta$ :

$$-\frac{1}{2}\beta = \mathscr{P}^3\left\{\frac{n^2}{8(n-1)^2} - \frac{n}{8(n+2)} + \frac{1}{2n(n+2)}\left[\frac{(n+2)(2n+1)}{2(n+1)} - (n+1)\right]^2\right\},$$

whence

$$\beta = -\frac{n^3}{(n^2-1)^2}\mathscr{P}^3. \tag{20}$$

For $n = 1.5$ this gives

$$\beta = -\frac{54}{2}\mathscr{P}^3, \tag{21}$$

so that by (19) and (10)

$$729b_1 + b_2 = -540. \tag{22}$$

This may be satisfied, for example, by taking $b_1 = b_2 = -0.74$.

Instead of eliminating coma, one may first attempt to reduce the spherical aberration as much as possible. It is then found that, if the focal length is positive, the spherical aberration may be removed completely only if the object lies in a certain limited range; for $n = 1.5$ this range extends from $0.36f$ to $0.44f$ behind the lens. For any other position of the object a certain amount of spherical aberration will always be present.

In order to affect the curvature of field, the lens must be used with a stop placed in an appropriate position, which may be determined from the general formulae (13). For a lens with a positive focal lens and with $n = 1 \cdot 5$, the curvature may be eliminated only if the object lies within a range extending from one focal length in front of the lens to half of the focal length behind it.

For a system which consists of a number of lenses, the calculations become more complicated. The removal of the chromatic aberration is then of primary importance. The primary aberrations of an achromatic telescope objective which consists of two cemented thin lenses may still be investigated with relative ease, and it is found that all the primary aberrations except astigmatism and curvature of field may be eliminated; it may therefore be used over a narrow field only (at most up to about 3° for an $f/10$ system).

## 9.6.7 5.7 The Chromatic Aberration of a General Centred Lens System

The chromatic aberration has already been briefly discussed in § 4.7. Here explicit expressions will be derived for the chromatic change in the position of the image, and in the magnification, using the general formalism of the preceding sections.* The system will be assumed to be perfectly corrected for monochromatic light. This is a permissible idealization if effects of the leading order only are being considered; for changes in the monochromatic aberrations associated with a small change in the wavelength may be assumed to be small in comparison with the magnitudes of the aberrations themselves; in general these changes will therefore be of the same order of magnitude as the terms neglected in the Seidel theory.

According to § 5.5 (15) and § 5.5 (14),

$$n_{i-1}\left(\frac{1}{r_i}-\frac{1}{s_i}\right)=n_i\left(\frac{1}{r_i}-\frac{1}{s_i'}\right)=K_i, \tag{1}$$

$$s_{i+1}=s_i'-d_i. \tag{2}$$

If $\delta$ denotes the change in a quantity associated with a small change $\delta\lambda$ in the wavelength, we have from (1) and (2),

$$n_i\frac{\delta s_i'}{s_i'^2}-n_{i-1}\frac{\delta s_i}{s_i^2}=-\left(\frac{\delta n_i}{n_i}-\frac{\delta n_{i-1}}{n_{i-1}}\right)K_i, \tag{3}$$

$$\delta s_{i+1}=\delta s_i'. \tag{4}$$

Multiplying (3) by ${h_i}^2$ and using the relation

$$\frac{h_{i+1}}{h_i}=\frac{s_{i+1}}{s_i'}, \tag{5}$$

we find that

$$n_i\left(\frac{h_{i+1}}{s_{i+1}}\right)^2\delta s_{i+1}-n_{i-1}\left(\frac{h_i}{s_i}\right)^2\delta s_i=-{h_i}^2K_i\left(\frac{\delta n_i}{n_i}-\frac{\delta n_{i-1}}{n_{i-1}}\right). \tag{6}$$

0

- We again follow substantially the analysis of K. Schwarzschild, Abh. Königl. Ges. Wiss. Göttingen, Math.-phys. Kl., 4 (1905-1906), No. 3. Reprinted in Astr. Mitt. Königl. Sternwarte Göttingen (1905).

Let $\alpha$ denote the number of surfaces in the system. Then by adding all the equations of the form (6) it follows that

$$n_\alpha \left(\frac{h_{\alpha+1}}{s_{\alpha+1}}\right)^2 \delta s_{\alpha+1} = n_\alpha \left(\frac{h_\alpha}{s'_\alpha}\right)^2 \delta s'_\alpha = n_0 \left(\frac{h_1}{s_1}\right)^2 \delta s_1 - \sum_{i=1}^{\alpha} {h_i}^2 K_i \left(\frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}}\right).$$

Since the position of the object is independent of the wavelength, $\delta s_1 = 0$, and we obtain the following expression for the chromatic change $\delta s'_\alpha$ in the position of the image:

$$\delta s'_\alpha = -\frac{1}{n_\alpha} \left(\frac{s'_\alpha}{h_\alpha}\right)^2 \sum_{i=1}^{\alpha} {h_i}^2 K_i \left(\frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}}\right). \tag{7}$$

Next consider the chromatic change in the magnification $M$ of the system. We have

$$M = \frac{l_1}{l_0}\frac{l_2}{l_1}\frac{l_3}{l_2} \cdots \frac{l_\alpha}{l_{\alpha-1}}. \tag{8}$$

Now by § 5.5 (8),

$$\frac{l_i}{l_{i-1}} = \frac{n_{i-1}}{n_i}\frac{s'_i}{s_i},$$

so that

$$M = \frac{n_0}{n_\alpha} \left(\frac{s'_1}{s_1}\frac{s'_2}{s_2}\frac{s'_3}{s_3} \cdots \frac{s'_\alpha}{s_\alpha}\right). \tag{9}$$

By logarithmic differentiation,

$$\begin{aligned} \frac{\delta M}{M} &= \frac{\delta n_0}{n_0} - \frac{\delta n_\alpha}{n_\alpha} + \sum_{i=1}^{\alpha} \left(\frac{\delta s'_i}{s'_i} - \frac{\delta s_i}{s_i}\right) \\ &= \left(\frac{\delta n_0}{n_0} - \frac{\delta n_\alpha}{n_\alpha}\right) + \frac{\delta s'_\alpha}{s'_\alpha} - \frac{\delta s_1}{s_1} - \sum_{i=2}^{\alpha} \delta s_i \left(\frac{1}{s_i} - \frac{1}{s'_{i-1}}\right) \end{aligned} \tag{10}$$

The sum entering this expression may be rewritten as follows

$$\sum_{i=2}^{\alpha} \delta s_i \left(\frac{1}{s_i} - \frac{1}{s'_{i-1}}\right) = \sum_{i=2}^{\alpha} \frac{d_{i-1}\delta s_i}{s_i s'_{i-1}} = \sum_{i=2}^{\alpha} \frac{d_{i-1}}{{s_i}^2}\frac{h_i}{h_{i-1}}\delta s_i.$$

But by § 5.5 (18),

$$\frac{d_{i-1}}{h_{i-1}} = (k_i - k_{i-1})\, h_i n_{i-1},$$

so that

$$\sum_{i=2}^{\alpha} \delta s_i \left(\frac{1}{s_i} - \frac{1}{s'_{i-1}}\right) = \sum_{i=2}^{\alpha} {h_i}^2 n_{i-1} \left(\frac{k_i - k_{i-1}}{{s_i}^2}\right) \delta s_i,$$

or, re-arranging according to the $k_i$ 's,

$$\sum_{i=2}^{\alpha} \delta s_i \left( \frac{1}{s_i} - \frac{1}{s'_{i-1}} \right) = -n_0 \left( \frac{h_1}{s_1} \right)^2 k_1 \delta s_1 + n_\alpha \left( \frac{h_{\alpha+1}}{s_{\alpha+1}} \right)^2 k_\alpha \delta s_{\alpha+1} - \sum_{i=1}^{\alpha} k_i \left[ n_i \left( \frac{h_{i+1}}{s_{i+1}} \right)^2 \delta s_{i+1} - n_{i-1} \left( \frac{h_i}{s_i} \right)^2 \delta s_i \right].$$

Rewriting the second term on the right with the help of (4) and (5), and the last term with the help of (6), we obtain

$$\sum_{i=2}^{\alpha} \delta s_i \left( \frac{1}{s_i} - \frac{1}{s'_{i-1}} \right) = -n_0 \left( \frac{h_1}{s_1} \right)^2 k_1 \delta s_1 + n_\alpha \left( \frac{h_\alpha}{s'_\alpha} \right)^2 k_\alpha \delta s'_\alpha + \sum_{i=1}^{\alpha} h_i^2 k_i K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right).$$

On substituting this expression into (10) and on using the fact that $\delta s_1 = 0$, we obtain

$$\frac{\delta M}{M} = \left( \frac{\delta n_0}{n_0} - \frac{\delta n_\alpha}{n_\alpha} \right) + \frac{\delta s'_\alpha}{s'_\alpha} \left( 1 - \frac{n_\alpha}{s'_\alpha} h_\alpha^2 k_\alpha \right) - \sum_{i=1}^{\alpha} h_i^2 k_i K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right).$$

But, by § 5.5 (22) and § 5.5 (15),

$$\begin{aligned} \frac{\delta s'_\alpha}{s'_\alpha} \left[ 1 - \frac{n_\alpha}{s'_\alpha} h_\alpha^2 k_\alpha \right] &= \frac{\delta s'_\alpha}{s'_\alpha} \left[ 1 - \frac{n_\alpha}{s'_\alpha (L_\alpha - K_\alpha)} \right] \\ &= \frac{\delta s'_\alpha}{s'_\alpha - t'_\alpha} \end{aligned}$$

so that we finally obtain the following expression for the chromatic change $\delta M$ in the magnification

$$\frac{\delta M}{M} = \left( \frac{\delta n_0}{n_0} - \frac{\delta n_\alpha}{n_\alpha} \right) + \frac{\delta s'_\alpha}{s'_\alpha - t'_\alpha} - \sum_{i=1}^{\alpha} h_i^2 k_i K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right). \tag{11}$$

If the refractive indices of the object and the image space are equal, as is usually the case, the first two terms on the right-hand side cancel out. If the system is achromatized with respect to the position of the image, $\delta s'_\alpha$ (given by (7)) also vanishes, and we then obtain the following two conditions for the achromatization of the position of the image and of the magnification

$$\sum_{i=1}^{\alpha} h_i^2 K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right) = 0 \tag{12}$$

and

$$\sum_{i=1}^{\alpha} h_i^2 k_i K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right) = 0. \tag{13}$$

For a thin lens in air, we obtain, on using the appropriate formulae § 5.6, the following expressions for the two sums:

$$\Sigma h_i^2 K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right) = h^2 \mathscr{P} \frac{\delta n}{n-1} \tag{14}$$

and

$$\Sigma h_i^2 k_i K_i \left( \frac{\delta n_i}{n_i} - \frac{\delta n_{i-1}}{n_{i-1}} \right) = h^2 k \mathscr{P} \frac{\delta n}{n-1} \tag{15}$$

where $h, k$, and the power $\mathscr{P}$ of the lens are given by §5.6 (4) and §5.6 (5). The corresponding expressions for the chromatic change in the position of the image and in the magnification are then obtained by substituting from (14) into (7) and from (15) into (11) respectively.

## 9.7 Image-Forming Instruments

The three preceding chapters give an account of the geometrical theory of optical imaging, using for the main part the predictions of Gaussian optics and of the Seidel theory. An outstanding instance of the invaluable service rendered by this branch of optics lies in its ability to present the working principles of optical instruments in an easily visualized form. Although the quality of optical systems cannot be estimated by means of Gaussian theory alone, the purpose served by the separate optical elements can be indicated in this way, so that a simple, though somewhat approximate, picture of the action of the system can often be obtained without entering into the full intricacy of the techniques of optical design.

The development of optical instruments in the past has proceeded just as fast as technical difficulties have been overcome. It is hardly possible to give a step-by-step account of the design of optical systems, for two reasons. Firstly, the limitations of a given arrangement are not indicated by the predictions of the simple theory; in particular cases this needs to be supplemented by a fuller analysis often involving tedious calculations. Secondly, difficulties of a practical nature may prevent an otherwise praiseworthy arrangement from being used. It is not intended in this account to discuss the theoretical and practical limitations in individual cases; only the basic principles underlying the arrangement of some of the more important optical instruments will be given, in order to provide a framework for some of the later chapters which deal with the more detailed theories of optical image formation.*

### 9.7.1 6.1 The Eye

Perhaps the simplest of optical instruments is that consisting of a single convergent lens forming a real image of an object upon a light-sensitive surface. Examples of such an optical system are found in the photographic camera and in the eye (Fig. 6.1). The back inner surface of the eye, † called the retina, consists of a layer of lightsensitive cells. Inasmuch as the eye forms an integral part of many optical systems, an understanding of its characteristics is an essential part of instrumental optics. The present section will include a description of a few of its features.

Refraction in the eye takes place mostly at the first surface or cornea of the eye, but also at the surfaces of the lens of the eye situated behind a contractible iris.

---

0

- For fuller accounts of optical instruments see, for example, Dictionary of Applied Physics, ed. R. Glazebrook (London, Macmillan, 1923), Vol. IV; L. C. Martin, An Introduction to Applied Optics, Vol. II (London, Pitman, 1950); G. A. Boutry, Optique Instrumentale (Paris, Masson, 1946). References to books and articles dealing with particular types of instruments will be found in D. A. Jacobs, Fundamentals of Optical Engineering (New York, McGraw-Hill, 1943).
  † For a detailed account of physiological optics the reader is referred to the standard work of H. Helmholtz, Handbuch der Physiologischen Optik (Hamburg and Leipzig); L. Voss, 3 Aufl., I (1909), II (1911), III (1910). English translation edited by J. P. C. Southall (published by the Optical Society of America, 1924); and to J. P. C. Southall, An Introduction to Physiological Optics (Oxford University Press, 1937).

The latter gives an aperture-diameter which may be varied by unconscious muscular action, according to the total flux of light entering the eye, between approximately 1.5 mm and 6 mm . The lens is doubly convex, and its density, and hence its refractive index, is not uniform but decreases towards the outer zones. This serves partly to correct for spherical aberration (see §5.3). The average value of the refractive index of the lens is $1 \cdot 44$, while that of the vitreous humour between the lens and the retina is approximately $1 \cdot 34$, near to that of water.

The radius of curvature of the lens surfaces may be altered by muscular contraction, to serve the purpose of focusing. This power of adjustment is called accommodation, and the least distance of distinct vision of a normal or emmetropic eye is around 25 cm . The accommodation of the eye varies greatly from one person to another and, in each person, with age.* The correction of excessive departures from normality by the introduction of supplementary eye-lenses forms the subject of ophthalmics. There are three principal abnormalities which occur in the

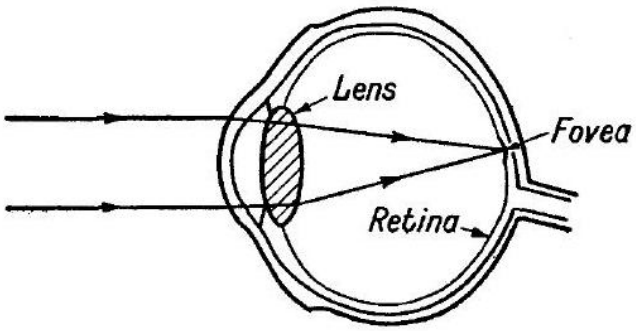



Fig. 6.1. The human eye.

eye:
(a) Myopia or short sight, in which the rays from an infinitely distant object-point reach a focus in front of the retina. This is corrected by means of a divergent lens placed in front of the eye.
(b) Hypermetropia or long sight, in which the true focus of an infinitely distant object lies behind the retina. Correction is obtained by means of a convergent lens.
(c) Astigmatism, in which the power of the eye differs in different planes containing its optical axis. This defect is corrected by means of a cylindrical or toroidal surface on the lens of the eye-glass.

The retina contains two principal types of light-sensitive cells, the rods and the cones. The latter, according to the usual theory of colour-vision, may be subdivided into the three types of colour-sensitivity. They predominate and are of greatest concentration in the fovea, the central part of the retina. The field of most distinct vision defined by this region and determining the fixation, or direction of "seeing", is slightly less than one degree in angular extent. The diameter of the cones in the fovea varies between 0.0015 mm and 0.005 mm . The rods, which are not so highly selective in colour-range, predominate in the outer regions and are the most responsive to low-intensity sources. This is the cause of the phenomenon of "averted vision" in scanning, for instance, the night sky.

There is an ultimate limit to the clarity of detail which one can attain with any given optical system. In the last analysis this limit arises, as will be seen in § 8.6.2, from the wave nature of light. The practical limit down to which two object points separated by a small distance may be distinguished by the eye, its visual acuity, is clearly a quantity of the greatest interest to the optical designer, for upon it depend the tolerances, both optical and mechanical, to which a visual instrument ought to be made. For a normal eye, this limiting angular separation is about one minute of arc in angular measure, corresponding to separation of about 0.0045 mm

on the retina, if a focal length of 15 mm is assumed. When the aperture of the eye is increased beyond 5 mm , the effect of chromatic aberration and spherical aberration will decrease the resolving power. In designing a visual instrument one therefore usually assumes that

a pencil not more than 4-5 mm in diameter is to enter the eye. Visual instruments are seldom, if ever, designed with a view to compensating the defects of the eye, because the amounts present in the eye vary from person to person.

## 9.7.2 6.2 The Camera

In the photographic camera (Fig. 6.2), a real inverted image of an object is formed by a lens or combination of lenses upon the surface of a photographic film or plate. The object may be stationary or in motion relative to the camera. In the latter case short exposures are necessary, and the aperture must therefore be as large as possible, in order to collect sufficient light. It follows from § 4.8 (24) that the light reaching the image per unit area (the illumination $E$ ) from an extended object is proportional to the ratio $d^2/f^2$, $d$ being the diameter of the entrance pupil and $f$ the focal length, i.e.

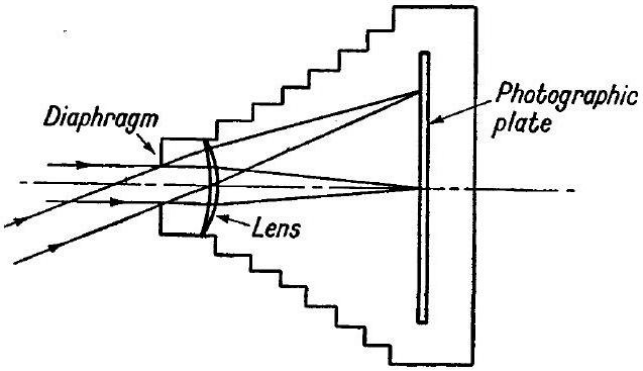



Fig. 6.2. The photographic camera.

it is inversely proportional to the square of the $F$ number (nominal focal ratio). This may usually be varied by means of a variable diaphragm.

In the absence of aberrations, the image of a distant object which subtends an angle $\delta\theta$ at the first nodal point is of linear dimensions $f\delta\theta$. Hence, in order to produce a large image, the focal length must be large. Thus the two main characteristics of a camera are the focal length of its optical system and the range of the focal ratios at which it can operate. Another possible requirement of a camera is that it should cover a large angular field. Cameras are often designed to cover a field between ten and ninety degrees, using a focal length in the range from one inch to six feet. Greater focal lengths ( $6 - 60$ft ) are needed in astronomical photography, where fields of much smaller angular extent ( $\sim 1^\circ$ ) containing effective point sources must often be reproduced on a large scale.

The simplest of small cameras use a single convergent lens, usually of meniscus form (concavo-convex) with a stop somewhat displaced from the lens, as shown in Fig. 6.2. No correction is made for chromatic effects and therefore such lenses were more satisfactory when photographic emulsions were generally sensitive to only a small region at the blue end of the

0

- At 10 years the power of the eye can cover 14 dioptres, at 25 years 10 dioptres, and at 50 years 2.5 dioptres.

spectrum ($4000 - 4800$Å). Such lenses are often used in cheaper types of camera at a focal ratio $f/11$ over a 45° field.

The "landscape lens", as this meniscus form was once called, originated about 1812 and was followed, soon after the introduction of photography, by the achromatic doublet of Chevalier (Fig. 6.3(a)). The aim of the achromatic combination was to make the focus of the photographically active light coincident with that of the visual light (brightest around $\lambda = 5800$Å ); good images could not be obtained at focal ratios lower than $f/16$. The need for a much "faster" lens (i.e. one with a lower focal
ratio) for portrait photography resulted in the widespread use of the Petzval portrait lens invented about 1840 (Fig. 6.3(b)) and comprising two separated doublets. This arrangement gave a field-surface with considerable curvature but good definition was obtainable over a field of a few degrees at focal ratios down to $f/4$. A development of the Petzval lens by Dallmeyer has been used extensively in cin-projectors.

It has been known since about 1841 that a symmetrical lens would automatically be free from distortion. This is strictly true only for the case of unit magnification, but is substantially true for other working distances. This principle has been the basis of

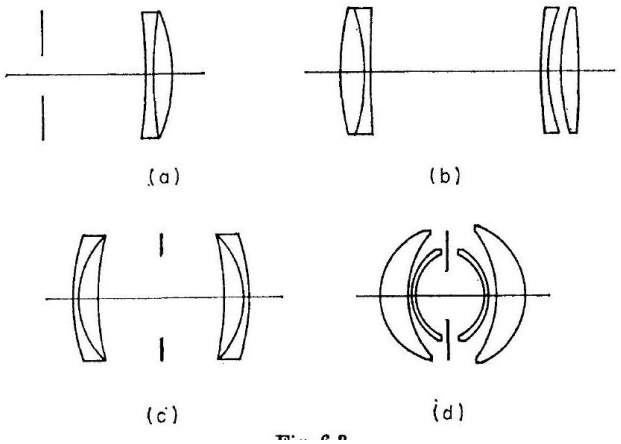



Fig. 6.3.

(a) The Chevalier lens.
(b) The Petzval lens.
(c) The Rapid-Rectilinear lens.
(d) The Topogon lens.

design of several long lines of successful photographic lenses. The Rapid-Rectilinear of Dallmeyer (Fig. 6.3(c)) dates from 1866. It was used in many cameras and eventually at focal ratios down to $f/8$ (as late as 1920), over fields of about 45°. Many firms manufactured such lenses and they became known under the general term of "aplanat". The same principle of symmetry, leading to freedom from distortion, has been continuously followed in the design of lenses intended to cover very wide fields. The original lens of this kind, the Goerz Hypergon of 1900, could, at a focal ratio around 25, cover a flat field up to 150°, although it has more often been sufficient to cover 90° at $f/16$. The four-element Topogon (Fig. 6.3 (d)) of 1933 which will cover a field of 90° at $f/6 \cdot 3$, represents the basic design for modern wide-angle lenses.

During most of the nineteenth century, lens-designers were restricted to the use of two main types of glass; boro-silicate crown, and dense flint. Around 1890 the Jena Company introduced a radically new type of glass-the barium crown-having a high refractive index but low dispersive power. Incorporating this additional type of glass into lens systems permitted higher degrees of correction to be obtained. Rudolph of Jena, whose name is associated with the origin of

many types of photographic lenses, coined the word "anastigmatic" to describe a lens for which the astigmatism at one off-axis angle could be reduced to zero. Using the older types of glass only, Rudolph had produced an anastigmat which later became known as the Zeiss Protar (Fig. 6.4(a)), but using three types of glass he was able to make an anastigmatic triplet. Two such lenses used symmetrically constituted a highly corrected combination. Such a lens
was produced by Zeiss under the name Triple Protar and by Goerz as the Dagor or double anastigmat (Fig. 6.4(b)). A typical lens of this type might work over a 50° field at $f/4.5$. Rudolph found further that if each half were formed of four cemented

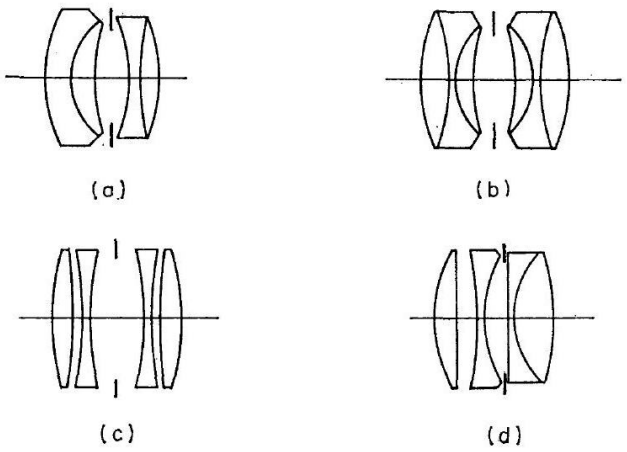



Fig. 6.4.

(a) The Protar lens.
(b) The Dagor lens.
(c) The Celor lens.
(d) The Tessar lens.
components the system could be made "convertible", i.e. the halves, having different focal lengths, could be used separately or in combination. This was produced as the Double Protar of 1894.

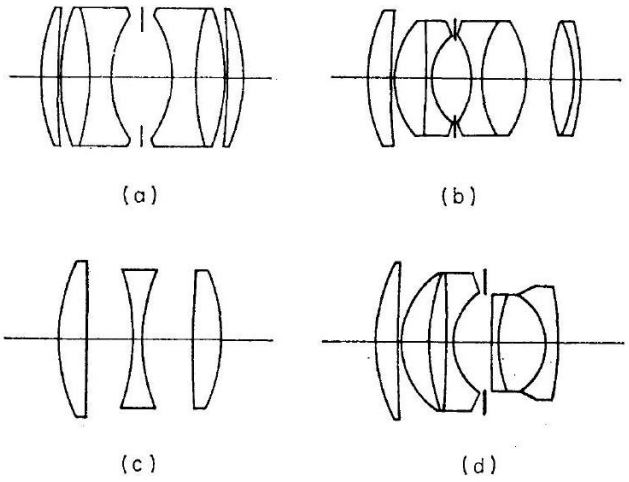



Fig. 6.5.

(a) The Planar lens.
(b) The Zeiss Biotar lens.

(c) The Cooke Triplet.
(d) The Sonnar lens.

A lens consisting of four separate elements was developed by Goerz in 1898 and called the Celor (Fig. 6.4(c)). This became the basis of design of several important lenses at focal ratios down to $f/4 \cdot 5$, such as the Unifocal of Steinheil and the Cooke Aviar lens for aerial photography. Rudolph was working along similar lines and his Tessar of 1902 (Fig. 6.4(d)) is still well known as a relatively inexpensive lens giving
excellent images over fields up to 60° and at focal ratios as low as $f/3 \cdot 5$. It has been constructed with a great variety of focal lengths.

An earlier symmetrical design by Rudolph, the Planar of 1895, incorporated two meniscus-type lenses as the first and last elements (Fig. 6.5(a)). Since 1920 many very fast lenses have been developed in which a meniscus lens is used as the first element, although in most cases the last element is a cemented doublet as for example in Mert's Zeiss Biotar (Fig. 6.5(b)).

A separate stream in lens design was initiated by the development of the threelens anastigmat, otherwise known as the Cooke Triplet (Fig. 6.5(c)), by H. D. Taylor in 1895. Full advantage was taken of the various parameters which can be altered in a system of three separated elements. This "anastigmatic" design largely replaced

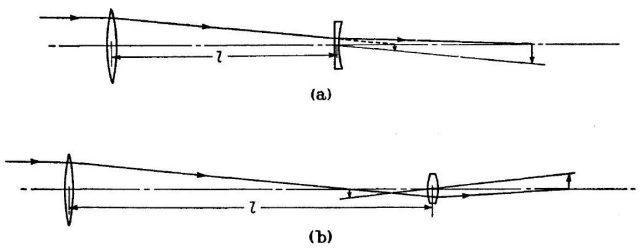



Fig. 6.6.
(a) The telephoto lens.
(b) The photoheliograph.

the older "aplanatic" lenses of the rapid-rectilinear type in low- to medium-priced cameras. Types of high-speed lens whose design developed from the Cooke Triplet (e.g. the Sonnar, $f/1 \cdot 5$, Fig. 6.5 (d)), have been used widely in cin-cameras and miniature cameras. In the Sonnar the central negative component and the second positive component are both triplets.

In cameras used for certain purposes which require long effective focal lengths and correspondingly small fields an optical arrangement consisting essentially of an objective and an enlarging lens may be used. In the telephoto lens, the enlarging lens is a divergent element placed anterior to the primary image (Fig. 6.6(a)). In the photoheliograph, an instrument used for photographing the sun on a large scale, the enlarging lens is usually a convergent system (Fig. 6.6(b)). In these systems the effective focal length $f$ is approximately given by the expression §4.7 (3):

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{l}{f_1 f_2},$$

where $f_1$ and $f_2$ are respectively the focal lengths of the objective and the enlarging lens, and $l$ is their separation. The effective focal length may be greater than the overall length from the objective to the focal plane by a factor up to fifteen.*

0

• Fuller discussions of photographic lenses will be found in W. Mert, R. Richter and M. von Rohr, Das Photographische Objectiv (Vienna, J. Springer, 1932). More recent designs are described in articles

Finally it is to be noted that the $F$ number of a photographic objective is an approximate measure of the "speed of action" of the system only where the object is an extended one. With a point-source object (as is effectively the case in astronomical photography), the light reaching the image plane would, under ideal conditions, be concentrated in a vanishingly small area, so that the square of the aperture diameter, rather than $F$ number, would characterize its light-gathering power. In reality the situation is more complicated, as several factors contribute to a spreading of the light over a finite (though often a very small) area. Chief amongst these are diffraction (see Chapter VIII), the granularity of the photographic emulsion, and the unsteadiness of the atmosphere.

## 9.7.3 6.3 The Refracting Telescope*

The telescope is an optical system by means of which an enlarged image of a distant object may be viewed. The principle of the astronomical refracting telescope is shown

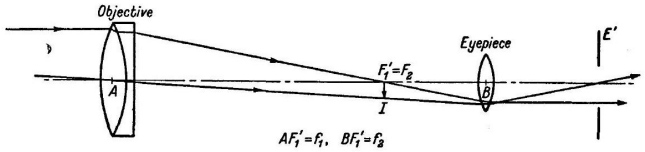



Fig. 6.7. The astronomical refracting telescope.

in Fig. 6.7. It consists of two convergent lenses the first of which, the objective, is usually an achromatic doublet forming a real inverted image $I$, which is examined using the second lens, the eyepiece. In the normal state of adjustment the second focal plane of the objective coincides with the first focal plane of the eyepiece, so that an incident pencil of parallel rays emerges as a parallel pencil. The image may be erected by the use of an auxiliary lens.

The entrance pupil of the system coincides with the objective; its image by the remainder of the system, the exit pupil, is denoted by $E'$ in Fig. 6.7. The eye should be so placed that its entrance pupil coincides with $E'$; then all the light entering the objective at different off-axis angles will reach the eye.

The magnifying power of an instrument used for examination of objects at infinity is defined as the angular magnification at the pupils. By the Smith-Helmholtz formula, §4.4 (49), this is the reciprocal of the linear magnification at the pupil planes, so that the magnifying power is equal to the ratio of the radius of the entrance pupil to that of the exit pupil. An expression for the magnifying power is immediately obtained from §4.3 (31) where, because the angles are assumed to be small, $\gamma$ and $\gamma'$ may be written in place of $\tan\gamma$ and $\tan\gamma'$. This gives

$$\frac{\gamma'}{\gamma} = \frac{f_1}{f_2'}.$$

Now the eye may be accommodated to distances other than infinity and the telescope may easily be adjusted to form a virtual image at, say, a distance $D$ from

by W. Taylor and H. W. Lee, Proc. Phys. Soc., 47 (1935), 502; H. W. Lee, Rep. Progr. Phys. (London, Physical Society), 7 (1940), 130; R. Kingslake, Proceedings of the London Conference on Optical Instruments (London, Chapman & Hall, 1951), 1 and C. G. Wynne, Rep. Progr. Phys. (London, Physical Society), 19 (1956), 298.

0

- For a fuller discussion of telescopes see, for example, A. Danjon and A. Couder, "Lunettes et Tlescopes" (Paris, Revue d'Optique, 1935).

the eye.* $D$ should, however, not be decreased below the value of about 25 cm which, as already mentioned, is the closest distance at which an object can be viewed distinctly without eye-strain with a normal eye.

The earliest telescope of which there exists definite knowledge is that of Galileo, constructed in 1609. In this system (Fig. 6.8) the objective was a convergent lens, but the eyepiece was a divergent lens, placed anterior to the primary image in such a way that the focal points of the two lenses coincided beyond the eyepiece. An erect image was then obtained at infinity, there being no intermediate image. In this type of telescope, known as the Galilean telescope, the aperture stop is not in the plane of the objective, for the image of the objective formed by the eyepiece lies between the two lenses and is therefore not accessible to the eye. The eye is placed close behind the eyepiece, so that the pupil of the eye becomes the aperture stop and the exit pupil. The principal ray for points at the limit of the field will pass very near to

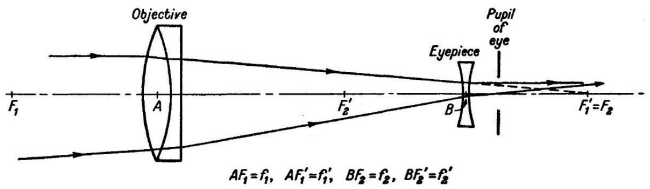



Fig. 6.8. The Galilean telescope.

the edge of the objective, so that the objective serves as a field stop rather than as an aperture stop.

A disadvantage of the Galilean telescope is that it is limited to small fields and that it does not possess a real image where cross-wires, or a graticule, may be placed. On the other hand, its small overall length and the fact that it gives an erect image make the system very suitable for use in opera glasses, usually with a magnification of two to three times.

In the binocular, which uses the optical arrangement similar to that of an astronomical telescope, an erect image is obtained by means of four reflections, as shown in Fig. 6.9. These reflections take place at an incidence angle of 45° at the glass-air surfaces of so-called Porro prisms; this angle being, of course, beyond the critical angle. A reflection of this kind forms a convenient and efficient device often used in small optical instruments. An alternative form of binocular uses the König erecting prism. This incorporates the device of the "roof" where two adjacent reflecting surfaces, at right angles to each other within a second of arc, are placed athwart the optical beam, as shown in Fig. 6.10. The larger spacing in the more usual type of binocular of the two objectives is of advantage, however, since it serves to enhance the stereoscopic effect. †

Terrestrial telescopes are generally similar to the astronomical telescope; but an erect image must be obtained, either by the help of erecting prisms, as in the binocular, or with the use of additional lenses. A typical system is shown in Fig. 6.11.

The telescope objective is usually an achromatic combination of two lenses in which the radii of curvature are such as to give, with spherical surfaces, little spherical

0

• The magnification is then changed to the value:

$$\frac{\gamma'}{\gamma} = \frac{f_1}{f_2'} + \frac{f_1}{D}.$$

† For a fuller account of binoculars see M. von Rohr, Die Binokularen Instrumente (Berlin, J. Springer, 1920).

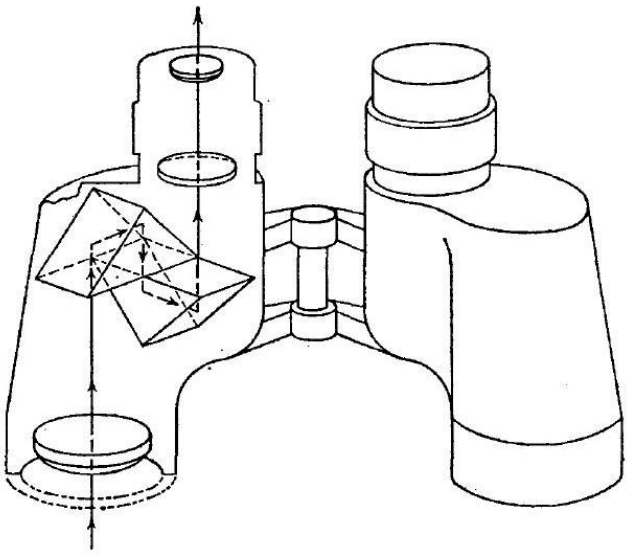


Fig. 6.9. The binocular telescope.

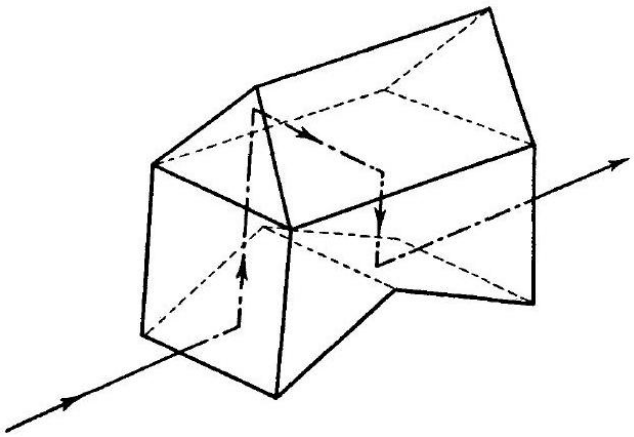


Fig. 6.10. The König erecting prism.

aberration. If a cemented doublet is to be used, a design in which the first (crown) component is equi-convex and the second (flint) component is plano-concave is often a suitable solution (Fig. 6.12(a)). If freedom from coma is desired as well as freedom from spherical aberration, as in a high-quality astronomical telescope, it is necessary to utilize the extra flexibility in design of two non-cemented components, and to choose the radii to give an aplanatic design. The design may then be approximately that

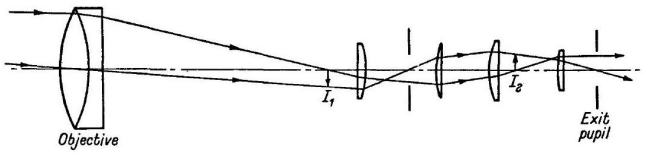



Fig. 6.11. The terrestrial telescope.

shown in Fig. 6.12(b). This includes a concavo-convex flint component. The steep concave curve of the first surface of this component can be reduced if a coma-free design with aspherical figuring on one surface to give axial stigmatism is adopted.

An objective with the crown lens in front is the usual form, originally known as the Fraunhofer objective. The Steinheil, or "flint-in-front", objective is less used because of the greater susceptibility of the flint glass (less stable than crown glass) to atmospheric attack.

The photo-visual objective, shown in Fig. 6.12(c), is the third type of objective commonly used in refracting telescopes. The use of three different glasses permits a considerable reduction of chromatic difference of focus over a given spectral range to
be obtained. This objective is almost equally good for "photographic" and "visual" wavelengths, and at one time was popular for astronomical use at apertures up to about 12 in . The steep curves of the centre component, making centering difficult, and the early use of unstable varieties of flint glass, have detracted, however, from the value of these lenses.

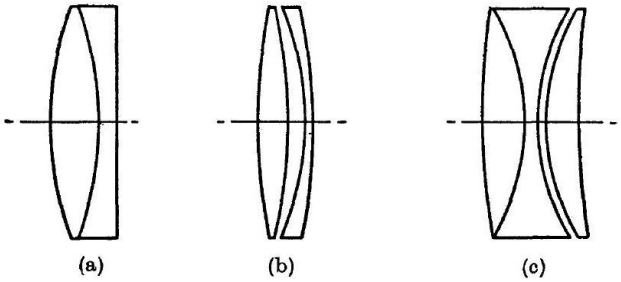



Fig. 6.12.

(a) Cemented achromat.
(b) Aplanatic achromat.
(c) Photo-visual objective.

The eyepiece of a visual telescope must, broadly, satisfy two requirements. Firstly, it must have a focal length which will give the required degree of magnification. Secondly, it must have sufficient aperture to collect the light from an extended object. These requirements can be achieved by means of a single lens, as shown in Fig. 6.13(a), but more conveniently two smaller lenses can be used as in Fig. 6.13(b), which shows the optical arrangement of the Ramsden eyepiece.

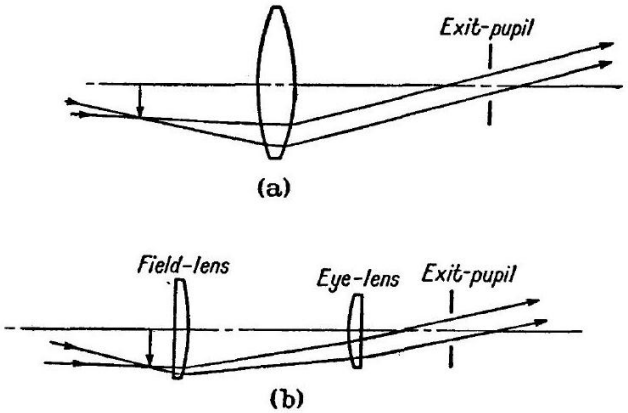



Fig. 6.13.

(a) Single-lens eyepiece.
(b) The Ramsden eyepiece.

The eyepiece should form an image of the aperture stop at a position convenient for placing the eye. The distance of this image behind the last surface of the eyepiece is called the eye-relief.

The first component of a two-component eyepiece is called the field-lens because it is often very near to the primary image. It is separated from this image partly to prevent any dust or imperfection on the lens being visible, and partly to render the plane of the image accessible to cross-wires, or a graticule. The second component is called the eye-lens.

The most common eyepiece in use is the Ramsden eyepiece, and the other chief type is the Huygenian eyepiece, shown in Fig. 6.14(a), which has the field lens anterior to the primary image. The disadvantages of the Huygenian eyepiece are firstly the rather short eye-relief obtainable, and secondly the fact that it cannot be used with

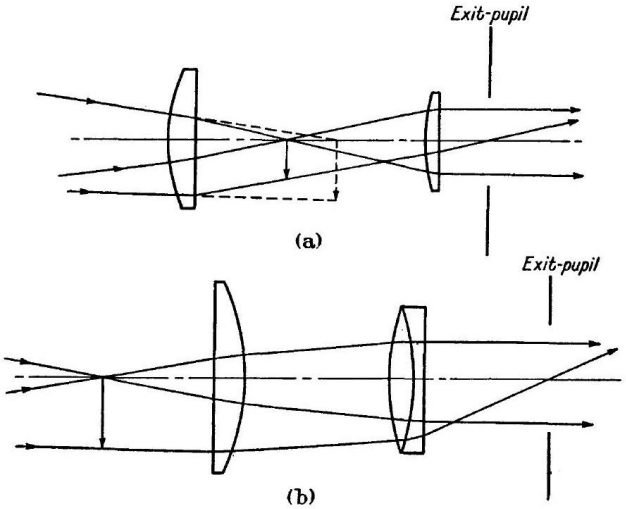



Fig. 6.14.
(a) The Huygenian eyepiece.
(b) The Kellner eyepiece.

cross-wires external to itself. It cannot give such a high correction for spherical aberration as the Ramsden eyepiece, but is free from lateral chromatic aberration and off-axis coma.

The use of two lenses (see Fig. 6.14) instead of one aids the removal of harmful chromatic aberrations, and correction in the Ramsden eyepiece can be greatly

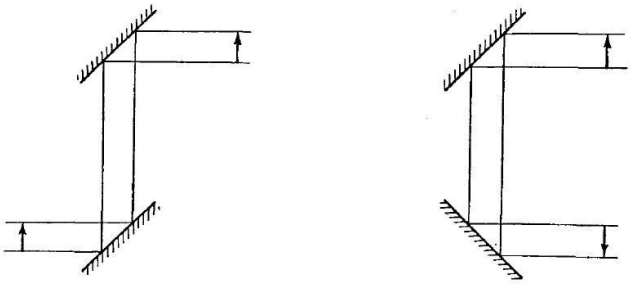


Fig. 6.15. Periscopic mirrors.

improved by using a doublet as the eye-lens. More complicated are the "orthoscopic" and "symmetrical" eyepieces, each using four components but generally similar to the Ramsden eyepiece. A higher degree of correction for aberrations is obtained, especially with regard to distortion, and longer eye-reliefs are available for a given power.

The principle of the periscope, i.e. an instrument used for viewing objects which are situated so that direct viewing is precluded by an obstacle, is the same as that of the

telescope, but two requirements make the designing of such an instrument more complicated. In the first place, the periscope must operate with a bent ray-path and

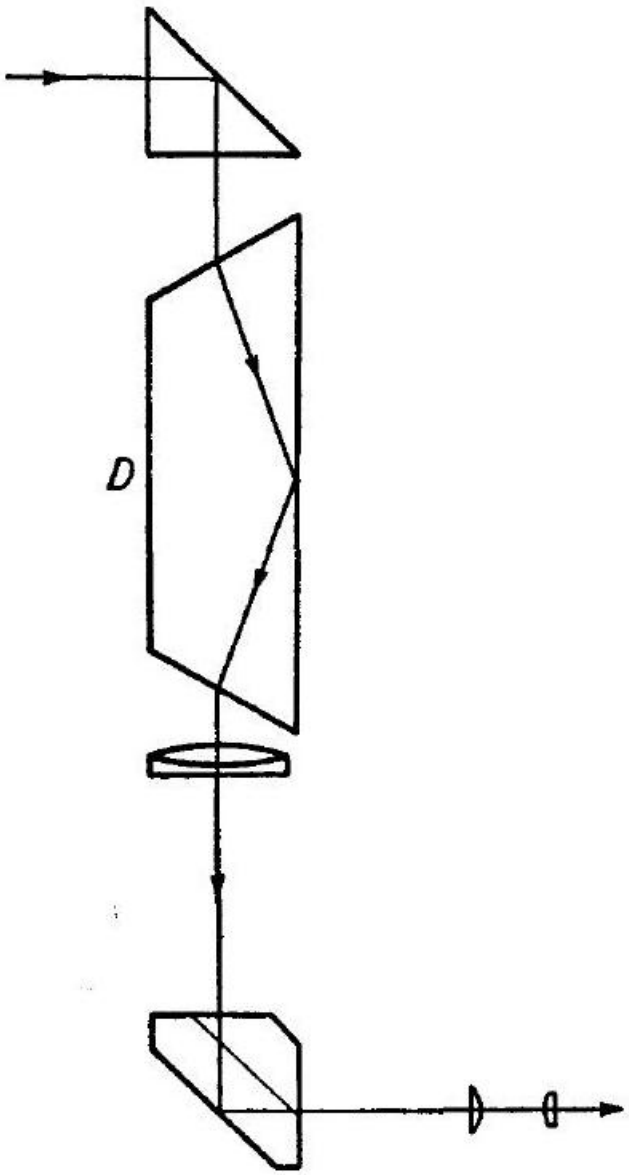



Fig. 6.16. Periscope with Dove prism.

give an erect non-reversed image. An even number of reflections, or none, are necessary to give a non-reversed image, and two simple arrangements are shown in Fig. 6.15, the first producing parallel and the second antiparallel directions of incident and emergent beams. The first must be used with an erecting telescopic system and the second with an inverting telescope. Another arrangement is shown in Fig. 6.16, where the direction of the emergent beam is fixed while the first prism rotates to give different directions of view. The Dove prism $D$, which incorporates a single reflection, rotates at half the speed of the top element in order to maintain an erect image. The fourth prism is of the "roof" type (Amici) so as to provide, in all, four reflections.

The second complication which exists in periscopes arises from the desire for a large angular field of view, even though the telescope is contained within a narrow tube. A tube eight inches in diameter and forty feet long permits, at maximum, a total field of view of only two degrees. The solution to this problem which is adopted makes use of a series of lenses passing the light

down the tube. Different spacings of the lenses are possible. In Fig. 6.17 three arrangements are shown of a system of unit magnification which employs three lenses.

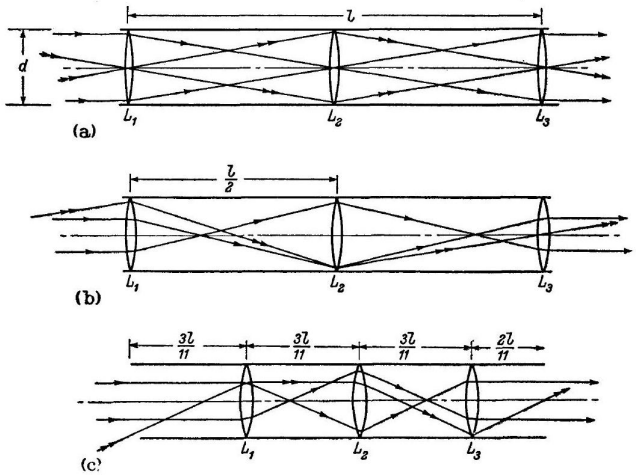



Fig. 6.17. Three-lens systems; unit magnification.

Table XIV

| | Focal lengths of lenses | | | Image | Aperture on-axis | Field (radians) | Vignetting at limit of field |
|---|---|---|---|---|---|---|---|
| | $L_1$ | $\boldsymbol{L_2}$ | $L_3$ | | | | |
| (a) | $l/2$ | l/4 | $\boldsymbol{l2}$ | Inverted | $d$ | $2d/l$ | 0 |
| (b) | $l/6$ | $l/6$ | $l/6$ | Erect | $d/2$ | $2d/l$ | 100% |
| (c) | l/11 | l/11 | l/11 | Erect | $d/2$ | $11d/2l$ | $\sim 86\%$ |

The range of fields of view of these three systems and their effective apertures are shown in Table XIV. It will be clear that, with a larger number of lenses, a wider range of possibilities exists.*

## 9.7.4 6.4 The Reflecting Telescope

In the astronomical reflecting telescope, the light from a celestial objective is received on a concave primary mirror. The mirror serves the same purpose as the objective in a refracting telescope, namely to form a real image of the object in its focal plane. This image is either received directly on a photographic plate, or is examined visually by an eyepiece. This type of arrangement is the most common form of astronomical telescope used to-day. As in the case of the refracting telescope, the magnifying power is equal to the ratio of the focal length of the objective to that of the eyepiece.

The first telescope of this type was made in 1668 by Newton. In the arrangement which he used a small plane mirror was placed in the path of the rays reflected by the primary mirror, so as to divert the rays to one side of the tube, where the image could be conveniently examined (Fig. 6.18). This arrangement is known as the Newtonian telescope.

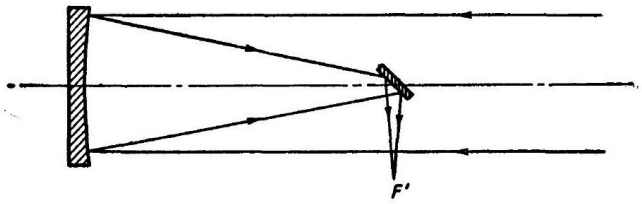



Fig. 6.18. The Newtonian telescope.

In order to have a stigmatic axial image, the figure of the mirror must be a paraboloid. The fact that off-axis images formed by a parabolic mirror suffer strongly from coma, together with the circumstance that large relative apertures are employed, makes the usable field very small. In the case of a 36 in . $f/6$ mirror the total field which can be covered is of the order of 20 minutes of arc, while in the case of the largest telescope in existence to-day, the Hale telescope, which has a primary mirror of diameter 200 in ., the field is about 45 minutes in angular diameter. †

Other types of reflecting telescope employ two principal mirrors, a primary concave mirror and a secondary mirror. In the Cassegrain telescope (Fig. 6.19) the secondary mirror is convex. In a large astronomical telescope, such an arrangement is often an alternative to the Newtonian arrangement. Different mirrors can be utilized to provide a variety of focal lengths. If the figure of the primary mirror is a paraboloid, the secondary mirror must have a hyperbolic figure. Other figures for the two mirrors can be chosen which give a stigmatic axial image and different amounts of off-axis coma. In §4.10 we mentioned a two-mirror system proposed by Schwarzschild, which is completely aplanatic; the figures of the two mirrors are, however, more complicated. The Cassegrain system is often employed in terrestrial telescopes used for viewing landscapes as it provides a compact and inexpensive instrument with a long focal length, and a high degree of enlargement can be obtained without the need for a short eye-relief.

The Gregorian telescope, less common than the Cassegrain, employs a concave secondary mirror placed beyond the prime focus of the primary mirror.

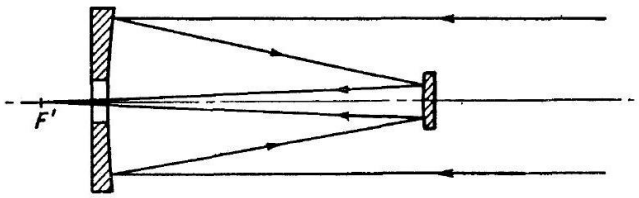



Fig. 6.19. The Cassegrain telescope.

In the largest reflecting telescopes, very great focal lengths, greater than that of the largest refractor, may be obtained by the use of a secondary mirror. In the Coud arrangement the beam is reflected by a third (plane) mirror down the polar axis (about which the telescope rotates), so that it can be made to feed a fixed spectrograph.

One of the principal advantages of the reflecting telescope is its complete freedom from chromatic aberration. This, as well as the fact that the curvature required in the mirror-surface is much less than that in a lens, enables reflectors to be built which have a much smaller focal ratio than refractors. As well as giving brighter images, this results in a much more compact optical system. Moreover, a mirror can be built in larger sizes than a lens, because optical inhomogeneities in a mirror-block are of no significance.

The disadvantage of the reflecting telescope for astronomical use lies principally in its sensitivity to thermal effects in the mirror-surface and in the telescope-tube, in its inability to cover more than a small angular field of view, and in the mechanical difficulties of making a sufficiently rigid mounting. It is this last aspect, in fact, which produced the most difficult of the problems which confronted the builders of the Hale telescope.

0

- Further details will be found in an article by T. Smith in the Dictionary of Applied Physics, Ed. R. Glazebrook (London, Macmillan, 1923) Vol. IV, p. 350.
  † F. E. Ross, Publ. Astr. Soc. Pac., 46 (1934), 342. For a general account of the 200 in. telescope see articles by J. A. Anderson, Publ. Astr. Soc. Pac., 60 (1948), 221; and B. Rule idem, p. 225.

The successful use of the reflecting telescope with photographic plates has produced the desire for covering larger fields, and one of the methods available for increasing the usable field consists in employing field lenses just in front of the photographic plate. These lenses are designed to introduce off-axis coma such as to compensate that of the main mirror. Fields of the order of $1\frac{1}{2}^\circ$ may be obtained with a 36 in . telescope.*

Much more successful in an attempt to photograph wide angular fields with large

reflecting systems has been the optical arrangement already referred to in §4.10, invented by B. Schmidt around 1930.* While a parabolic mirror gives perfect imaging for the axial rays and badly comatic images a short distance off-axis, a

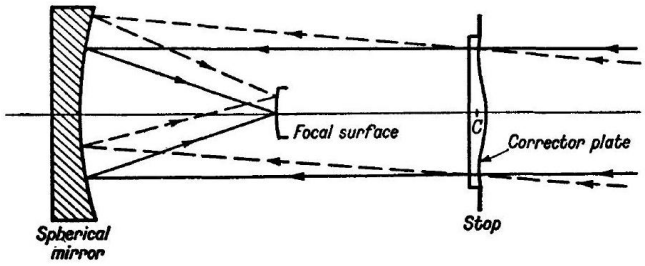



Fig. 6.20. The Schmidt camera.
(The figuring of the corrector plate is greatly exaggerated.)

spherical mirror with an aperture-stop at its centre of curvature $C$ would give uniform images over a wide spherical field-surface concentric with itself, each image suffering from a large amount of spherical aberration. Schimidt introduced into the aperture stop a thin, nearly plane-parallel plate, called the corrector plate, one face of which was plane, whilst the other was figured (Figs. 6.20, 6.21). The purpose of this plate was

0

• See F. E. Ross, Astrophys. J., 81 (1935), 156.

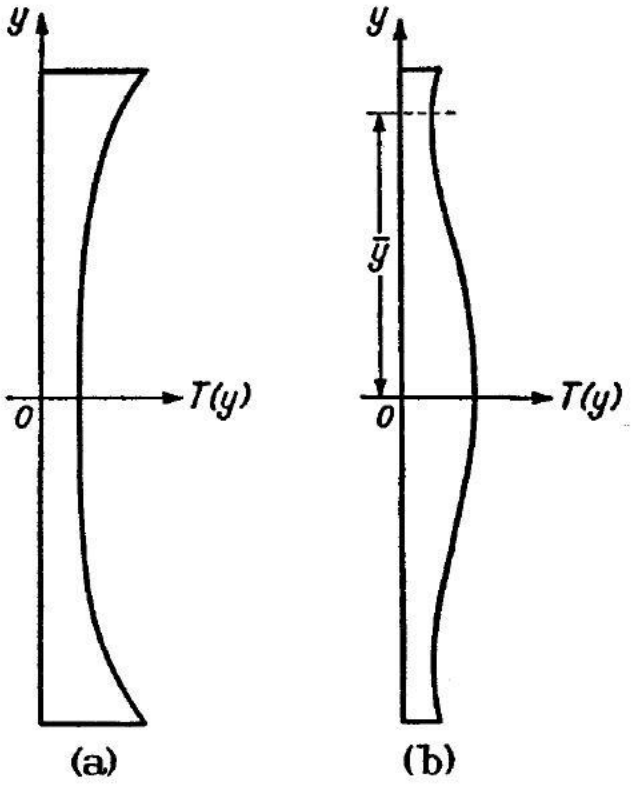



Fig. 6.21. Profiles of Schmidt corrector plates.

(The figurings are greatly exaggerated.)

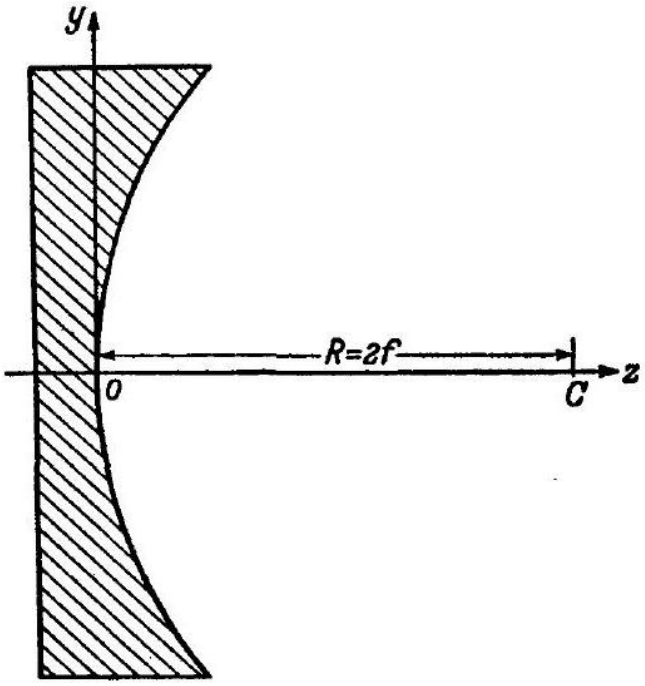



Fig. 6.22. Spherical mirror.

to "pre-correct" the plane wave-fronts entering the system in such a way as to compensate exactly the spherical aberration introduced by the spherical mirror.

We can easily derive an expression for the aspheric profile of the corrector plate of the Schmidt camera. For this purpose let us compare the expressions for a spherical

and a paraboloidal mirror. The equation of a spherical mirror of radius $R = 2f$ is given by (see Fig. 6.22) i.e.

$$\left(z^{(s)} - 2f\right)^2 + y^2 = (2f)^2$$
$$z^{(s)} = \frac{y^2}{4f} + \frac{y^4}{64f^3} + \dots$$

whilst for a paraboloid with the same paraxial radius ( $R = 2f$ ),

$$z^{(p)} = \frac{y^2}{4f}$$

Now a pencil of parallel rays incident in the direction of the axis would be rendered stigmatic by the paraboloid; hence at the zone of radius $y$, the amount of precorrection which has to be introduced in the waves incident upon the spherical mirror is given by

$$2\left(z^{(s)} - z^{(p)}\right) = \frac{y^4}{32f^3}$$

approximately, terms involving sixth and higher powers of $y$ being neglected.* After passage through the plate, the rays are still nearly parallel, so that it makes no appreciable difference whether $y$ is measured at the mirror or at the plate. Hence if $n$ is the refractive index of the plate, its thickness $T(y)$ for a zone of radius $y$ must exceed its axial thickness $T(0)$ by an amount given by

$$(n-1)[T(y) - T(0)] = 2\left(z^{(s)} - z^{(p)}\right)$$
$$T(y) - T(0) = \frac{y^4}{32(n-1)f^3} \tag{1}$$

The profile of such a corrector plate is shown in Fig. 6.21(a).
The asphericities required are very small. Consider for example an $f/3 \cdot 5$ Schmidt camera, with a corrector plate of aperture diameter $2y_0 = 40$ cm. Then $f = 140$ cm. With $n = 1.5$ the maximum asphericity $[T(y) - T(0)]_{\max}$ is, according to (1), equal to ${y_0}^4/32(n-1)f^3 = 0.0036$ cm.

A further improvement may be obtained by comparing the spherical mirror with a somewhat different "reference paraboloid". If $f'$ is the focal length of the paraboloid, then $z^{(p)} = y^2/4f'$, and one obtains in place of (1) the following expression for the plate profile:

$$T(y) - T(0) = \frac{y^4 - Ay^2}{32(n-1)f^3} \tag{2}$$

where

$$A = 16f^3\left(\frac{1}{f'} - \frac{1}{f}\right)$$

With a corrector plate given by (2), rays incident parallel to the axis will converge, after passing through the system, to a point at a distance $f'$ from the axial point of the mirror. This additional degree of freedom (choice of $f'$ ) may be used to minimize the chromatic aberration introduced by the plate. Since

[0]* B. Schmidt, Central-Zeitung f. Optik u. Mechanik, 52 (1931), Heft 2, and Mitt. Hamburg. Sternw. Bergedorf, 7 (1932), No. 36, 15. See also R. Schorr, Zeitschr. f. Instrumkde, 56 (1936), 336, Mitt. Hamburg. Sternw. Bergedorf, 7 (1936), Nr. 42, 175, and Astr. Nachr., 258 (1936), 45.

$$\frac{dT}{dy} = 0 \quad \text{when} \quad y\left(2y^2 - A\right) = 0$$

it follows that a ray parallel to the axis and incident at height $\bar{y} = \sqrt{A/2}$ is not deflected by the plate. The zone of radius $\bar{y} = \sqrt{A/2}$ is known as the neutral zone. In order to minimize the chromatic aberration, the neutral zone must be fairly close to the edge of the plate,* as shown in Fig. 6.21(b).

It is evident that the corrector plate does not impart to rays which pass obliquely through the plate precisely the same deviations as to those passing normally. In consequence, the system will not be free from off-axis aberrations. Judged by ordinary standards, the images formed on the spherical receiving surface concentric with the

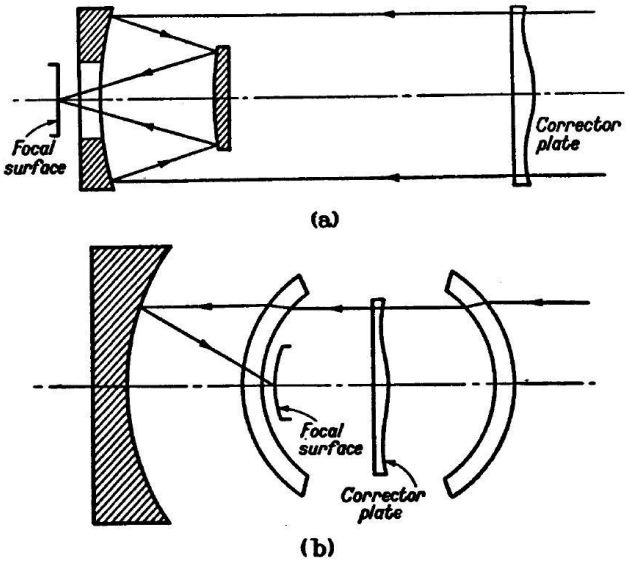



Fig. 6.23.
(a) The Schmidt-Cassegrain camera.
(b) The Baker Super-Schmidt camera.

mirror are, however, of excellent quality; † this is so even in a Schmidt camera of low focal ratio, working over a field many times larger than is possible with systems of a more conventional design.

More complex systems using the principle of the Schmidt camera, have been considered ‡ and in some cases built. Most important amongst these are the SchmidtCassegrain camera which employ two (spherical or aspheric) mirrors together with a

0

- This may be shown to be a permissible approximation if the focal ratio of the Schmidt camera is not lower than about $f/3$.

0

- More precisely, the diameter of the colour confusion circle may be shown to be least when the neutral zone is at about 0.87 of the full radius $y_0$ of the plate. This corresponds to a choice $A = \frac{3}{4}{y_0}^2$. For a discussion of this point see B. Strömgren, Vierteljahrsschrift der Astr. Ges., 70 (1935), 82, and E. H. Linfoot, Mon. Not. Roy. Astr. Soc., 109 (1949), 279. [Also his Recent Advances in Optics (Oxford, Clarendon Press, 1955) 182 and 192.]
  † This is partly due to the fact that field curvature may be shown to be the only primary aberration of the system. (B. Strömgren, loc. cit.) The fifth-order aberrations of the Schmidt camera were investigated

corrector plate (Fig. 6.23(a)) and the Baker Super-Schmidt camera, consisting of a spherical mirror, two menisci, and a corrector plate (Fig. 6.23(b)). The former possesses the advantage that it operates with a flat field surface, which can be made more accessible than in the Schmidt camera; also, the overall length of this system is shorter than that of the Schmidt camera with the same focal length. The Baker Super-Schmidt camera can operate with extremely low focal ratios.

## 9.7.5 6.5 Instruments of Illumination

The function of collecting as much light as possible from an artificial source and of deflecting it through the entrance pupil of an optical system, covering specified angles

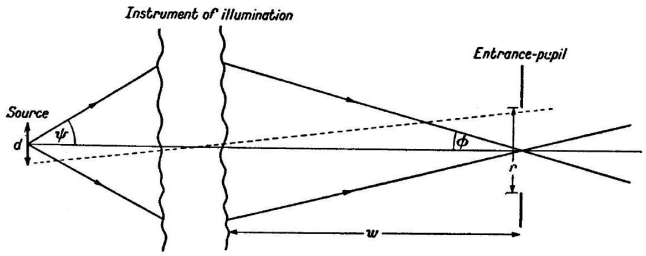



Fig. 6.24. Notation relating to an instrument of illumination.

of acceptance at that entrance pupil, is carried out by an instrument of illumination. Light can only be obtained from a source of finite size and so the design of the instrument depends upon the nature of the source as well as upon the specification of the illumination required.

In Fig. 6.24, the entrance pupil of diameter $r$ is shown, of an optical system requiring illumination over directions making angles with the axis up to a value $\phi$. Let $d$ be the diameter of the source, supposed circular. For greatest efficiency the illuminating system must work with magnification $r/d$, while its aperture must subtend an angle $2\phi$ from the entrance pupil. The working distance $w$ must be greater than a certain prescribed value. It is clear from a consideration of Fig. 6.24 that for lower values of $d$ the angle of collection, $2\psi$, will be greater, hence less light will be wasted. Also, large angles of illumination (large $\phi$ ) will often make possible smaller values of $w/r$, so that larger sources can then be used with reasonable efficiency. Thus it is clear that in any particular case there is a restriction on the largest source which will give reasonable efficiency and that this restriction is less stringent when large angles of illumination are required. In a photographic enlarger ( $\phi \sim 20°$ ), large sources (frosted bulbs or cold-cathode tubes) often provide reasonably efficient illumination. In a cinematograph projector ( $\phi \sim 5°$ ) requiring maximum brightness, an electric filament is suitable. With a searchlight, where a strong beam over only one or two degrees is required, especially small bright sources are the only possible ones, unless the size of the instrument is to be unduly great.

In many instruments, the illumination can be provided by means of a suitable source and a condenser lens. This lens must be sufficiently free from aberrations to be able to utilize the smallness of the source. Every condenser may be said to possess a

by C. Carathodory, Elementare Theorie des Spiegelteleskops von B. Schmidt (Leipzig, Teubner, 1940), and E. H. Linfoot, loc. cit.

‡ For a review of such systems see H. Slevogt, Zeitschr. f. Instrumkde, 62 (1942), 312; H. Köhler, Astr. Nachr., 278 (1949), 1; E. H. Linfoot, Mon. Not. Roy. Astr. Soc., 108 (1948), 81. [Also his Recent Advances in Optics (Oxford, Clarendon Press, 1955) Chapter IV.] For a fuller description of the Baker Super-Schmidt camera see F. L. Whipple, Sky and Telescope, 8 (1949), 90.

"focal sphere", which represents the minimum source size with which it can efficiently be used. A design which is usually sufficiently good to give satisfactory results is that of the double plano-convex condenser (Fig. 6.25(a)), though sometimes a triple condenser is used (Fig. 6.25(b)). In other cases a single lens may be adequate.

The illumination required by microscopes involves more complex consideration and will be briefly discussed later (§ 8.6.3 and § 10.5.2).

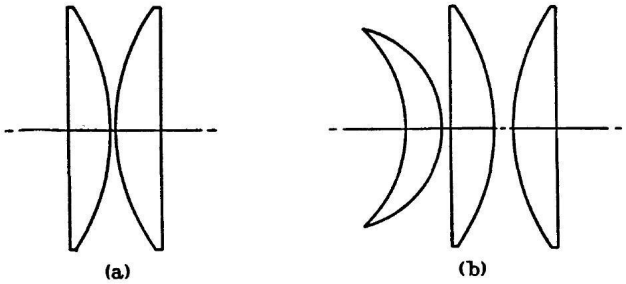



Fig. 6.25. Condenser lenses.

## 9.7.6 6.6 The Microscope

The apparent size of an object is determined by the size of its retinal image. If the eye is unaided, this apparent size depends on the angle which the object subtends at the eye. For a normal eye, the least distance of distinct vision is, as mentioned in §6.1, about 25 cm . This is the most favourable distance at which to examine the detail of any object. If a convergent lens is placed before the eye, the object may be brought much closer; for the lens will form an enlarged virtual image at a distance greater

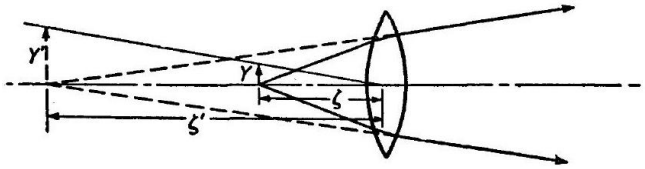



Fig. 6.26. The magnifier.

than the object distance (see Fig. 6.26), and it is the virtual image rather than the object itself, which is then being viewed.

The magnifying power of a visual instrument used to examine nearby objects is defined as the ratio of the angle subtended at the eye by the image of an object when the object is so placed that the image is at a standard distance (usually 25 cm ) from the eye, to the angle subtended by the object when placed at the standard distance from the eye and viewed directly. This is equivalent to defining the magnifying power $M$ as the ordinary linear magnification when the image is situated at the standard distance in front of the exit pupil of the instrument. For a simple lens, we have (see Fig. 6.26)

$$M = \frac{Y'}{Y} = \frac{\zeta'}{\zeta},$$

where $\zeta' = 25$ cm is the standard distance. By § 4.4 (31),

$$\frac{1}{\zeta} - \frac{1}{\zeta'} = \frac{1}{|f|}, \tag{2}$$

so that

$$M = 1 + \frac{\zeta'}{|f|} = 1 + \frac{25}{|f|}, \tag{3}$$

the focal length $f$ being measured in centimetres in the last expression. Since $f$ is usually small compared to 25 cm , the magnifying power may simply be written as

$$M \sim \frac{25}{|f|}. \tag{4}$$

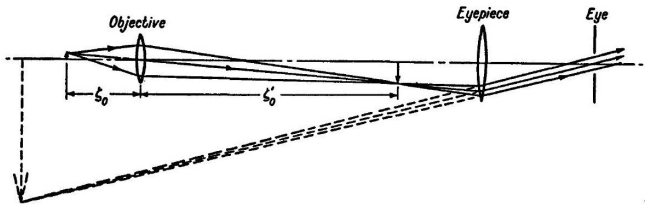



Fig. 6.27. Illustrating the principle of the microscope.

Various forms of magnifiers are used; the simple double-convex lens, or an achromatic doublet, is the most usual in pocket magnifiers or watchmaker's eye-glasses. A separated pair of plano-convex lenses similar to the Ramsden eyepiece (Fig. 6.13(b)) is often used. It is evident that the distance between the object and the eye will become inconveniently small if larger apparent fields of view are required with large magnification. For this reason the optical system of the microscope (invented by Galileo about 1610), employing an objective of short focal length and a magnifying eyepiece, was evolved, the magnification being achieved in two stages (Fig. 6.27).

The objective of the microscope forms an enlarged image of the object in a position suitable for viewing through the eyepiece. The magnification at the objective is given by - $\zeta_0'/\zeta_0$ where $\zeta_0$ and $\zeta_0'$ are the working distances of the objective, satisfying a relation of the form (2). If $f_1$ is the focal length of the eyepiece, measured in centimetres, the magnification at the eyepiece may, according to (4), be written as $25/|f_1|$. Hence the magnification of the system working as a whole is

$$M = -\frac{25}{|f_1|}\frac{\zeta_0'}{\zeta_0},$$

the negative value indicating that the image is inverted. The magnifications of the objective and eyepiece are usually stated separately in microscope specifications.

The other quantity of interest in the general specification of a microscope objective is its numerical aperture (N.A.) already defined in § 4.8.2 as the product $n \sin\theta$, where $\theta$ is the angular semi-aperture on the object side (i.e. the semi-angle of the cone of rays from the axial object point which is received by the objective) and $n$ is the refractive index of the medium of the object space. This quantity is a measure not only of the
light-gathering power of the objective, but also, as will be shown in § 8.6.3, of its resolving power, i.e. of the ultimate limit to the clarity of detail which can be obtained with it. Here we shall describe how the highest numerical apertures are attained.

Microscope objectives must, in general, be highly corrected for spherical aberration and coma and for chromatic aberration, since they are to receive rays over as large an aperture as possible. In objectives of low magnification ( $\sim 10\times$ ) it is possible to employ two separated cemented achromats (Fig. 6.29(a)), the combination being corrected for spherical aberration

and coma. Such lens combinations are, however, not suitable for objectives of high magnification ( $\sim 50\times$ or more). One then uses different lens systems, which utilize the existence of aplanatic points of a spherical surface (cf. § 4.2.3) in the following way:

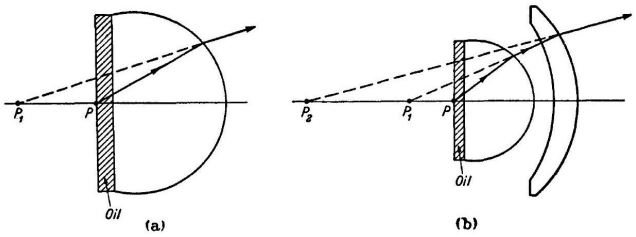



Fig. 6.28. Reduction of the angular divergence of rays in microscopes.

The object is placed at a point $P$ close to the plano-convex lens (Fig. 6.28(a)) and the space between the object and the lens is filled with oil, the refractive index $n$ of which is nearly equal to that of the lens. If $C$ is the centre of curvature of the curved portion of the lens, and $r$ the radius of curvature, then a point at distance $r_1/n_1$ from $C$ will give rise to a virtual image $P_1$ at distance $n_1 r_1$ from $C$, the imaging between the two points being aplanatic. In this way a higher numerical aperture may be employed, the angular divergence of the rays from $P$ being reduced within the system, without the introduction of monochromatic aberrations. Chromatic aberration is, however, introduced and this must be compensated by the rest of the system.

The angular divergence of the rays may be reduced still further by the use of a convergent meniscus* (Fig. 6.28(b)). The front surface of the meniscus has its centre of curvature at $P_1$, and the radius of its rear surface is such that $P$ is an aplanatic point with respect to it. The rays refracted from this surface form a virtual image at the other aplanatic point $P_2$. Adding further menisci, it is possible to produce successive virtual images $P_3, P_4, \ldots$ lying further and further away from $P$, thus reducing the angular divergence of the rays more and more. However, more than two lenses are seldom employed as otherwise the chromatic aberration cannot be adequately compensated.

The oil-immersion objectives just described admit a wider cone of rays than a dry objective of the same diameter would. For in a dry objective, the cone of rays emerging from the cover glass which protects the object pass into air, and consequently the rays are bent outwards by the refraction, the angular divergence being thus increased;

in an oil-immersion objective, on the other hand, they are not refracted on emergence from the cover glass. It is only with oil-immersion objectives that the highest numerical apertures ( $\sim 1 \cdot 4$ ) can be attained.

The demand has also arisen for microscope objectives which can operate with light which includes the near ultra-violet wavelengths ( $\lambda \sim 2500$Å ) and wavelengths of the visible and the near infra-red regions. For example, if a microscope is to be used in

0

- This method is due to G. B. Amict, Ann. de chim. et phys. (3), 12 (1844), 117.

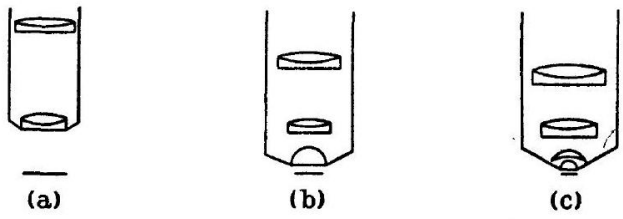



Fig. 6.29. Microscope objectives.
$\begin{array}{lll}(a) Low power. & (b) Medium power. & {(c) High power.}\end{array}$

conjunction with a spectrograph and a photometric instrument, a high degree of achromatism is necessary.

The design of achromatic microscope objectives was first systematically investigated by Burch,* who took as starting point Schwarzschild's analytical solution of the aplanatic two-mirror systems mentioned in §4.10. Burch found that, if the numerical aperture is to exceed 0.5 , at least one of the mirrors must be aspherical. Numerical apertures as high as those of the best conventional oil-immersion objectives cannot however be realized in a reflecting system. A typical Burch achromatic

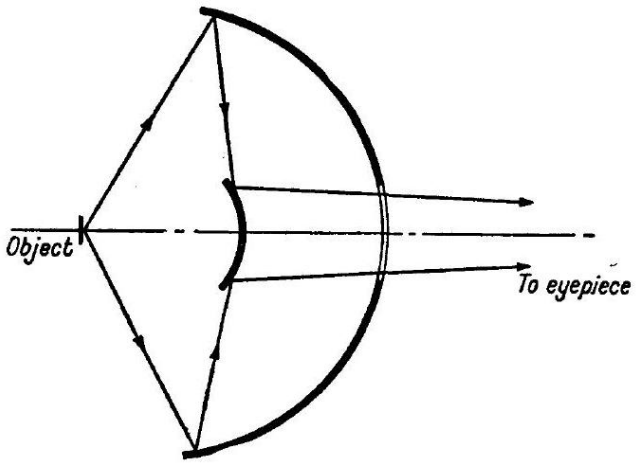



Fig. 6.30. The Burch reflecting microscope objective.

objective is shown diagrammatically in Fig. 6.30. Apart from achromatism, its chief advantage lies in its great working distance (distance between the object and the nearest surface of the objective), which may be as great as the focal length of the objective. A disadvantage is the effect of the obstruction due to the secondary mirror; in systems of lower numerical aperture which employ spherical surfaces, the obstruction may exceed 45 per cent.

Microscope objectives which combine reflection and refraction have also been

considered. Thus Bouwers* has constructed a system in which the reflections take place at concentric spherical silvered surfaces within glass, the silvering on the second surface being so distributed that the effective obstruction is reduced to 20 per cent. A two-mirror objective with low obstruction ratio was also described by Grey. † It uses quartz and fluorite components of considerable power which, whilst maintaining adequate colour correction throughout the

0

- C. R. Burch, Proc. Phys. Soc., 59 (1947), 41; ibid., 59 (1947), 47. The first paper contains also a review of earlier researches on achromatic microscope objectives.

ultra-violet and visible wavelengths, effectively replace the figuring which would otherwise be necessary on the mirrors.

The objects examined under a microscope are usually non-luminous and have therefore to be illuminated. In most cases the object has the form of a thin slice of transparent material. It is then illuminated from behind, or transilluminated. In systems of low numerical aperture (N.A. up to 0.25 ), diffuse sky-light reflected at oblique incidence from a concave mirror will suffice, but otherwise illumination from an artificial source is necessary. To obtain sufficient concentration of light, an auxiliary lens system, a condenser, is used. There are various methods of illumination, two of which will be described in Chapter X.

The present section has included only the basic principles of a microscope. An adequate discussion of the image formation in a microscope needs more refined methods (see § 8.6.3, §9.5 and § 10.5).

## 9.8 Elements of the Theory of Interference and Interferometers

### 9.8.1 7.1 Introduction

In Chapter III a geometrical model of the propagation of light was derived from the basic equations of Electromagnetic Theory, and it was shown that, with certain approximations, variations of intensity in a beam of light can be described in terms of changes in the cross-sectional area of a tube of rays. When two or more light beams are superposed, the distribution of intensity can no longer in general be described in such a simple manner. Thus if light from a source is divided by suitable apparatus into two beams which are then superposed, the intensity in the region of superposition is found to vary from point to point between maxima which exceed the sum of the intensities in the beams, and minima which may be zero. This phenomenon is called interference. We shall see shortly that the superposition of beams of strictly monochromatic light always gives rise to interference. However, light produced by a real physical source is never strictly monochromatic but, as we learn from atomistic theory, the amplitude and phase undergo irregular fluctuations much too rapid for the eye or an ordinary physical detector to follow. If the two beams originate in the same source, the fluctuations in the two beams are in general correlated, and the beams are said to be completely or partially coherent depending on whether the correlation is complete or partial. In beams from different sources, the fluctuations are completely uncorrelated, and the beams are said to be mutually incoherent. When such beams from different sources are superposed, no interference is observed under ordinary experimental conditions, the total intensity being everywhere the sum of the intensities of the individual beams. We shall see later (Chapter X) that the "degree of correlation" that exists between the fluctuations in two light beams determines, and conversely is revealed by, the "distinctness" of the interference effects to which the beams give rise on superposition.

There are two general methods of obtaining beams from a single beam of light, and these provide a basis for classifying the arrangements used to produce interference. In one the beam is divided by passage through apertures placed side by side. This method, which is called division of wave-front, is useful only with sufficiently small sources. Alternatively the beam is divided at one or more partially reflecting surfaces, at each of which part of the light is reflected and part transmitted. This method is called division of amplitude; it can be used with extended sources, and so the effects may be of greater intensity than with division of wave-front. In either case,

0

- A. Bouwers, Achievements in Optics (Amsterdam, Elsevier, 1946).
  † D. S. Grey, Proceedings of the London Conference on Optical Instruments (London, Chapman & Hall, 1951), 65.

it is convenient to consider separately the effects which result from the superposition of two beams (two-beam interference), and those which result from the superposition of more than two beams (multiple-beam interference).

Historically, interference phenomena have been the means of establishing the wave nature of light (see Historical Introduction) and to-day they have important practical uses, for example in spectroscopy and metrology. In this chapter we shall be concerned mainly with the idealized case of interference between beams from
perfectly monochromatic sources. The elementary monochromatic theory is adequate to describe the action of the apparatus used in the great majority of interference experiments. Where necessary, we shall, with the help of Fourier's theorem, take explicit account of the fact that real sources are not monochromatic; but this fact is always implied by our treatment of extended primary sources, which we consider to be made up of a large number of mutually incoherent point sources. Throughout the chapter we assume, whenever possible, that the individual beams obey the laws of geometrical optics, neglecting diffraction effects which, as briefly explained in § 3.1.4, arise in the neighbourhood of focal points and shadow boundaries. These effects will be considered in detail in Chapter VIII for the case of monochromatic light. The general case of interference and diffraction with polychromatic partially coherent light will be considered in Chapter X.

## 9.8.2 7.2 Interference of Two Monochromatic Waves

The intensity $I$ of light has been defined as the time average of the amount of energy which crosses in unit time a unit area perpendicular to the direction of the energy flow. For a plane wave, according to § 1.4 (8), (9),

$$I = v\langle w\rangle = \frac{c}{4\pi}\sqrt{\frac{\varepsilon}{\mu}}\,\langle E^2\rangle = \frac{c}{4\pi}\sqrt{\frac{\mu}{\varepsilon}}\,\langle H^2\rangle \tag{1}$$

and we have seen in Chapter III that these relations hold, at least as an approximation, for waves of more general type. Since we shall be comparing intensities in the same medium we may take the quantity $\langle \boldsymbol{E}^2\rangle$ as measure of intensity. We shall be concerned mainly with monochromatic fields, and represent the electric vector $\boldsymbol{E}$ in the form

$$\boldsymbol{E}(\boldsymbol{r},t) = \mathscr{R}\left\{A(\boldsymbol{r})e^{-i\omega t}\right\} = \frac{1}{2}\left[A(\boldsymbol{r})e^{-i\omega t} + \boldsymbol{A}^{\star}(\boldsymbol{r})e^{i\omega t}\right] \tag{2}$$

Here $A$ is a complex vector with Cartesian rectangular components

$$A_x = a_1(r)e^{ig_1(r)}, \quad A_y = a_2(r)e^{ig_2(r)}, \quad A_z = a_3(r)e^{ig_3(r)} \tag{3}$$

where $a_j$ and $g_j (j = 1, 2, 3)$ are real functions. For a homogeneous plane wave the amplitudes $a_j$ are constant, whilst the phase functions $g_j$ are of the form $g_j(r) = k.r - \delta_j$, where $k$ is the propagation vector and the $\delta_j$ 's are phase constants which specify the state of polarization.

From (2),

$$E^2 = \frac{1}{4}\left(A^2e^{-2i\omega t} + A^{\star 2}e^{2i\omega t} + 2A\cdot A^{\star}\right) \tag{4}$$

whence, taking the time average over an interval large compared with the period $T = 2\pi/\omega$,

$$\langle \boldsymbol{E}^2\rangle = \frac{1}{2}\boldsymbol{A}\cdot\boldsymbol{A}^{\star} = \frac{1}{2}\left(|A_x|^2 + |A_y|^2 + |A_z|^2\right) = \frac{1}{2}\left({a_1}^2 + {a_2}^2 + {a_3}^2\right). \tag{5}$$

Suppose now that two monochromatic waves $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are superposed at some point $P$. The total electric field at $P$ is

$$E = E_1 + E_2 \tag{6}$$

so that

$$E^2 = E_1^2 + E_2^2 + 2E_1.E_2. \tag{7}$$

Hence the total intensity at $P$ is

$$I = I_1 + I_2 + J_{12}, \tag{8}$$

where

$$I_1 = \left\langle E_1^2 \right\rangle, \quad I_2 = \left\langle E_2^2 \right\rangle \tag{9a}$$

are the intensities of the two waves, and

$$J_{12} = 2\left\langle E_1 \cdot E_2 \right\rangle \tag{9b}$$

is the interference term. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be the complex amplitudes of the two waves, where

$$A_x = a_1 e^{ig_1}, \ldots, \quad B_x = b_1 e^{ih_1}, \ldots \tag{10}$$

The (real) phases $g_j$ and $h_j$ of the two waves will in general be different, since the waves will have travelled to $P$ by different paths; but if the experimental conditions are such that the same phase difference $\delta$ is introduced between the corresponding components, we have

$$g_1 - h_1 = g_2 - h_2 = g_3 - h_3 = \delta = \frac{2\pi}{\lambda_0}\Delta\mathscr{S}, \tag{11}$$

where $\Delta\mathscr{S}$ is the difference between the optical paths for the two waves from their common source to $P$, and $\lambda_0$ is the wavelength in vacuum. In terms of $A$ and $B$,

$$\begin{aligned} \boldsymbol{E}_1 \cdot \boldsymbol{E}_2 &= \frac{1}{4}\left(Ae^{-i\omega t} + A^\star e^{i\omega t}\right) \cdot \left(Be^{-i\omega t} + B^\star e^{i\omega t}\right) \\ &= \frac{1}{4}\left(A \cdot Be^{-2i\omega t} + A^\star \cdot B^\star e^{2i\omega t} + A \cdot B^\star + A^\star \cdot B\right), \end{aligned} \tag{12}$$

so that

$$\begin{aligned} \boldsymbol{J}_{12} = 2\left\langle \boldsymbol{E}_1 \cdot \boldsymbol{E}_2 \right\rangle &= \frac{1}{2}\left(\boldsymbol{A} \cdot \boldsymbol{B}^\star + \boldsymbol{A}^\star \cdot \boldsymbol{B}\right) \\ &= a_1 b_1 \cos\left(g_1 - h_1\right) + a_2 b_2 \cos\left(g_2 - h_2\right) + a_3 b_3 \cos\left(g_3 - h_3\right) \\ &= \left(a_1 b_1 + a_2 b_2 + a_3 b_3\right)\cos\delta. \end{aligned} \tag{13}$$

This expression shows the dependence of the interference term on the amplitude components and on the phase difference of the two waves.

In the derivation of (13) we have made no use of electromagnetic theory, and in particular, no use of the fact that the vibrations are transverse. Now in the Historical Introduction it was mentioned that Fresnel and Arago found that two light beams polarized at right angles to each other do not interfere, from which they concluded that light vibrations must be transverse. This conclusion is easily verified from (13). Suppose the two waves are propagated in the $z$-direction, and let the electric vector of the first wave be in the $xz$-plane and that of the second wave in the $yz$-plane. Then

$$a_2 = 0, \quad b_1 = 0,$$

and from (13) the interference term is

$$J_{12} = a_3 b_3 \cos\delta.$$

Since the observations of Fresnel and Arago showed that no interference takes place under these circumstances, we must conclude that $a_3 = b_3 = 0$, i.e. that the electric vectors of the two waves must be perpendicular to the $z$-direction. Hence light waves must be transverse, in agreement with our earlier deduction from electromagnetic theory.

Let us consider the distribution of intensity resulting from the superposition of two waves which are propagated in the $z$-direction, and are linearly polarized with their $\boldsymbol{E}$ vectors in the $x$-direction. Then

$$a_2 = a_3 = b_2 = b_3 = 0$$

so that using (5), (9a) and (13),
and

$$\left.\begin{aligned} I_1 = \tfrac{1}{2}a_1^2, \quad I_2 = \tfrac{1}{2}b_1^2 \\ J_{12} = a_1 b_1 \cos\delta = 2\sqrt{I_1 I_2}\cos\delta. \end{aligned}\right\} \tag{14}$$

The total intensity is then given by (8) as

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta. \tag{15}$$

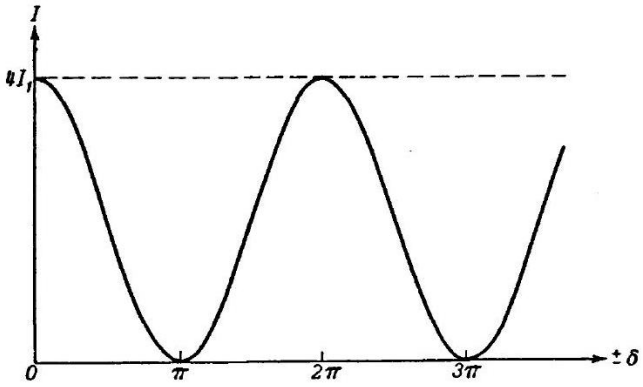



Fig. 7.1. Interference of two beams of equal intensity; variation of intensity with phase difference.

Evidently there will be maxima of intensity
when

$$\left.\begin{aligned} I_{\max} = I_1 + I_2 + 2\sqrt{I_1 I_2} \\ |\delta| = 0, 2\pi, 4\pi, \ldots, \end{aligned}\right\} \tag{16a}$$

and minima of intensity
when

$$\left.\begin{aligned} I_{\min} = I_1 + I_2 - 2\sqrt{I_1 I_2} \\ |\delta| = \pi, 3\pi, \ldots \end{aligned}\right\} \tag{16b}$$

In the special case when $I_1 = I_2$, (15) reduces to

$$I = 2I_1(1 + \cos\delta) = 4I_1 \cos^2 \frac{\delta}{2} \tag{17}$$

and the intensity varies between a maximum value $I_{\max} = 4I_1$, and a minimum value $I_{\min} = 0$ (Fig. 7.1).

The same formulae hold also for natural unpolarized light since, as will be seen later (§ 10.8.2), a beam of natural light may be represented as a superposition of two incoherent beams linearly polarized at right angles to each other (say in the $x$ and $y$ directions). The interference between the $x$-components and $y$-components may then be considered separately, and the total intensity is obtained by addition of the separate intensities. Since $\delta$ has the same value in each case, the above formulae are again obtained.

## 9.8.3 7.3. Two-Beam Interference: Division of Wave-Front

### 7.3.1 Young's experiment

The earliest experimental arrangement for demonstrating the interference of light is due to Young. Light from a monochromatic point source $S$ falls on two

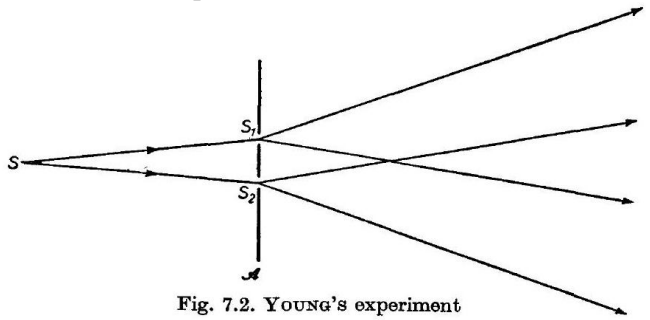



Fig. 7.2. Young's experiment

pinholes $S_1$ and $S_2$ which are close together in a screen $\mathscr{A}$ and equidistant from $S$ (Fig. 7.2). The pinholes act as secondary monochromatic point sources* which are in phase, and the beams from them are superposed in the region beyond $\mathscr{A}$. In this region an interference pattern is formed.

Suppose the pattern is observed over a plane $xOy$, normal to a perpendicular

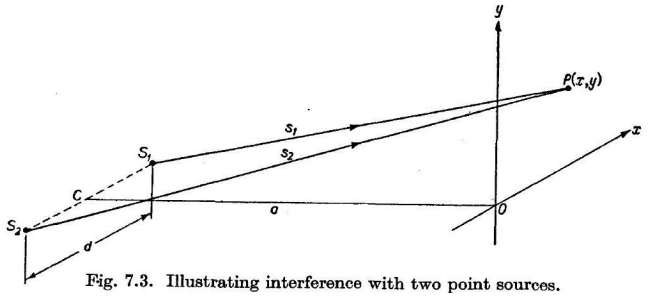



Fig. 7.3. Illustrating interference with two point sources.

bisector $CO$ of $S_1S_2$, and with the $x$-axis parallel to $S_1S_2$ (Fig. 7.3). Let $d$ be the separation of the pinholes, and $a$ the distance between the line joining the pinholes and the plane of observation. For a point $P(x, y)$ in the plane of observation,

$$s_1 = S_1P = \sqrt{a^2 + y^2 + \left(x - \frac{d}{2}\right)^2} \tag{1a}$$

$$s_2 = S_2P = \sqrt{a^2 + y^2 + \left(x + \frac{d}{2}\right)^2} \tag{1b}$$

so that

$$s_2{}^2 - s_1{}^2 = 2xd \tag{2}$$

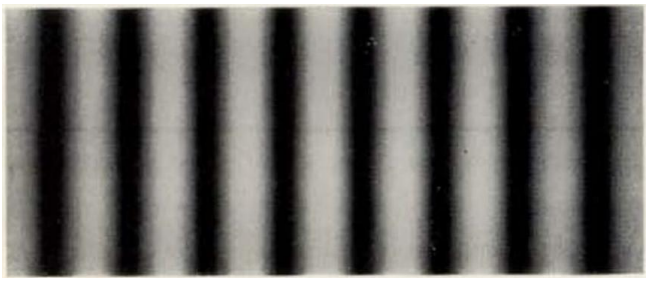


Fig. 7.4. Young's fringes.

The difference of geometrical path for light reaching $P$ from $S_2$ and $S_1$ may therefore be expressed in the form

$$\Delta s = s_2 - s_1 = \frac{2xd}{s_2 + s_1} \tag{3}$$

In practice, because of the short wavelength of visible light, the pattern can be observed conveniently only if $d$ is much smaller than $a$. Then provided $x$ and $y$ are also small compared with $a$,

$$s_2 + s_1 \sim 2a \tag{4}$$

so that apart from terms of the second and higher orders in $d/a, x/a, y/a$,

$$\Delta s = \frac{xd}{a} \tag{5}$$

If $n$ is the refractive index of the medium (assumed homogeneous) in which the experiment is made, the difference of optical path from $S_2$ and $S_1$ to $P$ is therefore

$$\Delta \mathscr{S} = n\Delta s = \frac{nxd}{a} \tag{6}$$

and the corresponding phase difference is

$$\delta = \frac{2\pi}{\lambda_0}\frac{nxd}{a} \tag{7}$$

Since the angle $S_1PS_2$ is very small, we can consider the waves from $S_1$ and $S_2$ to be propagated in the same direction at $P$, so that the intensity can be calculated from §7.2(15); according to (7) and §7.2 (16) there are maxima of intensity when

$$x = \frac{ma\lambda_0}{nd}, \quad |m| = 0, 1, 2, \ldots \tag{8a}$$

and minima of intensity when

0

- These secondary sources have directional properties which will be discussed in detail in the theory of diffraction (Chapter VIII).

$$x = \frac{ma\lambda_0}{nd}, \quad |m| = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{8b}$$

The interference pattern in the immediate vicinity of $O$ thus consists of bright and dark bands called interference fringes (Fig. 7.4), equidistant and running at right angles to the line $S_1S_2$ joining the two sources. The separation of adjacent bright fringes is $a\lambda_0/nd$. At any point of the interference pattern the number $m$ defined by

$$m = \frac{\delta}{2\pi} = \frac{\Delta\mathscr{S}}{\lambda_0} \tag{9}$$

is called the order of interference at the point: thus the bright fringes correspond to integral orders.

### 7.3.2 Fresnel's mirrors and similar arrangements

Young's experiment is of historical importance as a crucial step in establishing the wave theory of light. It also provides a method, though one of low precision, of
measuring the wavelength of monochromatic light with extremely simple apparatus; it is necessary to measure only $d, a$, and the fringe spacing, which in air ( $n \sim 1$ ) is $a\lambda_0/d$. However, the light from the primary source $S$ does not reach the region of superposition by uninterrupted rectilinear paths of the kind to which geometrical optics applies, and in order to show that this circumstance is not essential to the

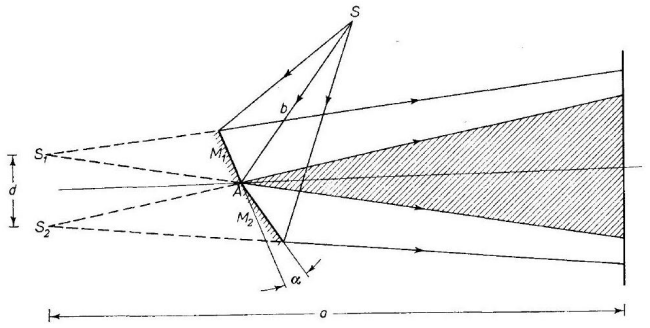



Fig. 7.5. Fresnel's mirrors.

production of interference effects a number of alternative ways of producing two coherent sources were subsequently devised.

One example is the arrangement known as Fresnel's mirrors (Fig. 7.5). Light from a point source $S$ is incident on two plane mirrors $M_1$ and $M_2$, mutually inclined at a small angle, and reflection at the mirrors gives rise to two virtual images $S_1, S_2$ of $S$

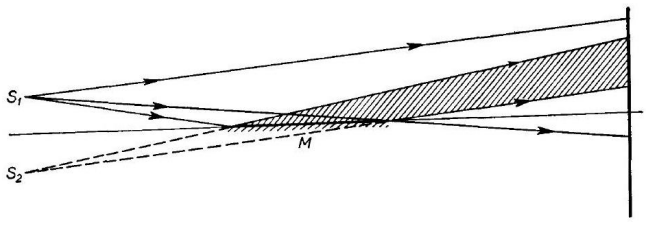



Fig. 7.6. Lloyd's mirror.

which act as coherent sources. The plane $SS_1S_2$ is evidently normal to the line of intersection of the mirrors, which it meets in $A$. If $SA = b$, then

$$S_1A = S_2A = b,$$

so that the perpendicular bisector of $S_1S_2$ also passes through $A$. The separation of $S_1$ and $S_2$ is

$$d = 2b \sin \alpha, \tag{10}$$

where $\alpha$ is the angle between the mirrors.
An even simpler arrangement is Lloyd's mirror (Fig. 7.6). A point source $S_1$ is placed some distance away from a plane mirror $M$ and close to the plane of the mirror
surface, so that light is reflected at nearly grazing incidence. The coherent sources are the primary source $S_1$ and its virtual image $S_2$ in the mirror. The perpendicular bisector of $S_1S_2$ then lies in the plane of the mirror surface.

We may mention two other similar devices. Fresnel's bi-prism (Fig. 7.7) is formed by two equal prisms of small refracting angle placed together base to base with their refracting edges parallel. A pencil of light from a point source $S$ is divided by refraction into two overlapping pencils. These refracted pencils are not strictly stigmatic,

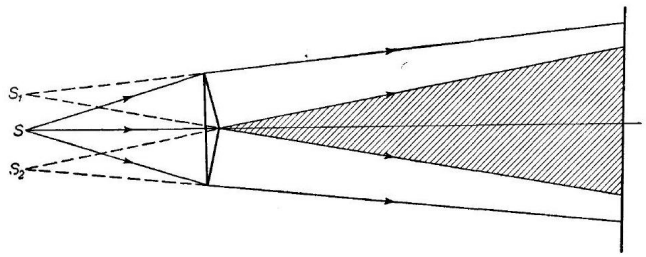



Fig. 7.7. Fresnel's bi-prism.

but because of the smallness of the refracting angle and of the angular aperture of the pencils used we may neglect this aberration and suppose the prisms to form two virtual images $S_1, S_2$ of $S$. Billet's split lens (Fig. 7.8) consists of a convex lens cut diametrically into two, the halves being separated by a small distance at right angles to the optic axis. Two real images $S_1, S_2$ are then produced from a single point source $S$.

With all these arrangements which use a primary point source, interference fringes are visible in monochromatic light over any plane in the region common to the diverging pencils from the sources $S_1$ and $S_2$ (shown hatched in Figs. 7.5 to 7.8).

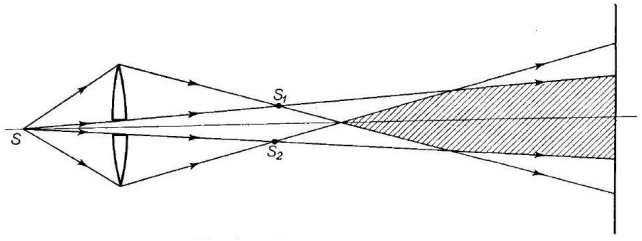



Fig. 7.8. Billet's split lens.

Such fringes are said to be non-localized. Denoting as before the distances $S_1P$ and $S_2P$ by $s_1$ and $s_2$, the locus of points $P$ for which the phase difference between waves from $S_2$ and $S_1$ is constant is the surface defined by

$$s_2 - s_1 = \text{ constant } . \tag{11}$$

Hence the maxima and minima of the resultant intensity form a family of hyperboloids of revolution about $S_1S_2$ as axis, and having $S_1$ and $S_2$ as common foci. The fringes in a plane normal to the perpendicular bisector $CO$ of $S_1S_2$ are sections of these hyperboloids, and are themselves hyperbolae; but near $O$ they approximate, as we
have seen in §7.3.1, to equidistant straight lines running at right angles to $S_1S_2$. In a plane normal to $S_1S_2$ the sections of the hyperboloids are concentric circles, but such fringes cannot be observed with the arrangements described above. They can however be seen in Meslin's experiment, in which the halves of Billet's split lens are separated along the optic axis instead of transversely (Fig. 7.9). The sources $S_1$ and $S_2$ are then at different points of the optic axis, and the corresponding pencils overlap in a region between them. Here fringes can be observed which in planes normal to $S_1S_2$ are

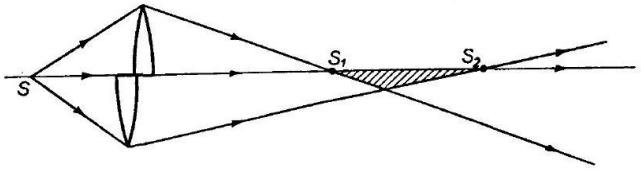



Fig. 7.9. Meslin's experiment.

concentric circles centred on the optic axis; the portions seen are semi-circular since the region common to the pencils is limited by a plane passing through the optic axis.

### 7.3.3 Fringes with quasi-monochromatic and white light

So far we have assumed that the primary point source is monochromatic. We now remove this restriction and suppose that a point source $S$ of polychromatic light is used, for example with Fresnel's mirrors (Fig. 7.5). As we shall see later (§7.5.8), such light can be represented by a combination of mutually incoherent monochromatic components extending over a range of frequencies. Each component produces an interference pattern as described above, and the total intensity is everywhere the sum of the intensities in these monochromatic patterns. Suppose that the components cover a wavelength range $\Delta\lambda_0$ around a mean wavelength $\bar{\lambda}_0$. The central maxima of all the monochromatic patterns, corresponding to equality of the paths from $S_1$ and $S_2$, coincide in passing through $O$, but elsewhere the patterns are mutually displaced because their scale is proportional to wavelength; the maxima of order $m$ are spread over a distance $\Delta x$ in the plane of observation where, by (8),

$$\Delta x = \frac{|m|a}{nd}\Delta\lambda_0. \tag{12}$$

Consider first the case when the wavelength range $\Delta\lambda_0$ is small compared to the mean wavelength $\bar{\lambda}_0$, i.e.

$$\frac{\Delta\lambda_0}{\bar{\lambda}_0} \ll 1. \tag{13}$$

We shall call light which satisfies this condition quasi-monochromatic light.*
If over the field of observation

$$|m| \ll \frac{\bar{\lambda}_0}{\Delta\lambda_0}, \tag{14}$$

or by (9), if

$$|\Delta\mathscr{S}| \ll \frac{\bar{\lambda}_0^2}{\Delta\lambda_0} \tag{15}$$

we can neglect $\Delta x$ compared with the mean separation $a\lambda_0/nd$ of adjacent maxima, and take the component patterns to be in coincidence. There are then fringes in the plane of observations which have the same appearance as those given by a strictly monochromatic source of wavelength $\bar{\lambda}_0$.

If the light is quasi-monochromatic but (15) is not satisfied, the fringes are less distinct than with monochromatic light, the total intensity depending on the distribution of intensity amongst the monochromatic components.

If the light is not quasi-monochromatic, i.e. if (13) is not satisfied, what is observed depends also on the spectral response of the radiation detector used. A case of practical importance is when the light is white and the observation is visual, so that the effective wavelength range extends from about 4000Å to about 7000Å and $\Delta\lambda_0/\bar{\lambda}_0$ is of the order of $\frac{1}{2}$. There is then a central white fringe in the position of the monochromatic fringe of order zero, with a few coloured maxima and minima on either side, and farther away what appears to the eye to be uniform white illumination. This light at some distance from the centre of the pattern is not normal white light. Thus at distance $x$ from the central fringe there are according to (8) maxima of intensity for

$$\lambda_0 = \frac{nd}{a}\frac{x}{m}, \quad |m| = 1, 2, 3, \ldots \tag{16a}$$

and minima of intensity for

$$\lambda_0 = \frac{nd}{a}\frac{x}{m}, \quad |m| = \frac{1}{2}, \frac{3}{2}, \frac{5}{2} \ldots \tag{16b}$$

Hence if the light enters a spectroscope with its slit parallel to the direction of monochromatic fringes, so that $x$ is constant for the light admitted, the spectrum is crossed by light and dark bands parallel to the slit-an example of a so-called channelled spectrum. The bands are equally spaced in spectroscopic wave-number $1/\lambda_0$, the separation of adjacent bright bands being $a/nd|x|$.

As will appear later, the white light fringe pattern is useful in interferometry because in particular cases it enables the monochromatic fringe corresponding to zero path difference to be recognized. We have referred to the central fringe obtained with Fresnel's mirrors as an intensity maximum, and this is the case also with Fresnel's bi-prism, Billet's split lens, and in Young's experiment. With Lloyd's mirror, however, the fringe which lies in the plane of the mirror surface is an intensity minimum, and with a white light source it appears black. This is because the phase of the wave reflected from the mirror is changed by $\pi$ on reflection (see § 1.5.2), irrespective of its wavelength. In Mesun's experiment also the centre of the pattern is

---

0

- Since $\lambda_0 = 2\pi c/\omega_0$, (13) may also be expressed in the form

$$\frac{\Delta\omega_0}{\bar{\omega}_0} \ll 1,$$

where $\Delta\omega_0$ is the effective frequency range and $\bar{\omega}_0$ the mean frequency.

an intensity minimum, in this case because, as we shall see later (§ 8.8.4), the wave suffers a phase change of $\pi$ on its passage through focus.

### 7.3.4 Use of slit sources; visibility of fringes

The foregoing discussion has been in terms of a primary point source, but since practical sources have finite extension we must examine how such extension affects the fringes. The description of a true physical source involves atomistic theory, which is outside the scope of this book, but for our purposes it is adequate to idealize the situation by regarding the source to be made up of a large number of point
sources which are mutually incoherent. The intensity at any point in the wave-field is then the sum of the intensities from the individual point sources.

With the arrangements we have described (Meslin's experiment excepted) the fringes are perpendicular to the plane which contains the primary point source $S$ and the derived sources $S_1$ and $S_2$, and it follows that if $S$ is displaced perpendicular to this plane the fringes are merely displaced parallel to their length. Hence a line source-or in practice a sufficiently narrow slit source-in this direction may be used without impairing the distinctness of the fringes, at least so far as their curvature is negligible. Similarly the pinholes of Young's experiment may be replaced by narrow

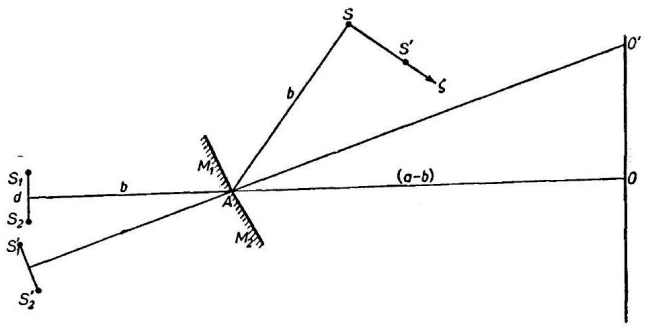



Fig. 7.10. Illustrating Fresnel's mirrors with slit source.

slits parallel to the slit source. In this way the intensity of the patterns can be increased, though there is an additional experimental difficulty in securing correct orientation of the source.

To obtain even more light in the patterns it is necessary to increase the width of the source slit, but then the fringes become less distinct. Consider for example the case of Fresnel's mirrors. If the source $S$ is displaced to $S'$ at right angles to $SA$ and in the plane $SS_1S_2$ (Fig. 7.10), the secondary sources $S_1, S_2$ are displaced to $S_1', S_2'$, and apart from terms of second order their separation $d$-and therefore the fringe spacing -is unchanged; but the central fringe is displaced from $O$ to $O'$, since it lies on the perpendicular bisector of $S_1'S_2'$ which passes through $A$. If $SS' = \zeta$ (taken positive in Fig. 7.10) then $S_1S_1' = S_2S_2' = \zeta$, so that

$$OO' = \frac{(a-b)\zeta}{b}. \tag{17}$$

Hence if at $P$ the optical path difference for light from the source $S$ is $\Delta\mathscr{S}$, the optical path difference for light from the source $S'$ is, by (6),

$$\Delta\mathscr{S}' = \Delta\mathscr{S} - \frac{nd}{a}OO' = \Delta\mathscr{S} - D\zeta, \tag{18}$$

where

$$D = \frac{(a-b)nd}{ab} \tag{19}$$

The corresponding phase difference is

$$\delta(\zeta, \Delta\mathscr{S}) = \frac{2\pi}{\lambda_0}(\Delta\mathscr{S} - D\zeta). \tag{20}$$

Suppose now that the source is a slit of width $e$ centred on $S$. We assume that the number of point sources which constitute the slit source is so large that we may treat it as effectively continuous, and imagine it divided into elementary strips perpendicular to the plane $SS_1S_2$. If $i_1 d\zeta$ is the intensity at $P$ from an elementary strip when the light reaches $P$ from one mirror only, the intensity at $P$ from the elementary strip through $S'$ is, by § 7.2 (17),

$$i(\zeta, \Delta\mathscr{S})d\zeta = 2i_1(1 + \cos\delta)d\zeta$$

and the total intensity at $P$ is given by

$$I(e, \Delta\mathscr{S}) = 2i_1 \int_{-e/2}^{e/2} (1 + \cos\delta)d\zeta \tag{21}$$

On substituting from (20) into (21) and evaluating the integral we obtain

$$I(e, \Delta\mathscr{S}) = 2I_1 \left\{ 1 + \left( \frac{\sin\frac{\pi De}{\lambda_0}}{\frac{\pi De}{\lambda_0}} \right) \cos\left( \frac{2\pi}{\lambda_0} \Delta\mathscr{S} \right) \right\} \tag{22}$$

where $I_1 = i_1 e$.

Following Michelson we take as measure of the distinctness of the fringes at $P$ their visibility $\mathscr{V}$ defined by

$$\mathscr{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}}, \tag{23}$$

where $I_{\max}$ and $I_{\min}$ are the maximum and minimum intensities in the immediate neighbourhood of $P$. Evidently $\mathscr{V}$ has a maximum value of unity when $I_{\min} = 0$, as is the case for fringes from two equal monochromatic point sources, and declines to zero when $I_{\max} = I_{\min}$ and the fringes disappear. In the present case

$$I_{\max} = 2I_1 \left( 1 + \frac{\left|\sin\frac{\pi De}{\lambda_0}\right|}{\frac{\pi De}{\lambda_0}} \right), \quad I_{\min} = 2I_1 \left( 1 - \frac{\left|\sin\frac{\pi De}{\lambda_0}\right|}{\frac{\pi De}{\lambda_0}} \right), \tag{24}$$

so that

$$\mathscr{V} = \frac{\left|\sin\frac{\pi De}{\lambda_0}\right|}{\frac{\pi De}{\lambda_0}} \tag{25}$$

The visibility $\mathscr{V}$ given by (25) is shown as a function of the source width $e$ in Fig. 7.11. We see that the visibility is greater than about 0.9 when $e$ does not exceed $\lambda_0/4D$, or from (20), when the range of phase difference at $P$ corresponding to the elements of the source does not exceed $\pi/2$. If we somewhat arbitrarily take this condition to define the maximum slit width which gives rise to satisfactory fringes, we have from (19) $e \leqslant \lambda_0 ab/4(a-b)nd$, or, using (10),

$$e \leqslant \frac{\lambda_0 a}{8(a-b)n\sin\alpha} \tag{26}$$

where, as before, $\alpha$ is the angle between the mirrors. Evidently the tolerable value of $e$ increases as $(\ a - b\ )$ decreases, i.e. as the plane of observation approaches the mirror junction. For example, with typical values $\alpha = 2$ minutes of arc, $a = 120$ cm, $b = 100$ cm, $\lambda_0 = 5500\text{Å}, n = 1$, we have $e \leqslant 0.7$ mm; the spacing of the bright fringes is according to (8) and (10) given by the formula $a\lambda_0/2bn \sin\alpha$; in the present case it is equal to 0.57 mm .

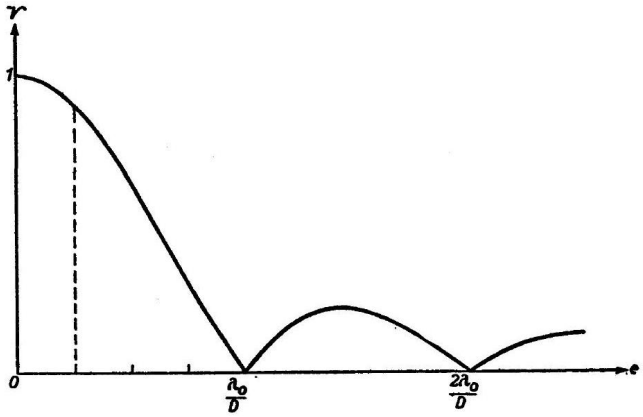



Fig. 7.11. Variation of fringe visibility with width of source slit. (Fresnel's mirrors.)

Similar considerations apply to the source width with Fresnel's bi-prism, Bulet's split lens, and in Young's experiment. With Lloyd's mirror the situation is different since displacement of the source $S$ normal to the mirror plane results in displacement of its image in the opposite direction. With a slit source of finite width the elementary patterns therefore have minima which coincide in the plane of the mirror but are of different spacing, so that the visibility of the fringes decreases with increasing distance from the mirror plane.

### 7.3.5 Application to the measurement of optical path difference: the Rayleigh interferometer

We will anticipate a result from the theory of diffraction by saying that the light from the secondary sources $S_1$ and $S_2$ in Young's experiment has greatest intensity

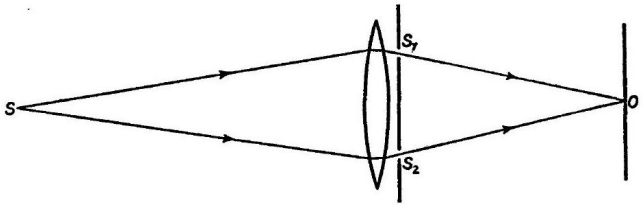



Fig. 7.12. Illustrating the use of a lens with Young's arrangement.

in the direction of the geometrical rays from the primary source $S$. In Young's experiment these rays diverge beyond $S_1$ and $S_2$, but if a lens is introduced in front of the apertures (Fig. 7.12) they may be brought together at $O$, conjugate to $S$ with respect to the lens. In this way the intensity in the interference pattern near $O$ is increased, and the fringes can be observed with the apertures $S_1$ and $S_2$ farther apart.

The separation of adjacent bright fringes is still $a\lambda_0/nd$, and by the principle of equal optical path, if the lens is stigmatic for $S$ the fringe of order zero is at $O$. If the lens is not stigmatic for $S$ the fringe of order zero will be displaced from $O$ by an amount which de pends on the

difference of optical path from $S$ to $O$ through the two apertures; for a difference of optical path $\Delta\mathscr{S}$, the displacement will be $\Delta m$ times the separation of adjacent bright fringes where

$$\Delta m = \frac{\Delta\mathscr{S}}{\lambda_0}. \tag{27}$$

Evidently the arrangement may be used as a quantitative test of the performance of the lens, as was done by Michelson.* If one aperture is fixed over the centre of the

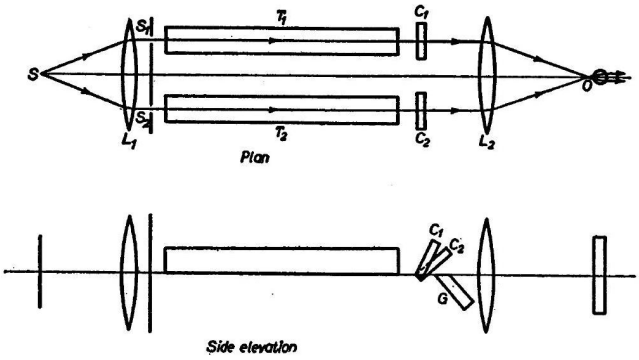



Fig. 7.13. The Rayleigh interferometer.

lens, measurements of $\Delta m$ for different radial positions of the other aperture give the departure from sphericity of the wave-front from $S$ after passage through the lens (the wave aberration).

Similarly, if a plate of transparent material of thickness $l$ and refractive index $n'$ is introduced into the path of the light from $S_2$, the optical path $[SS_2O]$ is increased by $(n'-n)\,l$ and there will be a change $\Delta m$ in the order of interference at $O$ given by

$$\Delta m = \frac{(n'-n)\,l}{\lambda_0}. \tag{28}$$

From measurement of $\Delta m, l$ and $\lambda_0$, the difference ( $n'-n$ ) between the refractive index of the plate and that of the surrounding medium can be determined, and used in this way the arrangement forms the basis of the Rayleigh interferometer, † which is used for the precision determination of the refractive indices of gases. A modern form of this instrument is shown in plan and side elevation in Fig. 7.13. Light from a slit source $S$ is collimated by the lens $L_1$ before falling on slit apertures $S_1, S_2$ which are parallel to $S$; the geometrical rays from $S_1$ and $S_2$ are then parallel in passing through separate gas chambers $T_1$ and $T_2$ before being re-combined by the lens $L_2$ in its focal

plane, where interference fringes are formed, parallel to the slits.* The presence of the gas chambers requires $S_1$ and $S_2$ to be widely separated, so that the fringes are closely spaced and a high magnification is needed to observe them; further, the restriction on the width of the source slit $S$ is correspondingly severe, so that the amount of light in the pattern is small. Since magnification is needed only in the direction at right angles to the fringes it is best obtained by a cylindrical eyepiece in the form of a thin glass rod parallel to the fringes. The pattern viewed in this way appears much brighter than if a spherical eyepiece were used, but the use

0

- A. A. Michelson, Astrophys. J., 47 (1918), 283.
† Lord Rayletgh, Proc. Roy. Soc., 59 (1896), 198. F. Haber and F. Löwe, Zeits f. angew Chem., 23 (1910), 1393.

of a cylindrical eyepiece has a further important advantage. It allows a second fixed system of fringes, with the same spacing as the main system but formed by light from $S_1$ and $S_2$ which passes beneath the gas chambers, to be used as a fiducial mark. By means of the glass plate $G$ this fiducial system is displaced vertically so that its upper edge meets the lower edge of the main system in a sharp dividing line which is the edge of $G$ as viewed through $L_2$. In this way the detection of displacement of the main fringe system due to changes in the optical path in $T_1$ and $T_2$ is made to depend on the vernier acuity of the eye, which is high, and displacements as small as $\frac{1}{40}$ order can be detected. Accidental displacements of the optical system also become less important in so far as both fringe systems are equally affected.

In the technical use of the instrument it is convenient to compensate the optical path difference, rather than to count fringes. To achieve this the light emerging from the gas chambers is passed through thin glass plates, of which one, $C_1$, is fixed while the other, $C_2$, can be rotated about a horizontal axis to give reproducible changes in the optical path of the light from $S_2$. The compensator is calibrated with monochromatic light to give the mean rotation corresponding to one order displacement of the main system. The fringe systems are then used as a null indicator of equality of the optical paths $[SS_1O]$ and $[SS_2O]$. In normal use the gas chambers are evacuated, and with white light the central fringes of the main and fiducial systems are brought into approximate coincidence by means of the compensator. An exact setting for coincidence of the zero orders is then made with monochromatic light. Next the gas under investigation is admitted to one chamber and again, first with white light and then with monochromatic light, the compensator is adjusted to bring the zero orders into coincidence. † The difference between the two settings of the compensator in terms of

$$m = [(n' - 1)\, l + \Delta\mathscr{S}]\, \frac{1}{\lambda_0}$$

so that

$$\frac{\partial m}{\partial \lambda_0} = -\left(n' - 1 - \lambda_0 \frac{\partial n'}{\partial \lambda_0}\right) \frac{l}{\lambda_0^2} - \left(\Delta\mathscr{S} - \lambda_0 \frac{\partial}{\partial \lambda_0}(\Delta\mathscr{S})\right) \frac{1}{\lambda_0^2}$$

Hence the achromatic fringe is at $O$ when

$$\left(\Delta\mathscr{S} - \lambda_0 \frac{\partial}{\partial \lambda_0}(\Delta\mathscr{S})\right)_{\lambda_0} = -l \left(n' - 1 - \lambda_0 \frac{\partial n'}{\partial \lambda_0}\right)_{\bar{\lambda}_0}$$

its calibration gives the order displacement $\Delta m$ of the main system due to the introduction of the gas, and its refractive index $n'$ is obtained from (28):

$$(n' - 1) = \frac{\lambda_0}{l} \Delta m, \tag{29}$$

where $l$ is the length of the gas chamber. With typical values $l = 100$ cm, $\lambda_0 = 5500$Å and settings correct to $\frac{1}{40}$ order, the detectable change of $(n' - 1)$ is about $10^{-8}$.*

0

- The exact distribution of intensity in the focal plane of $L_2$ can be found by the theory of Fraunhofer diffraction (§ 8.5).
  † The optical paths from $S_1$ and $S_2$ to the main pattern involve media of differing dispersion, so that, unlike in the simple case considered in § 7.3.3, the zero orders for light of different wavelengths do not in general coincide, and with white light there is no perfectly white fringe. The fringe showing least colour is that for which $\partial m/\partial \lambda_0 = 0$ at some mean wavelength $\lambda = \bar{\lambda}_0$ in the visible spectrum which depends on the wavelength response of the eye, and this is called the achromatic fringe by analogy with the term achromatic as applied to a lens. If the compensator introduces an optical path difference $\Delta\mathscr{S}$, the order of interference at $O$ is

Higher sensitivity can in principle be obtained by increasing $l$, but this is limited in practice by difficulties of temperature control. For the same reason, versions of the instrument intended for the measurement of differences of refractive index of liquids employ only short chambers. Further, the path difference that can be compensated is limited, so that if the difference of refractive index in the two chambers is large the length of the chambers must be proportionately reduced.

### 7.3.6 Application to the measurement of angular dimensions of sources: the Michelson stellar interferometer

We have already seen (§ 7.3.4) that in Young's experiment the distinctness of the fringes is affected by extension of the source in the direction joining the apertures $S_1, S_2$. This effect is the basis of methods of measuring the angular dimensions of small sources.

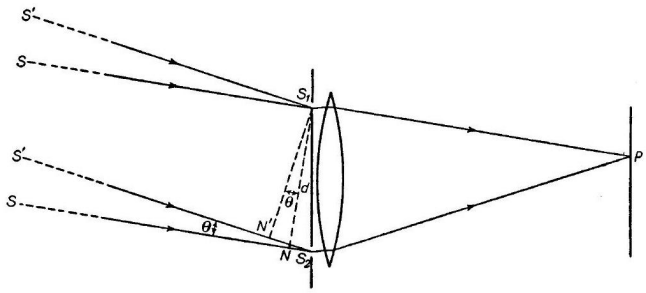



Fig. 7.14. Illustrating a telescope objective diaphragmed by two apertures and illuminated by two distant point sources.

Suppose a telescope objective, diaphragmed by two equal small apertures $S_1, S_2$ of separation $d$, is used to view two distant quasi-monochromatic point sources $S$ and $S'$, of effective wavelength $\lambda_0$, separated by angle $\theta$ in the direction joining the apertures (Fig. 7.14). $S$ and $S'$ each give an interference pattern in the focal plane with the same fringe spacing, and if $S$ and $S'$ are incoherent sources, the combined pattern
and with this setting of the compensator the zero order of the monochromatic pattern is not in general at $O$ since this requires

$$(\Delta\mathscr{S})_{\bar{\lambda}_0} = -l\,(n'-1)_{\bar{\lambda}_0}\,.$$

The effect may be large enough to make identification of the monochromatic fringe of zero order uncertain, in which case a preliminary measurement must be made at low pressure or with short chambers.

We note further that the achromatic fringe is recognizable only if, at the point of the pattern where $(\partial m/\partial\lambda_0)_{\overline{\lambda_0}} = 0$, the range of $m$ for wavelengths in the visible spectrum is sufficiently small. For this reason, in interferometers which are required to show fringes in white light, the interfering waves are arranged to have as far as possible equal paths in media of identical dispersion.

- The value of $(n'-1)$ for common gases is of the order of $10^{-4}$ (cf. Table I, p. 13).
  is formed by summing at each point the intensities in these two patterns. If $N$ is the foot of the perpendicular from $S_1$ to $SS_2, S_1N$ lies in a plane wave-front from $S$ so that the optical path difference for light from $S$ at a point $P$ in the focal plane is

$$\begin{aligned}\Delta\mathscr{S} &= [SS_2P] - [SS_1P] \\ &= [NS_2] + [S_2P] - [S_1P]\,. \end{aligned} \tag{30}$$

Similarly if $N'$ is the foot of the perpendicular from $S_1$ to $S'S_2$, the optical path difference at $P$ for light from $S'$ is

$$\Delta\mathscr{S}' = [N'S_2] + [S_2P] - [S_1P] \tag{31}$$

Hence the patterns from $S'$ and $S$ are mutually displaced through $\Delta m$ orders, where

$$\Delta m = \frac{|\Delta\mathscr{S}' - \Delta\mathscr{S}|}{\lambda_0} = \frac{|[N'S_2] - [NS_2]|}{\lambda_0} \sim \frac{\theta d}{\lambda_0} \tag{32}$$

for $n \sim 1$ and $\theta$ small. When

$$\Delta m = 0, 1, 2, \ldots, \tag{33a}$$

the intensity maxima from $S$ and $S'$ coincide and the fringes in the combined pattern are most distinct. On the other hand when

$$\Delta m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{33b}$$

the intensity maxima from $S$ coincide with intensity minima from $S'$; the fringes in the combined pattern are then least distinct, and disappear if the intensities from $S$ and $S'$ are equal. Thus the fringes show periodic variations of distinctness as the separation of the apertures is increased from zero. In particular there is a first minimum of distinctness when

$$d = \frac{\lambda_0}{2\theta} \tag{34}$$

and if this condition is observed, $\theta$ may be determined from $d$ and $\lambda_0$.

Let us consider now the more general case of a quasi-monochromatic primary source extended about an origin $S$. As before we may take such a source to be made up of mutually incoherent point sources. We may imagine it divided into elementary strips at right angles to the direction joining the apertures $S_1$ and $S_2$, and find the intensity at a point $P$ in the focal plane as the sum of the intensity contributions from these elements. Suppose that $\Delta\mathscr{S}$ is the optical path difference at $P$ for light from the element through the origin $S$, so that for given $d$ the position of $P$ is defined by $\Delta\mathscr{S}$. The optical path difference for light from the element at angle $\alpha$ from $S$ is, by (32), $\Delta\mathscr{S} + \alpha d$, and the corresponding phase difference is

$$\delta(\alpha, \Delta\mathscr{S}) = \frac{2\pi}{\lambda_0}(\Delta\mathscr{S} + \alpha d) \tag{35}$$

We again assume that the number of the point sources that make up the extended source is so large that we may treat the extended source as effectively continuous. The total intensity at $P$ is then by §7.2 (15)

$$I(d, \Delta\mathscr{S}) = \int i_1 d\alpha + \int i_2 d\alpha + 2\int \sqrt{i_1 i_2}\cos\delta d\alpha \tag{36}$$

where $i_1(\alpha, \Delta\mathscr{S})d\alpha$ and $i_2(\alpha, \Delta\mathscr{S})d\alpha$ are the intensities at $P$ due to an elementary strip when the light reaches $P$ through one of the apertures only, and the integration is taken over the range of values of $\alpha$ subtended by all the source elements.

We have already mentioned in §7.3.5 that, because of diffraction, the secondary sources formed by the apertures $S_1$ and $S_2$ have directional properties, but if we can neglect these* over the range of values of $\alpha$ involved in (36), we may write

$$i_1(\alpha, \Delta\mathscr{S}) = i_2(\alpha, \Delta\mathscr{S}) = i(\alpha) f(\Delta\mathscr{S}) \tag{37}$$

where $i(\alpha)$ is proportional to the intensity of the corresponding strip of the source, and $f(\Delta\mathscr{S})$ characterizes the directional properties of $S_1$ (or $S_2$ ). From (36), using (37) and (35) we then have

$$\begin{aligned} I(d, \Delta\mathscr{S}) &= 2f(\Delta\mathscr{S}) \int i(\alpha)\{1 + \cos\delta\} d\alpha \\ &= f(\Delta\mathscr{S}) \left\{ P + C(d) \cos\left(\frac{2\pi}{\lambda_0}\Delta\mathscr{S}\right) - S(d) \sin\left(\frac{2\pi}{\lambda_0}\Delta\mathscr{S}\right) \right\} \end{aligned} \tag{38}$$

where

$$\left.\begin{aligned} P &= 2\int i(\alpha) d\alpha \\ C(d) &= 2\int i(\alpha) \cos\left(\tfrac{2\pi}{\lambda_0}\alpha d\right) d\alpha \\ S(d) &= 2\int i(\alpha) \sin\left(\tfrac{2\pi}{\lambda_0}\alpha d\right) d\alpha \end{aligned}\right\} \tag{39}$$

If we further assume that the variation of $f(\Delta\mathscr{S})$ is slow compared with that of $\cos(2\pi\Delta\mathscr{S}/\lambda_0)$ and $\sin(2\pi\Delta\mathscr{S}/\lambda_0)$*, the positions of the maxima and minima of $I$ are given by

$$\frac{\partial I}{\partial(\Delta\mathscr{S})} = 0 = -\frac{2\pi}{\lambda_0} \left\{ C \sin\left(\frac{2\pi}{\lambda_0}\Delta\mathscr{S}\right) + S \cos\left(\frac{2\pi}{\lambda_0}\Delta\mathscr{S}\right) \right\}$$

i.e. by

$$\tan\left(\frac{2\pi}{\lambda_0}\Delta\mathscr{S}\right) = -\frac{S}{C} \tag{40}$$

From (38) and (40) the extreme values of $I$ are therefore

$$I_{\text{ext}} = f(\Delta\mathscr{S}) \left[ P \pm \sqrt{C^2 + S^2} \right] \tag{41}$$

and hence the visibility of the fringes, defined by (23), is

$$\mathscr{V}(d) = \frac{\sqrt{C^2 + S^2}}{P} \tag{42}$$

The functions $\mathscr{V}(d)$ for a number of different functions $i(\alpha)$ are shown in Fig. 7.15. Fig. 7.15(a) corresponds to the case of two point sources, which we treated simply at the beginning of this section. In Fig. 7.15(b) the source is of uniform intensity and rectangular, with sides parallel to the line joining the apertures $S_1$ and $S_2$; evidently $\mathscr{V}(d)$ in this case is the same as the visibility function of Fig. 7.11. In Fig. 7.15(c) the source is a circular disc with radial symmetry; the left-hand curve corresponds to uniform intensity, and the others to different degrees of darkening from centre to edge.

0

* This approximation is justified if the apertures are sufficiently narrow, as will be clear from § 8.5.

Conversely, if the positions of the fringes and their visibility are measured as functions of the aperture separation $d$, the functions $C$ and $S$ can be determined from (40) and (42), apart from the constant factor of proportionality $P$ and the sign; the latter may usually be fixed from considerations of physical plausibility. The intensity distribution $i(\alpha)$ of the source can then be obtained from (39) by the Fourier

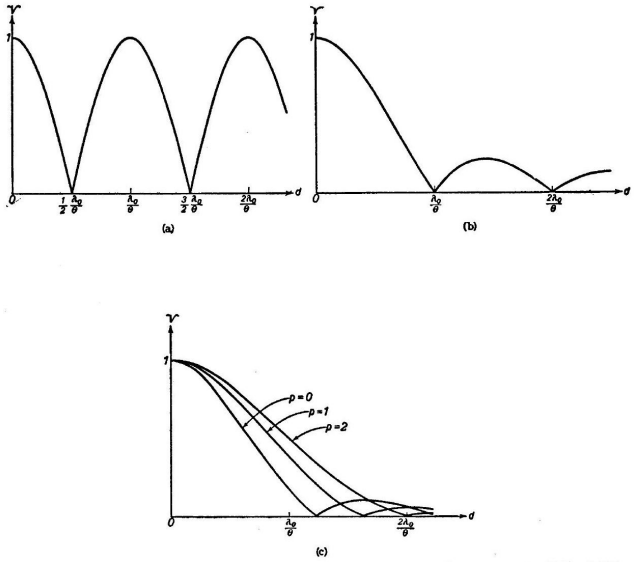



Fig. 7.15. Variation of fringe visibility with aperture separation (arrangement of Fig. 7.14).

(a) Two equal point sources with angular separation $\theta$ in the direction of the line joining the apertures.
(b) Uniform rectangular source with sides parallel to the line joining the apertures and of angular width $\theta$.
(c) Circular disc source of angular diameter $\theta = 2\beta_0$, with intensity distribution $I(\beta) \propto (\beta_0{}^2 - \beta^2)^p$, where $\beta$ is the angular radius from the centre. [After A. A. Michelson and F. G. Pease, Astrophys. J., 53 (1921), 256.]
inversion theorem. Measurements of this kind, though in principle possible, are difficult. However, if the source is known in advance to have one of the forms to which Fig. 7.15 refers, its angular dimensions can be determined simply by observing the smallest value of $d$ for which the visibility of the fringes is minimum. This condition occurs when

$$d = \frac{A\lambda_0}{0}, \tag{43}$$

where $A = 0.5$ for two point sources of angular separation $\theta$, as we have seen in (34); $A = 1.22$ for a uniform circular disc source of angular diameter $\theta$; and $A > 1.22$ for a circular disc source darker at the edges than at the centre.

This method was suggested first by Fizeaj* and later by Michelson † as a means of determining the angular dimensions of astronomical objects which are too small or distant to be measured with an undiaphragmed telescope (cf. § 8.6.2). Such objects emit white light, and because of intensity considerations the observations have to be made with white light fringes. ‡ It is therefore necessary to assume for $\lambda_0$ in (43) an effective wavelength which depends on the wavelength distribution of intensity of the light and the colour response of the eye. With this limitation the method was successfully used to measure the angular diameters of planetary satellites § and the angular

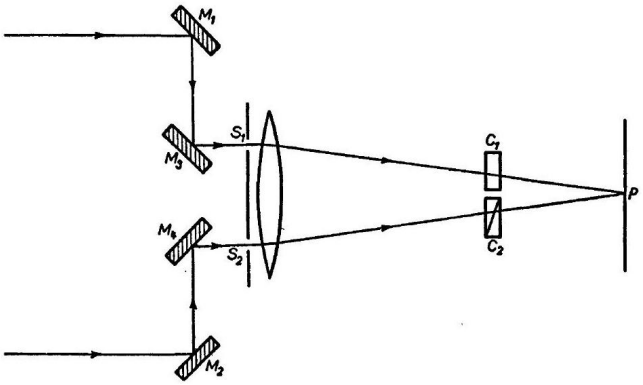



Fig. 7.16. Michelson's stellar interferometer.

separation of double stars whose diameters are small compared with their distance apart||, but attempts to apply it to single stars failed because of their small angular diameters; the fringes remained distinct with the largest aperture separation permitted by available telescopes. To overcome this restriction Michelson 9 constructed his stellar interferometer (Fig. 7.16). The apertures $S_1, S_2$ diaphragming the telescope are fixed, and light reaches them after reflection at a symmetrical system of mirrors $M_1, M_2, M_3$, and $M_4$, mounted on a rigid girder in front of the telescope. The inner mirrors $M_3$ and $M_4$ are fixed, but the outer mirrors $M_1$ and $M_2$ can be separated symmetrically in the direction joining $S_1$ and $S_2$. If the optical paths [ $M_1M_3S_1$ ] and [ $M_2M_4S_2$ ] are maintained equal, the optical path difference for light from a distant point source is the same at $S_1$ and $S_2$ as at $M_1$ and $M_2$, so that the outer mirrors play the part of the movable apertures in the Fizeaje method. The smallest angular diameter that can be measured with the arrangement is thus determined not by the

diameter of the telescope objective, but by the maximum separation of the outer mirrors. There is the further advantage that the fringe spacing, which depends on the separation of $S_1$ and $S_2$, remains constant as the separation of the movable apertures is varied. The interferometer was mounted on the large reflecting telescope (diameter 100 in .), of the Mount Wilson Observatory, which was used simply because of its mechanical strength. The apertures $S_1$ and $S_2$ were 114 cm apart, giving a fringe spacing of about 0.02 mm in the focal plane.

The maximum separation of the outer mirrors was 6.1 m , so that the smallest measurable angular diameter (with $\lambda_0 = 5500$Å ) was about 0.02 seconds of arc. Because of inevitable mechanical imperfections, two auxiliary devices were necessary to ensure correct adjustment in all positions of the outer mirrors. A plane parallel glass plate $C_1$, which could be inclined in any direction, was used to maintain the geometrical pencils from $S_1$ and $S_2$ in coincidence in the focal plane. A further plane parallel glass plate $C_2$, of variable thickness, was used to compensate inequalities of the optical paths [ $M_1M_3S_1$ ] and [ $M_2M_4S_2$ ]. This compensation, essential since the fringes in white light are visible only near order zero, was controlled by observing the channelled spectrum with a small spectroscope.

0

- H. Fizeaj, C. R. Acad. Sci. Paris, 66 (1868), 934.
  † A. A. Michelson, Phil. Mag. (5), 30 (1890), 1.
  ‡ The general problem of fringes produced by extended sources of finite wavelength range is elegantly treated by the theory of partial coherence (Chapter X). In § 10.4, p. 512, the action of Michelson's stellar interferometer is briefly discussed from the standpoint of this theory.
  § A. A. Michelson, Nature, Lond., 45 (1891), 160.
  || J. A. Anderson, Astrophys. J., 51 (1920), 263.
  If A. A. Michelson, Astrophys. J., 51 (1920), 257; A. A. Michelson and F. G. Pease, Astrophys. J., 53 (1921), 249.

The first star whose angular diameter was successfully measured was Betelgeuse ( $\alpha$ Orionis). The value found was 0.047 seconds of arc. With the distance of this star from the sun, determined trigonometrically, its linear diameter is then found as $4.1 \times 10^8$ km, which is about 300 times the diameter of the sun ( $1.4 \times 10^6$ km ) and exceeds the diameter of the earth's orbit ( $3 \times 10^8$ km ). Only a few other stars have been measured, all of them, like Betelgeuse, giant stars with linear diameters many times greater than the sun. In part the smallness of this number is due to the inherent difficulties of the measurements, which are hampered by the disturbing effects of atmospheric turbulence, though Michelson and Pease discovered these effects to be of less consequence than in normal observations with telescopes of large aperture. Variations of refractive index above the small apertures of the interferometer cause the interference pattern to move as a whole, and providing this motion is slow the fringes remain observable, whereas under the same conditions the star image formed by the full aperture telescope would be much impaired. But apart from these observational difficulties, a maximum separation of the outer mirrors of 6 m is insufficient to permit measurements of the great majority of stars, which have diameters not very different from that of the sun. At the distance of the nearest star the sun's disc would subtend an angle of only 0.007 seconds of arc, and to observe the first disappearance of the fringes a mirror separation of about 20 m would be necessary. The construction of such a large interferometer would be a difficult undertaking because of the requirement of rigid mechanical connection between the collecting mirrors and the eyepiece.

An instrument analogous to the Michelson stellar interferometer has been used in radio-astronomy to determine the angular size of celestial radio sources.* It consists of two separated aerials supplying signals to a common detector system. In this case also it is technically difficult to increase the separation of the aerials without introducing inconstant phase differences into the paths between aerials and detector. To overcome this difficulty Hanbury Brown and Twiss † have devised another form of radio interferometer, in which the signals at the aerials are detected separately and the angular size of the source is obtained from measurements of the correlation of the

intensity fluctuations of the signals as a function of aerial separation. They have also shown* that an equivalent arrangement may be used with visible light. The light from a star, collected by two concave mirrors, is focused on to two photoelectric cells, and the correlation of fluctuations in their photo-currents is measured as the mirror separation is varied. With this arrangement (called intensity interferometer) large mirror separations present no difficulty, so that measurements of stars of much smaller diameter than hitherto become possible. The principle of the method is best understood from the standpoint of the theory of partial coherence and is briefly discussed in § 10.4, p. 512.

## 9.8.4 7.4. Standing Waves

With the arrangements we have considered so far, the two waves which interfere are propagated in approximately the same direction at the point of observation. We now consider the interference effects of two waves propagated in different directions, taking as example the interference of the incident and reflected waves when a plane monochromatic wave falls on a highly reflecting plane surface.

Suppose the surface is the plane $z = 0$, with the positive direction of $z$ pointing into the

0

- See, for example, J. L. Pawsey and R. N. Bracewell, Radio Astronomy (Oxford, Clarendon Press, 1955), Chapter II; or R. N. Bracewell, "Radio Astronomy Techniques" in Encyclopedia of Physics, ed. S. Flgge (Berlin, Springer), 54 (1959), Chapter V.
  † R. Hanbury Brown and R. Q. Twiss, Phil. Mag. (7), 45 (1954), 663.

medium from which the wave is incident, and let $n_1$ and $n_2$ be the refractive indices of the two media (Fig. 7.17).

Let $\theta_i$ be the angle of incidence, and let the $xz$-plane be the plane of incidence. If $E^{(i)}$ is the electric vector of the incident wave, and if as in § 1.5.2 we denote by $A_{11}$ and $A_\perp$ the amplitudes

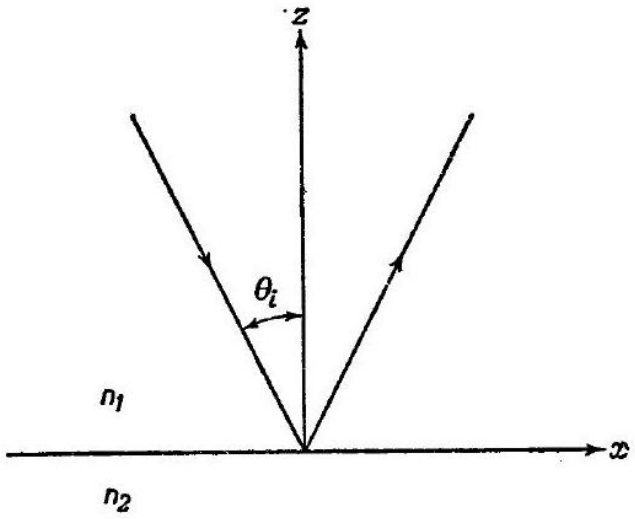



Fig. 7.17. Reflection of a plane wave at a plane surface.

$$E_x^{(i)} = -A_1 \cos\theta_i e^{-i\tau_i}, \quad E_y{}^{(i)} = A_\perp e^{-i\tau_i}, \quad E_z{}^{(i)} = -A_{11} \sin\theta_i e^{-i\tau_i}, \tag{1}$$

(as usual the real parts being understood), where

$$\tau_i = \omega\left(t - \frac{x\sin\theta_i - z\cos\theta_i}{v_1}\right), \tag{2}$$

and $v_1$ is the velocity of propagation in the first medium. The Cartesian components of the electric vector $E^{(r)}$ of the reflected wave are given by similar expressions (eqn. (16), § 1.5), which, if the law of reflection § 1.5 (7) is used, are

$$E_x^{(r)} = R_{11} \cos\theta_i e^{-i\tau_r}, \quad E_y^{(r)} = R_\perp e^{-i\tau_r}, \quad E_z^{(r)} = -R_{11} \sin\theta_i e^{-i\tau_r}, \tag{3}$$

where

$$\tau_r = \omega\left(t - \frac{x\sin\theta_i + z\cos\theta_i}{v_1}\right). \tag{4}$$

The amplitudes $R_\parallel, R_\perp$ of the reflected wave are related to the amplitudes $A_\parallel, A_\perp$ of the incident wave by the Fresnel formulae (21) of § 1.5, viz.

$$R_{\text{ll}} = \frac{n_2\cos\theta_i - n_1\cos\theta_t}{n_2\cos\theta_i + n_1\cos\theta_t} A_{\text{t}}, \quad R_\perp = \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t} A_\perp. \tag{5}$$

To simplify the discussion we suppose $n_2/n_1$ to be so large that the reflectivity may be taken as unity; (5) then gives, in the limit as $n_2/n_1 \to \infty$,*

$$R_\parallel = A_\parallel, \quad R_\perp = -A_\perp. \tag{6}$$

0

• R. Hanbury Brown and R. Q. Twlss, Nature, 178 (1956), 1046; Proc. Roy. Soc. A248 (1958), 199, 222.

By substituting from (6) into (3), and adding the expressions for the incident and reflected fields, we obtain the total field. Thus the $x$-component of the electric vector of the total field is

$$\begin{aligned} E_x &= E_x^{(i)} + E_x^{(r)} \\ &= 2A_{\text{il}} \cos\theta_i \left\{ \sin\left( \frac{\omega z \cos\theta_i}{v_1} \right) \right\} e^{-i\left[\omega\left(t - \frac{x\sin\theta_i}{v_1}\right) - \frac{\pi}{2}\right]} \end{aligned} \tag{7a}$$

Similarly,

$$E_y = -2A_\perp \left\{ \sin\left( \frac{\omega z \cos\theta_i}{v_1} \right) \right\} e^{-i\left[\omega\left(t - \frac{x\sin\theta_i}{v_1}\right) - \frac{\pi}{2}\right]} \tag{7b}$$

and

$$E_z = -2A_\parallel \sin\theta_i \left\{ \cos\left( \frac{\omega z \cos\theta_i}{v_1} \right) \right\} e^{-i\left[\omega\left(t - \frac{x\sin\theta_i}{v_1}\right)\right]} \tag{7c}$$

In a strictly analogous manner we obtain from eqs. (13) and (16) of § 1.5 the following expressions for the components of the magnetic vector of the total field:

$$H_x = -2A_\perp n_1 \cos\theta_i \left\{ \cos\left( \frac{\omega z \cos\theta_i}{v_1} \right) \right\} e^{-i\left[\omega\left(t - \frac{x\sin\theta_i}{v_1}\right)\right]} \tag{8a}$$

$$H_y = -2A_1 n_1 \left\{ \cos\left( \frac{\omega z \cos\theta_i}{v_1} \right) \right\} e^{-i\left[\omega\left(t - \frac{x\sin\theta_i}{v_1}\right)\right]} \tag{8b}$$

$$H_z = 2A_\perp n_1 \left\{ \sin\theta_i \sin\left( \frac{\omega z \cos\theta_i}{v_1} \right) \right\} e^{-i\left[\omega\left(t - \frac{x\sin\theta_i}{v_1}\right) - \frac{\pi}{2}\right]} \tag{8c}$$

where the Maxwell relation $n_1 = \sqrt{\varepsilon_1}$ has been used. Each of the expressions (7) and (8) represents a wave propagated in the $x$-direction with velocity $v_1/\sin\theta_i$. The amplitude of the wave is not constant, but varies periodically in the $z$-direction, with period $2\pi v_1/\omega\cos\theta_i = \lambda_0/n_1\cos\theta_i$, where $\lambda_0$ is the wavelength in vacuum.

The case of normal incidence ( $\theta_i = 0$ ) is of particular interest. If $A_x, A_y$ are the amplitude components of the electric vector, we may write, from (1), $A_\parallel = -A_x$, $A_\perp = A_y$; and (7) and (8) give

$$\left.\begin{aligned} E_x &= -2A_x \left\{ \sin\left(\tfrac{\omega z}{v_1}\right) \right\} e^{-i\left(\omega t - \frac{\pi}{2}\right)}, \\ E_y &= -2A_y \left\{ \sin\left(\tfrac{\omega z}{v_1}\right) \right\} e^{-i\left(\omega t - \frac{\pi}{2}\right)} \\ E_z &= 0, \end{aligned}\right\} \tag{9}$$

$$\left.\begin{aligned} H_x &= -2A_y n_1 \left\{ \cos\left(\tfrac{\omega z}{v_1}\right) \right\} e^{-i\omega t} \\ H_y &= 2A_x n_1 \left\{ \cos\left(\tfrac{\omega z}{v_1}\right) \right\} e^{-i\omega t} \\ H_z &= 0 \end{aligned}\right\} \tag{10}$$

We see that at each instant of time the phase is constant throughout the first medium. There is no finite velocity of propagation, and we speak of a standing wave. The amplitudes of the electric and magnetic vectors are periodic functions of $z$; the planes

0

- We exclude grazing incidence ( $\cos\theta_i \to 0$ ), for which the reflectivity approaches unity even when $n_2/n_1$ is not large; in this case (5) gives $R_1 = -A_1, R_\perp = -A_\perp$. This corresponds to the case of Lloyd's mirror (§ 7.3.2).

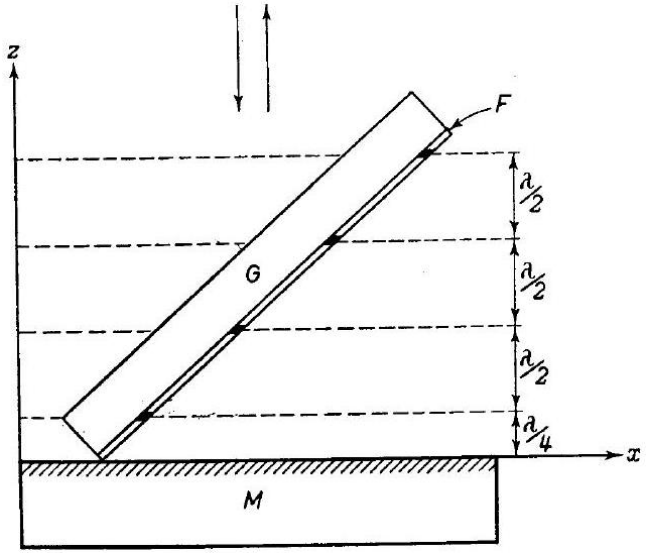



Fig. 7.18. Wiener's experiment on standing waves. (The inclination of the plate to the mirror is greatly exaggerated.)

of zero amplitude are called nodes, and the planes where the amplitude has extreme values are called anti-nodes. From (9), the nodes of the electric field are given by

$$z = \frac{m\pi v_1}{\omega} = \frac{m\lambda_0}{2n_1}, \quad m = 0, 1, 2, \ldots, \tag{11a}$$

and the anti-nodes by

$$z = \frac{m\lambda_0}{2n_1}, \quad m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{11b}$$

From (10), the nodes of the magnetic field coincide with the anti-nodes of the electric field, and vice versa. In particular, the reflecting surface is a node of the electric field and an anti-node of the magnetic field.

The existence of standing light waves was first demonstrated experimentally by Wiener.* His arrangement is illustrated in Fig. 7.18. A plane mirror $M$, silvered on the front surface, was illuminated normally by a parallel beam of quasi-monochromatic light. A film $F$ of transparent photographic emulsion, coated on the plane surface of a glass plate $G$ and less than 1/20 wavelength thick, was placed in front of $M$ and inclined to it at a small angle. On development, the emulsion was found to be blackened in equidistant parallel bands with transparent regions between. The

maxima of blackening corresponded to the intersection of $F$ with the anti-nodal planes of either the electric or the magnetic field. From further experiments, in which the emulsion coated plate was pressed in contact with a convex spherical reflecting surface, Wiener concluded that there was no blackening at the surface of the mirror, which we have seen to be an anti-node of the magnetic field.* The blackened regions therefore correspond to anti-nodes of the electric field, i.e. the photochemical action is directly related to the electric and not to the magnetic vector. This conclusion is, of course, to be expected from electron theory. The photographic process is an ionization process, in which an electron is removed from an atomic bond of silver

0

• o. Wiener, Ann. d. Physik, 40 (1890), 203.

halide, and the electromagnetic force on a charged particle at rest is proportional to the electric vector (cf. § 1.1(34)).

Similar experiments have been made using fluorescent films, † and photo-emissive films, ‡ as detectors of standing waves in place of the photographic emulsion used by

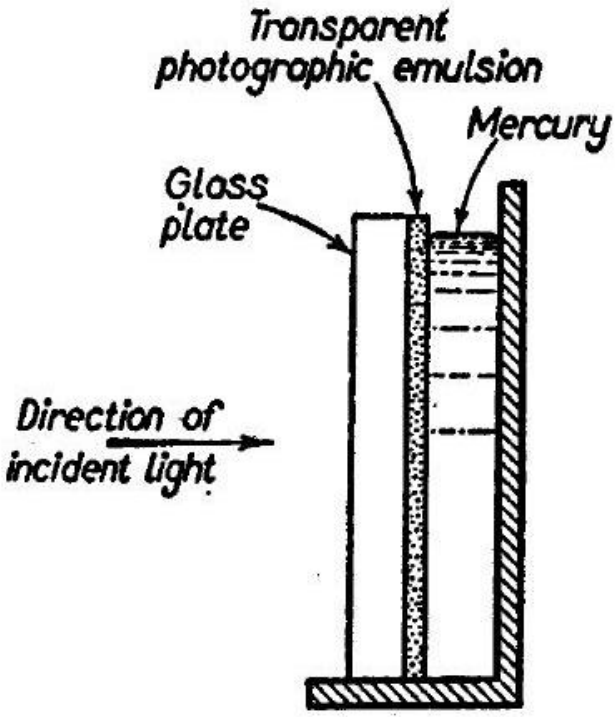



Fig. 7.19. Lippmann's arrangement for colour photography.

Wiener. In both cases, and again as is to be expected from electron theory, the maximum response was found at the anti-nodes of the electric field.

Standing light waves are the basis of a method of colour photography originated by Lippmann §. Plates coated with a transparent fine-grain photographic emulsion are exposed in the camera, with the emulsion side away from the incident light and in contact with a reflecting surface of mercury (Fig. 7.19). Suppose for simplicity that the plate is exposed to normally incident quasi-monochromatic light of wavelength $\lambda_0$. Since the photochemical action is maximum at the anti-nodes of the electric field, given by (11b), the silver in the developed plate forms a system of equidistant layers, parallel to the surface of the emulsion and with optical separation $\lambda_0/2$. If the plate is now illuminated normally with white light, these silver layers act as partially reflecting surfaces, so that the reflected light consists of a series of beams with optical path differences which are integral multiples of $\lambda_0$. We shall consider later (§7.6) the interference of such a series of beams. The analysis shows that there is a maximum of resultant intensity for wavelength $\lambda_0$, which is quite sharp if the number of beams is large. The Lippmann plate thus acts as a selective reflector for light of the wavelength used to prepare it.

Wiener|| also used his arrangement to examine the interference effects when the angle of incidence is 45° and the incident light is linearly polarized. He found that with the direction of electric vibrations in the incident light perpendicular to the plane of incidence, the emulsion was blackened in a system of equidistant parallel bands; but that with the direction of electric vibrations in the incident light in the plane of incidence, the blackening was uniform. This result again confirms that the photochemical effect is directly related to the electric and not to the magnetic field. For

0

when the direction of electric vibrations is perpendicular to the plane of incidence, $A_{\mathrm{II}} = 0$, and with $\theta_i = 45°$, (7) gives

$$\left.\begin{aligned} E_x &= 0 \\ E_y &= -2A_\perp \left\{\sin\left(\tfrac{\omega z}{\sqrt{2}v_1}\right)\right\} e^{-i\left[\omega\left(t-\frac{x}{\sqrt{2}v_1}\right)-\frac{\pi}{2}\right]} \\ E_z &= 0 \end{aligned}\right\} \tag{12}$$

so that the amplitude of the electric vector and also the time-averaged electric energy density vary periodically in the $z$-direction. When the direction of vibration is in the plane of incidence $A_\perp = 0$, and from (7),

$$\left.\begin{aligned} E_x &= \sqrt{2}A_1 \sin\left(\tfrac{\omega z}{\sqrt{2}v_1}\right) e^{-i\left[\omega\left(t-\frac{x}{\sqrt{2}v_1}\right)-\frac{\eta}{2}\right]} \\ E_y &= 0 \\ E_z &= -\sqrt{2}A_1 \cos\left(\tfrac{\omega z}{\sqrt{2}v_1}\right) e^{-i\left[\omega\left(t-\frac{x}{\sqrt{2}v_1}\right)\right]} \end{aligned}\right\} \tag{13}$$

From (13) and § 1.4 (54) it follows that the time-averaged energy density is in this case equal to

$$\langle w_e \rangle = \frac{n_1^2}{16\pi} E.E^\star = \frac{n_1^2}{16\pi}\left(E_x E_x^\star + E_z E_z^\star\right) = \frac{n_1^2}{8\pi} A_{\mathrm{u}} A_{\mathrm{i}}^\star \tag{14}$$

and so on time average the electric energy density is independent of $z$. On the other hand it follows from (8) that for the magnetic field the situation is reversed; the time averaged magnetic energy density varies periodically with $z$ when the direction of magnetic vibrations in the incident light is in the plane of incidence, but is on time average independent of $z$ when the direction of magnetic vibrations in the incident light is perpendicular to the plane of incidence.

## 9.8.5 7.7 The Comparison of Wavelengths with the Standard Metre

The standard of length is the distance between two lines engraved on a bar of platinum-iridium alloy at a temperature of 0°C. This bar, which is kept in France, is called the International Prototype Metre, and, as we have already mentioned in §7.6.4, the relation of optical wavelengths to actual lengths rests on a comparison of the wavelength of the red line ( 6438Å ) of cadmium with this standard. The comparison was first made in 1892 by Michelson and Benoit ‡ using a form of the Michelson interferometer, but in 1905 the measurement was repeated with increased accuracy by Benoit, Fabry and Perot.§ They used five Fabry-Perot etalons of

lengths approximately $6 \cdot 25, 12 \cdot 5, 25, 50$ and 100 cm . The number of wavelengths of the cadmium red line in the length of the shortest etalon was determined by the method of excess fractions, which requires a knowledge of sufficiently accurate wavelength ratios (§7.6.4) but not actual wavelengths. The length of each etalon was then compared with that of the next longer by means of white light fringes. For this, the two etalons were set parallel and illuminated by white light, and the transmitted light was passed through a thin wedge of air formed between two halfsilvered plane surfaces. Under these circumstances, fringes similar to those described in § 7.6.8(b) are localized in the wedge, the central fringe being along the line where the thickness

0

- C. Fabry and H. Buisson, Journ. de Phys. (5), 9 (1919), 189.
  † H. Barrell and J. E. Sears, Phil. Trans. Roy. Soc., A., 238 (1939), 1.
  ‡ A. A. Michelson and J. R. Benoit, Trav. et Mem. Bur. Int. Poids et Mes., 11 (1895), 1.
  § J. R. Benoit, C. Fabry and A. Perot, Trav. et Mem. Bur. Int. Poids et Mes., 15 (1913), 1.

of the wedge is equal to the difference between the length of the longer etalon and twice the length of the shorter etalon. This difference was obtained in wavelengths of the red cadmium line from a previous calibration of the wedge thickness, and so after four inter-comparisons the number of wavelengths in the length of the 100 cm etalon was determined. Finally, the difference between the length of this etalon and a copy of the prototype metre was measured. This part of the experiment involved making settings with high-quality travelling microscopes on the terminal graduation lines on the metre and on similar graduation lines engraved on the edges of the end plates of the etalon, the distance between these latter lines and the reflecting surfaces of the etalon being determined by a subsidiary experiment. The final result obtained for the wavelength of the red cadmium line in dry air at 15°C and 760 mm Hg pressure was $6438.4696 \times 10^{-10}$ m, with a probable error for the interferometric measurements of about 1 part in $10^7$.

Further re-determinations* of the relationship between the wavelength of the red cadmium line and the metre were made in a number of standardizing laboratories with methods the same as, or similar in principle to, that of Benoit, Fabry and Perot. The experiment of Sears and Barrell † is of interest because they were able to make direct measurements of the wavelength in vacuum. They used only three Fabry-Perot etalons, the longest being slightly over a metre in length, and the others approximately one-third and one-ninth of this. The spacers were invar cylinders with optically flat chromium-plated ends to which the etalon plates were wrung, and the joint so formed was airtight so that the etalons could be evacuated. The number of wavelengths in the length of the shortest etalon was determined by the method of excess fractions, and the inter-comparison of the etalons was made with the fringes of superposition described in §7.6.8. The longest etalon was large enough to accommodate a steel end gauge, nominally a metre long, and the distances in wavelengths between the polished ends of the gauge and the reflecting surfaces of the etalon were determined from observations of fringes at infinity with reflected light. In this way the length of the end gauge was obtained in terms of the wavelength of the red cadmium line with an accuracy of about 2 parts in $10^8$. Finally, the end gauge was compared with a copy of the metre by well-established procedure for comparison of an end standard with a line standard. ‡

When the results of all determinations are reduced to the wavelength in standard air (air at 15°C and 760 mm Hg pressure, containing no water vapour and 0.03 per cent of carbon dioxide), the mean value is found to be $6438.4696 \times 10^{-10}$ m, which by chance is the value used for the definition of the ngström. The greatest departure of any one determination from the mean is 1 part in about $3 \times 10^6$, or about

$3 \times 10^{-4}$ mm in a metre. This is much less than the width of the graduation lines on the Prototype Metre and on the copies of it which have been used in the experiments, and there is no doubt that it represents the uncertainty in the comparisons made with travelling microscopes. The interferometric measurements can be made with much higher precision, so that it is natural to consider the replacement of the material standard by a definition of the metre in terms of the wavelength of some spectral line.

The International Committee of Weights and Measures accepted a proposal to this effect in 1954,* with the proviso that, in order to preserve as far as possible continuity in the standard of length, the re-definition should be consistent with the value $6438.4696 \times 10^{-10}$ m for the wavelength of the red cadmium line. In 1958 the Committee decided † that the best available choice of spectral line is the line of approximate wavelength 6056Å corresponding to the tran-

0

• For references see H. Barrell, Proc. Roy. Soc., A, 186 (1946), 164.
† J. E. Sears and H. Barrell, Phil. Trans. Roy. Soc., A, 231 (1932), 75; 233 (1934), 143.
‡ For details see J. E. Sears, Proc. Roy. Soc., A, 186 (1946), 152.

sition between levels $2p_{10}$ and $5d_5$ of the krypton atom of mass number 86 ; and on the basis of measurements in five different laboratories it proposed that the metre should be defined as exactly $1,650,763 \cdot 73$ wavelengths in vacuum of this radiation. This definition was approved unanimously in 1960 by the 11th General Conference of Weights and Measures. ‡ The primary standards of wavelength and the primary standard of length are therefore now identical, and the Angström is exactly $10^{-10}$ m.

## 9.9 7.5. Two-Beam Interference: Division of Amplitude

### 9.9.1 7.5.1 Fringes with a plane parallel plate

Suppose a plane parallel plate of transparent material is illuminated by a point source $S$ of quasi-monochromatic light (Fig. 7.20). Whatever its position, a point $P$ on the same side of the plate as $S$ is reached by two rays-one reflected at the upper surface and the other at the lower surface of the plate-so that there is a non-localized interference pattern on the same side of the plate as $S$. From considerations of symmetry, the fringes in planes parallel to the plate are circular about the normal $SN$ to the plate as axis, so that at any position of $P$ they run perpendicular to the plane $SNP$. We may expect from the discussion of §7.3.4 that the visibility of these fringes will be reduced if the source is extended parallel to the plane $SNP$, but to this there is an important exception when $P$ is at infinity, i.e. when the fringes are observed with a relaxed eye, or in the focal plane of a telescope objective. Under these conditions the two rays from $S$ to $P$, namely $SADP$ and $SABCEP$ (Fig. 7.21), derive from the same incident ray and are parallel after leaving the plate. The difference of the optical paths along them is

$$\Delta\mathscr{S} = n'(AB + BC) - nAN \tag{1}$$

where $n', n$ are the refractive indices of the plate and the surrounding medium, and $N$ is the foot of the perpendicular from $C$ to $AD$. If $h$ is the thickness of the plate, and $\theta, \theta'$ are the angles of incidence and refraction at the upper surface, we have

$$AB = BC = \frac{h}{\cos\theta'} \tag{2}$$

$$AN = AC\sin\theta = 2h\tan\theta'\sin\theta \tag{3}$$

$$n'\sin\theta' = n\sin\theta \tag{4}$$

0

• Proc. Verb. Com. Int. Poids et Mes., 24 (1954) 2.
† Proc. Verb. Com. Int. Poids et Mes. (2), 26-B (1958) M.30.
‡ C.R. 11$^{\text{me}}$ Conf. Gn. Poids et Mes. (Paris, Gauthier-Villars, 1960) pp. 51, 85.

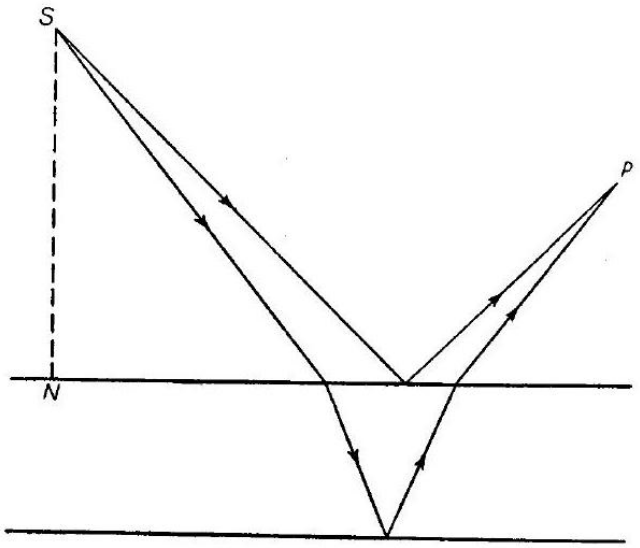



Fig. 7.20. Plane parallel plate with point source.

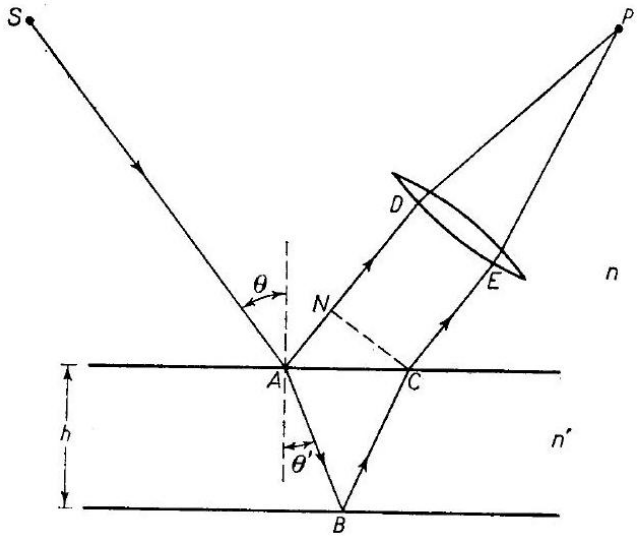



Fig. 7.21. Plane parallel plate: illustrating formation of fringes localized at infinity.

Hence from (1), (2), (3), and (4),

$$\Delta\mathscr{S} = 2n'h\cos\theta', \tag{5}$$

and the corresponding phase difference is

$$\delta = \frac{4\pi}{\lambda_0} n'h\cos\theta'. \tag{6}$$

We must also take into account the phase change of $\pi$ which, according to the Fresnel formulae of §1.5 (21a), occurs on reflection at either the upper or lower surface. The total phase difference at $P$ is therefore

$$\delta = \frac{4\pi}{\lambda_0} n' h \cos\theta' \pm \pi \tag{7a}$$

$$= \frac{4\pi h}{\lambda_0} \sqrt{n'^2 - n^2 \sin^2\theta} \pm \pi \tag{7b}$$

Now $\theta$ is determined only by the position of $P$ in the focal plane of the telescope, so that $\delta$ is independent of the position of $S$. It follows that the fringes are just as distinct with an extended source as with a point source. Since this is true only for one particular plane of observation, the fringes are said to be localized-in this case-localized at infinity.

The intensity in the pattern varies according to the relation § 7.2 (15); from (7) and § 7.2 (16) there are bright fringes when

$$2n'h\cos\theta' \pm \frac{\lambda_0}{2} = m\lambda_0, \quad m = 0, 1, 2, \ldots \tag{8a}$$

and dark fringes when

$$2n'h\cos\theta' \pm \frac{\lambda_0}{2} = m\lambda_0, \quad m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{8b}$$

A given fringe is thus characterized by a constant value of $\theta'$ (and therefore $\theta$ ), and so is formed by light incident on the plate at a particular angle. For this reason the fringes are often called fringes of equal inclination. When the axis of the telescope objective is normal to the plate, the fringes are concentric circles about the focal point for normally reflected light ($\theta = \theta' = 0$). The order of interference is highest at the centre of the pattern, where it has the value $m_0$ given by

$$2n'h \pm \frac{\lambda_0}{2} = m_0\lambda_0 \tag{9}$$

$m_0$ is not necessarily an integer, and we may write

$$m_0 = m_1 + e \tag{10}$$

where $m_1$ is the integral order of the innermost bright fringe, and $e$, which is less than unity, is called the fractional order at the centre. For the $p$ th bright fringe from the centre, of angular radius $\theta_p$, the order of interference is $m_p$, where by (8a),

$$2n'h\cos\theta'_p \pm \frac{\lambda_0}{2} = m_p\lambda_0$$

$$= [m_1 - (p-1)]\,\lambda_0 \tag{11}$$

and from (9), (10), and (11),

$$2n'h\left(1 - \cos\theta'_p\right) = (p - 1 + e)\lambda_0. \tag{12}$$

Now if $\theta'_p$ is small, we have from the law of refraction $n' \sim n\theta_p/\theta'_p$, and $1 - \cos\theta'_p \sim \theta'^2_p/2 \sim n^2\theta_p^2/2n'^2$; hence from (12),

$$\theta_p \sim \frac{1}{n}\sqrt{\frac{n'\lambda_0}{h}}\sqrt{p - 1 + e} \tag{13}$$

The angular scale of the pattern is thus proportional to $\sqrt{1/h}$, and if $e = 0$, so that there is an intensity maximum at the centre, the radii of the bright fringes are proportional to the square roots of the positive integers.

Similar fringes localized at infinity may be obtained with a parallel plate of air, bounded by the inner plane surfaces of two transparent plates (Fig. 7.22). With this arrangement the fringes can be observed as the thickness $h$ is varied continuously by separating the plates. As $h$ increases, the fringes expand from the centre of the pattern, and a new fringe appears there each time $h$ increases by $\lambda_0/2$ (taking $n' = 1$ for air). To avoid disturbing effects of the reflections at the outer surfaces, the plates are made slightly wedge-shaped. The difference of optical path from $S$ to $P$ for light reflected at the outer surfaces then varies with the position of $S$, so that with an extended source this light gives, on average, uniform intensity over the focal plane. The fringes formed by light reflected from the inner parallel surfaces are superposed on this background.

Unless the plate is very thin, the fringes correspond to high orders of interference and so are not visible in white light. For example, if $h = 1$ cm and $n' = 1 \cdot 5$, the order of interference at the centre is about 75,000 for $\lambda_0 = 4000$Å and about 43,000 for $\lambda_0 = 7000$Å; hence in the visible spectrum there are about 32,000 wavelengths giving intensity maxima at the centre. It is clear from (5) and §7.3 (15) that if the fringes are to be distinct, the departure from strict monochromatism of the source becomes more severely restricted as the optical thickness of the plate is increased.

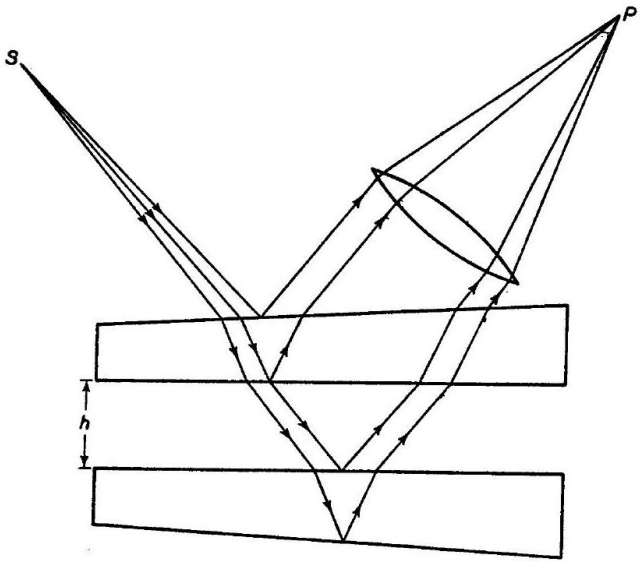



Fig. 7.22. Illustrating formation of fringes localized at infinity with plane parallel plate of air.

As we shall see later (§7.5.8), there is an upper limit to the optical thickness of a plate that will yield fringes with available sources.

So far we have assumed that the optical thickness of the plate is everywhere the same. In practice this assumption may be justified by using a suitable diaphragm to limit the illuminated area of the plate. From (8), a change of optical thickness $\Delta(n'h)$ results in a displacement of the pattern through $\Delta m$ orders, where

$$\Delta m = \frac{2\cos\theta'}{\lambda_0}\Delta(n'h), \tag{14}$$

and at the centre, where $\theta' = 0$,

$$\Delta m = \frac{2}{\lambda_0}\Delta(n'h). \tag{15}$$

Thus if the plate is moved relative to the diaphragm, variations of $n'h$ may be determined from changes of the order of interference at the centre. The method has been used in optical workshops for testing plates which are required to be of uniform optical thickness.*

We have considered so far only the light reflected from the plate, but evidently similar considerations apply also to the transmitted light. In this case (Fig. 7.23) two rays from $S$ reach the focal plane of the telescope at $P$, one directly transmitted and the other after two internal reflections. The difference of optical path along them is found, in a way similar to the derivation of (5), to be

$$\Delta\mathscr{S} = 2n'h\cos\theta', \tag{16}$$

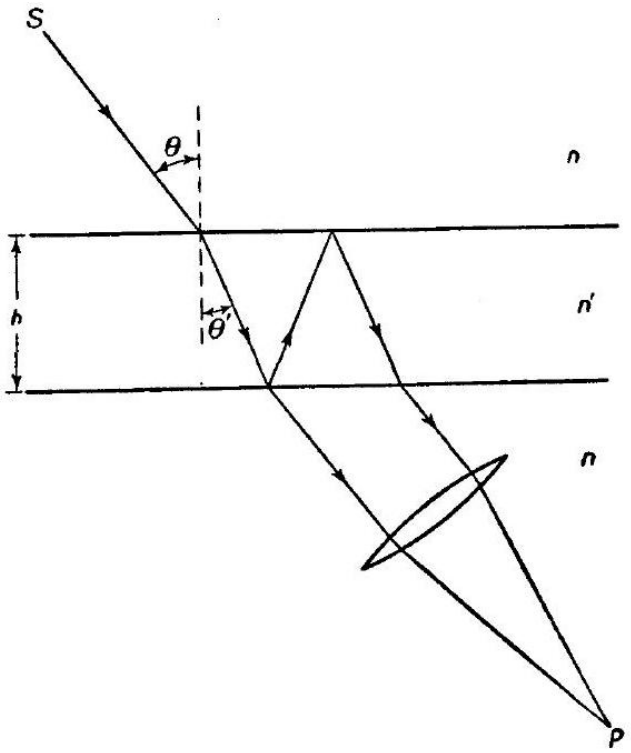



Fig. 7.23. Plane parallel plate: illustrating the formation of fringes localized at infinity in transmitted light.

so that the corresponding phase difference is

$$\delta = \frac{\mathbf{4}\pi}{\lambda_0} n'h\cos\theta'. \tag{17}$$

There is no additional phase difference from phase changes on reflection, since the internal reflections take place at the two surfaces under identical conditions. With an extended source there is again an interference pattern localized at infinity, and comparison of (17) and (7a) shows that this pattern in transmitted light and the pattern in reflected light are complementary, in the sense that the bright fringes of the one and the dark fringes of the other are in the same angular positions relative to the plate normal. However, if the reflectivity of the plate surfaces is low (as for example with glass-air boundaries, for which the reflectivity at normal incidence is

0

* F. Twyman, Priem and Lens Making (London, Hilger and Watts, Ltd., 2nd edition, 1952), p. 388.

about 0.04 ), the two beams which form the transmitted pattern are of very different intensity; by §7.2 (16), the difference of intensity between maxima and minima is small, and the fringes are of low visibility.

The foregoing discussion is only approximate because we have ignored the effect of multiple internal reflections in the plate. In reality a series of beams reaches $P$ from $S$, and not just two as we have supposed. So long as the reflectivity of the plate surfaces is low, the approximation is good because the beams after the first two carry negligible energy. From the more exact discussion given later (§ 7.6) we shall see that the positions of the maxima and minima are given correctly by (8); but if
the reflectivity of the plate surfaces is high, the multiple reflections greatly modify the distribution of intensity in the fringes.

## 9.9.2 7.5.2 Fringes with thin films; the Fizeau interferometer

Suppose a transparent film with plane reflecting surfaces, not necessarily parallel, is illuminated by a point source $S$ of quasi-monochromatic light. Two rays* from $S$, namely $SAP$ and $SBCDP$ (Fig. 7.24), reach any point $P$ on the same side of the film as $S$, so that there is a non-localized interference pattern in this region. The difference between these two optical paths from $S$ to $P$ is

$$\Delta\mathscr{S} = n(SB + DP - SA - AP) + n'(BC + CD), \tag{18}$$

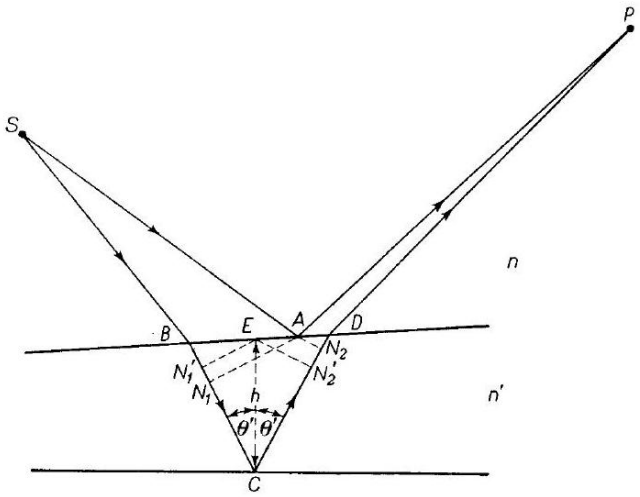



Fig. 7.24. Thin film with point source.

where $n', n$ are respectively the refractive indices of the film and of the surrounding medium. The exact value of $\Delta\mathscr{S}$ may be difficult to calculate, but if the film is sufficiently thin, $B, A$, and $D$ are close together on the upper surface, so that
and

$$nSA \sim nSB + n'BN_1, \tag{19a}$$

$$nAP \sim nDP + n'N_2D, \tag{19b}$$

where $AN_1, AN_2$ are respectively perpendicular to $BC, CD$. From (18) and (19),

$$\Delta\mathscr{S} \sim n'\left(N_1C + CN_2\right). \tag{20}$$

Further, if the angle between the surfaces of the film is sufficiently small,

$$N_1C + CN_2 \sim N_1'C + CN_2', \tag{21}$$

where $N_1', N_2'$ are respectively the feet of perpendiculars from $E$ to $BC, CD$; and $E$ is the intersection with the upper surface of the normal to the lower surface at $C$. Now

$$N_1'C = CN_2' = h\cos\theta', \tag{22}$$

where $h = CE$ is the thickness of the film at $C$, measured normal to the lower surface, and $\theta'$ is the angle of reflection in the film. Hence for a thin film of small angle we may write, from (20), (21), and (22)

$$\Delta\mathscr{S} = 2n'h\cos\theta', \tag{23}$$

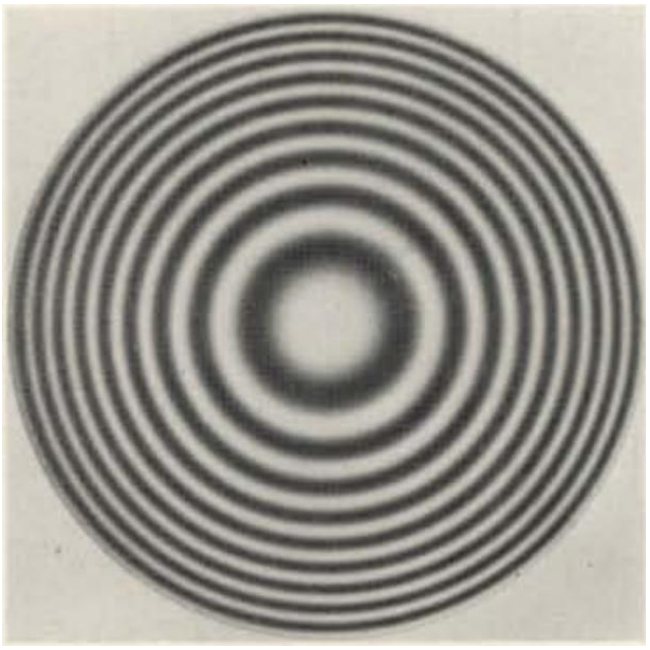


Fig. 7.28. Newton's rings.

0

• We again neglect multiple reflections; their effect is considered later (§ 7.6).

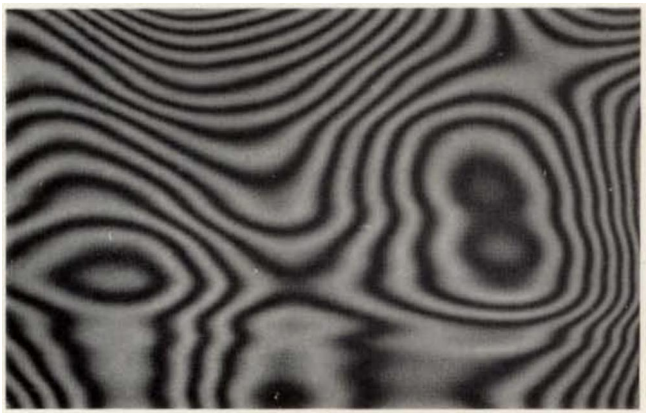


Fig. 7.26. Fringes of equal thickness given by a thin sheet of glass.

Such fringes may be observed in a thin air film between two reflecting surfaces of two transparent plates. Near normal incidence, the condition (25) for a dark fringe then becomes, with $\overline{\cos\theta'} = 1$, and the wavelength $\lambda = \lambda_0/n$ in air,

$$h = \frac{m\lambda}{2}, \quad m = 0, 1, 2, \ldots \tag{26}$$

The fringes are thus contours of the film at thickness intervals $\lambda/2$. If the film is of constant thickness the intensity over it is uniform; this effect is commonly used to test the figure of an optical surface by observing the film between the surface and a reference surface (proof plate) of equal and opposite curvature. When the air film is a wedge formed between plane surfaces, the fringes are equidistant and

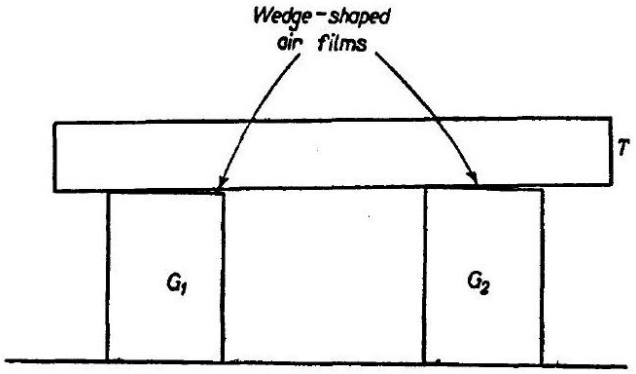



Fig. 7.27. Interferometric comparison of end gauges.

parallel to the apex of the wedge. The linear separation of adjacent bright fringes is $\lambda/2\alpha$, where $\alpha$ is the wedge angle; for example with $\alpha = 1$ minute of arc, $\lambda = 5500$Å, the separation is about 1.9 mm , showing that the wedge angle must be very small if the fringes are to be reasonably spaced. Wedge fringes are used in testing end gauges which serve as standards of length in mechanical workshops. The gauge $G_1$ (Fig. 7.27) is a steel block with two opposite surfaces, which define its length, polished plane and parallel. One of these surfaces, and one surface of a reference gauge $G_2$ of the same nominal length, is wrung in contact with a plane steel surface, and the plane surface of a transparent plate $T$ is allowed to rest on the upper surfaces of the gauges. In general there are wedge-shaped air films between the plate and the gauges, and fringes can be observed in them with monochromatic light. The difference in the lengths of the gauges can be found from their distance apart and the fringe spacing.*

The fringes called Newton's rings (Fig. 7.28), which are of historical interest in connection with Newton's views on the nature of light, are another example of fringes of equal thickness. They are observed in the air film between the convex spherical surface of a lens and a plane glass surface in contact (Fig. 7.29). The fringes are circles about the point of contact $C$. If $R$ is the radius of curvature $OC$ of the convex surface, the thickness of the film at distance $r$ from $C$ is

$$h = R - \sqrt{R^2 - r^2} \sim \frac{r^2}{2R} \tag{27}$$

and the corresponding phase difference at $P$ is

$$\delta = \frac{4\pi}{\lambda_0} n' h \cos\theta'. \tag{24}$$

In general, for a given $P$, both $h$ and $\theta'$ vary with the position of $S$, and a small extension of the source makes the range of $\delta$ at $P$ so large that the fringes disappear. There is, however, a special case when $P$ is in the film, as when observations are made with a microscope focused on the film, or with the eye accommodated for it. Under these circumstances $h$ is practically the same for all pairs of rays from an extended source reaching $P'$, conjugate to $P$ (Fig. 7.25), and differences of $\delta$ at $P'$ are due mainly to differences of $\cos\theta'$. If the range of values of $\cos\theta'$ is sufficiently small, the range of $\delta$ at $P'$ may be much less than $2\pi$, even with a source of appreciable extension, and distinct fringes are then visible, apparently localized in the film. In

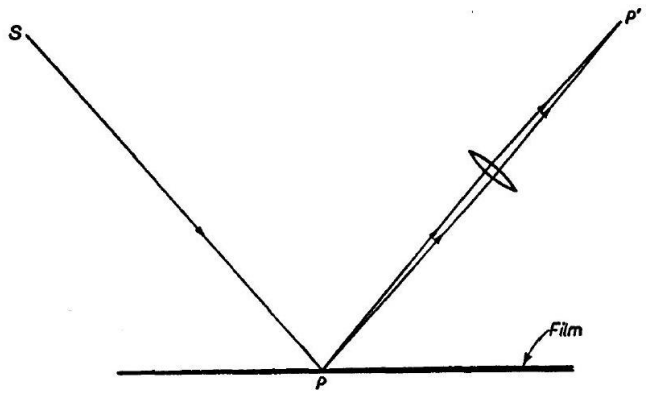



Fig. 7.25. Thin film: illustrating formation of fringes localized in the film.

practice, the condition of a small range of $\cos\theta'$ can be satisfied by observing near normal incidence, and also by restricting the entrance pupil, though the pupil of the unaided eye may itself be sufficiently small. Taking into account the phase change of $\pi$ on reflection at one of the surfaces of the film, there are, by (24) and §7.2(16) maxima of intensity at $P'$ (and so apparently at $P$ ) when

$$2n'h\overline{\cos\theta'} \pm \frac{\lambda_0}{2} = m\lambda_0, \quad m = 0, 1, 2, \ldots \tag{25a}$$

and minima of intensity when

0

- The phase change on reflection at the metallic surface is not strictly $\pi$, as assumed in (25), but this does not affect the fringe spacing.

$$2n'h\overline{\cos\theta'} \pm \frac{\lambda_0}{2} = m\lambda_0, \quad m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{25b}$$

where $\overline{\cos\theta'}$ is a mean value of $\cos\theta'$ for the points of the source which contribute light to $P'$. The quantity $n'h$ which appears in these relations is the optical thickness of the film at $P$; as far as our approximations are justified, the state of interference at $P$ is unaffected by the film thickness elsewhere. It follows that (25) holds even if the bounding surfaces of the thin film are not plane, so long as the angle between them remains small. Then if $\overline{\cos\theta'}$ is effectively constant, the fringes are loci of points in the film at which the optical thickness is constant, and for this reason they are often called fringes of equal thickness (Fig. 7.26).
if we neglect terms of the fourth order. With normal incidence, the condition for a dark fringe is therefore, by (26) and (27),

$$r = \sqrt{mR\lambda}, \quad m = 0, 1, 2, \ldots, \tag{28}$$

so that the radii of the dark fringes are proportional to the square roots of the positive integers. If the lens and plate are separated, the points of the film with given $h$ move inwards and the fringes collapse towards the centre, where one disappears each time

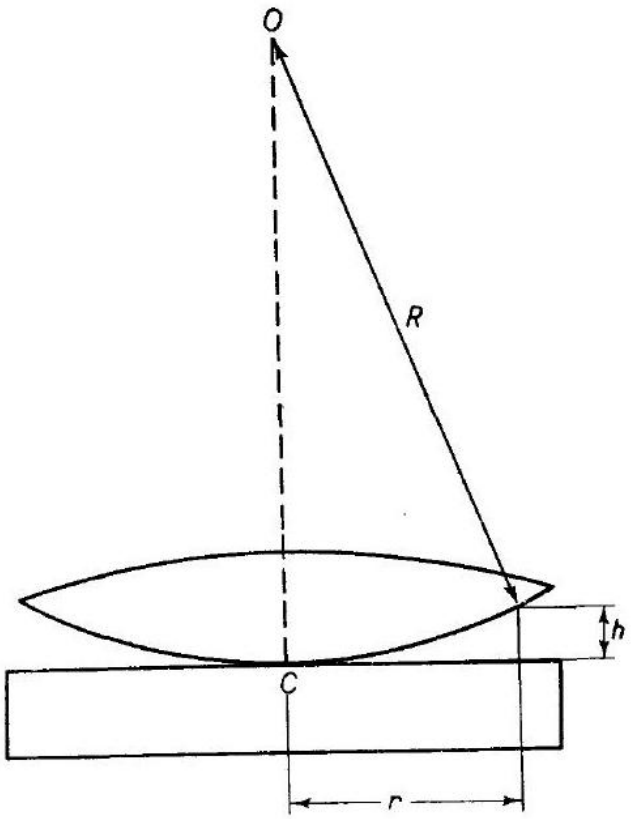



Fig. 7.29. Illustrating the formation of Newton's rings.

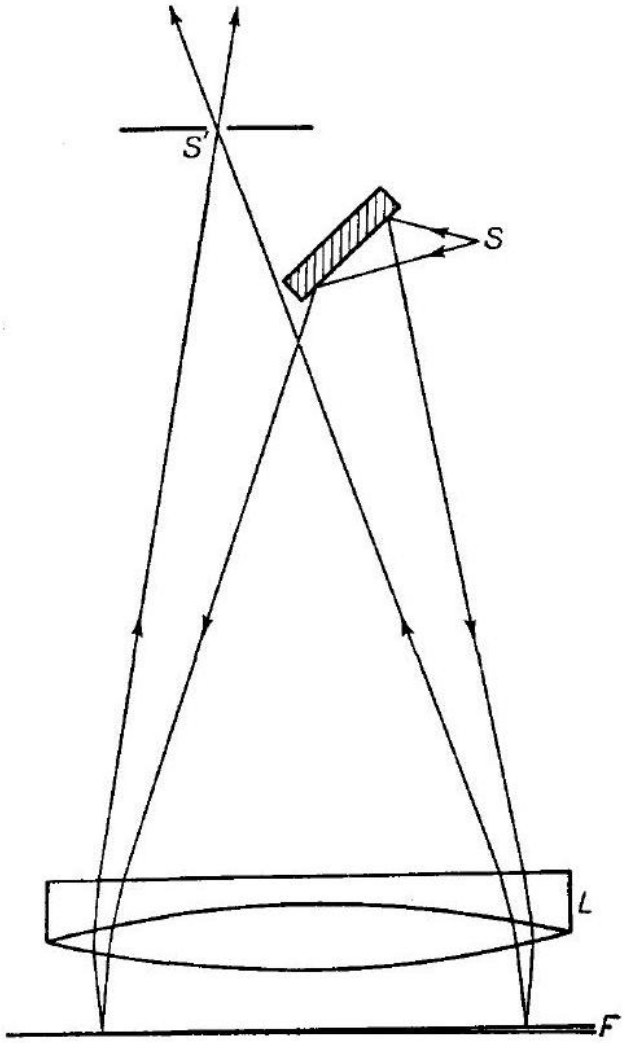



Fig. 7.30. The Fizeau interferometer.

the separation increases by $\lambda/2$. It is interesting to note that, like Young's experiment, the arrangement provides a means of determining the approximate wavelength of light with very simple apparatus.

If the film thickness is only a few half-wavelengths, the orders of interference in the monochromatic pattern are very low, and fringes are visible with a white light source. The reflected colours exhibited by soap bubbles, and by oil films on water, are examples of such fringes. With the arrangement for observing Newton's rings (Fig. 7.29 ), the phase difference at the centre is $\pi$ for all wavelengths when the lens and plate are in contact, so that with white light there is a black central spot. Away from this the patterns from the different monochromatic components of the source become increasingly out of step; to visual observation there are coloured rings immediately surrounding the centre, in a characteristic sequence known as Newton's colours, and farther out is what appears to the eye to be uniform white illumination (cf. §7.3.3). Similarly, with a wedge-shaped air film, there is in white light a black fringe defining the apex of the wedge.

We have been concerned so far with light reflected from the film, but an interference pattern localized in the film is visible also in transmitted light. As in the case of a plane parallel plate, the patterns with reflected and transmitted light are
complementary; the bright fringes of one appear at the same points of the film as the dark fringes of the other. With surfaces of low reflectivity, the visibility of the fringes in the transmitted

pattern is low because of the disparity of intensities of the interfering beams.

We have seen that if the fringes are to be distinct the range of values of $\cos\theta'$ corresponding to each point of the film must be restricted, and that the fringes follow lines of equal optical thickness only if $\overline{\cos\theta'} \sim 1$. These conditions are satisfied simultaneously over a large area of the film in the Fizeau interferometer* (Fig. 7.30).

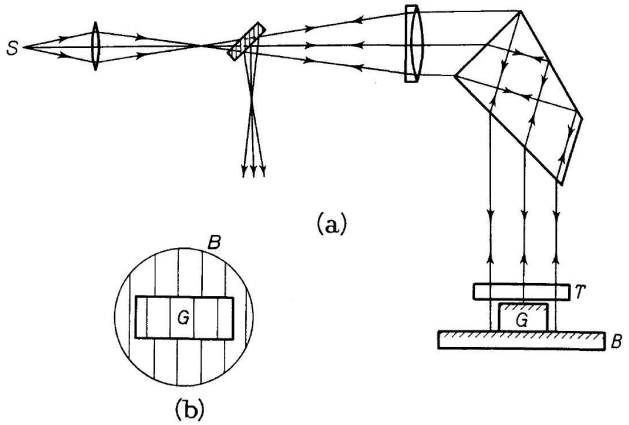



Fig. 7.31. N.P.L. gauge-measuring interferometer: (a) arrangement; (b) field of view.

Light from a quasi-monochromatic source $S$, after reflection at a small mirror, is collimated by the lens $L$ and falls on the film $F$ at nearly normal incidence. The light reflected from the surfaces of the film returns through $L$ and converges to an aperture $S'$ in the focal plane of $L$. To an eye placed immediately behind $S'$ and accommodated for the film, fringes are visible over the whole area illuminated by $L$, following lines of equal optical thickness. These fringes are often called Fizeau fringes. As we shall see later (§7.5.3), Fizeau fringes may be obtained with films which are not thin providing the source is sufficiently small, and the interferometer is used in optical workshops to test the uniformity of optical thickness of plane parallel transparent plates. It is also used in the National Physical Laboratory (N.P.L.) in Great Britain to measure the lengths of end gauges. † The arrangement for this is shown in Fig. 7.31(a). One surface of the gauge $G$ is wrung in contact with the plane surface of a polished steel plate $B$ so that the upper surfaces of $G$ and $B$ are parallel, and a transparent plate $T$, whose lower surface is plane, is mounted above them. In general there are wedge-shaped air films between $T$ and $G$, and between $T$ and $B$, in each of which an interference pattern can be observed. The fringes are parallel straight

lines with the same spacing in each pattern, and by adjusting the inclination of $T$ the fringes are arranged to run at right angles to one edge of the gauge; the appearance is then as in Fig. 7.31(b). If at a point, on the edge of the gauge the air films over $G$ and $B$ are of thickness $h_1$ and $h_2$ respectively, and if the corresponding orders of interference are $m_1, m_2$, the length $d$ of the gauge is given by (26) as

$$d = h_2 - h_1 = (m_2 - m_1)\frac{\lambda}{2} = \Delta m\frac{\lambda}{2}, \tag{29}$$

where $\Delta m$ is the order displacement of the two patterns. Thus we may write

0

• H. Fizead, Ann. Chim. Phys., (3), 66 (1862), 429.
† F. H. Rolx, Engineering, 144 (1937), 162.

$$d = (x + e)\frac{\lambda}{2} \tag{30}$$

where $x$ is an unknown integer, and $e < 1$ is the fractional order displacement, which can be measured by interpolation. The value of $x$ is found by the method of excess fractions, which was first used in interferometry by Benoit.* The fractional order displacement of the patterns is measured with light of four known wavelengths, which are conveniently isolated in turn from a suitable source by rotation of the constant deviation prism $D$. There are then relations of the form

$$d = (x_1 + e_1)\,\frac{\lambda_1}{2} = (x_2 + e_2)\,\frac{\lambda_2}{2} = (x_3 + e_3)\,\frac{\lambda_3}{2} = (x_4 + e_4)\,\frac{\lambda_4}{2} \tag{31}$$

where $x_1, x_2, x_3$, and $x_4$ are unknown integers; and these relations define a set of tetrads of possible values of $x_1, x_2, x_3$, and $x_4$. An approximate measurement of $d$ with a micrometer is sufficient to decide which of these tetrads is correct, and hence a precise value of $d$ is obtained. With proper precautions, measurements of gauges up to about 10 cm in length can be made; if the fractional order is correct to 0.1 , the accuracy is $\pm 2.5 \times 10^{-6}$ cm for $\lambda = 5000$Å.

It is evident that there is a limiting case of the Fizeau arrangement when the source $S$ (Fig. 7.30) is reduced to a point, so that $\theta'$ has a unique value at each point of the film. Under these circumstances, however, at least if the film surfaces are plane, the fringes must be non-localized, i.e. the fringes over any plane common to the reflected beams are just as distinct as those in the film. We are therefore led to examine more closely the concept of fringe localization, and its relation to the extension of the source.

## 9.9.3 7.5.3 Localization of fringes †

The arrangements for producing interference effects that we have considered so far may be represented generally as devices by which light from a source is made to reach points in a region of space by two different paths. Let $P$ be a point in this region, and suppose first that the light originates in a quasi-monochromatic point source $S$, of wavelength $\lambda_0$. If $SA_1B_1P, SA_2B_2P$ are the two rays from $S$ to $P$ (Fig. 7.32), the phase difference at $P$ is

$$\delta_0 = \frac{2\pi}{\lambda_0}\,\{[SA_2B_2P] - [SA_1B_1P]\} \tag{32}$$

The value of $\delta_0$ depends on the position of $P$ but it is uniquely determined for all $P$, so that interference fringes, following the loci of points for which $\delta_0$ is constant,

are formed over any plane in the region common to the two paths from $S$. We say that the fringes are non-localized; their visibility depends only on the relative intensities of the two waves. Such fringes are always obtained with a point source.

Suppose now that the source is a quasi-monochromatic primary source extended about $S$. As in §7.3.4 we assume that such a source is made up of incoherent point sources, each of which gives rise to a non-localized interference pattern; the total intensity is then, at each point, the sum of the intensities in these elementary patterns. If the phase difference at $P$ is not the same for all points of the extended source, the elementary patterns are mutually displaced in the vicinity of $P$, and the visibility of the fringes at $P$ is less than with a point source. In general, as we shall see later,

0

- J. R. Benoit, Journ. de Phys. (3), 7 (1898), 57.
  † For a more extensive discussion of the localization of fringes see J. Mac de Lpinay and C. Fabry, Journ. de Phys. (2), 10 (1891), 5.

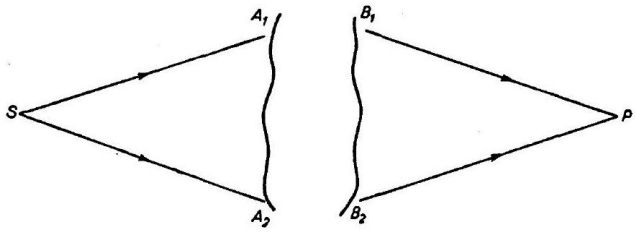



Fig. 7.32. Illustrating interference of two beams from a point source.

the mutual displacement-and hence the reduction of visibility-increases with the extension of the source, but at a rate which depends on the position of $P$. Thus as the source is extended about $S$, the visibility at some positions of $P$ may remain at or near its value with a point source, when elsewhere it has fallen effectively to zero. We then say that the fringes are localized; such fringes are characteristic of an extended source.

The reduction of visibility associated with a given source extension is in general difficult to calculate, since it depends on the relative intensities of the elementary patterns as well as their mutual displacement. However, we can easily recognize two extreme cases. A mutual displacement of the elementary patterns which is small compared with one order will obviously have little effect on the fringes; and in the special circumstances to which Fig. 7.11 refers we have seen that the loss of visibility remains inappreciable, at least for visual observations, so long as

$$\delta_{\max} - \delta_{\min} \lesssim \frac{\pi}{2}, \tag{33}$$

where $\delta_{\max}, \delta_{\min}$ are the maximum and minimum phase differences at $P$ for points of the extended source. This criterion is valid also for the source distributions of Fig. 7.15 (b), (c), and we may assume it to be generally satisfactory. On the other hand, when

$$\delta_{\max} - \delta_{\min} \gg \pi, \tag{34}$$

the elementary patterns are mutually displaced through many orders, and the visibility of the fringes will be very small.

Let us now examine how $\delta$ for a point $S'$ of the extended source depends on the position of $S'$ relative to $S$. The points $S$ and $P$ are "conjugate" in the sense that a point source of wavelength $\lambda_0$ at $P$ would give rise to two waves at $S$ with phase difference $\delta_0$. If $W_1, W_2$ are the wave-fronts through $S$ for these two waves (Fig. 7.33), $W_1, W_2$ are normal to $SA_1, SA_2$ respectively; their curvatures depend on the
positions of $S$ and $P$, and on the optical properties of the particular arrangement. Let the normals from $S'$ to $W_1, W_2$ meet them in $N_1, N_2$ respectively. The optical path from $N_1$ to $P$ is equal to $[SA_1B_1P]$, and similarly the optical path from $N_2$ to $P$ is equal to $[S_{A_2}B_2P]$. The phase difference at $P$ corresponding to $S'$ is therefore

$$\delta = \frac{2\pi}{\lambda_0}\left\{[S'N_2] + [SA_2B_2P] - [S'N_1] - [SA_1B_1P]\right\}, \tag{35}$$

or by (32),

$$\delta - \delta_0 = \frac{2\pi n}{\lambda_0}(s_2 - s_1), \tag{36}$$

where $s_1 = S'N_1, s_2 = S'N_2$, and $n$ is the refractive index of the medium surrounding the source.

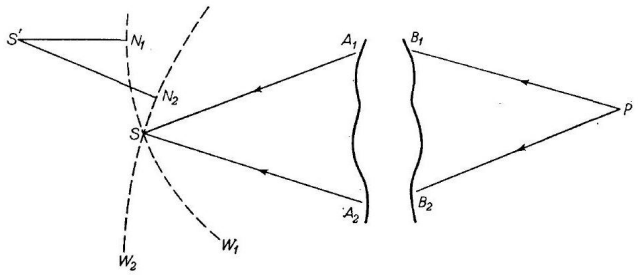



Fig. 7.33. Illustrating interference of two beams from an extended source.

Let $S$ be the origin of rectangular coordinate axes $SX, SY, SZ$, with $SZ$ and $SX$ respectively internal and external bisectors of the angle $A_1SA_2$ (Fig. 7.34). If $W_1, W_2$ have radii of curvature $R_1, R_2$ in the planes $S'SA_1, S'SA_2$ respectively, the corresponding centres of curvature $C_1, C_2$ are at $\{R_1 \sin(\beta/2), 0, R_1 \cos(\beta/2)\ )$, $(\ -R_2 \sin(\beta/2), 0, R_2 \cos(\beta/2)\}$ respectively, where $\beta$ denotes the angle $A_1SA_2$. If $(\ x, y, z\ )$ are the coordinates of $S'$, then

$$C_1S' = s_1 + R_1 = \sqrt{\left(x - R_1 \sin\frac{\beta}{2}\right)^2 + y^2 + \left(z - R_1 \cos\frac{\beta}{2}\right)^2}$$
$$= R_1\sqrt{1 - \frac{2}{R_1}\left(x \sin\frac{\beta}{2} + z \cos\frac{\beta}{2}\right) + \frac{(x^2 + y^2 + z^2)}{R_1^2}}. \tag{37}$$

When the source has linear dimensions small compared with $R_1$ we may neglect powers of $x/R_1, y/R_1, z/R_1$ higher than the second, so that

$$s_1 \sim -x \sin\frac{\beta}{2} - z \cos\frac{\beta}{2} + \frac{1}{2R_1}\left\{x^2 + y^2 + z^2 - \left(x \sin\frac{\beta}{2} + z \cos\frac{\beta}{2}\right)^2\right\}. \tag{38}$$

Similarly,

$$s_2 \sim x \sin\frac{\beta}{2} - z \cos\frac{\beta}{2} + \frac{1}{2R_2}\left\{x^2 + y^2 + z^2 - \left(x \sin\frac{\beta}{2} - z \cos\frac{\beta}{2}\right)^2\right\}. \tag{39}$$

In most cases of practical interest $\beta$ is small and terms in the expansions of (38) and (39) which involve $\beta^2$ and $\beta z/R$ may be neglected. From (36), (38), and (39) we then obtain

$$\delta - \delta_0 \sim \frac{2\pi n}{\lambda_0}\left\{\beta x + \frac{1}{2}\left(\frac{1}{R_2} - \frac{1}{R_1}\right)(x^2 + y^2)\right\}. \tag{40}$$

When the term of (40) independent of $\beta$ can be neglected,

$$\delta - \delta_0 \sim \frac{2\pi n}{\lambda_0}\beta x \tag{41}$$

With this approximation, $\delta = \delta_0$ when $x = 0$, so that the phase difference (and hence reduction of the visibility) is negligible when the source is extended in the plane YSZ. By (33) and (41), the reduction of visibility remains inappreciable if the source

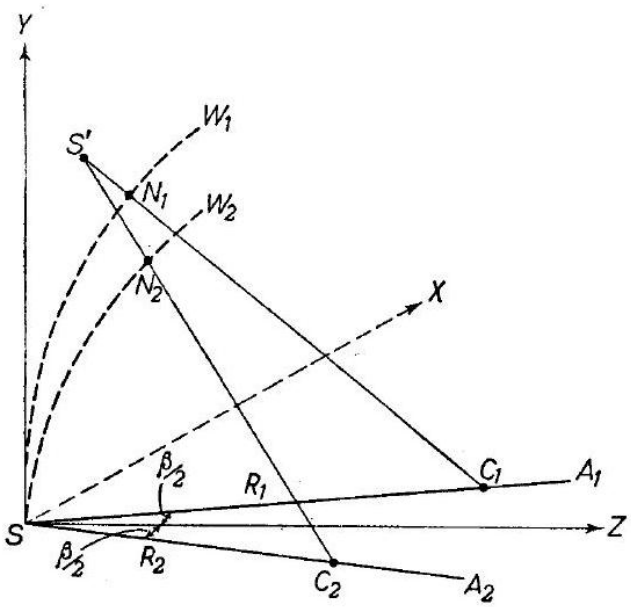



Fig. 7.34. Illustrating the discussion of fringe localization with a two-beam interference arrangement.

extension normal to this plane does not exceed $\lambda_0/4n\beta$. Now the orientations of $SA_1$ and $SA_2$, and hence of the coordinate axes we have chosen, depend in general on the position of $P$. If for all $P$ the planes $YSZ$ have a common line intersection, a line source (or in practice a slit source of width less than $\lambda_0/4n\beta$ ) along this intersection gives fringes which have everywhere the same visibility as with a point source; the fringes remain non-localized. This is so for the arrangements of §7.3.2 (except Meslin's experiment). With Fresnel's mirrors, for example, whatever the position of $P$, the plane $YSZ$ contains the line through $S$ parallel to the mirror junction, and a slit source oriented along this line gives non-localized fringes, as we have seen in §7.3.4. It is easy to verify that the expression of §7.3. (26) for the slit width is equivalent to $e \leqslant \lambda_0/4n\beta$. More generally, however, the planes $YSZ$ for all $P$ have no common intersection. A line source may lie in the plane $YSZ$ for points $P$ in a restricted region; but elsewhere the source extends outside this plane to distances large compared with $\lambda_0/2n\beta$, and according to (34) and (41) this means that the visibility is very small. Under these circumstances the fringes given by a slit source are localized, and the region of localization depends on the orientation of the slit.

The value of $\beta$ depends on the position of $P$, and by (40), if we neglect the dependence of $R_1$ and $R_2$ on $\beta, \delta - \delta_0$ for given $x$ decreases as $\beta$ decreases. It follows that, as the source extension normal to the plane $YSZ$ is increased, the fringes become localized in regions corresponding to sufficiently small values of $\beta$. In particular, $\beta = 0$ when $SA_1$ and $SA_2$ coincide; the region of localization includes those points $P$-if they
exist-which lie at the intersection of two rays derived from a single incident ray from $S$. In the vicinity of these points, $\beta$ is negligible, and we have from (40),

$$\delta - \delta_0 \sim \frac{\pi n}{\lambda_0}\left(\frac{1}{R_2} - \frac{1}{R_1}\right)\left(x^2 + y^2\right). \tag{42}$$

If $W_1, W_2$ are spherical, $R_1, R_2$ are independent of $x$ and $y$, and the reduction of visibility is independent of the direction in which the source is extended in the plane $YSZ$. For a circular source in this plane, the reduction of visability is, according to (33) and (42), inappreciable providing the radius of the source does not exceed about $\sqrt{\lambda_0 R_1 R_2 / 2n\,|R_1 - R_2|}$.

In the case of a plane parallel plate observed with a telescope (§ 7.5.1), $\beta = 0$ when $P$ is in the focal plane of the telescope objective. $R_1$ and $R_2$ are then both infinite for

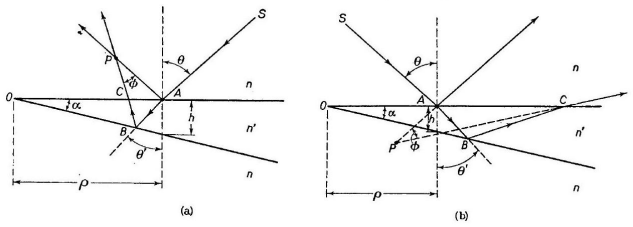



Fig. 7.35. Rays reflected and refracted at the surfaces of a plane wedge.

all positions of $S$, so that by (42), $\delta = \delta_0$ for all $x$ and $y$. Thus the fringes in the focal plane of the objective suffer no reduction of visibility whatever the position and extension of the source. In practice, the aperture of the telescope always sets an upper limit to the effective extension of the source. Hence if the fringes are observed outside the focal plane of the objective, as when the telescope is imperfectly adjusted, the reduction of visibility remains inappreciable so long as this effective extension of the source does not exceed the value allowed by (33) and (40).

As a further example*, let us consider the fringes in reflected light from a transparent wedge of refractive index $n'$, bounded by plane surfaces inclined at a small angle $\alpha$, and situated in a medium of refractive index $n$. Suppose $S$ is a quasi-monochromatic point source, and consider a ray in a principal section of the wedge (section at right angles to the apex), which meets the front surface of the wedge in $A$ (Fig. 7.35). This incident ray gives rise to a reflected ray $AP$, and a refracted ray $AB$ which, after reflection at $B$ on the rear surface and refraction at $C$ on the front surface, emerges along $CP$ to meet the first ray in $P$. When the source is extended about $S$, the fringes near the plane $SAP$ become localized in the vicinity of $P$.

Let $\rho$ be the distance of $A$ from the apex $O$ of the wedge; let $\theta, \theta'$ be respectively the angles of incidence and refraction at $A$; and let $\phi$ be the angle $APC$. When $B$ and $O$ are on the same side of the normal from $A$ to the rear surface of the wedge,

$P$ is real (Fig. 7.35 (a)); the angle of reflection at $B$ is $(\theta' - \alpha)$, the angle of emergence at $C$ is $(\theta - \phi)$, and we have by elementary geometry,

$$AB = \frac{\rho \sin \alpha}{\cos(\theta' - \alpha)} \tag{43}$$

$$\frac{BC}{AB} = \frac{\cos \theta'}{\cos(\theta' - 2\alpha)} \tag{44}$$

$$\frac{AC}{AB} = \frac{\sin 2(\theta' - \alpha)}{\cos(\theta' - 2\alpha)} \tag{45}$$

$$\frac{AP}{AC} = \frac{\cos(\theta - \phi)}{\sin \phi} \tag{46}$$

$$\frac{CP}{AC} = \frac{\cos \theta}{\sin \phi} \tag{47}$$

We also have, by Snell's law, for refraction at $A$ and $C$ respectively.

0

- For fuller treatments of the fringes given by a plane wedge see J. Mace de Lpinay, Journ. de Phys. (2), 9 (1890), 121; W. Feussner and L. Janicki, Handbuch der Physikalischen Optik, I (ed. E. Gehrcke, Leipzig, Barth, 1926), 396; G. F. C. Searle, Phil. Mag. (7), 37 (1946), 361.

$$n' \sin \theta' = n \sin \theta \tag{48}$$

$$n' \sin (\theta' - 2\alpha) = n \sin(\theta - \phi) \tag{49}$$

From (48) and (49), by subtraction and the use of the identity

$$\sin a - \sin b = 2 \cos \frac{a+b}{2} \sin \frac{a-b}{2}$$

we have

$$\sin \frac{\phi}{2} = \frac{n' \sin \alpha \cos (\theta' - \alpha)}{n \cos \left(\theta - \frac{\phi}{2}\right)}; \tag{50}$$

and providing $\theta$ is not too near $\pi/2$, we may write, for $\alpha$ small,

$$\phi \sim \frac{2n'\alpha \cos \theta'}{n \cos \theta} \tag{51}$$

From (43), (45), and (46),

$$AP = \frac{2\rho \sin \alpha \sin (\theta' - \alpha) \cos(\theta - \phi)}{\cos (\theta' - 2\alpha) \sin \phi} \tag{52}$$

For $\alpha$ small we may expand (52) in powers of $\alpha$, using (51). If we then retain only the leading term, we obtain, if we use also (48),

$$AP \sim \frac{\rho n^2 \sin \theta \cos^2 \theta}{n'^2 - n^2 \sin^2 \theta} \tag{53}$$

The quantity neglected in (53) is of the order of $\rho\alpha$, i.e. it is of the order of the thickness of the wedge $A$. When $S$ and $O$ are on the same side of the normal from $A$ to the rear surface of the wedge (Fig. 7.35 (b)), $P$ is virtual, but the approximate formula (53) holds.

If the wedge is an air film between glass plates, and if we ignore the effects of refraction in the plate on the front side, we may put $n = n' = 1$ in (53) and find $AP \sim \rho \sin \theta$, so that the angle $OPA \sim \pi/2$. The locus of $P$ is a circle on $OS_1$ as
diameter, where $S_1$ is the reflected image of $S$ in the front surface of the wedge (Fig. 7.36).

If $S$ is so far from the wedge that all the incident rays may be taken to be parallel* to $SA, \theta$ and hence $\theta'$ and $\phi$ are independent of $\rho$, and we see from (52) that $AP$ is

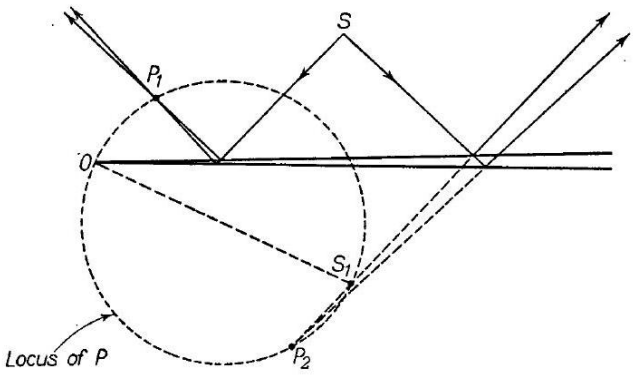



Fig. 7.36. Localization of fringes in the principal section of an air wedge (neglecting effects of refraction at the bounding surfaces).

proportional to $\rho$. The locus of $P$ is now a plane passing through the vertex of the wedge (Fig. 7.37); the angle $\gamma$ between this plane of localization and the plane normal to the reflected beam is found, using (53), and the relation $AN = \rho \sin\theta$, to be

$$\tan\gamma = \frac{AN - AP}{ON} \sim \left(\frac{n'^2 - n^2}{n'^2 - n^2\sin^2\theta}\right)\tan\theta. \tag{54}$$

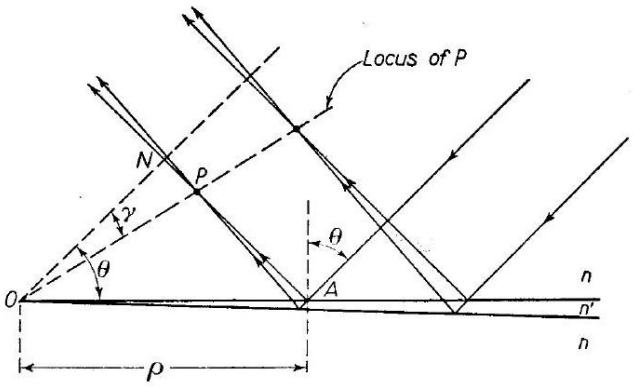



Fig. 7.37. Localization of wedge fringes when the source is at infinity.

Evidently when $\theta = 0, \gamma \sim 0$, so that for normally incident light the plane of localization effectively coincides with the front surface of the wedge.

Let us now consider the difference of optical path $\Delta\mathscr{S}$ at $P$. From Fig. 7.35 (a), when $P$ is real,

$$\Delta\mathscr{S} = n'(AB + BC) + n(CP - AP). \tag{55}$$

By (43) and (44), for $\alpha$ small and $\theta'$ not too near $\pi/2$,

$$AB + BC = \frac{2\rho\sin\alpha\cos\alpha}{\cos(\theta' - 2\alpha)} \sim \frac{2h}{\cos\theta'} \tag{56}$$

where

$$h = \rho\tan\alpha \tag{57}$$

is the thickness of the wedge at $A$, and in the approximation we have neglected terms involving powers of $\alpha$. Similarly, by (43), (45), (46), and (47),

$$CP - AP = \frac{-2\rho\sin\alpha\sin\left(\theta - \frac{\phi}{2}\right)\sin(\theta' - \alpha)}{\cos\frac{\phi}{2}\cos(\theta' - 2\alpha)} \sim -\frac{2h\sin\theta\sin\theta'}{\cos\theta'}. \tag{58}$$

0

- This condition is satisfied exactly in the Fizeau arrangement (§ 7.5.2).

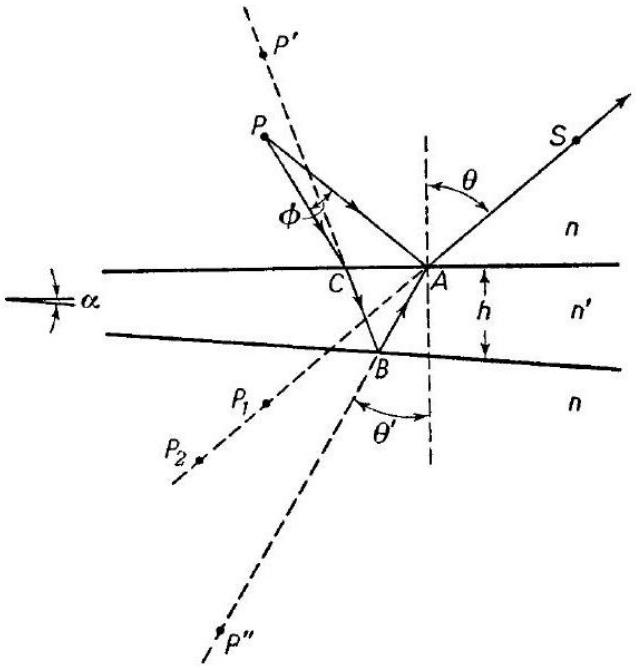



Fig. 7.38. Centres of curvature of wave-fronts from a point source after reflection and refraction at the surfaces of a plane wedge.

Hence by (55), (56), (58), and (48),

$$\Delta\mathscr{S} \sim 2n'h\cos\theta', \tag{59}$$

and the corresponding phase difference at $P$ is

$$\delta \sim \frac{4\pi}{\lambda_0} n'h\cos\theta' = \frac{4\pi h}{\lambda_0}\sqrt{n'^2 - n^2\sin^2\theta}. \tag{60}$$

In a similar way we find that this approximation holds also for the conditions of Fig. 7.35 (b), when $P$ is virtual. When the phase change $\pi$ which occurs on reflection at one of the wedge surfaces is taken into account, there are, by (60) and §7.2(16), maxima of intensity at $P$ when

$$2n'h\cos\theta' \pm \frac{\lambda_0}{2} = m\lambda_0, \quad m = 0, 1, 2, \ldots, \tag{61a}$$

and minima of intensity when

$$2n'h\cos\theta' \pm \frac{\lambda_0}{2} = m\lambda_0, \quad m = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{61b}$$

Evidently, for $\theta'$ constant, the fringes in the plane of localization are equidistant and parallel to the apex of the wedge.

Finally, let us see how far the source may be extended about $S$ without appreciably reducing the visibility of the fringes at $P$. For this we imagine a point source placed at $P$, and determine the radii of curvature $R_1, R_2$ of the wave-fronts $W_1, W_2$ for light reaching $S$ after reflection at the front and rear surfaces of the wedge (Fig. 7.38).
$W_1$ is spherical; its centre of curvature is at $P_1$ on $AS$, where $P_1$ is the reflected image of $P$ in the front surface of the wedge. Thus

$$R_1 = SP_1 = SA + AP \tag{62}$$

$W_2$ is in general not spherical. Consider first its radius of curvature in the principal section of the wedge. After refraction at $C$, the centre of curvature $P'$ is situated on $BC$ at a position given by formula (21) of §4.6. With the notation of §4.6 we have $n_0 = n, n_1 = n', r_y = \infty, \theta_0 = \theta - \phi, \theta_1 = \theta' - 2\alpha, d_0{}^{(t)} = CP, d_1{}^{(t)} = CP'$, and § 4.6 (21) gives

$$\frac{CP'}{CP} = \frac{n' \cos^2(\theta' - 2\alpha)}{n \cos^2(\theta - \phi)} \tag{63}$$

After reflection at $B$, the centre of curvature is at $P''$ on $AB$, where $P''$ is the reflected image of $P'$ in the rear surface of the wedge; and after refraction at $A$, the centre of curvature is at $P_2$ on $AS$, where according to §4.6(21) (with $n_0 = n'$, $n_1 = n, r_y = \infty, \theta_0 = \theta', \theta_1 = \theta, d_0{}^{(t)} = AP'', d_1{}^{(t)} = A_2P$),

$$\frac{AP_2}{AP''} = \frac{n \cos^2\theta}{n' \cos^2\theta'} \tag{64}$$

The radius of curvature $R_2$ of the section of $W_2$ in the principal section of the wedge is therefore

$$R_2 = SP_2 = SA + AP_2 \tag{65}$$

where according to (64), and the relation $AP'' = AB + BC + CP'$,

$$AP_2 = \frac{n \cos^2\theta}{n' \cos^2\theta'} (AB + BC + CP') \tag{66}$$

or, using (63),

$$AP_2 = \frac{n \cos^2\theta}{n' \cos^2\theta'}(AB + BC) + \frac{\cos^2\theta \cos^2(\theta' - 2\alpha)}{\cos^2\theta' \cos^2(\theta - \phi)} CP \tag{67}$$

Now for small $\alpha$ this reduces to

$$\begin{aligned} AP_2 &\sim \frac{n \cos^2\theta}{n' \cos^2\theta'}(AB + BC) + CP \\ &\sim \frac{n \cos^2\theta}{n' \cos^2\theta'} \cdot \frac{2h}{\cos\theta'} + AP - \frac{2h \sin\theta \sin\theta'}{\cos\theta'} \end{aligned} \tag{68}$$

where the relations (56) and (58) have been used. From (62), (65), and (68) we finally obtain

$$R_2 - R_1 \sim \frac{2h}{\cos\theta'} \left( \frac{n \cos^2\theta}{n' \cos^2\theta'} - \sin\theta \sin\theta' \right) \tag{69}$$

In a similar way we may derive the corresponding expression relating to the section of $W_2$ at right angles to the principal section of the wedge. In place of §4.6(21) we must now use §4.6 (22); this gives, instead of (63) and (64),

$$\frac{CP'}{CP} = \frac{n'}{n}, \quad \frac{AP_2}{AP''} = \frac{n}{n'} \tag{70}$$

and leads to the following formula for the difference in the radii of curvature of the sections of $W_1$ and $W_2$ at right angles to the principal section:

$$R_2 - R_1 \sim \frac{2h}{\cos\theta'} \left( \frac{n}{n'} - \sin\theta \sin\theta' \right). \tag{71}$$

As we have noted earlier, (33) and (42) imply that the reduction of visibility at $P$ is inappreciable if the radial extension of the source about $SA$ does not exceed $\sqrt{\lambda_0 R_1 R_2/2n\,|R_1 - R_2|}$. Now when $\theta$ is not too close to $\pi/2$ we see from (69) and (71) that $R_2 - R_1$ is of the order of $h$; then if $R_1$ is sufficiently large compared with $h$, as with the Fizeau arrangement, we may take $R_2/R_1 \sim 1$. With this approximation the tolerable angular radius of the source, as viewed from $P_1$, is

$$\varepsilon \sim \frac{1}{R_1}\sqrt{\frac{\lambda_0 R_1 R_2}{2n\,|R_1 - R_2|}} \sim \sqrt{\frac{\lambda_0}{2n\,|R_1 - R_2|}}. \tag{72}$$

In particular, near normal incidence, (69) and (71) give $R_2 - R_1 \sim 2hn/n'$, and (72) becomes

$$\varepsilon \sim \frac{1}{2n}\sqrt{\frac{\lambda_0 n'}{h}} \tag{73}$$

With typical values for a thin air film, $h = 0.01$ cm, $n' = n \sim 1, \lambda_0 = 5500$Å, (73) gives $\varepsilon \sim 2°$. Evidently the tolerable source extension is proportional to $\sqrt{1/h}$; for example, in the above case, but with $h = 1$ cm, $\varepsilon$ according to (73) is about 12 minutes of arc. Thus, as we have mentioned in discussing the Fizeau interferometer, fringes of high visibility may be obtained when the wedge is not thin, providing the source is sufficiently small.

## 9.9.4 7.5.4 The Michelson interferometer

In the arrangements of §7.5.1 and §7.5.2 the two beams are superposed except in the region between the bounding surfaces of the plate or film. For some purposes this is inconvenient; it may be avoided by using an auxiliary semi-reflecting surface to produce the beams, which may then be clearly separated before re-combination. Such an arrangement is the basis of the Michelson interferometer.*

The simplest form of the instrument is shown in Fig. 7.39. Light from an extended source $S$ is divided at the semi-reflecting surface $\mathscr{A}$ of a plane parallel glass plate $D$ into two beams at right angles. These are reflected at plane mirrors $M_1, M_2$, and return to $D$, where they are re-combined to enter the observing telescope $T$. $M_2$ is fixed, while $M_1$ is mounted on a carriage and can be moved towards or away from $D$ by means of a micrometer screw. The beam reflected from $M_1$ traverses the dispersive material of $D$ three times before reaching $T$, compared with a single passage for the beam reflected from $M_2$. To remove this asymmetry, which would otherwise prevent the use of white light fringes, a compensating plate $C$, of material and thickness identical to $D$ and parallel to it, is introduced between $D$ and $M_2$.

Suppose $M_2'$ is the image of $M_2$ in the beam divider. The optical path between $S$ and the point $P$ along a ray $SI_1JI_2P$, transmitted at $\mathscr{A}$ and reflected at $M_2$, is equal to the optical path between $S$ and $P$ along the ray $SI_1KI_2P$, reflected at $\mathscr{A}$ and reflected at the virtual surface $M_2'$. The interference pattern observed with the telescope may therefore be considered to arise from an air film bounded by the real reflecting surface $M_1$ and the virtual reflecting surface $M_2'$, providing we associate

with the latter a phase change $\phi$ equal to the difference between the phase changes for external and internal reflection at $\mathscr{A}$. The value of $\phi$ depends on the nature of the semi-reflector $\mathscr{A}$.

When $M_1$ and $M_2'$ are parallel, the fringes given by a quasi-monochromatic source are circular and localized at infinity. They differ from the fringes of equal inclination considered

0

- A. A. Michelson, Amer. J. Sci. (3), 22 (1881), 120; Phil. Mag. (5), 13 (1882), 236.

in §7.5.1 only in that here there are no multiple reflections, so that the intensity distribution is strictly in accordance with § 7.2 (15). If $M_1$ is moved so that it approaches $M_2'$, the fringes contract towards the centre, but the angular scale of

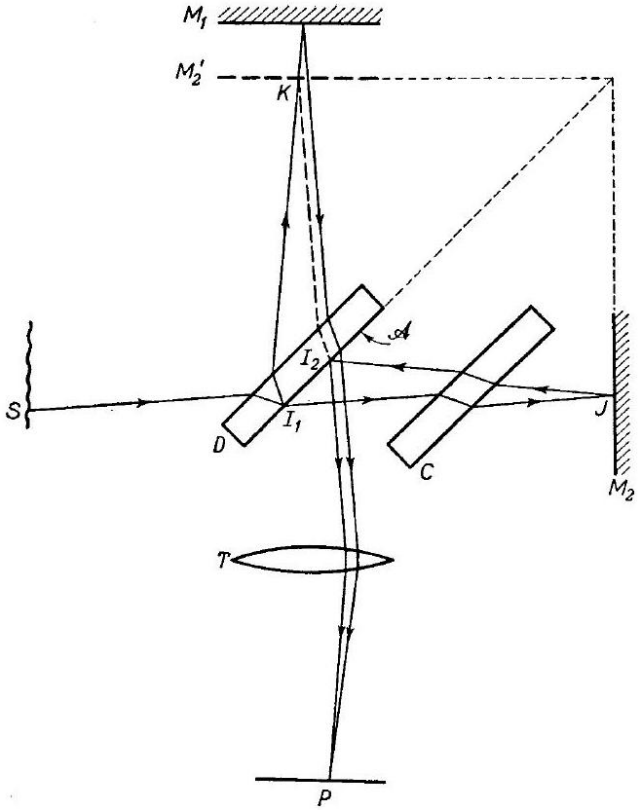



Fig. 7.39. The Michelson interferometer.

the pattern increases until, when $M_1$ coincides with $M_2'$, the illumination over the field of view is uniform at a level which depends on $\phi$. The mirrors $M_1$ and $M_2$ are then said to be in optical contact. When $M_1$ and $M_2'$ are close together but mutually inclined to form a wedge of small angle, there are fringes localized at or near the surface of this wedge. If the separation of $M_1$ and $M_2'$ is sufficiently small, these fringes are fringes of equal thickness and so are equidistant straight lines parallel to the apex of the wedge. As the separation increases, however, the range of incidence angle corresponding to each point of the field of view, and the variation of mean incidence angle over the field of view, cease to be negligible; the visibility of the fringes decreases, and they become curved with convex side towards the wedge apex.

Whether $M_1$ and $M_2'$ are parallel or inclined, a change $\Delta m.\lambda_0$ of the optical path in either arm of the instrument results in a displacement of the pattern through $\Delta m$ orders. Displacements can be estimated visually to about 1/20 th order, but in certain circumstances displacements as small as 1/1000 order can be detected* by a method due to Kennedy. †

When the separation of $M_1$ and $M_2'$ is only a few wavelengths, fringes are visible with white light. They are used to recognize a reference fringe in the monochromatic pattern. If $|\phi| = \pi$, the central fringe of the white light pattern is black and defines the intersection of $M_1$ and $M_2'$, so that it is in the same position as the monochromatic fringe with $|m| = \frac{1}{2}$; otherwise the

0

• K. K. Illingworth, Phys. Rev. (2), 30 (1927), 692.
† R. J. Krennedy, Proc. Nat. Acad. Sci., 12 (1926), 621.

achromatic fringe does not in general coincide with a bright or dark fringe of the monochromatic pattern, but this presents no difficulty if the transfer between white light and monochromatic patterns is made consistently.

Except in modified forms using collimated illumination (§ 7.5.5), the Michelson interferometer is now obsolete, but it is famous because of its use by Michelson in three important experiments: the Michelson-Morley ether-drift experiment;* the first systematic study of the fine structure of spectral lines; and the first direct comparison of the wavelength of spectral lines with the standard metre. † In this book we are not concerned with the first of these experiments, since we confine our attention to the optics of stationary media; nor shall we describe the third, since more precise measurements have subsequently been made by other methods (§ 7.7). Michelson's method of analysing spectral lines has also been superseded by more direct methods, but because of its considerable theoretical interest and because of its connection with the theory of partial coherence we shall discuss it in detail later (§7.5.8).

## 9.9.5 7.5.5 The Twyman-Green and related interferometers

If a Michelson interferometer is illuminated by a point source $S$ of quasi-monochromatic light at the focus of a well-corrected lens $L_1$, and the light emerging from the interferometer is collected by a second well-corrected lens $L_2$, the arrangement becomes equivalent to the Fizenu interferometer, but with the beams having clearly separated paths (Fig. 7.40). Let $W_1$ be a plane wave-front in the beam returning from $M_1, W_2$ the corresponding plane wave-front in the beam returning from $M_2$; and let $W_1'$ be the virtual plane wave-front returning from $M_2$ which would emerge from the beam divider coincident and co-phasal with $W_1$. The difference of optical path between the emergent rays which intersect virtually at a point $P$ on $W_2$ is then

$$\Delta\mathscr{S} = nh, \tag{74}$$

where $h = PN$ is the normal distance from $W_1'$ to $P$, and $n$ is the refractive index of the medium between $W_1'$ and $W_2$. The corresponding phase difference is

$$\delta = \frac{2\pi}{\lambda_0} nh \tag{75}$$

and by (75) and §7.2 (16), an eye placed in the focal plane of $L_2$ and focused on $W_2$ (with the aid of an auxiliary lens if necessary) will see at $P$ a bright fringe if

$$nh = m\lambda_0, \quad |m| = 0, 1, 2, \ldots, \tag{76a}$$

and a dark fringe if

$$nh = m\lambda_0, \quad |m| = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{76b}$$

Thus there are in general straight line fringes, parallel to the apex of the wedge formed by $W_1'$ and $W_2$; if the latter are made parallel by suitably adjusting the orientation of $M_1$, the field of view is uniformly illuminated. If the source were a point, the fringes would be non-localized, but in practice, because of intensity considerations,

the extension of the source is not negligible. Since the paths of the emergent rays correspond to reflection from a wedge formed by $M_2$ and $M_1'$, where $M_1'$ is the virtual image of $M_1$ in the

0

• A. A. Michelson and E. W. Morley, Phil. Mag. (5), 24 (1887), 449.
† A. A. Michelson and J. R. Benoit, Trav. et Mem. Int. Bur. Poids et Mes., 11 (1895), 1.

beam divider, the fringes with an extended source are virtually localized in the vicinity of this wedge, as in the case of Frzeau fringes ( p .297 ); and the tolerable source extension is greatest when $M_1'$ and $M_2$ coincide. On the other hand, because of the departure from monochromatism of the source, the fringes are visible only if the optical paths in the two arms are sufficiently near to equality. It is important to note that the conditions of equality of optical paths, and coincidence

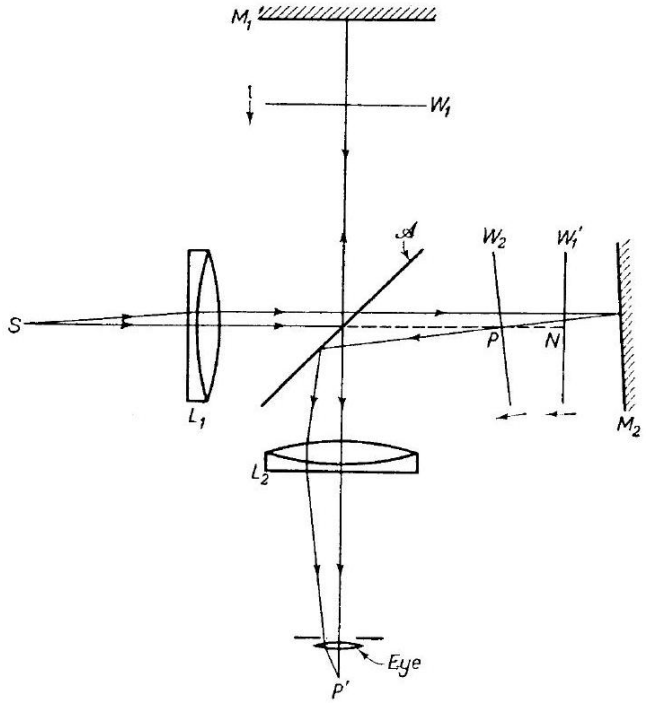



Fig. 7.40. The Michelson interferometer with collimated illumination.

of $M_1'$ and $M_2$, are not in general satisfied simultaneously if the arrangement is asymmetrical about $\mathscr{A}$.

This modification of the Michelson interferometer was introduced by Twyman and Green* for testing optical elements. The element is inserted in the arm $M_2$ in such a way that, if it were perfect, the returning wave-front $W_2$ would be plane. Since by (76) the bright fringes may be regarded as contours of $W_2$ defined by planes parallel to $W_1'$ at intervals of $\lambda_0$ (taking $n = 1$ for air), deformation of $W_2$ resulting from the double passage of light through the element may be measured. The sign of the deformation can be determined from the direction of motion of the fringes when the distance of $M_1$ from the beam divider is increased. The arrangement † for testing a prism at minimum deviation is shown in Fig. 7.41 (a). The fringes are observed coincident with one of the prism surfaces, and may be traced there by the observer as a guide to subsequent local polishing. In this way internal variations in refractive index of the prism material can be compensated. Fig. 7.41 (b) shows the arrangement used for testing a camera lens. ‡$M_2$ is a spherical convex mirror with its centre

of curvature at the focus of the lens $C$ to be tested. The latter can be rotated about a line perpendicular to its axis to allow tests at different obliquities, and a mechanical linkage ensures

0

- F. Twyman and A. Green, British Patent No. 103832 (1916).
  † F. Twyman, Phil. Mag. (6), 35 (1918), 49.
  ‡ F. Twyman, Phil. Mag. (6), 42 (1921), 777.

that the centre of curvature of $M_2$ remains in the focal plane as the lens is rotated. There is some uncertainty in relating defects shown by the fringes

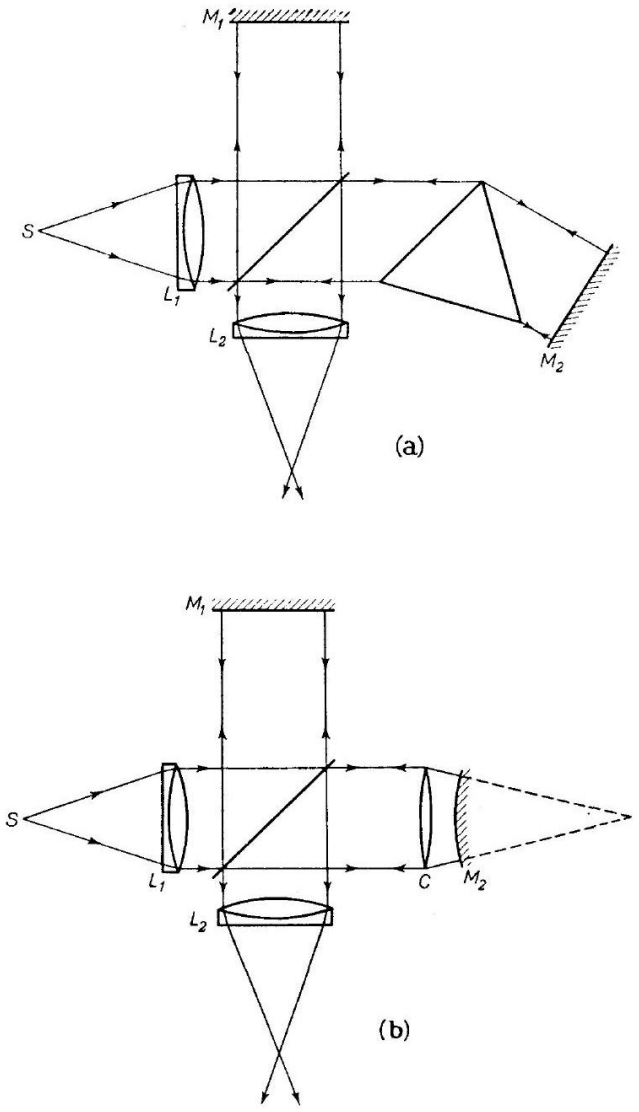



Fig. 7.41. The Twyman-Green interferometer.
(a) Arrangement for testing a prism.
(b) Arrangement for testing a camera lens.

to points of the lens aperture because, when the lens is imperfect, the outward and return paths of a ray through it do not coincide. This uncertainty is diminished by having the radius of curvature of $M_2$ as large as possible, and further by using a suitable optical system to observe the virtual fringe pattern on the surface of $M_2$. Photographs of such fringe patterns, together with the corresponding computed patterns, are shown in Fig. 7.42.

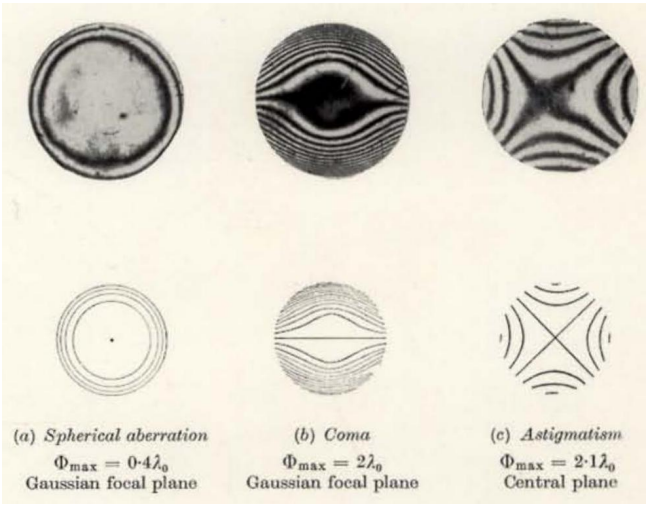



Fig. 7.42. Twyman-Green interference patterns from lenses showing primary aberrations. $\Phi_{\max }$ is the maximum wave aberration in the exit pupil. The patterns above are observed, those below are calculated.
(After R. Kingslake, Trans. Opt. Soc., London, 27 (1927), 94.)

The Twyman-Green arrangement is applied to the measurement of the length of end gauges in the Kösters interferometer* (Fig. 7.43 (a)). The gauge $G$ is wrung to the mirror $M_2$; and the mirror $M_1$ is positioned so that the equivalent virtual reflecting surface $M_1'$ is about midway between $M_2$ and the upper surface of $G$, and inclined to them at a suitable small angle. With a quasi-monochromatic source, parallel equidistant fringes are then visible over $M_2$ and $G$, running at right angles to one edge of $G$ (Fig. 7.43 (b)). The length of $G$ is found by the method of excess

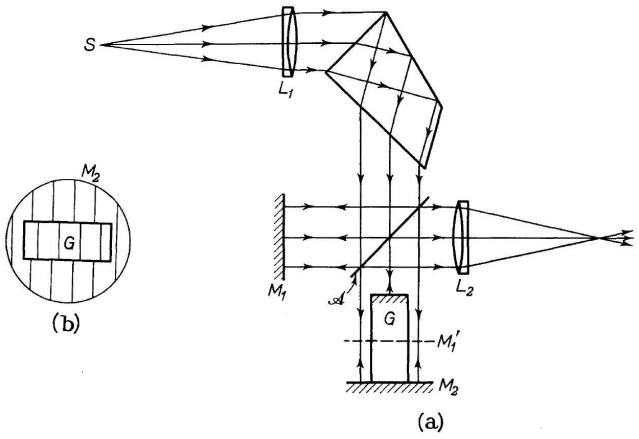



Fig. 7.43. The Kösters interferometer: (a) arrangement; (b) field of view.

fractions, as with the N.P.L. gauge measuring interferometer already described (§ 7.5.2). There is evidently a close resemblance between these two instruments. †

We may mention also a similar interferometer-the Dowell end-gauge interferometer, ‡ (Fig. 7.44 (a))-which enables the lengths of two end gauges to be compared without need for wringing contacts. Suppose $A_1', A_2'$ are the virtual images in $M_1$ of the end surfaces $A_1, A_2$ of the gauges $G_1, G_2$; and $B_1', B_2'$ are the virtual images in $M_2$ and the beam divider of the end surfaces $B_1, B_2$. It is arranged that the plane virtual reflecting surfaces $A_1', A_2', B_1', B_2'$ appear to overlap

(Fig. 7.44 (b)), and the interference effects may be considered to arise from the wedge-shaped air films between them. By suitably

adjusting $G_1$, a pattern (i) of horizontal fringes is obtained with monochromatic light over the region common to $A_1'$ and $B_1'$, and the zero order fringe, which can be recognized in white light, is brought to the centre of the field. $G_2$ is next adjusted until $A_2'$ is co-planar with $A_1'$. Then since $B_2'$ is parallel to $B_1'$, there are similar patterns (ii), (iii) of horizontal fringes over the regions common to $B_2'$ and $A_1', A_2'$, with zero orders in coincidence. In this condition, the distance $d$ between $B_2'$ and $B_1'$, which is the difference in

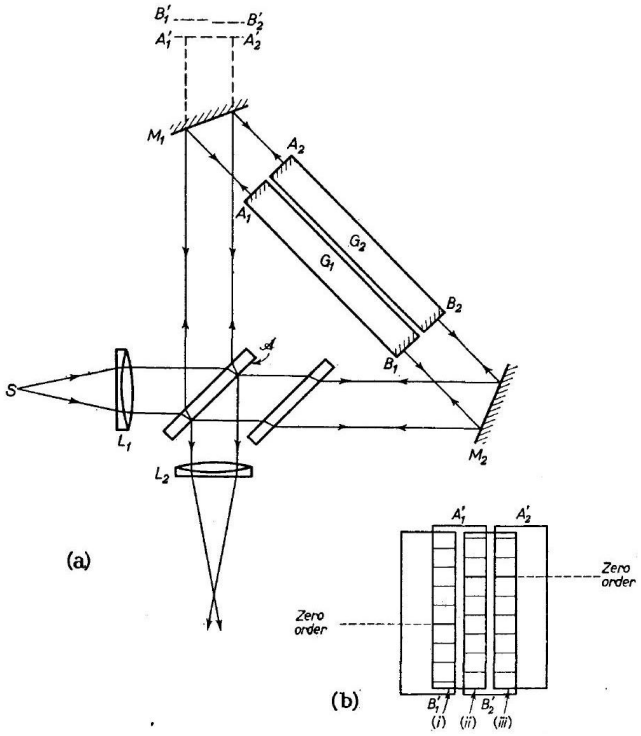



Fig. 7.44. The Dowell end-gauge interferometer: (a) arrangement, (b) field of view.

the lengths of the two gauges, is given by $d = \lambda_0 \Delta m/2n$, where $\Delta m$ is the displacement in orders of patterns (i) and (ii). The instrument may also be applied, using the method of excess fractions, to measure the length of a single gauge.

## 9.9.6 7.5.6 Fringes with two identical plates: the Jamin interferometer and interference microscopes

Suppose that light from a quasi-monochromatic point source $S$ is incident on two transparent plane parallel plates, each of thickness $h$ and refractive index $n'$, placed one behind the other

0

- W. Kösters, Handbuch der Physikalischen Optik, Vol. I (Ed. by E. Gehrcke, Leipzig, Barth, 1927), p. 484.
  † With both arrangements there is an upper limit to the separation of reflecting surfaces for which fringes of adequate visibility can be observed; this limit depends on the angular size of the source (§ 7.5.3), and its departure from monochromatism (§ 7.5.8). Since the Kösters arrangement allows the auxiliary reflecting surface $M_1'$ to be positioned midway between the ends of the gauge, it can be used with a given source to measure gauges about twice as long as the maximum with the N.P.L. instrument.
  ‡ J. H. Dowell, British Patent No. 555672, (1942).

and mutually inclined at a small angle $\alpha$ (Fig. 7.45). If we neglect rays which have suffered more than two reflections at the plate surfaces, there are two groups of parallel transmitted rays derived from an incident ray $SA$. The first group consists of the directly transmitted ray, and the rays reflected at both surfaces of either plate. The second group-SABDH, SABCFI, $SABDEJ$ and SABCFGK-consists of rays reflected at one surface of each plate; these rays traverse the gap between the plates three times, and are inclined at angle $2\alpha$ to the rays of the first group.

The rays of the second group may be re-combined at point $P$ in the focal plane of a lens $L$, and if we denote the optical paths between $S$ and $P$ along them by $\mathscr{S}_1, \mathscr{S}_2$, $\mathscr{S}_3$, and $\mathscr{S}_4$ respectively, we have, using §7.5 (5),

$$\mathscr{S}_2 - \mathscr{S}_1 = \Delta\mathscr{S}_{21} = 2n'h\cos\theta_1' \tag{77a}$$

$$\mathscr{S}_3 - \mathscr{S}_2 = \Delta\mathscr{S}_{32} = 2n'h\,(\cos\theta_2' - \cos\theta_1') \tag{77b}$$

$$\mathscr{S}_4 - \mathscr{S}_3 = \Delta\mathscr{S}_{43} = 2n'h\cos\theta_1' \tag{77f}$$

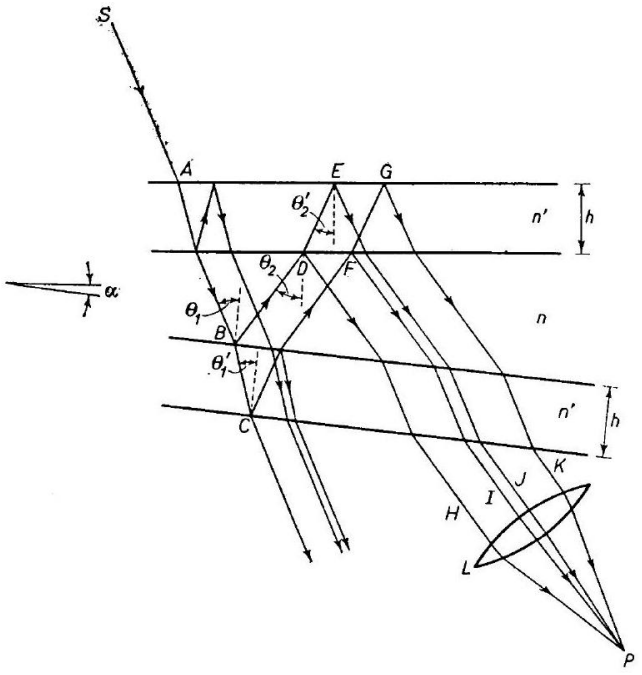



Fig. 7.45. Two identical plane parallel plates: illustrating formation of fringes localized at infinity.

where $\theta_1'$ and $\theta_2'$ are respectively the angles of refraction in the second plate at $B$ and the first plate at $D$. Since, for given $P$, the angles $\theta_1'$ and $\theta_2'$ depend only on the orientation of the plates, the optical path differences $\Delta\mathscr{S}$ are independent of the position of $S$, and an interference pattern is formed in the focal plane of $L$ with an extended source. Providing the range of angles of incidence is not too large, this pattern is not overlapped by light from rays of the first group. Further, since $\cos\theta_1' \sim \cos\theta_2'$, $\Delta\mathscr{S}_{21}$ and $\Delta\mathscr{S}_{43}$ are large compared with $\Delta\mathscr{S}_{32}$ when $h$ is sufficiently large, so that we may employ a source for which the condition §7.3 (15) for distinct fringes holds with respect to $\Delta\mathscr{S}_{32}$ but not with respect to $\Delta\mathscr{S}_{21}$ and $\Delta\mathscr{S}_{43}$. In these circumstances we may consider the fringe pattern to be associated only with rays such as $SABCFI$ and $SABDEJ$; the remainder of the light produces an effectively uniform background which merely reduces the visibility of the fringes.

From (77b), the phase difference at $P$ corresponding to $\Delta\mathscr{S}_{32}$ is

$$\delta = \frac{4\pi}{\lambda_0} n' h \left(\cos\theta_2' - \cos\theta_1'\right); \tag{78}$$

i.e. using the law of refraction,

$$\delta = \frac{4\pi h}{\lambda_0}\left(\sqrt{n'^2 - n^2\sin^2\theta_2} - \sqrt{n'^2 - n^2\sin^2\theta_1}\right) \tag{79}$$

where $\theta_1$ and $\theta_2$ are the angles of incidence at $B$ and $D$ respectively, and $n$ is the refractive index of the medium surrounding the plates. To find the form of the fringes, consider rectangular coordinate axes $OX, OY, OZ$, with origin $O$ at the second nodal point of the lens $L$, and $OZ$ parallel to the apex of the wedge between the plates (Fig. 7.46). Let $N_1$, in the plane $XOY$, be the focal point for light reflected normally at $B, (\theta_1 = 0)$; let $N_2$, also in the plane $XOY$, be the focal point for light reflected

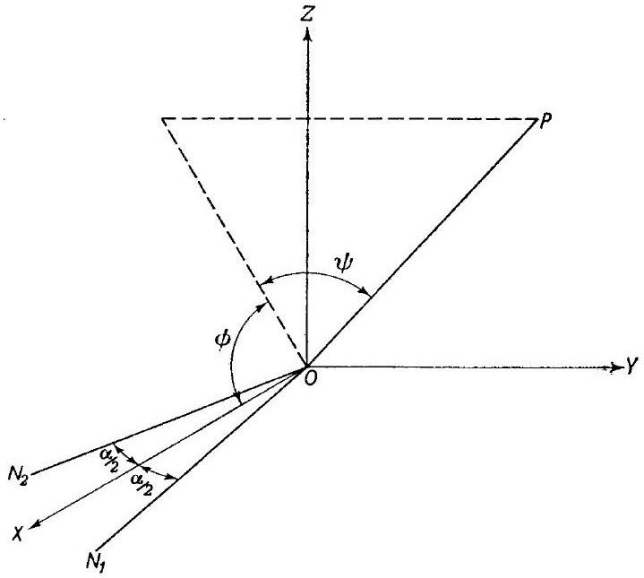



Fig. 7.46. Illustrating the discussion of fringes at infinity given by two mutually inclined plane parallel plates.

normally at $D\,(\theta_2 = 0)$; and let $OX$ be the internal bisector of $ON_1$ and $ON_2$. Then if the wedge opens out in the direction $OY$, the direction cosines of $ON_1$ and $ON_2$ are respectively $\{\cos(\alpha/2), \sin(\alpha/2), 0\}, \{\cos(\alpha/2), -\sin(\alpha/2), 0\}$. Let $OP$ make angle $\psi$ with the plane $XOZ$, and let the projection of $OP$ on this plane make angle $\phi$ with $OX$. The direction cosines of $OP$ are ( $\cos\psi\cos\phi, \sin\psi, \cos\psi\sin\phi$ ), and since the angles $PON_1 = \theta_1, PON_2 = \theta_2$, we have

$$\cos\theta_1 = \cos\frac{\alpha}{2}\cos\psi\cos\phi + \sin\frac{\alpha}{2}\sin\psi \tag{80a}$$

$$\cos\theta_2 = \cos\frac{\alpha}{2}\cos\psi\cos\phi - \sin\frac{\alpha}{2}\sin\psi \tag{80b}$$

For $\alpha$ small we may neglect terms involving the second and higher powers of $\alpha$ in the expansion of $\cos(\alpha/2)$ and $\sin(\alpha/2)$, so that from (80),

$$\cos^2\theta_1 = \cos^2\psi\cos^2\phi + \alpha\sin\psi\cos\psi\cos\phi, \tag{81a}$$

$$\cos^2\theta_2 = \cos^2\psi\cos^2\phi - \alpha\sin\psi\cos\psi\cos\phi. \tag{81b}$$

Then from (79) and (81),

$$\delta = \frac{4\pi h}{\lambda_0}\left\{\sqrt{n'^2 - n^2\left(1 - \cos^2\psi\cos^2\phi + \alpha\sin\psi\cos\psi\cos\phi\right)}\right.$$
$$\left. - \sqrt{n'^2 - n^2\left(1 - \cos^2\psi\cos^2\phi - \alpha\sin\psi\cos\psi\c\right.}\right. \tag{82}$$

and after expansion in powers of $\alpha$, and again neglecting terms in powers of $\alpha$ higher than the first, we obtain

$$\delta = -\frac{4\pi h}{\lambda_0}\frac{n^2\alpha\sin\psi\cos\psi\cos\phi}{\sqrt{n'^2 - n^2\left(1 - \cos^2\psi\cos^2\phi\right)}} \tag{83}$$

Thus according to (83) and §7.2(16), there is a bright fringe at $P$ when

$$\frac{\sin\psi\cos\psi\cos\phi}{\sqrt{n'^2 - n^2\left(1 - \cos^2\psi\cos^2\phi\right)}} = \frac{m\lambda_0}{2n^2h\alpha}, \quad |m| = 0, 1, 2, \ldots \tag{84a}$$

and a dark fringe when

$$\frac{\sin\psi\cos\psi\cos\phi}{\sqrt{n'^2 - n^2\left(1 - \cos^2\psi\cos^2\phi\right)}} = \frac{m\lambda_0}{2n^2h\alpha}, \quad |m| = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{84b}$$

When $P$ is close to the plane $XOZ$, so that $\psi$ is small, the order of interference $m$ is low, and in this region fringes can be observed in white light. The central white fringe ( $m = 0$ ) lies in the plane $XOZ$ and is formed by light for which $\theta_1 = \theta_2$.

In the special case of observation near normal incidence, $\psi$ and $\phi$ are both small, and neglecting terms of the second and higher powers in $\psi$ and $\phi$, (84a) reduces to

$$\psi = \frac{mn'\lambda_0}{2n^2h\alpha}, \quad |m| = 0, 1, 2, \ldots \tag{85}$$

The fringes in quasi-monochromatic light are thus equidistant straight lines, parallel to the apex of the wedge between the plates. The angular separation of adjacent bright fringes is proportional to the refractive index of the plates, and inversely proportional to their thickness and the angle between them. Such fringes were first observed by Brewster, and are called Brewster's fringes.

Similar fringes are used in the Jamin interferometer,* which was at one time widely used for measurements of the refractive index of gases, though it is now superseded by the Rayleigh interferometer (§7.3.5). The instrument consists essentially of two equally thick plane parallel glass plates of the same refractive index, opaquely silvered on the surfaces $M_1, M_2$, and arranged as shown in Fig. 7.47. A beam of light from an extended source, incident on one plate at about 45°, gives rise to two beams, one reflected from the front surface of the first plate and the rear surface of the second, the other reflected from the rear surface of the first plate and the front surface of the second; and the two are recombined to give an interference pattern in the focal plane of the telescope $T$. The thickness of the plates is such that, when suitably diaphragmed, the beams are clearly separated between the plates, and here are inserted gas chambers $G_1, G_2$ and a compensator $C_1, C_2$ similar to that described in § 7.3.5. In use, the plates are inclined so that they make a wedge of small angle with apex parallel to the plane of Fig. 7.47, which we take to be horizontal. The plate surfaces are approximately vertical, so that, with the notation of Fig. 7.46, the plane $XOZ$ is

0

* J. Jamin, C. R. Acad. Sci. Paris, 42 (1856), 482.

approximately horizontal; the region of observation then corresponds to $\psi$ small, and from (84a) the bright fringes are given by

$$\frac{\cos\phi}{\sqrt{n'^2 - n^2\sin^2\phi}}\psi = \frac{m\lambda_0}{2n^2h\alpha}, \quad |m| = 0, 1, 2, \ldots \tag{86}$$

Now $\phi \sim 45°$, and over a small angular field the variation of the term in $\phi$ is negligible. The fringes therefore follow loci of constant $\psi$, i.e. they are horizontal, and are equidistant. They are of low order, and by suitable adjustment the zero order fringe

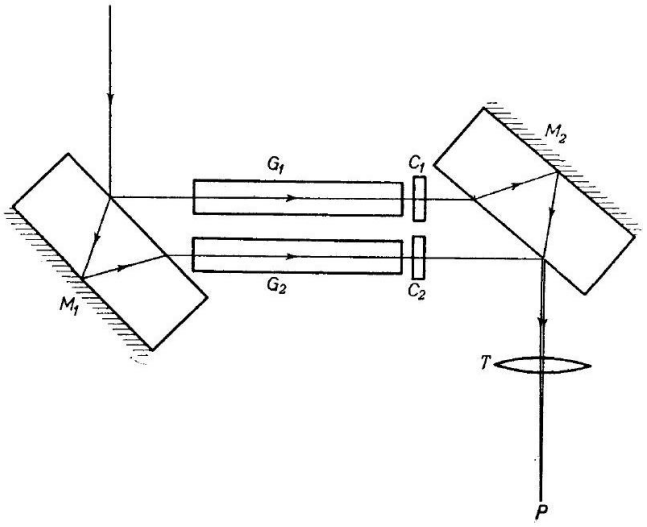



Fig. 7.47. The Jamin interferometer.

( $\psi = 0$ ), recognizable in white light, can be brought to the centre of the field of view.*

The interferometer is used for measurements in a manner similar to that already described for the Rayleigh interferometer, except that there is no second system of fringes to serve as a fiducial mark, and settings are made on a cross wire in the telescope. In consequence the instrument is more sensitive to disturbances of the optical system than the Rayleigh instrument, and the precision of measurement is lower.

A variant of the Jamin interferometer, using plates which are slightly wedgeshaped instead of plane parallel, was developed by Sirks, † and later by Pringsheim, ‡ for the measurement of the refractive index of small objects. The plates are set with the wedge apexes anti-parallel and the inner unsilvered surfaces approximately

parallel (Fig. 7.48), and are illuminated with collimated light. Corresponding to an incident ray $SA$ in a principal section of the wedges, the two rays $SABCG, SADEF$ which leave the second plate intersect virtually at a point $P$ behind the second plate; and with a quasi-monochromatic source of not too large extension there are fringes apparently localized in the

0

• If the plates are inclined so that they form a wedge with apex vertical, the plane $XOY$ is horizontal. The region of observation then corresponds to $\phi$ small, and the intensity maxima are given by (84a) as

$$\frac{\cos\psi\sin\psi}{\sqrt{n'^2 - n^2\sin^2\psi}} = \frac{m\lambda_0}{2n^2h\alpha}, \quad |m| = 0, 1, 2, \ldots$$

The fringes follow loci of constant $\psi$, which now means that they are vertical, but with $\psi \sim 45°$ the order of interference is not near zero. White light fringes cannot be obtained with this orientation of the plates.

† J. A. Sirks, Hd. Ned. Nat. en Geneesk. Congr., Groningen (1893), p. 92.

‡ E. Pringsheim, Verh. Phys. Ges. Berlin, 17 (1898), 152.

vicinity of $P$, which can be observed with the microscope $M$. The fringes at $P$ run at right angles to the plane defined by the two emergent rays, i.e. parallel to the wedge apexes. The object $O$ to be examined is placed between the plates in the path of the ray $CG$. The image $P'$ of $P$ in the front surface of the second plate also lies on $CG$, at a position which depends on the inclination of the plates. According to Pringsheim (loc. cit.), $P'$ is approximately midway between

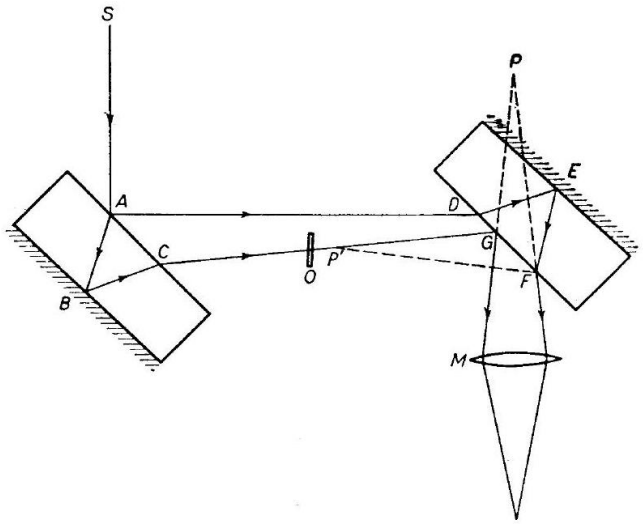



Fig. 7.48. The Sirks-Pringsheim interferometer.

the plates when the inner plate surfaces are parallel, but a small rotation of either plate about an axis parallel to the wedge apexes results in an appreciable displacement of $P'$. By means of such an adjustment, $P'$ can be made to lie in $O$, and the object and fringe system then appear superposed in the field of the microscope. The change in the order of interference at $P'$ which results from the introduction of the object can be determined by means of a compensator, and hence the refractive index of the object at $P'$ may be found if its thickness is known.

More recently, a similar combination of microscope and interferometer has been developed by Dyson* (Fig. 7.49). The object to be examined, mounted on a glass slide, is placed at $O$ between two identical glass plates $G_1, G_2$ which are plane wedges of small angle with their apexes anti-parallel. The upper surface of the lower plate $G_1$ is coated with a partially transparent silver film, and its lower surface has a small opaquely silvered central spot $C$ somewhat larger than the field of view of the microscope. The upper plate $G_2$ is coated on each side with a partially transparent silver film. Spaces between the object slide and the glass plates are filled with medium of the same refractive index as the glass. The arrangement is illuminated by light from the microscope condenser converging to an image of the source in the plane of $O$. Part of this light-the object beam-passes through $O$ and emerges from $G_2$ after reflection at the upper and lower surfaces. Another part-the reference beam-is reflected at

the upper surface of $G_1$ and converges to $C$; from here it is reflected, and after passing outside $O$, which lies in the shadow of $C$, it is directly transmitted through $G_2$. A glass block, with spherical upper surface $R$ opaquely silvered except for a small clear aperture $A$ on the axis, is cemented to $G_2$ and is so designed that, after reflection at $R$ and the upper surface of $G_2$, the object and reference beams converge to $A$. Near $A$ the reference beam then forms a real image $\sigma_1$ of the source, and the object beam forms a real image $\sigma_2$ of the source with a

0

• J. Dyson, Proc. Roy. Soc., A, 204 (1950), 170.

real image Π of the object plane superposed. The latter is observed by means of a conventional microscope.*

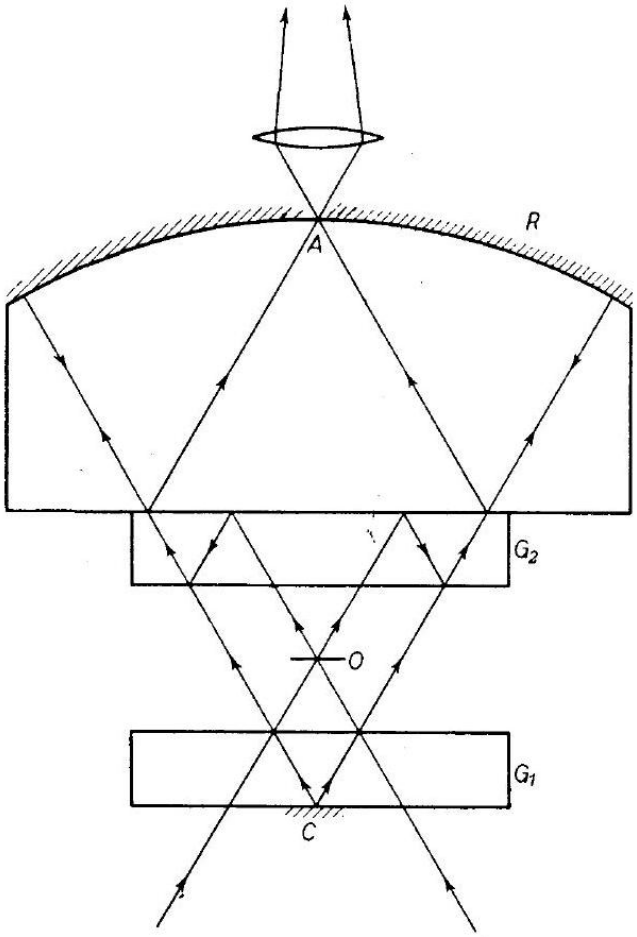



Fig. 7.49. The Dyson interferometer microscope.

Corresponding points of $\sigma_1$ and $\sigma_2$ are images of the same point of the source, and so act as mutually coherent secondary sources. The arrangement is adjusted so that such points coincide. In the absence of an object, the optical paths of the object and reference beams to any point $P$ of Π are then equal, and if we neglect small phase differences due to differing conditions of reflection at silvered surfaces, the order of interference is zero over all Π-a condition that can be recognized in white light. When an object is introduced at $O$, the optical path of the object beam to $P$ is increased by approximately $(n' - n)\, l$, where $n'$ and $l$ are respectively the refractive index and thickness of the object at the point $P'$ conjugate to $P$, and $n$ is the refractive index of the surrounding medium. Variations of $n'$ and $l$ over the object thus give rise to variations of intensity over Π in quasi-monochromatic light, and to variations of colour in white light. If the plate $G_2$ is moved in a direction normal to the wedge vertex and the optical axis, the optical path of the reference beam is altered, and by this means the change of optical path of the object beam can be compensated. From calibration of this movement of $G_2$ with monochromatic light, $(n' - n)\, l$ can be determined, giving $n'$ if $n$ and $l$ are known. †

## 9.9.7 7.5.7 The Mach-Zehnder interferometer; the Bates wave-front shearing interferometer

In the Jamin interferometer (§7.5.6), the front surfaces of the two plates act as beam-dividers and the rear surfaces as plane mirrors, but these elements cannot be

adjusted independently, and the separation of the two beams is limited by the thickness of the plates. A much more versatile instrument, in which the beams may be widely separated, is obtained when the beam dividers and mirrors are separate elements. This is the basis of the Mach-Zehnder interferometer,* which is used to measure variations of refractive index, and hence of density, in compressible gas flows.

The arrangement is shown in Fig. 7.50. Light from a source $S$ in the focal plane of a well-corrected lens $L_1$ is divided at the semi-reflecting surface $\mathscr{A}_1$ of a plane parallel glass plate $D_1$ into two beams, which, after reflection at plane mirrors $M_1, M_2$, are recombined at the semi-reflecting surface $\mathscr{A}_2$ of a second identical plane parallel plate $D_2$, and emerge to a well-corrected collecting lens $L_2$. The four reflecting surfaces are usually arranged to be approximately parallel, with their centres at the corners of a parallelogram. Suppose the source is a point source of quasi-monochromatic light. Let $W_1$ be a plane wave-front in the beam between $M_1$ and $D_2, W_2$ the corresponding plane wave-front in the beam between $M_2$ and $D_2$, and $W_1'$ the virtual plane wave-front between $M_2$ and $D_2$ which would emerge from $D_2$ coincident and co-phasal with $W_1$. At a point $P$ on $W_2$, the virtual phase difference between the emergent beams is then

$$\delta = \frac{2\pi}{\lambda_0} nh, \tag{87}$$

where $h = PN$ is the normal distance from $P$ to $W_1'$, and $n$ is the refractive index of the medium between $W_2$ and $W_1'$. At the point $P'$ in the emergent beams, conjugate to $P$, there will by §7.2 (16) be a bright fringe if

$$nh = m\lambda_0, \quad |m| = 0, 1, 2, \ldots, \tag{88a}$$

and a dark fringe if

$$nh = m\lambda_0, \quad |m| = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots \tag{88b}$$

When $W_1'$ and $W_2$ are parallel, the intensity is the same for all $P$, and under these circumstances an extended source would give fringes at infinity (i.e. in the focal plane of $L_2$ ) similar to those of the Jamin interferometer. In general, however, $W_1'$ and $W_2$ are mutually inclined, and the fringes are straight lines parallel to their intersection. It is these wedge fringes that are normally used in the examination of gas flows, and because of intensity considerations it is desirable to form them with the largest source extension possible without loss of visibility. As we have seen in § 7.5.3, the fringes then become localized in the region where intersecting rays have the smallest angular separation on leaving $S$. The position of this region of localization

0

- For beams other than those mentioned, $A$ is completely covered by the shadow of $C$, so that light from these beams does not enter the microscope.
† In practice, the adjustment for coincidence of $\sigma_1$ and $\sigma_2$ is never perfect. Further, the introduction of a refracting object, and the use of $G_1$ as a compensator, each results in a mutual displacement of $\sigma_1$ and $\sigma_2$ parallel to the optical axis. A satisfactory account of the interference effects in the region of $\sigma_1$ and $\sigma_2$ when corresponding points of these images do not coincide is beyond the scope of the elementary treatment of this chapter; it may be treated elegantly by the theory of partial coherence (Chapter X).

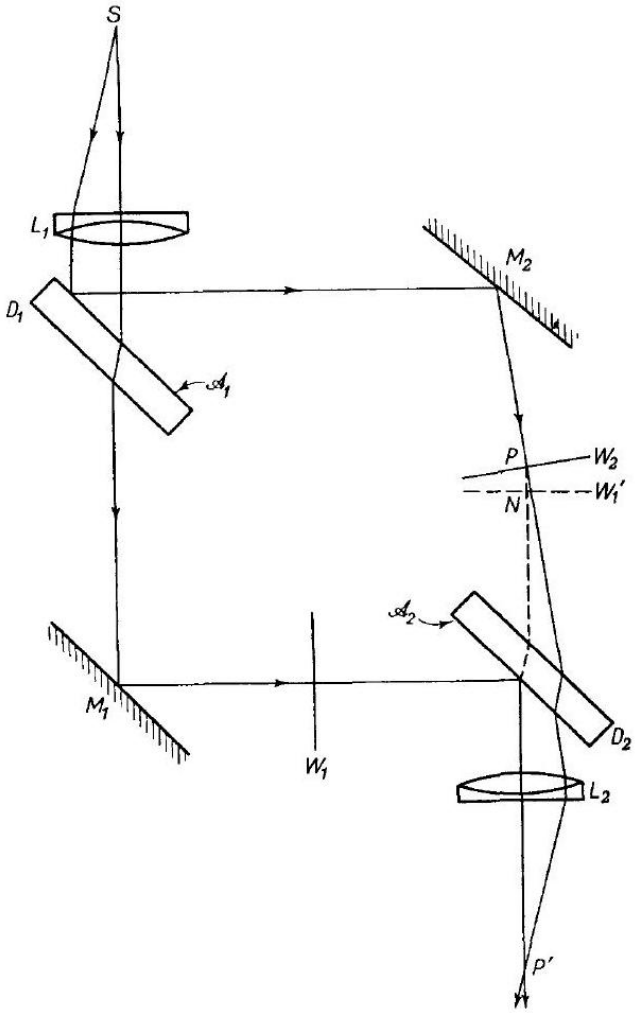



Fig. 7.50. The Mach-Zehnder interferometer.

can be varied by varying the combination of rotations of the elements used to produce
$nh = m\lambda_0, \quad |m| = 0, 1, 2, \ldots,$
and a dark fringe if
$nh = m\lambda_0, \quad |m| = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$
When $W_1'$ and $W_2$ are parallel, the intensity is the same for all $P$, and under these circumstances an extended source would give fringes at infinity (i.e. in the focal plane of $L_2$ ) similar to those of the $J_{\text{AMIN}}$ interferometer. In general, however, $W_1'$ and $W_2$ are mutually inclined, and the fringes are straight lines parallel to their intersection. It is these wedge fringes that are normally used in the examination of gas flows, and because of intensity considerations it is desirable to form them with the largest source extension possible without loss of visibility. As we have seen in § 7.5.3, the fringes then become localized in the region where intersecting rays have the smallest angular separation on leaving $S$. The position of this region of localization can be varied by varying the combination of rotations of the elements used to produce

the mutual inclination of $W_1'$ and $W_2$. For example, if the reflecting surfaces are initially parallel, and if for simplicity we consider the case of rotations about axes perpendicular to the plane of centres, the virtual region of localization is near $M_2$ when $M_2$ is rotated (Fig. 7.51

0

• L. Zehnder, Zeitschr. f. Instrkde, 11 (1891), 275. L. Mach, Zeitschr. f. Instrkde, 12 (1892), 89.

(a)), but lies between $M_2$ and $D_2$ when both $M_2$ and $D_2$ are rotated (Fig. 7.51 (b)). This property distinguishes the wedge fringes of the MachZehnder interferometer from those given by the Michelson interferometer with collimated light (§7.5.5), which are virtually localized in the vicinity of the mirrors.

In the technical use of the instrument, the region $C_1$ where the gas flow is to be examined-commonly the working section of a wind tunnel or shock-wave tubeand a compensating chamber $C_2$, are in opposite arms of the interferometer, which is

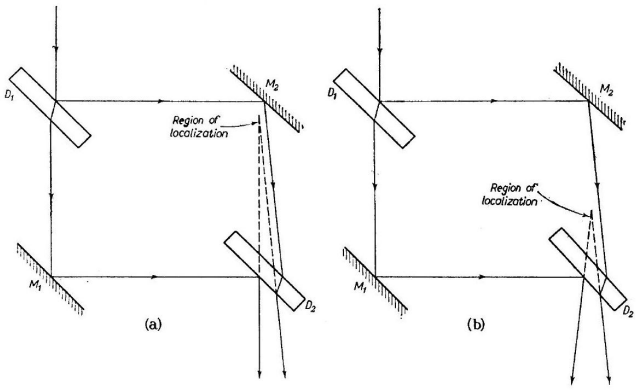



Fig. 7.51. Illustrating fringe localization in the Mach-Zehnder interferometer.

adjusted to give fringes near zero order, of desired orientation and spacing, and virtually localized near a chosen plane inside $C_1$, normal to the direction of the incident light. This plane is imaged on a photographic plate by means of $L_2$ and a highly-corrected camera lens. Photographs of the fringe pattern are obtained with and without gas flow, and the order displacement $\Delta m$ of the two patterns at selected points $P'$ of the image plane is measured, white light fringes being used if necessary to identify corresponding orders. If $n$ is the refractive index of the undisturbed gas in $C_1$, and $n'$ the refractive index under flow conditions, we have

$$\Delta m = \frac{1}{\lambda_0} \int (n' - n)\, ds \tag{89}$$

where the integration is taken along the path of the ray passing through $C_1$ and reaching $P'$. Let $Ox, Oy, Oz$ be rectangular coordinate axes in $C_1$, with origin $O$ in the chosen plane in $C_1$ and $OZ$ in the direction of the incident light; and let $P$, coordinates $(x, y, 0)$, be the point conjugate to $P'$ when there is no flow in $C_1$. Then if refractive deviations of rays caused by the flow are negligible, (89) may be written

$$\Delta m(x, y) = \frac{1}{\lambda_0} \int_0^z \{n'(x, y, z) - n\}\, dz, \tag{90}$$

where $s$ is the length of $C_1$; and when the flow satisfies certain symmetry conditions* (90) may be solved for ( $n' - n$ ) in terms of the measured $\Delta m$. The density change $\Delta\rho$ due to the flow may then be determined since $\Delta\rho$ is proportional to ( $n' - n$ ).

In a modified form due to Bates, † the Mach-Zehnder arrangement may also be used to measure the asphericity of convergent wave-fronts without the need for a substantially error-free reference wave-front, such as is required for the Twyman-Green method (p. 303). The arrangement, which is particularly advantageous for testing the performance of large aperture systems, is shown in Fig. 7.52 (a). The convergent

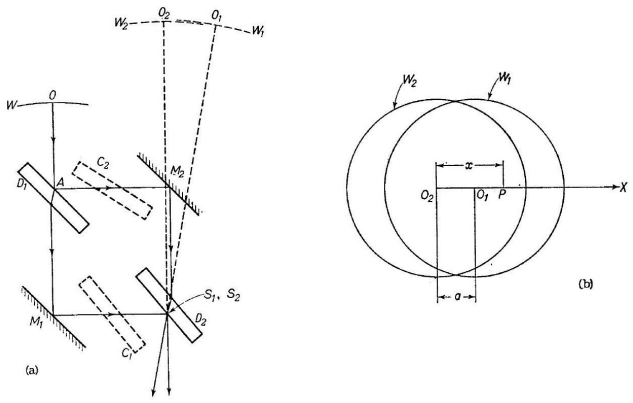



Fig. 7.52. The Bates wave-front shearing interferometer: (a) arrangement, (b) field of viow with sheared wave-fronts.

beam to be tested, with principal axis $OA$ which we take to be horizontal, is divided at $D_1$ into two beams converging to images $S_1, S_2$ of a sufficiently small quasi-monochromatic source. Initially the four reflecting surfaces are vertical and parallel, and positioned so that $S_1$ and $S_2$ coincide on the semi-reflecting surface of $D_2$. Corresponding to an incident wave-front $W$, the virtual emergent wave-fronts $W_1$ and $W_2$, with principal axes $O_1S_1$ and $O_2S_2$, are then exactly superposed, and an eye placed behind $D_2$ will see the field of view uniformly illuminated. $D_1$ and $M_1$ are now rotated as a whole about an axis parallel to $OA$, so that $O_1S_1$ and $O_2S_2$ are given a small vertical separation. This is substantially equivalent to tilting $W_2$ relative to $W_1$ about a horizontal axis, and the field is crossed by equidistant horizontal fringes, which are visible in white light.

Suppose now that $D_2$ is rotated about a vertical axis through $S_1$ and $S_2$; $O_1S_1$ then rotates about $S_1$ in a horizontal plane, i.e. $W_1$ is sheared relative to $W_2$. When $W_1$ and $W_2$ are perfectly spherical, the fringe pattern observed in the region where they

overlap is unaffected by the shear, but otherwise the fringes are displaced by an amount which depends on the asphericity of $W$. Thus if $O_2X$ is a coordinate axis in the shear direction, with origin $O_2$, the order displacement $\Delta m(x;a)$ at the point $P$, coordinate $x$, is

$$\Delta m(x;a) = \frac{1}{\lambda_0}\{\mathscr{S}(x) - \mathscr{S}(x-a)\}, \tag{91}$$

where $\mathscr{S}$ is the optical path between $W_2$ and a sphere with centre $S_2$ and radius $O_2S_2$, and $a$ is the shear distance (Fig. 7.52 (b)). From (91), since $\mathscr{S}$ is zero at $O_2$ ( $x = 0$ ),

$$\Delta m(a;a) = \frac{1}{\lambda_0}\mathscr{S}(a); \tag{92a}$$

and similarly

$$\Delta m(2a;a) = \frac{1}{\lambda_0}\{\mathscr{S}(2a) - \mathscr{S}(a)\},$$

so that

0

• See, for example, R. Ladenburg and D. Bershader, Interferometry in High Speed Aerodynamics and Jet Propulsion, Vol. IX, Physical Measurements in Gas Dynamics and Combustion (London, Oxford University Press, 1955), Article A.3.
† W. J. Bates, Proc. Phys. Soc., 59 (1947), 940.

$$\Delta m(a;a) + \Delta m(2a;a) = \frac{1}{\lambda_0}\mathscr{S}(2a). \tag{92b}$$

In a similar way we may find expressions $\mathscr{S}(3a), \mathscr{S}(4a)\ldots$ etc., and we see that $\mathscr{S}(x)$ can be determined at intervals of $a$ from measurements of $\Delta m$. Alternatively, when $a$ is not too large, we have from (91)

$$\Delta m(x;a) \sim \frac{a}{\lambda_0}\frac{d\mathscr{S}(x)}{dx} \tag{93}$$

showing that $\Delta m$ is proportional to the angular aberration $d\mathscr{S}/dx$ of the ray leaving $P$. Evidently when there is no symmetry of revolution the whole wave-front may in principle be examined by varying the shear direction.

When the wave-fronts are sheared, the emergent rays which intersect virtually at $P$ traverse the beam dividers at different angles, and if the fringe displacement on shearing is to depend only on the asphericity, this difference must be compensated. For this purpose, two compensating plates, identical to the plates used for the beam dividers, are introduced into the interferometer arms. One of them, $C_1$, is fixed parallel to $D_2$ and rotates with it during shearing; the other, $C_2$, is connected to $D_2$ by a mechanical linkage, which causes it to rotate at twice the rate of $D_2$ in the opposite sense. More recently Drew* has developed a simpler form of the interferometer in which compensating plates are unnecessary.

## 9.9.8 7.5.8 The coherence length; the application of two-beam interference to the study of the fine structure of spectral lines

When a gas, for example cadmium vapour, is excited under suitable conditions by an electrical discharge, it emits light whose spectrum consists of sharp bright lines separated by dark regions-a so-called emission line spectrum. If the light of one of these lines is isolated and used to illuminate, for example, a Michelson interferometer adjusted to give circular fringes, it is found that the fringes are distinct if the optical paths of the two interfering beams are nearly equal; but that as the difference of optical path is increased, the visibility of the fringes decreases (in general not monotonically), and they eventually disappear.

We can account for this disappearance of the fringes by supposing that the light of the spectral line is not strictly monochromatic, but is made up of wave trains of finite length, of which a large number pass at random time intervals during the time required to make an observation. Let us assume for the moment that all these wave trains are identical. Each one entering the interferometer is divided into two trains of equal length; and when the difference of optical path in the arms of the interferometer is greater than this length, one of these two wave trains has passed the point of observation $P$ before the other arrives. There is then no interference at $P$ due to pairs of wave trains derived from the same incident wave train; the wave trains superposed at $P$ at any instant are derived from different incident wave trains, and because these arrive at random, and in rapid succession, their contributions to the interference term average to zero over the relatively long time required to make an observation.

We can put this explanation into another form, which is mathematically more convenient for describing the variation of fringe visibility with path difference, by using the method of

0

- R. L. Drew, Proc. Phys. Soc. B, 64 (1951), 1005.

Applications of shearing interferometry are discussed in an article by O. Bryngdahl in Progress in Optics, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 37.

Fourier integral analysis. Let $F(t)$ be the light disturbance at a point at time $t$, due to a single wave train. We assume $F$ to be zero for $|t| \geqslant t_0$, and express it as a Fouring integral

$$F(t) = \int_{-\infty}^{\infty} f(v) e^{-2\pi i v t} dv, \tag{94}$$

where by the Fourier inversion theorem,

$$f(v) = \int_{-\infty}^{\infty} F(t) e^{2\pi i v t} dt \tag{95}$$

If $N$ such wave trains pass the point during the time required to make an observation, the total light disturbance involved in the observation may be written

$$V(t) = \sum_{n=1}^{N} F\left(t - t_n\right), \tag{96}$$

where the $t_n$ 's denote the times of arrival of the wave trains. The light intensity averaged over the time interval $2T$ needed to make an observation is

$$I = \frac{1}{2T} \int_{-T}^{T} |V(t)|^2 dt \sim \frac{1}{2T} \int_{-\infty}^{+\infty} |V(t)|^2 dt \tag{97}$$

if $T$ is large compared to the half-duration $t_0$ of each wave train. Now from (94) and (96),

$$V(t) = \int_{-\infty}^{+\infty} v(v) e^{-2\pi i v t} dv \tag{98}$$

where

$$v(v) = f(v) \sum_{n=1}^{N} e^{2\pi i v t_n}; \tag{99}$$

so that by Parseval's theorem,

$$\int_{-\infty}^{+\infty} |V(t)|^2 dt = \int_{-\infty}^{+\infty} |v(v)|^2 dv = \int_{-\infty}^{+\infty} |f(v)|^2 \sum_{n=1}^{N} \sum_{m=1}^{N} e^{2\pi i v (t_n - t_m)} dv. \tag{100}$$

We have

$$\begin{aligned} \sum_{n=1}^{N} \sum_{m=1}^{N} e^{2\pi i v (t_n - t_m)} &= N + \sum_{n \neq m} e^{2\pi i v (t_n - t_m} \\ &= N + 2 \sum_{n<m} \cos 2\pi v \left(t_n - t_m\right). \end{aligned} \tag{101}$$

Now since the $t_n$ 's are distributed at random, there is an equal likelihood for each cosine term to be positive or negative. Hence the average value of the double sum in (100) in a large number of similar experiments is $N$, and it follows from (97) and (100) that the mean intensity is given by

$$I = \frac{N}{2T} \int_{-\infty}^{+\infty} |f(v)|^2 dv \tag{102}$$

i.e. it is proportional to the integral of the intensities $i(\nu) = |f(\nu)|^2$ (incoherent superposition) of the monochromatic components of which a single wave train is made up.* With the

interferometer, each monochromatic component produces an interference pattern as described in §7.5.4, and as the path difference is increased from zero, these component patterns show increasing mutual displacement because of the difference of wavelength. The visibility of the fringes therefore decreases, and they disappear altogether when the optical path difference is sufficiently large.

These two ways of interpreting the absence of fringes at sufficiently large path differences-in terms of either a random succession of finite wave trains, or a superposition of monochromatic components distributed over a range of frequency-are for most practical purposes equivalent; and from the discussion above we must expect that the longer the wave trains, the narrower the frequency range over which the Fourier components have appreciable intensity. We may illustrate this relationship by a simple example. Suppose the wave trains are all of duration $\Delta t$, during which $F(t)$ is simply periodic with frequency $v_0$, i.e.

$$\left.\begin{array}{rll} F(t) & = f_0 e^{-2\pi i v_0 t} & \text{when} \quad |t| \leqslant \frac{\Delta t}{2}, \\ & = 0 \quad \text{when} & |t| > \frac{\Delta t}{2}, \end{array}\right\} \tag{103}$$

where $f_0$ is constant. Then from (95) and (103),

$$\begin{aligned} f(\nu) &= f_0 \int_{-\frac{\Delta t}{2}}^{\frac{\Delta t}{2}} e^{2\pi i(\nu-\nu_0)t} dt \\ &= f_0 \Delta t \left[\frac{\sin\left\{\pi\left(\nu-\nu_0\right)\Delta t\right\}}{\pi\left(\nu-\nu_0\right)\Delta t}\right]. \end{aligned} \tag{104}$$

The function $\left[\sin\left\{\pi\left(\nu-\nu_0\right)\Delta t\right\}/\pi\left(\nu-\nu_0\right)\Delta t\right]^2$, which governs the intensity distribution of the Fourier components of (103), is shown in Fig. 7.53. The frequency interval

$\nu_0 - \Delta\nu/2 \leqslant \nu \leqslant \nu_0 + \Delta\nu/2$ over which the intensity may be said to be appreciable is somewhat arbitrary, but since the first zero (which occurs when the argument of the sine term is equal to $\pi$ ) corresponds to $v - v_0 = \pm 1/\Delta t$, it is clear that

$$\Delta v \sim \frac{1}{\Delta t}. \tag{105}$$

Thus the effective frequency range of the Fourier spectrum is of the order of the reciprocal of the duration of a single wave train.

This example, in which the wave trains are all identical and of simple form, is only an idealization of the light from real sources. According to atomic theory, the loss of energy by atoms during emission results in damping of the wave trains. Further, the

0

- A rigorous formulation of the result is given by a theorem due to N. Campbell, well known in the analysis of random noise, especially in connection with the shot effect (fluctuations in intensity of a stream of electrons in vacuum tubes). Cf. S. O. Rice, Bell Tech. J., 23 (1944), 282. [Reprinted in Selected Papers on Noise and Stochastic Processes (ed. N. Wax, New York, Dover Publications, 1954), 133.]

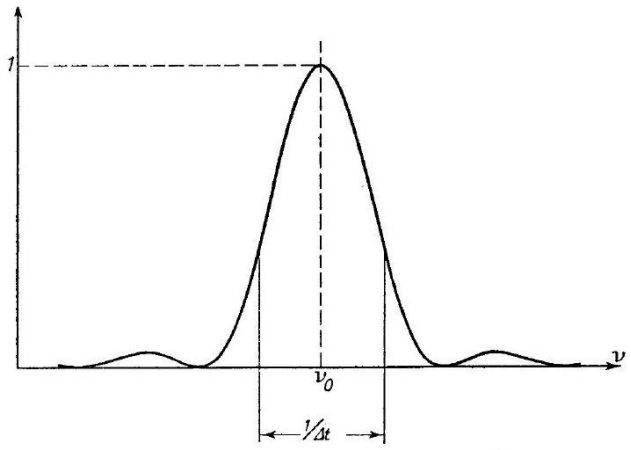



Fig. 7.53. The function $\left[\frac{\sin \left{\pi\left(\nu-\nu_{0}\right) \Delta t\right}}{\pi\left(\nu-\nu_{0}\right) \Delta t}\right]\hat{}{2}$.

atoms are in random thermal motion relative to the observer, so that observed spectra are modified by the Doppler effect. Again, emitting atoms are disturbed by their neighbours, with the result that the wave trains are irregularly modified. For these reasons we cannot expect with real light to give simple meanings to the terms "duration of the wave trains" and "frequency range of the Fourier spectrum". However, for any light disturbance $V(t)$, and its Fourier inverse $v(v)$, it is possible to define quantities $\Delta t$ and $\Delta\nu$ which may respectively be regarded as the average duration of the wave trains of which $V$ is composed, and the effective frequency range of the Fourier spectrum; and it can be shown that these averages always satisfy the reciprocity relation

$$\Delta t \Delta \nu \geqslant \frac{1}{4\pi}. \tag{106}$$

This inequality, which in a sense is analogous to the Heisenberg uncertainty relation in quantum mechanics, is derived and discussed in § 10.7.3. Here we remark only that in most cases of practical interest the inequality in (106) may be replaced by the order of magnitude sign.

The time $\Delta t$ of (106) is known as the coherence time of the light; and if $\overline{\lambda_0}$ is the mean wavelength, the length $\Delta l$ defined by

$$\Delta l = c\Delta t \sim \frac{c}{\Delta v} = \frac{\left(\bar{\lambda}_0\right)^2}{\Delta \lambda_0} \tag{107}$$

is known as the coherence length. By comparison of (107) and §7.3(15) we see that our earlier restriction on the path difference between quasi-monochromatic beams implies that the path difference must be small compared to the coherence length of the light. When the difference of optical path is of the order of, or much greater than, the coherence length, interference effects are no longer appreciable.

It is clear from the foregoing that observations of the variation of fringe visibility with optical path difference in a suitable interference experiment must yield information about the spectral intensity distribution of the light used. The first observations of this kind were made by Fizeau.* Using his interferometer (§ 7.5.2) illuminated by the yellow light of a sodium flame, he obtained Newton's rings and observed their appearance as the separation of the lens and plate was increased. He found that while the rings were distinct when the lens and plate were in contact, they were almost invisible near the position of the 490 th ring, but had regained almost their original distinctness near the 980 th ring; and he was able to follow such periodic variations

of visibility through 52 cycles each of about 980 rings. From these observations Fizeau correctly inferred that the yellow sodium light has two components of approximately equal intensity. The maxima of fringe visibility occur when the path difference is an integral multiple of the wavelength of each component, so that their wavelengths are in the approximate ratio 981/980. He was able to confirm this conclusion by direct observation with a prism spectroscope.

Later, more elaborate observations were made by Michelson, † who measured the visibility of circular fringes formed in his interferometer by comparing them with a set of circular fringes of known variable visibility. In this way he was able to construct visibility curves, showing fringe visibility as a function of path difference, for the light of a large number of spectral lines.

Let us examine how the visibility curve is related to the spectral intensity distribution. For simplicity we assume that the intensities of the two interfering beams are equal. For optical path difference $\Delta\mathscr{S}$, the phase difference is

$$\delta\left(k_0, \Delta\mathscr{S}\right) = k_0 \Delta\mathscr{S} \tag{108}$$

where $k_0 = 2\pi/\lambda_0$ is the wave number; so that by §7.2(17), the intensity due to the components in the elementary wave number range $dk_0$ is

$$i\left(k_0, \Delta\mathscr{S}\right) dk_0 = 2i_1\left(k_0\right)\left\{1 + \cos\left(k_0 \Delta\mathscr{S}\right)\right\} dk_0 \tag{109}$$

where $i_1\left(k_0\right)$ represents the spectral distribution of intensity of either beam. Since, as already shown, the different spectral components add incoherently, the total intensity in the interference pattern is

$$I(\Delta\mathscr{S}) = 2\int i_1\left(k_0\right)\left\{1 + \cos\left(k_0 \Delta\mathscr{S}\right)\right\} dk_0 \tag{110}$$

For the light of a spectral line, $i_1\left(k_0\right)$ will be negligible except in a small range of $k_0$ about some mean wave number $\bar{k}_0$. If then we put

$$\left.\begin{aligned} x &= k_0 - \bar{k}_0 \\ j(x) &= i_1\left(\bar{k}_0 + x\right), \end{aligned}\right\} \tag{111}$$

we may write in place of (110)

$$I = 2\int j(x)\left\{1 + \cos\left[\left(\bar{k}_0 + x\right)\Delta\mathscr{S}\right]\right\} dx \tag{112}$$

From this point the analysis is similar to that of p. 272. Thus (112) may be written (cf. § 7.3 (38))

$$I(\Delta\mathscr{S}) = P + C(\Delta\mathscr{S})\cos\left(\bar{k}_0 \Delta\mathscr{S}\right) - S(\Delta\mathscr{S})\sin\left(\bar{k}_0 \Delta\mathscr{S}\right), \tag{113}$$

where

$$\left.\begin{aligned} P &= 2\textstyle\int j(x)dx \\ C(\Delta\mathscr{S}) &= 2\textstyle\int j(x)\cos(x\Delta\mathscr{S})dx \\ S(\Delta\mathscr{S}) &= 2\textstyle\int j(x)\sin(x\Delta\mathscr{S})dx \end{aligned}\right\} \tag{114}$$

Since, for a spectral line, $j(x)$ has non-zero values only for $|x| \ll \bar{k}_0$, the variation of $C$ and $S$ is negligible compared with that of $\cos\left(\bar{k}_0 \Delta\mathscr{S}\right)$ and $\sin\left(\bar{k}_0 \Delta\mathscr{S}\right)$, and it follows that the positions of the extrema of $I$ are to a good approximation given by

0

• H. Fizeau, Ann. Chim. Phys. (3), 66 (1862), 429.
† A. A. Michelson, Phil. Mag. (5), 31 (1891), 338; ibid., 34 (1892), 280.

$$\frac{dI}{d(\Delta\mathscr{S})} = -\bar{k}_0 \left[C \sin\left(\bar{k}_0 \Delta\mathscr{S}\right) + S \cos\left(\bar{k}_0 \Delta\mathscr{S}\right)\right] = 0$$

i.e. by

$$\tan\left(\bar{k}_0 \Delta\mathscr{S}\right) = -\frac{S}{C}. \tag{115}$$

From (113) and (115), the extreme values of $I$ are therefore

$$I_{\text{ext}} = P \pm \sqrt{C^2 + S^2} \tag{116}$$

Hence the visibility curve is given by

$$\mathscr{V}(\Delta\mathscr{S}) = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{\sqrt{C^2 + S^2}}{P} \tag{117}$$

We note that (113) may also be written as

$$I = P\left[1 + \frac{\sqrt{C^2 + S^2}}{P} \cos\left(\phi + \bar{k}_0 \Delta\mathscr{S}\right)\right], \tag{118}$$

where $\tan\phi = S/C$, so that the visibility curve is the envelope of the normalized intensity curve $I/P$.

The calculated visibility curves for a number of assumed spectral distributions of intensity are shown in Fig. 7.54. But in practice we have the converse problem of determining the spectral distribution from the observed visibility curve. If $j(x)$ is symmetrical, $S = 0$ and (117) reduces to

$$\mathscr{V} = \frac{|C|}{P} \tag{119}$$

In this case the visibility curve determines $C$ (apart from the constant factor of proportionality $P$, and the sign, which may usually be fixed from considerations of physical plausibility); and $j(x)$ can be obtained from (114) by Fourter inversion. In general, however, the visibility curve determines only $\sqrt{C^2 + S^2}$, and alone it is in general insufficient to determine $j(x)$ since this requires both $C$ and $S$ to be known separately. It was pointed out by Rayleigh* that both $C$ and $S$ can be determined if measurements are made not only of the visibility of the fringes but also of their

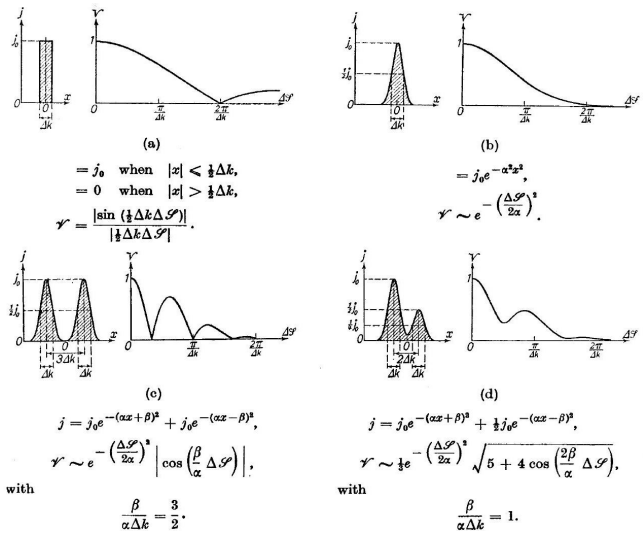



Fig. 7.54. Visibility curves corresponding to various spectral distributions. (The analytical expressions for $\mathscr{V}$ were derived on the assumption that the light is quasi-monochromatic.)
In figures (b), (c), (d), $\Delta k = 2\sqrt{\log_g 2}/\alpha \sim 1.66/\alpha$.

position, since the latter would by (115) give the ratio $C/S$; but such measurements are very difficult to make. †

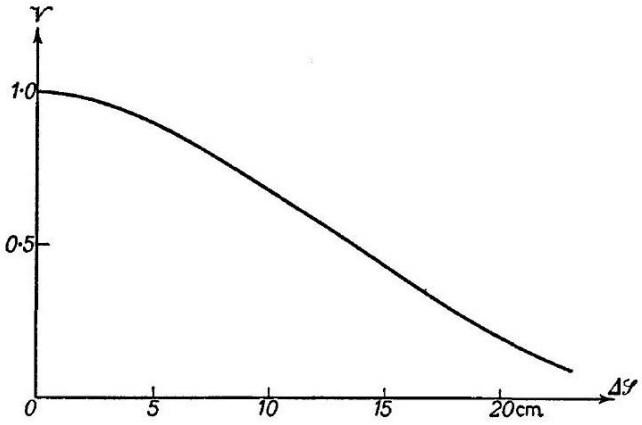
In spite of this limitation, Michelson was able to deduce structures for simple spectral lines which have been broadly confirmed by later work. In particular, he found the red line ( $\lambda = 6438$Å ) of cadmium to be the most nearly monochromatic of all those he examined: its visibility curve (Fig. 7.55) corresponded to a symmetrical spectral distribution of Gaussian form with a width of only 0.013Å at half peak intensity, and he was able to observe fringes with a path difference exceeding 500,000 wavelengths ( $\sim 30$ cm ).

This method of analysing spectra by means of two-beam interference, which is of historical importance as the first use of interference in spectroscopy, was superseded by methods involving multiple-beam interference (§7.6). More recently, however, interest in the use of the two-beam method has been revived partly because for technical reasons it has advantages in the infra-red region of the spectrum.*
(After A. A. Michelson, Phil. Mag. (5), 4 (1892), 291.)

0

- Lord Rayleigh, Phil. Mag. (5). 34 (1892), 407; see also E. Wolf, Proc. Phys. Soc., 80 (1962), 1269.
† A number of such measurements was reported by A. Prard, Rev. d'Optique, 7 (1928), 1; Runions l'Institut d'Optique (Paris: Revue d'Optique), 6 me anne, (1935), 10.

Fig. 7.55. Visibility curve for the red cadmium line.




## 9.10 7.6 Multiple-Beam Interference

When a beam of light is incident on a transparent plate, there are multiple reflections at the plate surfaces, with the result that a series of beams of diminishing amplitude emerges on each side of the plate. In discussing the interference effects produced by such plates (§7.5.1, §7.5.2, §7.5.6) we neglected the contribution to the resultant intensity made by those beams which suffer more than two reflections-an approximation which is justified when the reflectivity of the surfaces is low. We will now take into account all the reflected beams: we shall see that if the reflectivity of the surfaces is high, the intensity distribution in the fringe patterns is modified in a way which has important practical uses.

### 9.10.1 7.6.1 Multiple-beam fringes with a plane parallel plate

Consider a plane parallel transparent plate of refractive index $n'$, surrounded by a medium of refractive index $n$, and suppose that a plane wave of monochromatic light is incident upon the plate at angle $\theta$. Let the ray $SB_1$ (Fig. 7.56) represent the direction of propagation of the incident wave. At the first surface this wave is divided into two plane waves, one reflected in the direction $B_1C_1$ and the other transmitted into the plate in the direction $B_1D_1$. The latter wave is incident on the second surface at angle $\theta'$ and is there divided into two plane waves, one transmitted in the direction $D_1E_1$, the other reflected back into the plate in the direction $D_1B_2$; and the process of division of the wave remaining inside the plate continues as indicated in the figure.

Let $A^{(i)}$ be the amplitude of the electric vector of the incident wave, which we assume to be linearly polarized, with the electric vector either parallel or perpendicular to the plane of incidence. As in § 1.5.2 we take $A^{(i)}$ to be complex, with its phase equal to the constant part of the phase of the wave function. For each member of

either the reflected or the transmitted set of waves, the variable part of the phase of the wave function differs from that of the preceding member by an amount which corresponds to a double traversal of the plate. According to §7.5 (6) this phase difference is

0

- See, for example P. Fellgett, Journ. de Phys., 19 (1958), 187, 237. See also the review article on Fourier spectroscopy by G. A. Vanasse and H. Sakai in Progress in Optics, Vol. 6, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1967), p. 259.

$$\delta = \frac{4\pi}{\lambda_0} n' h \cos\theta' \tag{1}$$

where $h$ is the thickness of the plate and $\lambda_0$ is the wavelength in vacuum. For a wave travelling from the surrounding medium into the plate, let $r$ be the reflection coefficient (ratio of reflected and incident amplitudes), and $t$ the transmission coefficient

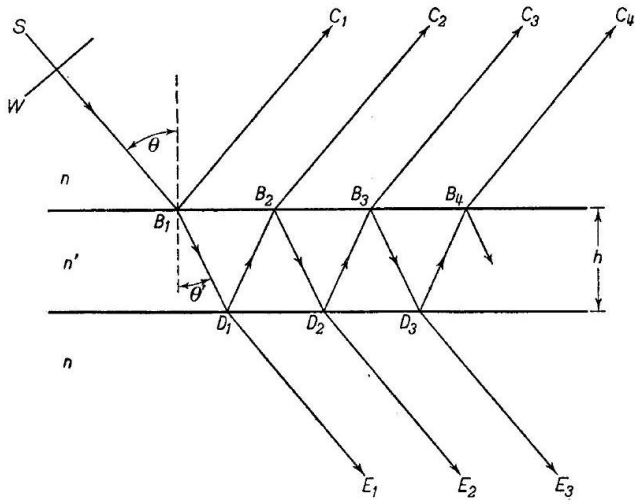



Fig. 7.56. Reflection of a plane wave in a plane parallel plate.

(ratio of transmitted and incident amplitudes); and let $r', t'$ be the corresponding coefficients for a wave travelling from the plate to the surrounding medium. The complex amplitudes of the waves reflected from the plate are then

$$rA^{(i)}, \quad tt'r'A^{(i)}e^{i\delta}, \quad tt'r'^3A^{(i)}e^{2i\delta}, \ldots\ldots tt'r'^{(2p-3)}A^{(i)}e^{i(p-1)\delta}, \ldots$$

Similarly, the complex amplitudes of the waves transmitted through the plate are, apart from an unimportant constant phase factor,

$$tt'A^{(i)}, \quad tt'r'^2A^{(i)}e^{i\delta}, \quad tt'r'^4A^{(i)}e^{2i\delta}, \ldots\ldots tt'r'^{2(p-1)}A^{(i)}e^{i(p-1)\delta}, \ldots$$

The quantities $r, r', t, t'$ are given in terms of $n, n', \theta, \theta'$ by the Fresnel formulae of § 1.5.2. For our present purposes we do not need these explicit expressions, but only relations between them. Thus for either polarized component, we see from relations (20a) and (35) of § 1.5 that

$$tt' = \mathscr{T}; \tag{2}$$

similarly using equation ( 21 (a)) of § 1.5 we have

$$r = -r', \tag{3}$$

whence by § 1.5 (33),

$$r^2 = r'^2 = \mathscr{R}, \tag{4}$$

where $\mathscr{R}$ and $\mathscr{T}$, respectively the reflectivity and transmissivity of the plate surfaces, are related by

$$\mathscr{R} + \mathscr{T} = 1. \tag{5}$$

If the first $p$ reflected waves are superposed, the amplitude $A^{(r)}(p)$ of the electric vector of the reflected light is given by the expression

$$\begin{aligned} A^{(r)}(p) &= \left\{ r + tt'r'e^{i\delta}\left(1 + r'^2e^{i\delta} + \ldots + r'^{2(p-2)}e^{i(p-2)\delta}\right)\right\} A^{(i)} \\ &= \left\{ r + \left(\frac{1 - r'^{2(p-1)}e^{i(p-1)\delta}}{1 - r'^2e^{i\delta}}\right) tt'r'e^{i\delta}\right\} A^{(i)} \end{aligned} \tag{6}$$

If the plate is sufficiently long, the number of reflected waves is large; and in the limit as $p \to \infty$, we have from (6), if (3) is also used,

$$A^{(r)} \equiv A^{(r)}(\infty) = -\frac{r'\left\{1 - (r'^2 + tt')\, e^{i\delta}\right\}}{1 - r'^2e^{i\delta}} A^{(i)}. \tag{7}$$

From (2), (4) and (5) we hence obtain

$$A^{(r)} = \frac{\left(1 - e^{i\delta}\right)\sqrt{\mathscr{R}}}{1 - \mathscr{R}e^{i\delta}} A^{(i)}, \tag{8}$$

so that the intensity $I^{(r)} = A^{(r)}A^{(r)\star}$ of the reflected light is

$$I^{(r)} = \frac{(2 - 2\cos\delta)\mathscr{R}}{1 + \mathscr{R}^2 - 2\mathscr{R}\cos\delta} I^{(i)} = \frac{4\mathscr{R}\sin^2\frac{\delta}{2}}{(1 - \mathscr{R})^2 + 4\mathscr{R}\sin^2\frac{\delta}{2}} I^{(i)}, \tag{9}$$

where $I^{(i)} = A^{(i)}A^{(i)\star}$ is the intensity of the incident light.
In a similar way we obtain the following expression for the amplitude $A^{(t)}$ of the transmitted light:

$$\begin{aligned} A^{(t)}(p) &= tt'\left(1 + r'^2e^{i\delta} + \ldots + r'^{2(p-1)}e^{i(p-1)\delta}\right) A^{(i)} \\ &= \left(\frac{1 - r'^{2p}e^{ip\delta}}{1 - r'^2e^{i\delta}}\right) tt'A^{(i)} \end{aligned} \tag{10}$$

In the limit as $p \to \infty$, (10) reduces to

$$A^{(t)} \equiv A^{(t)}(\infty) = \frac{tt'}{1 - r'^2e^{i\delta}} A^{(i)}. \tag{11}$$

Hence, using (2) and (4), we have

$$A^{(t)} = \frac{\mathscr{T}}{1 - \mathscr{R}e^{i\delta}} A^{(i)}, \tag{12}$$

and the corresponding intensity $I^{(t)} = A^{(t)}A^{(t)\star}$ of the transmitted light is

$$I^{(t)} = \frac{\mathscr{T}^2}{1 + \mathscr{R}^2 - 2\mathscr{R}\cos\delta} I^{(i)} = \frac{\mathscr{T}^2}{(1 - \mathscr{R})^2 + 4\mathscr{R}\sin^2\frac{\delta}{2}} I^{(i)}. \tag{13}$$

The formulae (9) and (13), which are known as Airy's formulae, are in agreement with results already derived from the general theory of propagation in stratified media (§ 1.6); for if in equations (60) of § 1.6 we put $r_{12} = r, r_{23} = r', t_{12} = t, t_{23} = t'$, $2\beta = \delta$, and use the relations (2) and (3) above, the expressions (9) and (13) follow.*

Suppose now that plane waves of equal intensity are incident over a range of angles, and the transmitted light is collected by a lens $L$ (Fig. 7.57). At a point $P$ in the focal plane of $L$, the intensity is in the ratio $I^{(t)}/I^{(i)}$ to the intensity at $P$ when the plate is

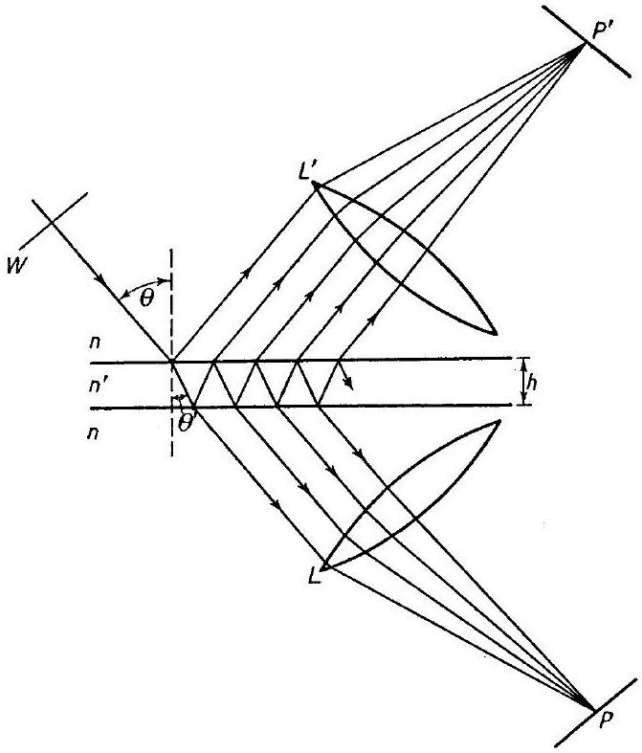



Fig. 7.57. Illustrating formation of multiple beam fringes of equal inclination with a plane parallel plate.

removed; hence according to (13) there will, when the plate is present, be maxima of intensity at $P$ when the order of interference $m$, defined by

$$m = \frac{\delta}{2\pi} = \frac{2n'h\cos\theta'}{\lambda_0}, \tag{14}$$

has integral values $1, 2, \ldots$, and minima when it has half-integral values $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$ Evidently there are fringes of equal inclination in the focal plane of $L$, following loci of constant $\theta'$ (and therefore $\theta$ ). Similarly, if the light reflected from the plate is collected by a lens $L'$, there are fringes of equal inclination in the focal plane of $L'$; and from (9) we find that the intensity maxima of this reflected pattern correspond to half-integral values $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \ldots$ of the order of interference $m$, and minima to integral values $1, 2, \ldots$ Thus the fringes of both patterns are in the positions given by the approximate treatment of §7.5.1, in which we took into account only the first two beams from each side of the plate.

The intensity distributions of the reflected and transmitted patterns are given by (9) and (13), which, using (5), we may write as

$$\frac{I^{(r)}}{I^{(i)}} = \frac{F\sin^2\frac{\delta}{2}}{1 + F\sin^2\frac{\delta}{2}}, \tag{15a}$$

$$\frac{I^{(t)}}{I^{(i)}} = \frac{1}{1 + F\sin^2\frac{\delta}{2}}, \tag{15b}$$

0

- When this comparison is made it must be remembered that, unlike here, the symbols $\mathscr{R}$ and $\mathscr{T}$ denote in § 1.6 the reflectivity and transmissivity of the whole plate.

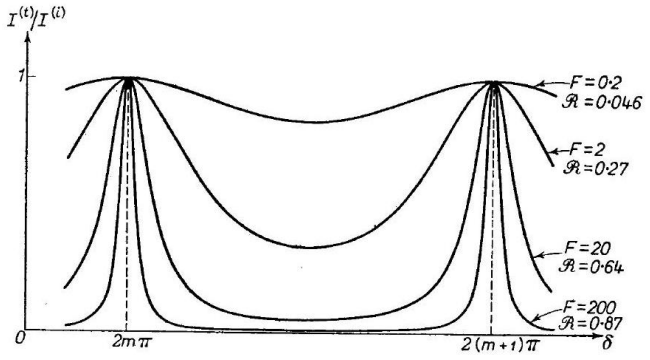



Fig. 7.58. Multiple beam fringes of equal inclination in transmitted light: ratio $I\hat{}{(t)} / I\hat{}{(i)}$ of transmitted and incident intensities as a function of phase difference $\delta$ ( $m$ is an integer).

where the parameter $F$ is defined by the formula

$$F = \frac{4\mathscr{R}}{(1-\mathscr{R})^2}. \tag{16}$$

Evidently the two patterns are complementary, in the sense that

$$\frac{I^{(r)}}{I^{(i)}} + \frac{I^{(t)}}{I^{(i)}} = 1. \tag{17}$$

The behaviour of $I^{(t)}/I^{(i)}$ as a function of the phase difference $\delta$ is shown for various values of $F$ in Fig. 7.58. When $\mathscr{R}$ is small compared with unity, $F$ also is small compared with unity, so that we may expand $1/\left(1 + F\sin^2\delta/2\right)$ in (15), and retain terms up to only the first power in $F$. This gives

$$\frac{I^{(r)}}{I^{(i)}} \sim F\sin^2\frac{\delta}{2} = \frac{F}{2}(1-\cos\delta), \tag{18a}$$

$$\frac{I^{(t)}}{I^{(i)}} \sim 1 - F\sin^2\frac{\delta}{2} = 1 - \frac{F}{2}(1-\cos\delta), \tag{18b}$$

i.e. the intensity variations are of the form of §7.2(15), characteristic of two interfering beams. If $\mathscr{R}$ is increased, the intensity of the minima of the transmitted pattern falls, and the maxima become sharper until, when $\mathscr{R}$ approaches unity so that $F$ is large, the intensity of the transmitted light is very small except in the immediate neighbourhood of those maxima. The pattern in transmitted light then consists of
narrow bright fringes on an almost completely dark background. Similarly the pattern in reflected light becomes one of narrow dark fringes on an otherwise nearly uniform bright background. The sharpness of the fringes is conveniently measured by their half-intensity width, or half-width, which, in the case of the pattern in transmitted light, is the width between the points on either side of a maximum where the intensity has fallen to half its maximum value. The ratio of the separation of adjacent fringes and the half-width we shall call the finesse $\mathscr{F}$ of the fringes. For the fringe of integral order $m$, the points where the intensity is half its maximum value are at

$$\delta = 2m\pi \pm \frac{\varepsilon}{2} \tag{19}$$

where by (15b)

$$\frac{1}{1+F\sin^2\frac{\varepsilon}{4}}=\frac{1}{2} \tag{20}$$

and when $F$ is sufficiently large, $\varepsilon$ is so small that we may put $\sin(\varepsilon/4)=\varepsilon/4$ in (20) and obtain the half-width as

$$\varepsilon=\frac{4}{\sqrt{F}}. \tag{21}$$

Since the separation of adjacent fringes corresponds to a change $2\pi$ of $\delta$, the finesse is then

$$\mathscr{F}=\frac{2\pi}{\varepsilon}=\frac{\pi\sqrt{F}}{2}. \tag{22}$$

So far we have assumed the light to be strictly monochromatic. With quasimonochromatic light, the intensity distribution is the sum of intensity distributions of the form of (15) due to each monochromatic component; and if these components cover a wavelength range $\Delta\lambda_0$ about mean wavelength $\bar{\lambda}_0$, the maxima of order $m$ spread over a distance which corresponds to $|\Delta\delta|$ in the pattern of wavelength $\bar{\lambda}_0$, where from (14), if we neglect dependence of $n'h$ on wavelength, $|\Delta\delta|=2\pi m\Delta\lambda_0/\overline{\lambda_0}$. We can take these component patterns to be in coincidence, and the resultant intensity distribution to be the same as that from a strictly monochromatic source of wavelength $\bar{\lambda}_0$, providing $|\Delta\delta|$ is negligible compared with the half-width $\varepsilon$ of a monochromatic fringe. From (22) this condition is $m\mathscr{F}\ll\overline{\lambda_0}/\Delta\lambda_0$, i.e.

$$\mathscr{F}\Delta\mathscr{S}\ll\frac{\bar{\lambda}_0^2}{\Delta\lambda_0}, \tag{23}$$

where

$$\Delta\mathscr{S}=\frac{\overline{\lambda_0}}{2\pi}\delta=m\bar{\lambda}_0 \tag{24}$$

is the difference of optical path of successive interfering beams; (23) may be compared with the analogous inequality §7.3 (15) relating to two-beam interference. The quantity on the right of (23) will be recognized as the coherence length of the light (cf. § 7.5 (107)).

From the foregoing it is evident that as the reflectivity of the surfaces-and hence $\mathscr{F}$-is increased, the intensity distributions become more favourable for the
measurement of fringe positions, and fringes due to different monochromatic components may become clearly separated in the transmitted pattern. It is for these reasons that multiple-beam interferometry is of practical importance. At normal incidence the reflectivity of a surface between available dielectrics is low at optical wavelengths: for example, for an interface between air ( $n\sim 1$ ) and glass ( $n'\sim 1.5$ ), §1.5(37) gives $\mathscr{R}\sim 0.04$. Increased reflectivity can be obtained at oblique incidence, and as we have seen in § 1.5, $\mathscr{R}\to 1$ when the direction of propagation in the optically denser medium approaches the critical angle. Reflectivity approaching unity may also be obtained near normal incidence by coating the dielectric surface with a multilayer system of suitable dielectric films (as shown in § 1.6) or with a partially transparent film of metal. The theory of metal films on dielectric substrates is given in Chapter XIII, but here we may anticipate it by saying that such films absorb light, that the phase changes on reflection are no longer necessarily zero or $\pi$, and that the reflectivities and phase changes on the two sides of the film are different when the bounding dielectric media are of different refractive index. In consequence the analysis above is not valid for a plate coated with metal films. However, when the films on the two surfaces of the plate are identical, equations (12) and (13) hold providing we interpret $\mathscr{R}$ as the reflectivity for internal reflection, and replace $\delta$ as defined by (1) by

$$\delta = \frac{4\pi}{\lambda_0} n' h \cos\theta' + 2\phi \tag{25}$$

where $\phi$ is the phase change on internal reflection. If then we put

$$\mathscr{R} + \mathscr{T} + \mathscr{A} = 1, \tag{26}$$

where $\mathscr{A}$ is the fraction of light absorbed by the metal, we have from (13), using (16)'

$$\frac{I^{(t)}}{I^{(i)}} = \left(1 - \frac{\mathscr{A}}{1-\mathscr{R}}\right)^2 \frac{1}{1 + F\sin^2\frac{\delta}{2}}. \tag{27}$$

Comparing (27) and (15b) we see that, for given $F$, the effect of absorption is to diminish the intensity of the transmitted pattern by a factor $[(1 - \mathscr{A}/(1-\mathscr{R})]^{-2}$.* Near normal incidence, the effect of the phase change $\phi$ at reflection is equivalent to an increase $\phi\lambda_0/2\pi$ in the optical thickness of the plate: at oblique incidence there is the further complication that the phase change depends on the state of polarization.

## 9.10.2 7.6.2 The Fabry-Perot interferometer

The multiple beam interference fringes from a plane parallel plate illuminated near normal incidence are used in the Fabry-Perot interferometer. † This instrument consists essentially of two glass or quartz plates $P_1, P_2$ (Fig. 7.59) with plane surfaces. The inner surfaces are coated with partially transparent films of high reflectivity, and are parallel, so that they enclose a plane parallel plate of air. The plates themselves are made slightly prismatic, in order to avoid disturbing effects due to reflections at the outer uncoated surfaces. In the original form of the instrument, one plate

was fixed and the other was mounted on a screw-controlled carriage to allow continuous variation of the plate separation, but because of difficulties of mechanical construction this arrangement is obsolete. Instead the plates are separated by a fixed spacer $D$, which is commonly a hollow cylinder of invar or silica with three projecting studs at each end, and the plates are kept in place by the pressure of springs. The spacer is optically worked so that the planes defined by the studs are as nearly parallel as possible, and fine adjustments can be made by varying the spring pressure. This form of the interferometer, with fixed plate separation, is sometimes called a Fabry-Perot etalon.

As we have seen in §7.6.1, light from an extended quasi-monochromatic source $S$ which satisfies (23) forms narrow bright fringes of equal inclination in the focal plane of the lens $L$. From (25) the order of interference is given by

$$m = \frac{\delta}{2\pi} = \frac{2n'h\cos\theta'}{\lambda_0} + \frac{\phi}{\pi}, \tag{28}$$

0

- With multilayer dielectric reflecting films also, the intensity distribution of the transmitted pattern is found to be of the form of (27); in this case the reduction of intensity is due principally to scattering: see P. Glacomo, Rev. d'Opt., 35 (1956), 442.
  † C. Fabry and A. Perot, Ann. Chim. Phys. (7), 16 (1899), 115.

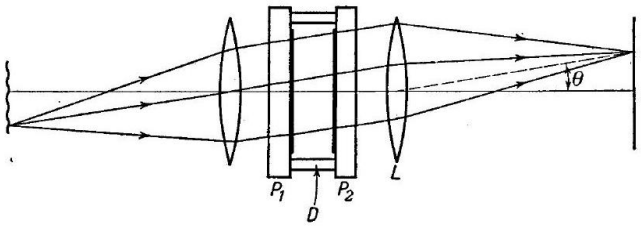



Fig. 7.59. The Fabry-Perot interferometer.

where $n'$ is the refractive index of the air between the plates, $h$ is the separation of the reflecting surfaces, $\theta'$ is the angle of reflection, and $\phi$ is the phase change. The axis of the lens is usually normal to the plates, and the bright fringes, corresponding to integral values of $m$, are then circles with common centre at the focal point for normally transmitted light (Fig. 7.60). At this point, $m$ has its maximum value $m_0$ given by

$$m_0 = \frac{2n'h}{\lambda_0} + \frac{\phi}{\pi}. \tag{29}$$

In general $m_0$ is not an integer, and we may write

$$m_0 = m_1 + e, \tag{30}$$

where $m_1$ is the integral order of the innermost bright fringe, and $e$, which is less than unity, is the fractional order at the centre. From (28), (29) and (30), in an exactly similar way to the derivation of equation (13) of §7.5, we obtain the angular radius $\theta_p$ of the $p$ th bright fringe from the centre, when $\theta_p$ is not too large, as

$$\theta_p = \frac{1}{n}\sqrt{\frac{n'\lambda_0}{h}}\sqrt{p-1+e}, \tag{31}$$

where $n$ is the refractive index of the air outside the plates. The diameter $D_p$ of this fringe is therefore given by

$$D_p^2 = (2f\theta_p)^2 = \frac{4n'\lambda_0 f^2}{n^2 h}(p-1+e), \tag{32}$$

where $f$ is the focal length of the lens $L$.



Fig. 7.60. Fabry-Perot fringes.

The characteristics of the Fabry-Perot interferometer which are of importance in practice are the finesse $\mathscr{F}$, already defined as the ratio of fringe separation and halfwidth; the peak-transmission

$$\tau = \left(\frac{I^{(t)}}{I^{(i)}}\right)_{\max}; \tag{33}$$

and the contrast factor

$$\mathscr{C} = \left(\frac{I^{(t)}}{I^{(i)}}\right)_{\max} / \left(\frac{I^{(t)}}{I^{(i)}}\right)_{\min}, \tag{34}$$

where $I^{(t)}$ is the intensity in the fringe pattern, and $I^{(i)}$ is the corresponding intensity when the interferometer is removed. If we assume for the moment that the inner surfaces of the plates are plane and parallel, and neglect the reflections at the outer surfaces, $I^{(t)}/I^{(i)}$ is given by (27). Comparing (27) with (15b), we see that the finesse is given by (22), and using (16) we have

$$\mathscr{F} = \frac{\pi\sqrt{\mathscr{R}}}{1-\mathscr{R}}. \tag{35}$$

From (27) the peak-transmission is

$$\tau = \left(1 - \frac{\mathscr{A}}{1-\mathscr{R}}\right)^2; \tag{36}$$

and also from (22) and (16), the contrast factor is

$$\mathscr{C} = 1 + F = \left(\frac{1+\mathscr{R}}{1-\mathscr{R}}\right)^2 = 1 + \frac{4\mathscr{F}^2}{\pi^2}. \tag{37}$$

As we have mentioned in §7.6.1, the plate coatings may be either metal films, usually of silver or aluminium, or dielectric films, formed of alternate layers each $\lambda_0/4$ in optical thickness, of high and low refractive index media-for example, zinc sulphide and cryolite.* Both kinds of film

are prepared by thermal evaporation in vacuum. At a given wavelength, $\mathscr{R}$ in general increases with increasing film thickness, in the case of metals, and with increasing number of layers, in the case of dielectrics (cf. Table III, § 1.6). However, it is found with both types of coating that at the high values of $\mathscr{R}$ which are of practical interest, an increase of $\mathscr{R}$ is accompanied by a increase of $\mathscr{A}/(1-\mathscr{R})$, i.e. from (35) and (36), $\tau$ decreases with increasing $\mathscr{F}$. Thus high values of both peak-transmission and finesse (or contrast factor) are incompatible requirements, and in practice some compromise must be made between them. Figure 7.61 shows values of $\tau$ and $\mathscr{F}$ derived from measurements on typical films. We see that, except in the extreme red, the highest values of finesse for given peak-transmission are obtained with dielectric films. We must note, however, that the reflection coefficient of a multilayer dielectric film is high only in a limited wavelength region about the wavelength $\lambda_0$ for which the layers have optical thickness $\lambda_0/4$, so that dielectric coatings are unsuited to applications where a single interferometer is to be used with light covering a wide range of wavelengths. At wavelengths in the visible spectrum, silver gives sharper fringes for given peak-transmission than aluminium, but below about 4000Å the converse is true, and aluminium coatings have been used for work in the ultra-violet as far as 2000Å.

The discussion above refers to an ideal interferometer in which the reflecting

surfaces are perfectly plane and parallel.* In practice the surfaces of interferometer plates cannot be worked perfectly plane, and in consequence the plate separation $h$ always varies over the aperture. The effect of such variations has been considered by Dufour and Picca † and by Chabbal. ‡ They have shown that the finesse and peak-transmission are always less than the values given by (35) and (36), and that

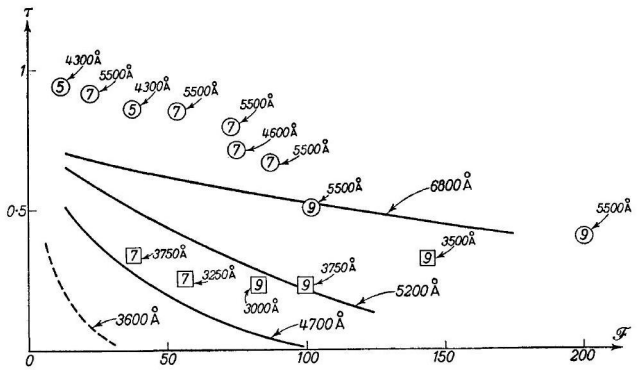



Fig. 7.61. Peak-transmission $\tau$ and finesse $\mathscr{F}$ with various reflecting coatings, assuming perfectly plane and parallel surfaces.

- Fresh silver films (From H. Kuhn and B. A. Wilson, Proc. Phys. Soc. B, 63 (1950), 745.)
- Aluminium films (From J. C. Burridge, H. Kuhn and A. Pery, Proc. Phys. Soc. B, 66 (1953), 963.)

◯ Zinc sulphide-cryolite films (From P. Glacomo, Rev. d'Opt., 35 (1956), 317.
J. Ring and W. L. Wilcock, Nature, 171 (1953), 648; ibid. 173 (1954), 994.)

□ Lead chloride-magnesium fluoride films (From S. Penselin and A. Steudel, Z. Phys., 142 (1955), 21.
The number of layers and the optimum wavelength for the dielectric films are indicated.
as $\mathscr{R} \to 1$, the finesse approaches a limit $\mathscr{F}_d$ which depends only on the defects of the plates,

0

• Combinations of metal films and dielectric films have also been used: see C. Dufour, Ann. de Physique (12), 6 (1951), 5.

i.e. with given plates there is an upper limit of fringe sharpness which cannot be exceeded whatever the reflectivity of the coatings. The value of $\mathscr{F}_d$ depends on the form and magnitude of the departure from plane parallelism. For the particular case where the defect is a slight spherical curvature of the plates, such that $n'h$ changes by $\lambda_0/q$ between centre and edge of the interferometer aperture, $\mathscr{F}_d = q/2$. As an example we may take the case of an interferometer with $q = 100$ (i.e. $\mathscr{F}_d = 50$ ), coated with fresh silver films and used at $\lambda \sim 5200$Å. The finesse and peak-transmission, compared with values taken from Fig. 7.61, are shown in

Table XV. We see that when $\mathscr{F} \gtrsim \mathscr{F}_d$, increase of reflectivity results in a reduction of peak-transmission with little compensating increase of the finesse. The example also illustrates the high degree of flatness of the plates that would be required to take full advantage of attainable reflectivities.

Table XV Fabry-Perot interferometer with fresh silver films, used at wavelength 5200Å. Finesse $\mathscr{F}$ and peak-transmission $\tau$ for (a) plane plates and (b) spherically curved plates such that nh differs by 52Å between centre and edge.
The values for (a) are from Fig. 7.61. The values for (b) were obtained from theoretical curves given by R. Chabbal, J. Rech. Cent. Nat. Rech. Sci., Labs. Bellevue (Paris), No. 24 (1953), 138.

| (a) Plane plates | | (b) Spherically curved plates ($\mathscr{F}_d = 50$) | |
|---|---|---|---|
| $\mathscr{F}$ | $\tau$ | $\mathscr{F}$ | $\tau$ |
| 25 | 0.59 | 22 | 0.55 |
| 50 | 0.44 | 36 | 0.34 |
| 75 | 0.30 | 42 | 0.20 |
| 100 | 0.20 | 45 | 0.11 |
| 125 | 0.13 | 46 | 0.06 |

## 9.10.3 7.6.3 The application of the Fabry-Perot interferometer to the study of the fine structure of spectral lines

When a Fabry-Perot interferometer is illuminated by quasi-monochromatic light which does not satisfy the condition (23), the form of the intensity distribution of the transmitted light differs from that given by (27), and yields some information about the spectral distribution of the light used. In particular, suppose the light has two monochromatic components. If we imagine that their wavelength difference is gradually increased, and providing they do not differ too greatly in intensity, their presence will eventually be evident from the presence of two mutually displaced sets of maxima in the interference pattern. The components are then said to be resolved by the interferometer. In this way Fabry and Perot* were able to observe directly the fine structure of spectral lines which Michelson could only infer (§ 7.5.8), and the Fabry-Perot interferometer has since played a dominant role in this branch of spectroscopy.

In order to compare the powers of different instruments to resolve spectral structure, it is convenient to consider the case when the two components are of equal intensity, and to fix somewhat arbitrarily a displacement of maxima at which the components may be said to be

0

- A modified version of the Fabry-Perot interferometer in which the plane mirrors are replaced by spherical ones of equal radii of curvature and with coincident foci was described by P. Connes, Rev. d'Opt., 35 (1956), 37; Journ. de Phys., 19 (1958), 262. Interferometers of this type are used, for example, as resonators for optical masers. (Cf. A. G. Fox and T. LI, Bell Tech. J., 40 (1961), 453, and G. D. Boyd and J. P. Gordon, ibid, 40 (1961), 489.)
† C. Dufour and R. Picca, Rev. d'Opt., 24 (1945), 19.
‡ R. Chabbal, J. Rech. Cent. Nat. Rech. Sci., Labs. Bellevue (Paris), No. 24 (1953), 138.

"just resolved." If $\lambda_0 \pm \frac{1}{2}\Delta\lambda_0$ are the wavelengths of the two components, the quantity $\lambda_0/\Delta\lambda_0$ is called the resolving power of the instrument. Such a criterion of resolution was first introduced by Lord Rayleigh, † in connection with prism and grating spectroscopes, where the intensity distribution for monochromatic light is of the form $I(\delta) = \{\sin(\delta/2)/(\delta/2)\}^2 I_{\max}$. Rayleigh proposed that, in this case, two components of equal intensity should be considered to be just resolved when the principal intensity maximum of one coincides with the first intensity minimum of the other (Fig. 7.62); in the combined distribution, the ratio of the intensity at the mid-point to that at the maxima is then $8/\pi^2 = 0 \cdot 811$.

Let us adopt this saddle-to-peak intensity ratio as a criterion of resolution in the

case of the Fabry-Perot interferometer.* If the reflectivity of the plates is not so high that their imperfections are important, the intensity distribution $I(\delta)$ due to a single monochromatic component may be written, from (27), in the form

$$I(\delta) = \frac{I_0}{1 + F\sin^2\frac{\delta}{2}} \tag{38}$$

and the total intensity $I_{\text{tot}}(\delta, \varepsilon)$ resulting from the superposition of two such components, whose relative displacement corresponds to a change of $\delta$ by an amount $\varepsilon$, is

$$\begin{aligned} I_{\text{tot}}(\delta, \varepsilon) &= I\left(\delta + \frac{1}{2}\varepsilon\right) + I\left(\delta - \frac{1}{2}\varepsilon\right) \\ &= \frac{I_0}{1 + F\sin^2\frac{\left(\delta+\frac{1}{2}\varepsilon\right)}{2}} + \frac{I_0}{1 + F\sin^2\frac{\left(\delta-\frac{1}{2}\varepsilon\right)}{2}} \end{aligned} \tag{39}$$

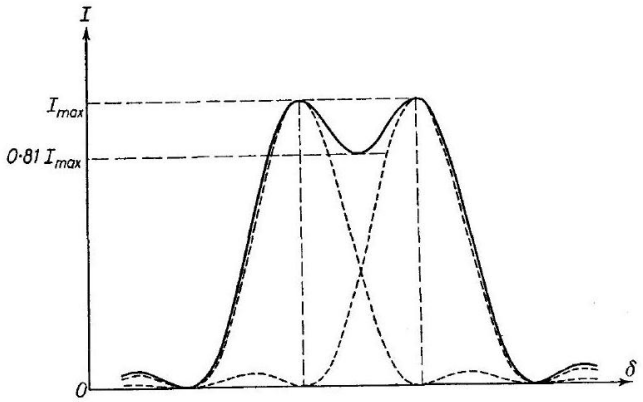



Fig. 7.62. Illustrating two monochromatic spectral components just resolved according to the Rayleigh criterion.

The total intensity at the point mid-way between the intensity maxima of the two components is $I_{\text{tot}}(2m\pi, \varepsilon)$ where $m$ is an integer; and if, when the components are just resolved, we take the maxima of total intensity to coincide with the intensity maxima of the components, the total intensity at the maxima is $I_{\text{tot}}\left(2m\pi \pm \frac{1}{2}\varepsilon, \varepsilon\right)$. Hence, with the criterion we have chosen, the two lines are just resolved when $\varepsilon$ is such that

0

- C. Fabry and A. Perot, Ann. Chim. Phys. (7), 16 (1899), 115.
† Lord Rayleigh, Phil. Mag. (5), 8 (1879), 261.

$$\frac{2I_0}{1+F\sin^2\frac{\varepsilon}{4}} = 0.81\left\{I_0 + \frac{I_0}{1+F\sin^2\frac{\varepsilon}{2}}\right\}. \tag{40}$$

If the finesse of the fringes is high, $\varepsilon$ in (40) is small compared with $\pi/2$ and we may put $\sin\varepsilon = \varepsilon$; (40) then reduces to

$$F^2\varepsilon^4 - 15\cdot 5F\varepsilon^2 - 30 = 0$$

which gives

$$\varepsilon = \frac{4\cdot 15}{\sqrt{F}} = \frac{2\cdot 07\pi}{\mathscr{F}} \tag{41}$$

because of (22).

Now from (28), if we take $n'$ to be independent of wavelength, and if $h$ is so large that we can neglect $\phi$ compared with $\delta$, we have

$$|\Delta\delta| = \frac{4\pi n'h\cos\theta'}{\lambda_0^2}\Delta\lambda_0 = 2\pi m\frac{\Delta\lambda_0}{\lambda_0} \tag{42}$$

At the limit of resolution, $|\Delta\delta|$ is equal to the value $\varepsilon$ given by (41), so that the resolving power of the interferometer is

$$\frac{\lambda_0}{\Delta\lambda_0} = 0.97m\mathscr{F} \tag{43}$$

By analogy with the expression (14) of §8.6 for the resolving power of a diffraction grating, where one is concerned with a finite number of interfering beams of equal intensity, the factor $0.97\mathscr{F}$ is sometimes called the effective number of beams. Near normal incidence, $m \sim 2n'h/\lambda_0$ and we may take the resolving power to be

$$\frac{\lambda_0}{\Delta\lambda_0} \sim \frac{2\mathscr{F}n'h}{\lambda_0} \tag{44}$$

Thus the resolving power of the interferometer is proportional to the finesse and to the optical separation of the plates. As an example we may take $\mathscr{F} = 30(\mathscr{R} \sim 0.9)$, a value which can easily be attained in the visible spectrum; with $n'h = 4$ mm, the resolving power for $\lambda_0 = 5000$Å is then about $5\times 10^5$, which is of the order of the resolving power attainable with the largest line gratings. For spectroscopic purposes, it is also convenient to use the interval $\Delta\kappa_0$ of spectroscopic wave number ( $\kappa_0 = 1/\lambda_0$ ) corresponding to the least resolvable wavelength difference $\Delta\lambda_0$ :

$$\Delta\kappa_0 = \frac{\Delta\lambda_0}{\lambda_0^2} \sim \frac{1}{2\mathscr{F}n'h} \tag{45}$$

The quantity $\Delta\kappa_0$ of (45) is sometimes called the resolving limit of the interferometer. For the above example, the resolving limit is about 0.04 cm$^{-1}$.

If the wavelength separation of two components is sufficiently large, the displacement between the two patterns is greater than the distance between adjacent maxima of either, and there is said to be "overlapping" of orders. The wavelength difference $(\Delta\lambda_0)_{\text{S.R.}}$ corresponding

0

- No special physical significance is to be attached to the Rayleigh criterion, and from time to time other criteria of resolution have been proposed. For purposes of comparison of different instruments the particular choice of criterion is of little importance.

to a displacement of one order ($|\Delta\delta| = 2\pi$) is called the spectral range of the interferometer; near normal incidence, from (42),

$$(\Delta\lambda_0)_{S.R.} = \frac{\lambda_0}{m} \sim \frac{\lambda_0^2}{2n'h} \tag{46}$$

In terms of spectroscopic wave-number, the spectral range is

$$(\Delta\kappa_0)_{S.R.} = \frac{(\Delta\lambda_0)_{S.R.}}{\lambda_0^2} \sim \frac{1}{2n'h} \tag{47}$$

It is seen that the spectral range is inversely proportional to the plate separation, so that an increase of resolving power obtained by increasing the plate separation is accompanied by a proportionate reduction of the spectral range. Comparing (44) and (46), we see also that the spectral range $(\Delta\lambda_0)_{S.R.}$ is approximately $\mathscr{F}$ times the least resolvable wavelength difference. With the values of $\mathscr{F}$ that can be reached in practice, the spectral range is very small when the resolving power is high. For the
typical case considered above $\left(\mathscr{F} = 30, n'h = 4 \text{ mm}, \lambda_0 = 5000\text{Å}\right), (\Delta\lambda_0)_{\text{S.R.}} \sim 0.3\text{Å}$. In §7.6.8 we shall see that a greater spectral range can be obtained by using two Fabry-Perot interferometers in series. Nevertheless, when complex spectra are to be examined, it is necessary to separate the patterns of all but the closest lines by means of auxiliary apparatus, as we shall describe later.

When two close components are resolved, the interval of spectroscopic wavenumber between them is easily determined from measurements of the ring diameters and the plate separation. From (29) and (30), for spectroscopic wave numbers $\kappa_0, \kappa_0'$, we may write

$$m_1 + e_1 = 2n'h\kappa_0 + \frac{\phi}{\pi}, \quad m_1' + e_1' = 2n'h\kappa_0' + \frac{\phi}{\pi} \tag{48}$$

where $m_1, m_1'$ are the integral orders of the first bright rings, $e_1, e_1'$ are the fractional orders at the centre, and we have assumed $\phi$ to be constant over the interval from $\kappa_0$ to $\kappa_0'$. Hence, by subtraction

$$\kappa_0' - \kappa_0 = \frac{(m_1' - m_1) + (e_1' - e_1)}{2n'h}. \tag{49}$$

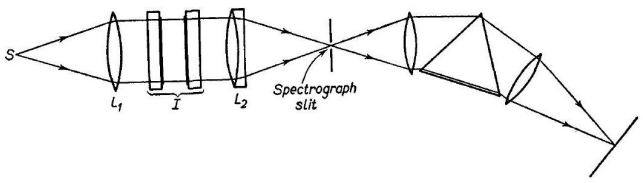



Fig. 7.63. A Fabry-Perot interferometer crossed with a prism spectrograph.

For each line, the fractional orders may be obtained from measurements of the diameters of two bright rings. Thus from (32) the diameters $D_p, D_q$ of the $p$ th and $q$ th rings are related by

$$\frac{D_p^2}{D_q^2} = \frac{p - 1 + e}{q - 1 + e}$$

whence

$$e = \frac{(q-1)D_p^2 - (p-1)D_q^2}{D_q^2 - D_p^2} \tag{50}$$

If measurements are made on more than two rings, a mean value of $e$ can be found by the least squares method. The value of the integer $(m_1' - m_1)$, which is the number of orders of overlap between the two patterns, is decided from observations of the change in the patterns when the plate separation is reduced. Usually the accuracy of the measured order separation $(\ m_1' - m_1\ ) + (e_1' - e_1)$ does not exceed 1 part in $10^3$; since plate separations less than about 1 mm are rarely used, $h$ is then obtained accurately enough by means of a micrometer, and $n'$ for air can be taken as unity.

We have already mentioned that because of the small spectral range of the interferometer, the patterns from all but very close lines must be separated. When the observations are photographic, this separation is usually achieved by "crossing" the interferometer with a stigmatic prism or grating spectrograph. One arrangement is shown in Fig. 7.63. The interferometer $I$ is illuminated with light from the source $S$ in the focal plane of the lens $L_1$, and the interference pattern is projected on to the
plane of the spectrograph slit by means of the well-corrected lens $L_2$.* The orientation of the interferometer is adjusted so that the centre of the pattern coincides with the centre of the slit. The slit then selects a diametrical section from the ring system of each line present, and the action of the spectrograph is to separate these sections. When the spectrum to be examined consists of emission lines, the slit may usually be relatively wide. The appearance in the focal plane of the spectrograph is then as in Fig. 7.64, the images of the slit formed in the light of each line being crossed by short arcs of the bright fringes. A similar appearance is obtained with absorption lines, provided the continuous background associated with each line extends over a wavelength range less than the spectral range of the interferometer. The absorption lines appear as dark lines across the broader bright fringes from the background (Fig. 7.65).

When the continuous background extends over a wide wavelength range, as for example in the solar spectrum, the situation is more complicated. † To understand it, let us suppose for the moment that the width of the spectrograph slit is negligibly small, and that the resolving power of the spectrograph is then infinite. Let $Ox, Oy$ be rectangular coordinate axes in the focal plane of the spectrograph, with origin $O$ on the line defined by light passing through the centre of the slit, and $Oy$ parallel to the slit. Along a line $x =$ constant, the wavelength $\lambda_0$ is constant, and the maxima of intensity are in positions determined by (28), which we may write

$$\left(m - \frac{\phi}{\pi}\right)\lambda_0 = 2n'h\cos\theta' = 2n'h\left(1 - \frac{\theta'^2}{2}\right) = 2h\left(1 - \frac{\theta^2}{2}\right), \quad m = 1, 2, \ldots \tag{51}$$

where we have assumed $\theta'$ to be small, and put $n' = n = 1$ for air. The angle $\theta$ is the angle of emergence from the interferometer; for $\theta$ small, it is related to the coordinate $y$ by $y = M^f\theta$, where $f$ is the focal length of the lens $L_2$ used to project the fringes on the slit, and $M$ is the magnification introduced by the spectrograph. Further, $\lambda_0$ is related to the coordinate $x$, say $F(x) = \lambda_0$, where the function $F$ is characteristic of the dispersion of the spectrograph. Using these relations to eliminate $\lambda_0$ and $\theta$ from (51), we obtain the loci of intensity maxima when the source emits a continuous spectrum:

$$\left(m - \frac{\phi}{\pi}\right)F(x) = 2h\left(1 - \frac{y^2}{2M^2f^2}\right), \quad m = 1, 2, \ldots \tag{52}$$

Between these maxima the relative intensity distribution is characteristic of the interferometer; the spectrum in the focal plane of the spectrograph is thus channelled with narrow bright fringes separated by broader dark intervals. The channels are symmetrical about the line $y = 0$, convex on the side towards longer wavelengths, and for $y =$ constant their wavelength separation is evidently equal to the spectral range of the interferometer. In the special case when $F(x)$ is a linear function of $x$, the channels are parabolic.

When absorption lines are present, the bright channels are interrupted by dark spots where they intersect the absorption lines, i.e., these dark spots have exactly the same positions as bright maxima would have if they were formed by emission lines of the same wavelength as the absorption lines. If the slit is opened symmetrically,

the dark spots extend in the $x$-direction; at the same time, the bright channels similarly broaden in the $x$-direction, until eventually they touch, and coalesce. In this condition, there is what appears to be a continuous spectrum containing dark "fringe" patterns from the absorption lines (Fig. 7.66), which may be measured in the same way as bright patterns from emission lines. In practice, because of diffraction, the resolving power of the spectrograph is not infinite and in consequence the light of any given wavelength is initially spread over a finite distance in the $x$-direction. The effect of this is similar to that of opening the slit, i.e. the bright channels are broadened in the $x$-direction, and they will overlap unless the spectrograph is able to resolve clearly wavelengths whose separation is that of adjacent bright channels. We have seen that this separation is equal to the spectral range of the interferometer, which decreases with increasing plate separation. The finite resolving power of the auxiliary spectrograph thus sets an upper limit to the plate separation that can be employed.

The Fabry-Perot interferometer may also be used for spectroscopy in conjunction with photoelectric detectors.* Light of the spectral line to be examined, after isolation by means of a preliminary monochromator, is fed to the interferometer, and the interference pattern is projected on to an annular aperture concentric with the rings, which allows the light in a small fraction of an order to reach the photocell. The pattern is explored by varying the optical separation of the interferometer plates, which causes the rings to expand or contract over the aperture. The arrangement has considerable technical importance because, as Jacquinot † has shown, the FabryPerot interferometer can pass much greater light flux for given resolving power than conventional prism and grating monochromators.

## 9.10.4 7.6.4 The application of the Fabry-Perot interferometer to the comparison of wavelengths

Measurements of wavelengths in spectra produced by prism and grating spectrographs are carried out by interpolation, the wavelength of any line being expressed in terms of the known wavelengths of neighbouring standard lines. Since with large spectrographs such comparisons can be made with an accuracy of about 1 part in $10^6$, it is necessary to know the relative wavelengths of the standard lines with an accuracy at least equal to this, and they must be sufficiently numerous to allow convenient interpolation. Further, since wavelengths are actual lengths, it is desirable to refer them to the metre. The establishment of a satisfactory system of wavelengths for spectroscopic purposes thus involves (a) the comparison of the wavelength of a chosen primary standard line with the material standard which represents the metre, and (b) the comparison of the wavelength of the primary standard with the wavelengths of other lines which are to serve as secondary standards throughout the spectrum. The primary standard, chosen in 1907, ‡ is the red line ( 6438Å ) of cadmium, excited under prescribed conditions.§ The measurement of the wavelength of this line in terms of the metre is a metrological process

0

- Other arrangements of interferometer and spectrograph have been used. The choice of a suitable arrangement for a particular purpose depends on such considerations as compactness and expense, the size, intensity and uniformity of the source, and the importance of "ghost" patterns which arise from unwanted reflections at surfaces in the system. [See S. Tolansky, High Resolution Spectroscopy, London, Methuen, 1947, Chapt. 9.]

† It was first discussed by C. Fabry and H. Buisson, Journ. de Phys. (4), 9 (1910), 197.

requiring special methods, which we shall describe in §7.7. The value found is $6438.4696 \times 10^{-10}$ metre, and is taken to be 6438.4696Å. This is the definition of the Angström; it does not differ from $10^{-10}$ metre by more than 1 part in $3 \times 10^6$. With the primary standard of wavelength so defined, the comparison of other wavelengths with it is a purely optical process which has been one of the important applications of the Fabry-Perot interferometer.

The interferometer is illuminated (preferably simultaneously) with light of the primary standard and unknown lines, and the patterns are separated and photographed by means of a suitable arrangement. Let $\lambda_s$ be the wavelength of the standard line in the air of refractive index $n'$ between the interferometer plates, and similarly let $\lambda_1, \lambda_2, \ldots$ be the wavelengths of the unknown lines. Then from (29) and (30), if we neglect the variation of $\phi\lambda$ with wavelength, we have

$$(m_{1s} + e_s)\,\lambda_s = (m_{11} + e_1)\,\lambda_1 = (m_{12} + e_2)\,\lambda_2 = \ldots\ldots = 2H, \tag{53a}$$

$$2H = 2h + \frac{\phi\lambda}{\pi}, \tag{53b}$$

where $m_{18}, m_{11}, m_{12}, \ldots$ are the integral orders of the first bright rings, and $e_3, e_1,\ e_2, \ldots$ are the fractional orders at the centre. For each line, the fractional orders may be obtained from measurements of the ring diameters, as we have already described (p. 337). The integers $m_{18}, m_{11}, m_{12}, \ldots$ are found by the method of excess fractions, to which we referred briefly on p. 291. For this, the starting point is a measurement of the separation $h$ of the plates by means of a micrometer. The integer $m'_{18}$ nearest $2h/\lambda_8$ is then an approximate value of $m_{18}$, and we may write

$$m_{1s} = m'_{1s} + x, \tag{54}$$

where $x$ is an unknown integer. If the uncertainty of the measurement of $h$ is $\Delta h$, then

$$|x| \lesssim \frac{2\Delta h}{\lambda_s} \tag{55}$$

For example, with $\Delta h \sim 0.01$ mm and $\lambda_s = 6438$Å, $|x| \lesssim 30$. Next, with known approximate values $\lambda'_1, \lambda'_2, \ldots$ of $\lambda_1, \lambda_2, \ldots$, the approximate orders of interference for the unknown lines corresponding to order ( $m'_{18} + e_8$ ) of $\lambda_8$ are calculated from (53a) as

$$\left.\begin{aligned} m'_{11} + e'_1 &= (m'_{1s} + e_s)\,\tfrac{\lambda_s}{\lambda'_1} \\ m'_{12} + e'_2 &= (m'_{18} + e_s)\,\tfrac{\lambda_s}{\lambda'_2} \ldots\ldots \end{aligned}\right\} \tag{56}$$

If we put $\lambda_1 = \lambda'_1 + \Delta\lambda'_1$, we have, from (53a) and (54),

$$\begin{aligned} (m_{11} + e_1)\,(\lambda'_1 + \Delta\lambda'_1) &= (m_{1s} + e_s)\,\lambda_s \\ &= (m'_{1s} + x + e_s)\,\lambda_s, \end{aligned} \tag{57}$$

whence, using (56)

---

0

- P. Jacquinot and C. Dufour, J. Rech. Cent. Nat. Rech. Sci., Labs. Bellevue (Paris), No. 6 (1948), 91.
  † P. Jacquinot, J. Opt. Soc. Amer., 44 (1954), 761.
  ‡ Trans. Int. Union for Co-operation in Solar Research, 2 (Manchester Univ. Press, 1908), p. 17.
  § Proc. Verb. Com. Int. Poids et Mes. (2), 17 (1935), 91.

$$m_{11} + e_1 = m'_{11} + x + \left[e'_1 + \left(\frac{\lambda_8 - \lambda'_1}{\lambda'_1}\right) x\right] - (m_{11} + e_1)\,\frac{\Delta\lambda'_1}{\lambda'_1} \tag{58}$$

The fractional part contributed by the expression in square brackets on the R.H.S. of (58) is now calculated for the possible values of $x$ allowed by (55). For the correct $x$, this calculated fraction must agree with the measured fraction $e_1$ on the L.H.S. of (58) within the uncertainty $(m_{11} + e_1)\,\Delta\lambda'_1/\lambda'_1$, which is small compared with unity if the plate separation is sufficiently small; for example, with approximate wavelengths obtained from grating measurements, $\Delta\lambda'_1/\lambda'_1 \sim 10^{-5}$, so that $(m_{11} + e_1)\,\Delta\lambda'_1/\lambda'_1 \sim 0.1$ for $m_{11} \sim 10^4$, which corresponds to $h \sim m_{11}\lambda_1/2 \sim 3$ mm for
wavelengths in the visible spectrum. The comparison of calculated and measured fractions then shows some values of $x$ to be inadmissible, and similar calculations with approximate wavelengths $\lambda'_2, \lambda'_3, \ldots$ enable $x$ to be determined unambiguously. Usually three lines suffice, if their wavelengths are suitably spaced, with perhaps a fourth to serve as a check.

With $x$ known, $m_{18}, m_{11}, m_{12}, \ldots$ are given by (54) and (58), and from (53a) we have

$$\lambda_1 = \frac{m_{18} + e_8}{m_{11} + e_1}\lambda_8, \quad \lambda_2 = \frac{m_{18} + e_8}{m_{12} + e_2}\lambda_8, \ldots. \tag{59}$$

Since the integral orders are not in doubt, the uncertainties $\Delta\lambda_1, \Delta\lambda_2, \ldots$ of these wavelengths derive only from the uncertainties of the measured fractions $e_8, e_1, e_2, \ldots$ If the latter are within $\pm 0.01$, a precision which is not difficult to attain with high reflectivity coatings, and with $m_{18}, m_{11}, m_{12}, \ldots$ of the order of $10^4$ as in the example above, $\Delta\lambda_1/\lambda_1, \Delta\lambda_2/\lambda_2, \ldots$ are of the order of $10^{-6}$, i.e. the measurements have increased the accuracy with which $\lambda_1, \lambda_2, \ldots$ are known by a factor of the order of ten. The improved wavelengths may now be used in a repetition of the experiment with larger plate separation, which in turn yields still more accurate wavelength values; and the process may be continued as far as is justified by the sharpness of the lines to be measured.

We have assumed above that, apart from the primary standard, only relatively inaccurate values of wavelength are available.* Once a number of secondary standards have been measured with an accuracy approaching that with which the primary standard is defined, the determination of further wavelengths is simplified. The standard wavelengths are used with the method of excess fractions † to determine $2H$, which need only be small enough to allow $m_{11}, m_{12}, \ldots$ to be determined unambiguously from the approximate wavelengths $\lambda'_1, \lambda'_2, \ldots$ Thus from (53a) and (57),

$$m_{11} + e_1 = \frac{2H}{\lambda'_1} - (m_{11} + e_1)\,\frac{\Delta\lambda'_1}{\lambda'_1}, \tag{60}$$

and $m_{11}$ is obtained unambiguously if $(m_{11} + e_1)\,\Delta\lambda'_1/\lambda'_1 \sim 0.3$ say. If as before $\Delta\lambda'_1/\lambda'_1 \sim 10^{-5}$, this implies $m_{11} \sim 3 \times 10^4$, which corresponds to $H \sim 10$ mm for wavelengths in the visible spectrum; and with the measured fraction $e_1$ within $\pm 0.01$, the uncertainty $\Delta\lambda_1/\lambda_1$ of the wavelength $\lambda_1 = 2H/(m_{11} + e_1)$ is about $3 \times 10^{-7}$, i.e. the single application of the interferometer increases the accuracy by a factor of about thirty.

In practice, the assumption that $\phi\lambda$ is independent of wavelength may not be justified if the lines to be compared are widely separated in the spectrum. Systematic errors from this cause can be eliminated if measurements are made with two values

---

0

- For their earliest comparisons of wavelengths with that of the red cadmium line, Fabry and Perot (Ann. Chim. Phys., 16 (1899), 289) employed the sliding plate form of their interferometer. They made use of accurate values of the wavelengths of the green ( 5086Å ) and blue ( 4800Å ) cadmium lines obtained by two-beam interferometry (A. A. Michelson and J. R. Benoit, Trav. et Mem. Bur. Int. Poids et Mes., 11



Fig. 7.64. Fabry-Perot fringes from lines in the emission spectrum of helium (photographic negative).

[After K. W. Meissner, J. Opt. Soc. Amer., 31 (1941), 416.]

(1895), 1), and determined integral orders by the socalled "method of coincidences". In principle this is the same as the method of excess fractions we have described, but it requires visual observation and is laborious.

† If the standard wavelengths are sufficiently close together, it is possible to calculate quickly an approximate value of $x$, and so reduce the amount of computation (see, for example, C. V. Jackson, Phil. Trans. Roy. Soc. A., 236 (1936), 1). We may mention also that G. R. Harrison (J. Opt. Soc. Amer., 36 (1946), 644) has described a machine by means of which wavelengths are determined directly from interferometer patterns, the necessary computations being carried out automatically.

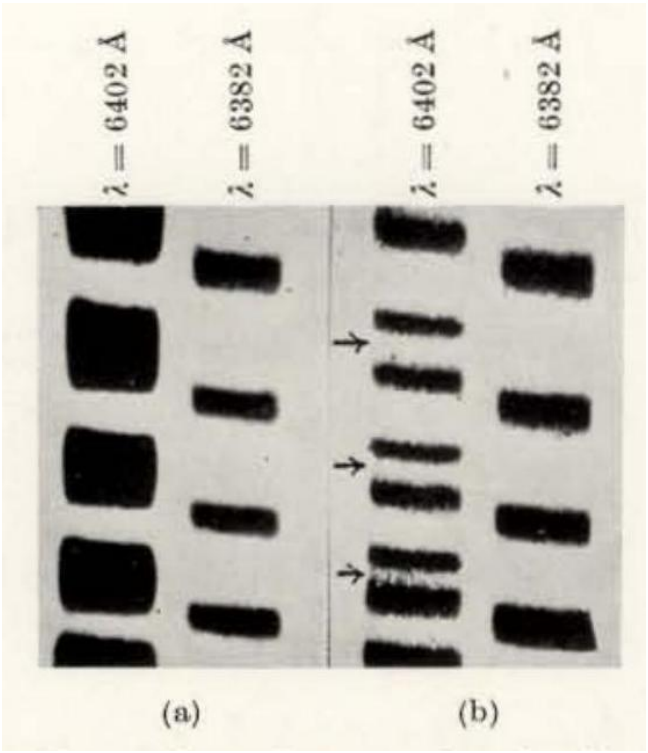



Fig. 7.65. Fabry-Perot fringes from two lines in the spectrum of neon (photographic negative): (a) emission only; (b) central absorption of the line of wavelength 6402Å, indicated by arrows.

[After K. W. Meissner, J. Opt. Soc. Amer., 32 (1942), 191.]

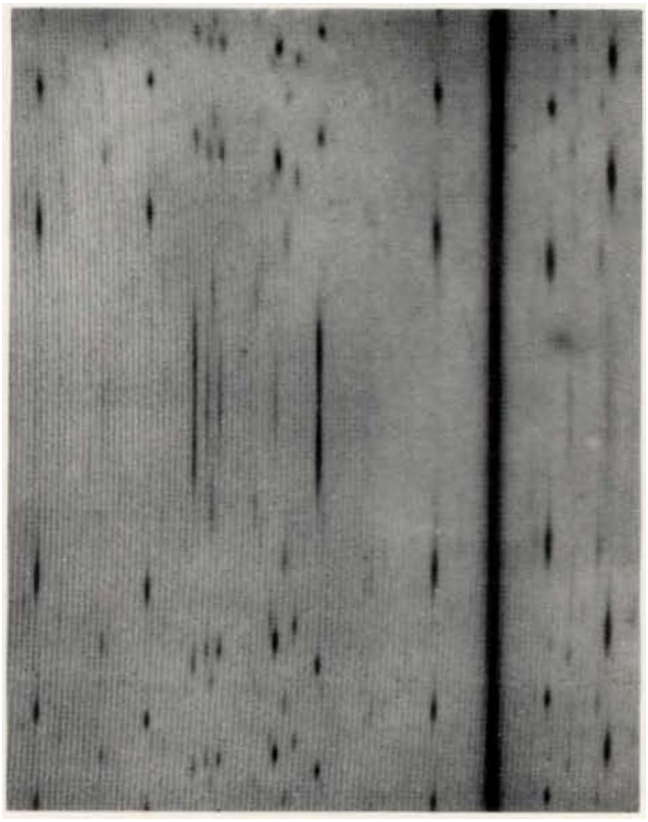


Fig. 7.66. Fabry-Perot fringes from absorption lines in the solar spectrum.
[From H. D. Babcock, Astrophys. J., 65 (1927), 140.]

$h_{\mathrm{I}}, h_{\mathrm{II}}$ of $h$. For then, from (29) and (30), with the same notation as in (53), we may write

$$\left\{(m_{1s}+e_s)_{\mathrm{I}}-\frac{\phi_s}{\pi}\right\}\lambda_s=\left\{(m_{11}+e_1)_{\mathrm{I}}-\frac{\phi_1}{\pi}\right\}\lambda_1=\ldots=2h_{\mathrm{I}}$$
$$\left\{(m_{1s}+e_s)_{\mathrm{II}}-\frac{\phi_s}{\pi}\right\}\lambda_s=\left\{(m_{11}+e_1)_{\mathrm{II}}-\frac{\phi_1}{\pi}\right\}\lambda_1=\ldots=2h_{\mathrm{II}}$$

whence

$$\lambda_1=\frac{(m_{1s}+e_s)_{\mathrm{II}}-(m_{1s}+e_g)_{\mathrm{I}}}{(m_{11}+e_1)_{\mathrm{II}}-(m_{11}+e_1)_{\mathrm{I}}}\lambda_g \tag{61}$$

The wavelengths obtained after correction for phase change must be reduced to their values under the standard conditions of dry air at 15°C and 760 mm Hg pressure for which the primary standard is defined. For this reduction it is necessary to know the dispersion of air. If the wavelengths are to be further reduced to values in vacuum, the refractive index of air must also be known. Values of the refractive index and dispersion of air sufficiently accurate for this purpose have been obtained by multiplebeam methods (§ 7.6.8).

## 9.10.5 7.6.5 The Lummer-Gehrcke interferometer

We have already mentioned in §7.6.1 that the reflection coefficient of a dielectric interface approaches unity when the angle of incidence in the denser medium nears the critical angle. This effect is used in the multiple beam interferometer devised by Lummer and Gehrcke,* which consists essentially of a long plane parallel plate of glass or crystalline quartz (Fig. 7.67). A beam of light from a source on the long axis of the plate is admitted through the prism $P$ fixed to one end of the plate, and meets the plate surfaces at an angle slightly smaller than the critical angle. A series of beams then leaves each side of the plate near grazing emergence, and these are collected by the lens $L$ to form an interference pattern in its focal plane.

For monochromatic light of wavelength $\lambda_0$, the phase difference $\delta$ between successive beams is given by (1), viz.

$$\delta=\frac{4\pi}{\lambda_0}n'h\cos\theta'=\frac{4\pi h}{\lambda_0}\sqrt{n'^2-n^2\sin^2\theta} \tag{62}$$

where $h$ and $n'$ are the thickness and refractive index of the plate, $n$ is the refractive index of the surrounding air, $\theta'$ is the angle of reflection in the plate and $\theta$ is the angle of emergence. The corresponding order of interference $m$ is

$$m=\frac{\delta}{2\pi}=\frac{2h}{\lambda_0}\sqrt{n'^2-n^2\sin^2\theta} \tag{63}$$

The fringes, corresponding to $\theta$ constant, form a family of hyperbolae, and near the centre $O$ of the pattern they approximate to straight lines parallel to the plate surfaces. From (63), the order of interference $m_0$ at the centre ($\theta=\pi/2$) is

$$m_0=\frac{2h}{\lambda_0}\sqrt{n'^2-n^2} \tag{64}$$

and,

0

- O. Lummer, Verh. Deutsch. Phys. Ges., 3 (1901), 85. O. Lummer and E. Gehrcke, Ann. d. Physik, (4), 10 (1903), 457.

$$m - m_0 = \frac{2h}{\lambda_0}\left(\sqrt{n'^2 - n^2\sin^2\theta} - \sqrt{n'^2 - n^2}\right) \tag{65}$$

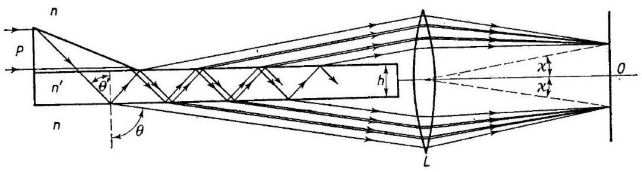



Fig. 7.67. The Lummer-Gehrcke interferometer.

If we put $\chi = (\pi/2) - \theta, \chi$ is small near the centre, so that we may take $\sin^2\theta = 1 - \chi^2$; then after expanding (65), and neglecting powers of $\chi$ higher than the second, we obtain

$$m - m_0 = \frac{n^2}{\sqrt{n'^2 - n^2}}\frac{h\chi^2}{\lambda_0}. \tag{66}$$

The angle $\chi_q$ corresponding to the $q$ th bright fringe from the centre, of integral order $m_q$, is given by

$$\chi_q = \frac{\sqrt[4]{n'^2 - n^2}}{n}\sqrt{\frac{\lambda_0}{h}}\sqrt{m_a - m_0} = \frac{\sqrt[4]{n'^2 - n^2}}{n}\sqrt{\frac{\lambda_0}{h}}\sqrt{q - e}, \tag{67}$$

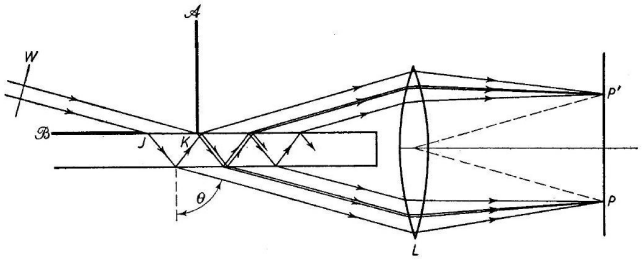



Fig. 7.68. Illustrating multiple reflections in a plane parallel plate when the externally reflected beam is suppressed.

where $e$ is the fractional order at the centre. The angular scale of the pattern is thus proportional to $\sqrt{\lambda_0/h}$, and if $e = 0$ the distances of the bright fringes from the axis are proportional to the square roots of the positive integers.

The intensity distribution in the pattern differs from that considered in §7.6.1 because there is now no contribution from light externally reflected at the plate surfaces. To simplify discussion of this point it is convenient to consider the arrangement of Fig. 7.68, in which light is refracted into the plate through an aperture $JK$ bounded by the diaphragm $\mathscr{A}\mathscr{B}$. It is evident that, as far as the action of the plate is concerned, the prism of Fig. 7.67 behaves as such an aperture, but with negligible
reflection coefficient.* Suppose $P$ and $P'$ are the focal points of $L$ for light emerging from the lower and upper surfaces of the plate at angle $\theta$, and let $A$ be the complex amplitude of the light at $P$ when the plate is removed. The complex amplitudes at $P$ of the beams of the lower set are (cf. §7.6.1), apart from an unimportant constant phase factor,

$$tt'A, \quad tt'r^2Ae^{i\delta}, \ldots\ldots \quad tt'r'^{2(p-1)}Ae^{i(p-1)\delta}$$

and the amplitude at $P$ resulting from the superposition of the first $p$ beams is therefore

$$\begin{aligned} A^{(t)}(p) &= \left(1 + r^2 e^{i\delta} + \ldots\ldots + r^{2(p-1)} e^{i(p-1)\delta}\right) tt' A \\ &= \frac{1 - \mathscr{R} e^{ip\delta}}{1 - \mathscr{R} e^{i\delta}} \mathscr{T} A. \end{aligned} \tag{68}$$

If we take the limit as the number of beams $p \to \infty$, we obtain

$$A^{(t)} \equiv A^{(t)}(\infty) = \frac{\mathscr{T}}{1 - \mathscr{R} e^{i\delta}} A \tag{69}$$

and the intensity $I^{(t)} = A^{(t)} A^{(t)\star}$ at $P$ is then

$$I^{(t)} \equiv I^{(t)}(\infty) = \frac{\mathscr{T}^2}{1 + \mathscr{R}^2 - 2\mathscr{R}\cos\delta} I = \frac{1}{1 + F \sin^2 \frac{\delta}{2}} I \tag{70}$$

where $I = AA^\star$ is the intensity at $P$ when the plate is removed. Similarly, the amplitudes at $P'$ of the beams of the upper set are

$$tt'r'A, \quad tt'r'3e^{i\delta}A, \quad \ldots\ldots\ldots tt' \left(r'\right)^{2p-1} e^{i(p-1)\delta} A \ldots\ldots\ldots$$

so that the amplitude at $P'$ resulting from the superposition of the first $p$ beams is

$$\begin{aligned} A^{(r)}(p) &= \left(1 + r'^2 e^{i\delta} + \ldots + r'^{2(p-1)} e^{i(p-1)\delta}\right) tt'r'A \\ &= \frac{1 - \mathscr{R} p e^{ip\delta}}{1 - \mathscr{R} e^{i\delta}} \sqrt{\mathscr{R}\mathscr{T}} A \end{aligned} \tag{71}$$

and in the limit as $p \to \infty$ the intensity at $P'$ is

$$I^{(r)} \equiv I^{(r)}(\infty) = \frac{\mathscr{R}^2}{1 + \mathscr{R}^2 - 2\mathscr{R}\cos\delta} I = \mathscr{R} I^{(t)} \tag{72}$$

From (70) and (72) it is evident that the maxima of the patterns above and below the plate are at the same angular positions relative to the plate normal.

Successive beams emerging from one side of the plate are evidently separated by a distance $2h \tan\theta'$ along the plate, so that for a plate of length $l$, the number of beams $p$ is given by

$$p \sim \frac{l}{2h} \cot\theta' = \frac{l}{2h} \sqrt{\frac{n'^2}{n^2 \sin^2\theta} - 1} \tag{73}$$

or, for $\theta \sim \pi/2$, and with $n = 1$ for air,

$$p \sim \frac{l}{2h} \sqrt{n'^2 - 1}. \tag{74}$$

As $\theta'$ approaches the critical angle, $\mathscr{R} \to 1$ and $\mathscr{R}^p$ ceases to be negligible, and the use of the limiting expressions (70), (72) with $p \to \infty$ is then not justified. In these

---

0

- A similar arrangement is realized in practice if light is admitted to a Fabry-Perot interferometer through a clear aperture in one of the reflecting coatings; its properties have been discussed by C. Dufour in Rev. d'Opt., 24 (1945), 11.

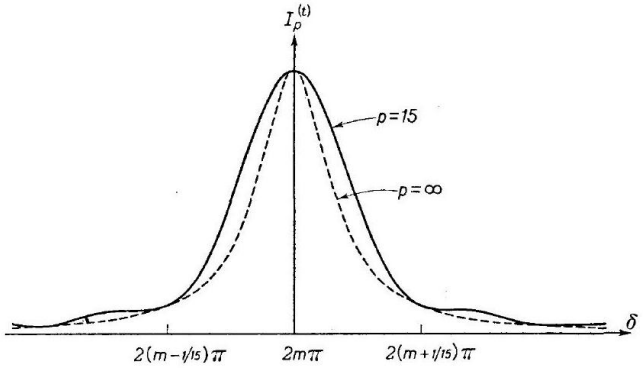



Fig. 7.69. Fringes with a Lummer-Gehrcke interferometer: intensity as a function of phase difference ( $\mathscr{R}=0.87$ ).

circumstances the intensity distribution in the upper pattern is obtained from (68) as

$$\begin{aligned} I^{(t)}(p) &= \frac{1+\mathscr{R}^{2p}-2\mathscr{R}^p\cos p\delta}{1+\mathscr{R}^2-2\mathscr{R}\cos\delta}\mathscr{T}^2 I \\ &= \frac{\left(1-\mathscr{R}^p\right)^2+4\mathscr{R}^p\sin^2\frac{p\delta}{2}}{(1-\mathscr{R})^2+4\mathscr{R}\sin^2\frac{\delta}{2}}\mathscr{T}^2 I \\ &= \frac{1+G_p\sin^2\frac{p\delta}{2}}{1+F\sin^2\frac{\delta}{2}}\left(1-\mathscr{R}^p\right)^2 I \end{aligned} \tag{75}$$

where

$$G_p = \frac{4\mathscr{R}^p}{\left(1-\mathscr{R}^p\right)^2}, \tag{76}$$

and $F$ is given by (16). Thus the intensity distribution is governed by the function $\left(1+G_p\sin^2 p\delta/2\right)/\left(1+F\sin^2\delta/2\right)$, which differs from the distribution function for $p\to\infty$ by the addition of the term $\left(G_p\sin^2 p\delta/2\right)/\left(1+F\sin^2\delta/2\right)$. This additional term does not affect the positions of the absolute maxima, which remain at $\delta/2 = m\pi$, where $m$ is an integer; but since $\left(G_p\sin^2 p\delta/2\right)/\left(1+F\sin^2\delta/2\right)$ is zero for $\delta/2 = (m+q/p)\pi, q = 0,1,2,\ldots,p$, and is otherwise positive, these absolute maxima are broader than those for $p\to\infty$, and there are secondary maxima between them (Fig. 7.69).

The Lummer-Gehrcke interferometer has been employed solely for the examination of the fine structure of spectral lines, for which purpose it is used, like the FabryPerot interferometer, in combination with an auxiliary dispersing instrument. The spectral range of the interferometer may be obtained from (63), which, taking $n=1$, we may write as

$$m^2\lambda_0^2 = 4h^2\left(n'^2-\sin^2\theta\right) \tag{77}$$

Hence, at a given angular position, the change $\Delta m$ in order corresponding to a change $\Delta\lambda_0$ in wavelength is, to first order of small quantities,

$$\Delta m = \frac{4h^2 n'\frac{dn'}{d\lambda_0}-m^2\lambda_0}{m\lambda_0^2}\Delta\lambda_0. \tag{78}$$

The spectral range which corresponds to a change of unity in $m$, is therefore

$$(\Delta\lambda_0)_{S.R.} = \frac{m\lambda_0^2}{\left|m^2\lambda_0-4h^2 n'\frac{dn'}{d\lambda_0}\right|}; \tag{79}$$

or, if we substitute for $m$ from (77), and put $\sin\theta \sim 1$ near grazing emergence,

$$(\Delta\lambda_0)_{S.R.} \sim \frac{\lambda_0^2}{2h}\frac{\sqrt{n'^2-1}}{\left|n'^2 - n'\lambda_0\frac{dn'}{d\lambda_0} - 1\right|}. \tag{80}$$

We see that $(\Delta\lambda_0)_{S.R.}$ is inversely proportional to the plate thickness. For quartz the factor $\sqrt{n'^2-1}/\left(n'^2 - n'\lambda_0\frac{dn'}{d\lambda_0} - 1\right)$ varies from about 0.6 at $\lambda_0 = 2000$Å to about 0.8 at $\lambda_0 = 6000$Å.

If we assume $\mathscr{R}$ constant, the points in the fringe pattern where $I^{(t)}(p)$ has fallen to half its maximum values correspond to $\delta = 2m\pi \pm \varepsilon/2$ ( $m$ integral), where from (75)

$$\frac{1 + G_p \sin^2\left(\frac{p\varepsilon}{4}\right)}{1 + F\sin^2\left(\frac{\varepsilon}{4}\right)} = \frac{1}{2},$$

i.e.

$$\sin^2\left(\frac{p\varepsilon}{4}\right) - \frac{F}{2G_p}\sin^2\left(\frac{\varepsilon}{4}\right) + \frac{1}{2G_p} = 0. \tag{81}$$

Evidently the finesse $\mathscr{F} = 2\pi/\varepsilon$ depends on both $\mathscr{R}$ and $p$, and it is shown as a function of $\mathscr{R}$ for several values of $p$ in Fig. 7.70. By (74), $p$ is determined essentially by the ratio $l/h$.$\mathscr{R}$ is of course a function of $\theta'$, and therefore of $\chi$ and for given $\chi, \mathscr{R}_\perp$ is greater than $\mathscr{R}_\parallel$(for $n' > n$ ); it is therefore advantageous to select the polarized component vibrating perpendicular to the plane of incidence. The limiting case $p \to \infty$ corresponds to a plate of infinite length, for which the intensity distribution is similar to that for a Fabry-Perot interferometer (cf. (70) and (27)); in this case, the finesse is given by (35). When $p$ is finite, Fig. 7.70 shows that $\mathscr{F}$ has the value corresponding
to $p = \infty$ so long as $\mathscr{R}$ is not too large. For such values of $\mathscr{R}, \mathscr{R}^p$ (and hence $G_p$ ) is negligible, and substantially all the light which enters the plate emerges to form the fringes. For larger values of $\mathscr{R}$ the value of $\mathscr{F}$ is smaller than that corresponding to $p = \infty$, and as $\mathscr{R}$ approaches unity, $\mathscr{F}$ approaches a maximum value

$$\mathscr{F}_l = \frac{2\pi}{\varepsilon_l}, \tag{82}$$

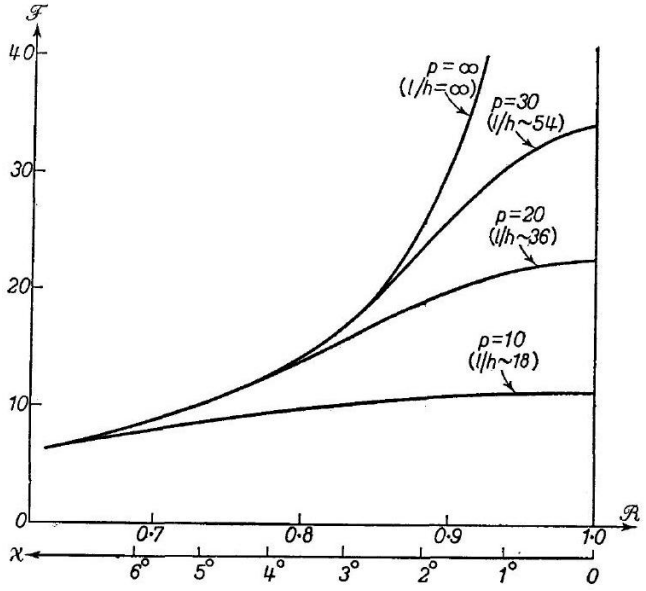



Fig. 7.70. Variation of finesse $\mathscr{F}$ with the reflectivity $\mathscr{R}$ and the number $p$ of reflections for a Lummer-Gehrcke interferometer. (For comparison, values of the angle $\chi$ from the centre of the pattern, and the ratio of plate length to plate thickness $l/h$ are shown for a quartz plate.)

where the limiting value $\varepsilon_l$ can be obtained with the help of (81). Thus setting $\mathscr{R} = 1-\mathscr{T}$, we have, for $\mathscr{T}$ small, $1-\mathscr{R}^p \sim p\mathscr{T}$, so that $F/G_p = (1-\mathscr{R}^p)^2/\mathscr{R}^{p-1}(1-\mathscr{R})^2 \sim p^2/(1-\mathscr{T})^{p-1} \to p^2$ as $\mathscr{T} \to 0$; and since $1/G_p \to 0$ as $\mathscr{T} \to 0$, (81) gives

$$\sin^2\left(\frac{p\varepsilon_l}{4}\right) - \frac{p^2}{2}\sin^2\left(\frac{\varepsilon_l}{4}\right) = 0. \tag{83}$$

Hence for $\varepsilon_l$ small,

$$\frac{\sin\frac{p\varepsilon_l}{4}}{\frac{p\varepsilon_l}{4}} = \frac{1}{\sqrt{2}}, \quad \frac{p\varepsilon_l}{4} = 0\cdot 45\pi, \quad \mathscr{F}_l = 1\cdot 1p. \tag{84}$$

However, (75) shows that as $\mathscr{R} \to 1$ the intensity $I^{(t)}(p)$ in the pattern tends to zero; an increasing proportion of the light which enters the plate remains inside after the last reflection and is lost at the far end of the plate. If, as is usual, any appreciable loss of light is unacceptable, the interferometer must be used in the region where $\mathscr{F}$ has the value corresponding to $p = \infty$, and it can be seen from Fig. 7.70 that the highest value of $\mathscr{F}$ is then approximately $2\mathscr{F}_l/3 \sim 0.7p$. In this region the form of
the relative intensity distribution is similar to that given by the Fabry-Perot interferometer, so that, as we have seen on p.335 we may take the smallest resolvable wavelength difference to be approximately $1/\mathscr{F}$ times the spectral range $(\Delta\lambda_0)_{\text{S.R.}}$ Using (74) and (80) we thus obtain the resolving power corresponding to $\mathscr{F} \sim 0.7p$ as

$$\frac{\lambda_0}{\Delta\lambda_0} \sim \frac{0.7p\lambda_0}{(\Delta\lambda_0)_{S.R.}} \sim 0.7\frac{l}{\lambda_0}\left|n'^2 - n'\lambda_0\frac{dn'}{d\lambda_0} - 1\right|. \tag{85}$$

The resolving power given by (85) depends only on the length of the plate, and not at all on its thickness, but we must remember that this resolving power is reached at a value of $\theta$ which approaches $\pi/2$ as $l$ increases. In practice, the length $l$ of the plate is limited by technical difficulties of manufacture.

When made of crystalline quartz the Lummer-Gehrcke interferometer is transparent down to about 2000Å, and for long it was the best available interferometer for high resolution spectroscopy in the ultra-violet region. However, the development of ultra-violet reflecting coatings suitable for use in the Fabry-Perot interferometer, and the successful construction of reflecting echelon gratings, have removed this advantage, and the Lummer-Gehrcke interferometer is now only rarely employed in research.

## 9.10.6 7.6.6 Interference filters

Suppose a parallel beam of white light is incident normally on a plane parallel plate having surfaces of high reflectivity. We see from (27) that there are maxima of the intensity of the transmitted light when the phase difference $\delta$ is an integral multiple of $2\pi$, or, by (25), when $\lambda_0 = \lambda_0{}^{(m)}$ where

$$\lambda_0^{(m)} = \frac{2n'h}{(m - \phi/\pi)}, \quad m = 1, 2, \ldots \tag{86}$$

and that at wavelengths on either side of these

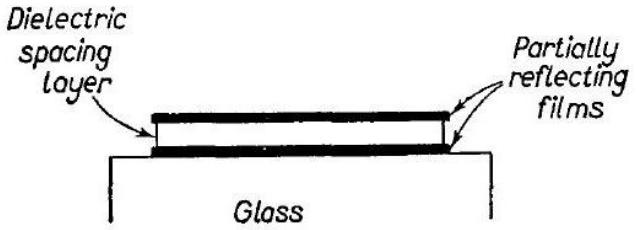



Fig. 7.71. Fabry-Perot type interference filter.

maxima the intensity of the transmitted light falls sharply to low values. The plate thus acts as a wavelength filter, with multiple transmission bands associated with integral values of the order $m$. In particular, if the optical thickness of the plate is only a few half-wavelengths of visible light, the transmission bands in the visible region are of low order and are widely separated in wavelength, so that it is usually possible to suppress the light in all but one of them, either by means of auxiliary absorbing filters, or by virtue of selective wavelength response of the detector used to make observations.

A filter of this kind may be prepared on the plane surface of a glass plate by depositing two reflecting films, separated by a spacing film of dielectric material (Fig. 7.71). If for normal incidence the filter is to have a transmission band of order $m$ at wavelength $\lambda_0{}^{(m)}$, the optical thickness $n'h$ of the spacing film must by (86) be $(m-\phi/\pi)\lambda_0{}^{(m)}/2$. With metal reflecting films, $\phi$ depends on $\lambda_0$ and the film thickness (and hence on the reflectivity $\mathscr{R}$ ). With multilayer dielectric reflecting films, $\mathscr{R}\left(\lambda_0{}^{(m)}\right)$ is maximum when the optical thickness of each layer is $\lambda_0{}^{(m)}/4$, and $\phi\left(\lambda_0{}^{(m)}\right)$ is then zero (cf. §1.6.); hence in this case the optical thickness of the spacer required is $m\lambda_0{}^{(m)}/2$.

Important characteristics of the filter are the peak-transmission $\tau$, defined as for the Fabry-Perot interferometer (eq.(33)); and the wavelength half-width $(\Delta\lambda_0)_{H.W.}$, which is defined as the interval between wavelengths in the transmission band at which
$I^{(t)}/I^{(i)}$ has fallen to half its maximum value. If the optical separation of the reflecting surfaces is constant over the aperture of the filter, $\tau$ is given by (36); and by (21), $(\Delta\lambda_0)_{\text{H.W.}}$ corresponds to a change of $\delta$ by an amount $4/\sqrt{F}$, providing we can neglect variations of $\mathscr{R}$ and $\mathscr{T}$ with wavelength over this wavelength interval. Now from (25), for normal incidence, and if we neglect

dispersion in the spacer, a change $\Delta\delta$ in $\delta$ corresponds to a change $\Delta\lambda_0$in wavelength given to the first order of small quantities by

$$\Delta\delta = \left(-\frac{4\pi}{\lambda_0^2}n'h + 2\frac{d\phi}{d\lambda_0}\right)\Delta\lambda_0; \tag{87}$$

hence near the transmission band of order $m$, we have, using (86),

$$\Delta\delta = \frac{2\pi}{\lambda_0(m)}\left\{-m + \frac{1}{\pi}\left[\frac{d}{d\lambda_0}(\phi\lambda_0)\right]_{\lambda_0(m)}\right\}\Delta\lambda_0. \tag{88}$$

The wavelength half-width of the filter, which as we have just seen corresponds to a change in $\delta$ of $4/\sqrt{F}$, is therefore

$$(\Delta\lambda_0)_{H.W.} = \frac{2{\lambda_0}^{(m)}}{\pi\sqrt{F}\Big|_{m-\frac{1}{\pi}\left[\frac{d}{d\lambda_0}(\phi\lambda_0)\right]_{\lambda_0(m)}}} = \frac{{\lambda_0}^{(m)}}{\mathscr{F}\left|\, m - \frac{1}{\pi}\left(\frac{d}{d\lambda_0}(\phi\lambda_0)\right)_{\lambda_0(m)}\right.}, \tag{89}$$

where the relation (22) has been used. When the reflecting films are of metal, $d\phi/d\lambda_0$ is so small that $d(\phi\lambda_0)/d\lambda_0 \sim \phi$. The phase change effects when the reflecting films involve dielectric multilayers have been discussed by Dufour,* who has shown that in the visible spectrum, with zinc sulphide and cryolite as the dielectric materials, $\frac{d}{d\lambda_0}(\phi\lambda_0)/\pi$ lies between about -1.0 and $-1.5$†.

We have seen that with available reflecting films $\tau$ decreases with increasing $\mathscr{F}$, so that for filters of given order prepared with given materials, smaller half-widths are associated with lower peak transmissions. In practice, as with the Fabry-Perot interferometer, an upper limit to the useful reflectivity is set by variations of the optical separation of the reflecting surfaces over the aperture. In this connection, however, departures from planeness of the glass surface on which the filter is prepared are not important, since the evaporated films follow the contours of the surface on which they are deposited. With a suitable arrangement for preparing the filter, the variation of optical separation of the reflecting surfaces can be made much smaller than is possible for the two independently worked surfaces of the Fabry-Perot interferometer, and correspondingly higher reflectivities may be employed. Typical performance data for filters with metal and metal-dielectric reflecting films are shown in Table XVI, and for all dielectric filters in Table XVII. It is seen that the highest peak-transmissions for given half-width are obtained with reflecting films which involve dielectric layers. We may note that, because of the wavelength dependence of the reflecting properties of such films, filters in which they are used have secondary transmission bands in addition to those given by (86).

Since filters are usually required to transmit a specified wavelength, the optical thickness of the spacer must be accurately controlled. From (86), a change $\Delta(n'h)$ in $n'h$ corresponds to a displacement $\Delta{\lambda_0}^{(m)}$ of the transmission band of order $m$ given

Table XVI Performance of interference filters with metal and metal-dielectric reflecting films. $M$ denotes a metal film (silver), $L$ denotes a quarter wave layer of low refractive index dielectric (magnesium fluoride), and $H$ denotes a quarter wave layer of high refractive index dielectric (zinc sulphide).

Performance of zinc sulphide-cryolite interference filters. $L$ denotes a quarter wave layer of cryolite, $\boldsymbol{H}$ denotes a quarter wave layer of zinc sulphide.

0

- C. Dufour, Rev. d'Opt., 31 (1952), p. 1.

† In discussing the resolving power of the Fabry-Perot interferometer, we neglected wavelength dependence of $\phi$ (cf. (42) and (88)). This is evidently justified when $m$ is sufficiently large.

From A. F. Turner, J. Phys., 11 (1950), 457.

| Type of filter | Wavelength $\lambda_0{}^{(m)}$ of maximum transmission | Peak transmission $\tau$ | Half-width |
|---|---|---|---|
| M - 2 L – M. . . | 5310A | 0.30 | 130 A |
| M-4L-M . . . | 5350 A | 0.26 | 70 A |
| MLH-2L-HLM | 5470 A | $0 \cdot 43$ | 48 A |
| MLHLH-2L-HLHLM | 6050 | 0.38 | 20 A |

From P. H. Lissberger and J. Ring, Optica Acta, 2 (1955), 45.

Table XVII

| Type of filter | Wavelength $\lambda_0{}^{(m)}$ of maximum transmission | Peak transmission |
|---|---|---|
| HLH-2L-HLH | 5185 A | 0.90 |
| HLHLH-2L-HLHLH | 4750 A | 0.85 |
| HLHLHL-2H-LHLHLH . | 6565 A | 0.90 |
| HLHLHLH-2L-HLHLHLH | 5200 A | $0 \cdot 70$ |
| HLHLHLHL-2H-LHLHLHLH | 6500 A | 0.80 |
| HLHLHLHLHL-2H-LHLHLHLHLH | 6600 A | 0.50 |

by $\Delta\lambda_0{}^{(m)} = \lambda_0{}^{(m)}\Delta\left(n'h\right)/n'h$. Thus with $\lambda_0{}^{(m)} = 5000$Å, for example, an error of 1 per cent in the optical thickness of the spacer results in an error of 50Å in the position of the pass-band, which is greater than the half-width of narrow band filters. Suitable methods for monitoring the thickness of the spacer during deposition have been described by Greenland and Billington,* Glacomo and Jacquinot † and Lissberger and Ring ‡. We see also from (25) that the pass-band may be adjusted to shorter wavelengths by tilting the filter so that the incidence is no longer normal, but the performance deteriorates as the tilt increases. In particular, when the reflecting films are of metal, the phase changes $\phi_0$ and $\phi_\perp$ for light polarized with its electric vector parallel and perpendicular to the plane of incidence become unequal, and the pass-bands for the two polarized components are at different wavelengths.

The wavelength displacement of the pass-band for given tilt depends on $\theta'$, the angle of reflection in the spacer, and so is lower the higher the refractive index of the spacer. For this reason it is advantageous to make the spacer of high refractive index material if the filter is to be used in a convergent beam.

In another form of interference filter-the so-called frustrated total reflection filter, first described by Leurgans and Turner*-each reflecting film is a thin layer of low refractive index material between media of higher refractive index. We have seen in § 1.5.4 that if such a layer is sufficiently thin, and is illuminated at an angle greater than the critical angle, the reflection is not total; some light passes through

0

- K. M. Greenland and C. Billington, Journ. de Phys., 11 (1950), 418.
  † P. Glacomo and P. Jacquinot, J. Phys., 13 (1952), 59A.
  ‡ P. H. Lissberger and J. Ring, Optica Acta 2 (1955), 42.

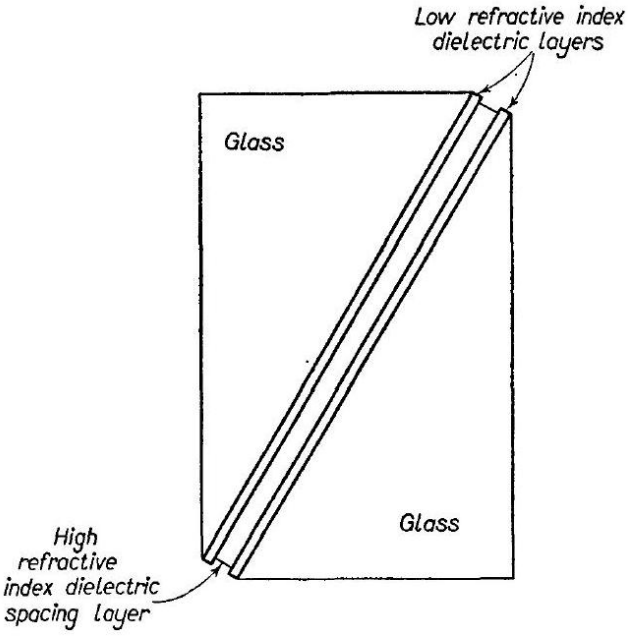



Fig. 7.72. A frustrated total reflection interference filter.

the layer, which thus acts as a non-absorbing reflector, and any desired reflectivity may be obtained by adjusting the layer thickness. The construction of the filter is shown in Fig. 7.72. The hypotenuse face of a prism of dense glass is coated with a

Table XVIII Performance data of some frustrated total reflection filters, with reflecting layers of magnesium fluoride on dense flint glass, and spacing layers of zinc sulphide.

From A. R. Turner, J. Phys., 11 (1950), 458.

| Order of interference | 1 | | 2 | | 1 | |
|---|---|---|---|---|---|---|
| Polarization | $\|\|$ | $\perp$ | $\|\|$ | $\perp$ | $\|\|$ | $\perp$ |
| Wavelength of maximum transmission $\lambda_0{}^{(m)}$ | 4600 A | 5300Å | 4630Å | 5080 A | - | 5110 A |
| Peak-transmission $\tau$ | $> 0.93$ | $> 0.93$ | 0.90 | 0.90 | - | 0.12 |
| Half-width $\|\Delta\lambda_0\|_{H,W}$. | 66Å | 120Å | 50Å | 49Å | - | 30Å |

layer of low refractive index, followed by a spacing layer of high refractive index and a further layer of low refractive index; and a second similar prism is attached to the first by means of a cement of the same refractive index as the glass. The prism angle and the refractive indices are chosen so that light incident normally on the prism base meets the low index layers at an angle greater than the critical angle. The wavelengths of the transmission bands depend on the optical thickness of the spacer, the angle of reflection in it, and the phase change on reflection; the reflectivity (and hence the half-width for given order), depends on the thickness of the low refractive index layers. The phase change depends on the state of polarization, and so the transmission bands of given order for the components vibrating parallel and perpendicular to the plane of incidence are at different wavelengths. Performance data for typical filters prepared for use in the visible spectrum are given in Table XVIII. Filters of this type have also been made for use with centimetre waves.*

0

- P. Leurgans and A. F. Turner, J. Opt. Soc. Amer., 37 (1947), 983.

## 9.10.7 7.6.7 Multiple-beam fringes with thin films

In §7.6.1 we have seen how the fringes of equal inclination from a plane parallel plate become much narrower when the reflectivity of the surfaces is increased. Increased reflectivity has a similar effect on the intensity distribution of the Fizenu fringes of equal thickness (§7.5.2) given by a thin film, with the result that finer details of the thickness contours of the film are revealed.

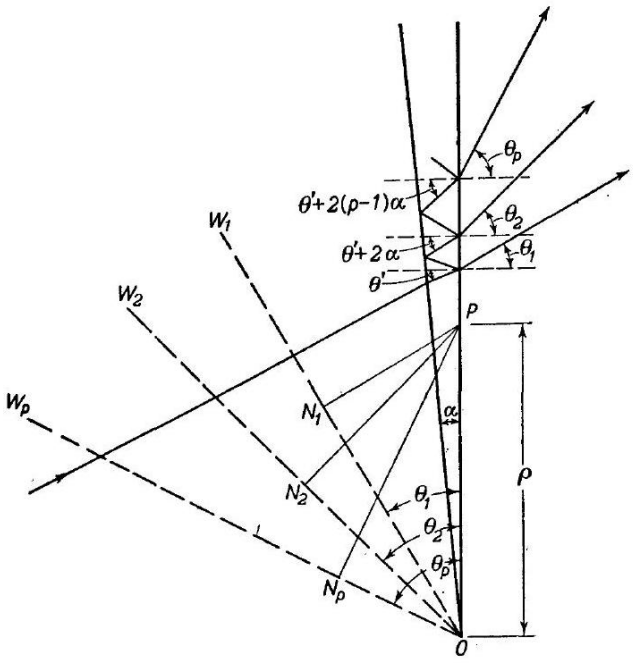



Fig. 7.73. Illustrating multiple reflections in a wedge.

For simplicity we consider first the case of a film in the form of a wedge with plane surfaces inclined at a small angle $\alpha$, illuminated with a plane wave of monochromatic light propagated at right angles to the wedge apex 0.† (Fig. 7.73.) Because of multiple reflections at the surfaces, the transmitted light consists of a set of plane waves propagated in different directions. If the wave transmitted through the first surface approaches the second surface from the side of the normal nearer the wedge apex

and is incident at angle $\theta'$, the $p$ th wave of the transmitted set emerges at angle $\theta_p$ where, from the laws of refraction and reflection,

$$n \sin \theta_p = n' \sin [\theta' + 2(p-1)\alpha] \tag{90}$$

$n'$ being the refractive index of the film, and $n$ that of the surrounding medium. The virtual transmitted wave-fronts $W_1, W_2, \ldots W_p, \ldots$, which contain the wedge apex would be co-phasal if there were no phase change at reflection. At a point $P$ on the second surface, at distance $\rho$ from $O$, the difference of optical path of the $p$ th wave and the wave which is directly transmitted ( $p = 1$ ) is therefore

$$\Delta \mathscr{S}_p = n\,(PN_p - PN_1) = n\rho\,(\sin \theta_p - \sin \theta_1)\,, \tag{91}$$

0

- B. H. Billings, J. Opt. Soc. Amer., 39 (1949), 634.
  † The analysis given here is essentially due to J. Brossel, Proc. Phys. Soc., 59 (1947), 224.

where $N_p, N_1$ are the feet of the perpendiculars from $P$ to $W_p, W_1$ respectively. If $\phi$ is the phase change arising from a single reflection at either surface of the wedge, the total phase difference $\delta_p$ between the $p$ th wave and the directly transmitted wave is, if (90) and (91) are used,

$$\begin{aligned}\delta_p &= \frac{2\pi}{\lambda_0}\Delta\mathscr{S}_p + 2(p-1)\phi \\ &= \frac{2\pi}{\lambda_0}n'\rho\left\{\sin\left[\theta' + 2(p-1)\alpha\right] - \sin\theta'\right\} + 2(p-1)\phi \\ &= \frac{4\pi}{\lambda_0}n'h\cos\theta'\frac{\sin(p-1)\alpha}{\tan\alpha}\left\{\cos(p-1)\alpha - \tan\theta'\sin(p-1)\alpha\right\} + 2(p-1)\phi \end{aligned} \tag{92}$$

where

$$h = \rho\tan\alpha$$

is the thickness of the film at $P$. If the wave transmitted through the first surface meets the second surface from the side of the normal away from the wedge apex, the wave-front $W_1$ is real. In this case $\delta_p$ is given by (92), but with the sign of the term in $\tan\theta'$ changed.

With notation similar to that of §7.6.1, the amplitudes of the transmitted waves are

$$tt'A^{(i)}e^{i\delta_1}, \quad tt'r'^2A^{(i)}e^{i\delta_2}, \ldots\ldots tt'r'^{2(p-1)}A^{(i)}e^{i\delta_p}, \ldots$$

and the amplitude at $P$ resulting from the superposition of the infinity of such waves is

$$A^{(t)} = A^{(i)}tt'\sum_{p=1}^{\infty} r'^{2(p-1)}e^{i\delta_p} = A^{(i)}\mathscr{T}\sum_{p=1}^{\infty}\mathscr{R}^{p-1}e^{i\delta_p} \tag{93}$$

The corresponding intensity is

$$I^{(t)} = I^{(i)}\mathscr{T}^2\left|\sum_{p=1}^{\infty}\mathscr{R}^{p-1}e^{i\delta_p}\right|^2 \tag{94}$$

Let us examine the consequences of retaining only a finite number of terms in the series in (93) and (94). $\left|A^{(t)}\right|$ cannot be greater than $\left|A^{(i)}\right|\mathscr{T}/(1-\mathscr{R})$, and if only the first $p$ terms are taken into account, the error $\left|\Delta A^{(t)}\right|$ in $A^{(t)}$ cannot exceed $\left|A^{(i)}\right|\mathscr{T}\mathscr{R}^p/(1-\mathscr{R})$. The corresponding relative error $\left|\Delta I^{(t)}\right|/I^{(t)}$ satisfies the relation

$$\frac{\left|\Delta I^{(t)}\right|}{I^{(t)}} \leqslant 2\left|\frac{\Delta A^{(t)}}{A^{(t)}}\right| \leqslant 2\left|\frac{A^{(i)}}{A^{(t)}}\right|\frac{\mathscr{T}\mathscr{R}^p}{(1-\mathscr{R})}. \tag{95}$$

We shall see later that in the conditions of most practical interest, $I^{(t)}$ is given nearly enough by the Arry formula (13), according to which the minimum intensity corresponds to $\left|A^{(t)}\right| = \left|A^{(i)}\right|\mathscr{T}/(1+\mathscr{R})$. Thus if for this particular value of $\left|A^{(t)}\right|$ we wish $\left|\Delta I^{(t)}\right|/I^{(t)}$ to be not greater than 0.01 say, which means that the relative error of intensities in the neighbourhood of maxima of $I^{(t)}$ will be much less than $0\cdot01$, we find from (95) that the number $p$ of terms which have to be retained in the series is given by the relation

$$\mathscr{R}^p \sim \frac{1}{200}\frac{1-\mathscr{R}}{1+\mathscr{R}}. \tag{96}$$

With $\mathscr{R} = 0.93$, for example, $p \sim 120$.

With only a finite number of terms in the summation of (94), and providing $\alpha$ is sufficiently small, we may expand (92) in powers of $\alpha$ and retain terms only up to the second power, whence

$$\delta_p = (p-1)\left\{\frac{4\pi}{\lambda_0} n'h\cos\theta' + 2\phi\right\} - (p-1)^2\alpha\frac{4\pi}{\lambda_0} n'h\sin\theta' - \frac{(p-1)\,(2p^2-4p+3)\,\alpha^2}{3}\cdot\frac{4\pi n'h\cos\theta'}{\lambda_0} \tag{97}$$

In particular, for normal incidence ( $\theta' = 0$ ),

$$\delta_p = (p-1)\left(\frac{4\pi n'h}{\lambda_0} + 2\phi\right) - \frac{(p-1)\,(2p^2-4p+3)\,\alpha^2}{3}\cdot\frac{4\pi n'h}{\lambda_0} \tag{98}$$

When the term of (98) in $\alpha^2$ is small compared with $\pi$, we may neglect it and take the phase difference $\delta_p - \delta_{p-1}$ of successive waves to be constant and equal to $\delta = 4\pi n'h/\lambda_0 + 2\phi$. $I^{(t)}$ is then given by the Airy formula (13), so that there are fringes in the plane of the wedge with an intensity distribution similar to that of the fringes at infinity from a plane parallel plate (Fig. 7.58). The positions of the intensity maxima, which correspond to $\delta = 2m\pi$, are given by

$$2n'h = \left(m - \frac{\phi}{\pi}\right)\lambda_0, \quad m = 1, 2, \ldots \tag{99}$$

Thus the fringes are parallel to the wedge apex; they are spaced at intervals $\lambda_0/2n'\alpha$. If the reflectivity is high ( $\mathscr{R} \gtrsim 0\cdot 9$ ), so that $p$ is large and we may replace the term $(p-1)\,(2p^2-4p+3)$ by $2p^3$, the condition that the term of (98) in $\alpha^2$ should be small compared with $\pi$ requires

$$n'h \ll \frac{3\lambda_0}{8p^3\alpha^2} \tag{100}$$

With values of $\alpha$ corresponding to a distance between adjacent fringes of the order of 1 cm or less, this condition represents a severe restriction of $n'h$, and the fringes have the Arry intensity distribution only in the region of the wedge near the apex where the optical thickness is a few wavelengths. For example, with $\alpha \sim 2.5\times 10^{-4}$ (which with $n' \sim 1, \lambda_0 = 5500$Å corresponds to about 1 fringe per mm), $p \sim 50$ ( $\mathscr{R} \sim 0.9$ ), (100) gives $n'h \ll 50\lambda_0$. Further from the wedge apex, the term of (98) in $\alpha^2$ becomes increasingly significant, and numerical calculations show (Fig. 7.74) that $I^{(t)}$ begins to be affected in three ways: the intensities of the maxima become less, and their half-widths greater, than the values given by the Airy formula; the maxima are displaced from the positions given by (99), away from the wedge apex; and the fringes become asymmetrical, with secondary maxima on the side away from the wedge apex.

According to (99) the fringes follow lines of equal optical thickness when the surfaces of the film are plane. However, the important practical application of the fringes is to the examination of films whose surfaces are not plane, and we must now consider how closely the fringes follow thickness contours in such circumstances. For the fringes to contour an irregular film exactly, the intensity at each point $P$ must be

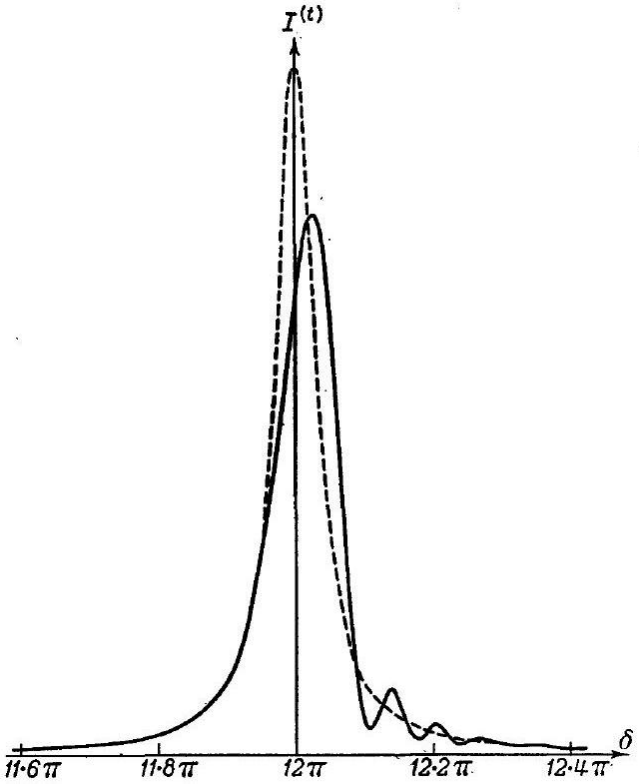



Fig. 7.74. Intensity distribution of multiple-beam Fizeau fringe (full line), compared with corresponding Arry distribution (chain line).

(After K. Kinosita, J. Phys. Soc. Japan, 8 (1953), 219.)

uniquely determined by the film thickness $h$ at $P$, and in general this is not so. A ray which reaches $P$ after $2(p-1)$ reflections (Fig. 7.75(a)) enters the film at $A_1$ and is reflected at $B_1, C_1, \ldots B_{p-1}, C_{p-1}$, and the corresponding contribution to the resultant amplitude at $P$ depends on the optical thickness of the film between $A_1$ and $P$. The fringes may thus be displaced from the true contours of optical thickness by an amount which depends on the shape of the film in the vicinity of $P$. The effect is evidently more important the higher the reflectivity, but for given reflectivity it is smallest where the film thickness is smallest, and when the incidence is near normal. We may estimate the extent of the region of the film which influences the intensity at $P$ if we assume that over it the surfaces are plane (Fig. 7.75(b)). Let $A_1', A_1'', \ldots A_{p-1}', A_{p-1}''$ be the images of $A_1$ corresponding to successive reflections at the second and first surfaces of the film. These images, which are in the principal section of the wedge through $A_1$, lie on a circle with centre $O$ at the wedge apex and radius $OA_1 = \rho'$; and the angle $A_1OA_{p-1}'' = 2(p-1)\alpha$. The image $A_{p-1}''$ is associated with the portion $C_{p-1}P$ of the ray so that $C_{p-1}P$ produced must pass through $A_{p-1}''$. If we take the ray at $B_1$ to be in a principal section of the wedge, and to be incident at angle $\theta'$ from the side
of the normal nearer $O$, the angle of incidence of the ray when it reaches $P$ is $\theta' + 2(p-1)\alpha$. In the triangle $A_{p-1}''PO$ we then have $OA_{p-1}'' = \rho'$, and the angles

$$A_p''PO = \pi/2 - [\theta' + 2(p-1)\alpha]\,, \quad PA_p''O = \pi/2 + (\theta' - \alpha)\,;$$

so that putting $OP = \rho$, we have $\rho'/\cos[\theta' + 2(p-1)\alpha] = \rho/\cos(\theta' - \alpha)$. Hence

$$\Delta\rho = \rho - \rho' = \frac{h}{\tan\alpha}\left\{1 - \frac{\cos[\theta' + 2(p-1)\alpha]}{\cos(\theta' - \alpha)}\right\}, \tag{101}$$

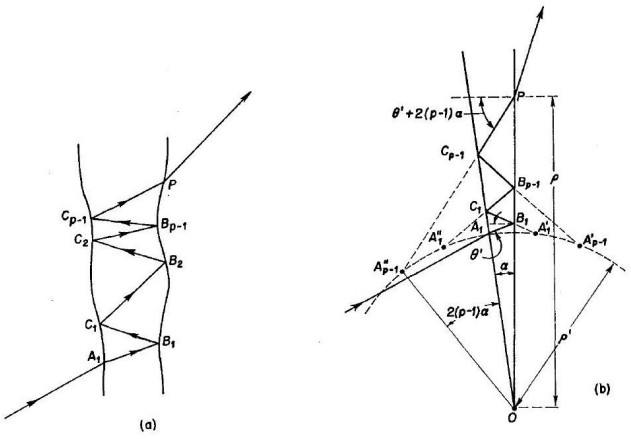



Fig. 7.75. Illustrating the displacement of a multiple reflected ray along a wedgeshaped film. (a) Irregular surfaces; (b) Plane surfaces.

and for $\theta' = 0$, and neglecting powers of $\alpha$ higher than the second, this reduces to

$$\Delta\rho = \frac{4p^2 - 8p + 3}{2} h\alpha. \tag{102}$$

Thus if there are $p$ significant beams we may expect the intensity at $P$ to be influenced by the film thickness over a distance of the order of $2p^2h\alpha$ about $P$. Even with high reflectivities this distance is small when $h$ is only a few wavelengths of visible light. For example, with $p \sim 50, h = 2.5 \times 10^{-4}$ cm (about five wavelengths), and $\alpha = 2.5 \times 10^{-4}$, it amounts to about $3 \times 10^{-3}$ mm. Providing small irregularities of the fringes are interpreted with caution, the fringes delineate optical thickness contours of irregular films when the film thickness is sufficiently small.

We have assumed so far that the source is a point, but in practice, because of intensity considerations, it is desirable to use an extended source of maximum possible size. According to (97), if we can neglect terms in $\alpha$ and $\alpha^2$, the intensity maxima for non-normal incidence occur when

$$2n'h\cos\theta' = \left(m - \frac{\phi}{\pi}\right)\lambda_0, \quad m = 1, 2, \ldots \tag{103}$$

and since this relation rests on neglect of the wedge angle it is valid for all planes of incidence. Hence if the source is extended radially, so that the angles of incidence cover the range $0 \leqslant \theta' \leqslant \varepsilon'$, the maxima of given order spread over a distance corresponding to $\Delta m$ orders, where from (103), for $\varepsilon'$ small,

$$\Delta m = \frac{2}{\lambda_0} n'h\,|\Delta(\cos\theta')| = \frac{n'h\varepsilon'^2}{\lambda_0} \tag{104}$$

The importance of this fringe broadening due to the source extension depends on the finesse $\mathscr{F}$ of the fringes formed by a single element of the source: an increase of $\varepsilon'$ will result mainly in increased intensity of the maxima when $n'h\varepsilon'^2/\lambda_0 \ll 1/\mathscr{F}$, but mainly in increased fringe width when $n'h\varepsilon'^2/\lambda_0 \gg 1/\mathscr{F}$. For a closer examination of the effect we may take advantage of its formal similarity to the broadening of Fabry-Perot fringes which results from spherical curvature of the interferometer plates (pp. 332-333). Appropriate modification of the results of Dufour and Picca* shows that there is little increase of intensity of the maxima once $n'h\varepsilon'^2/\lambda_0$ exceeds $1/\mathscr{F}$, and a reasonable criterion is that it should have half this value, the fringes then being about 10 per cent broader than with a point source. Taking into account refraction at

the first surface of the film, we may therefore take the tolerable angular radius $\varepsilon$ of the source to be

$$\varepsilon \sim \frac{n'\varepsilon'}{n} \sim \frac{1}{n}\sqrt{\frac{\lambda_0 n'}{2h\mathscr{F}}}, \tag{105}$$

which may be compared with the analogous relation § 7.5 (73) relating to two-beam interference. When $h$ is only a few wavelengths, $\varepsilon$ is not very small; for example, with $h \sim 5\lambda_0, \mathscr{F} = 30(\mathscr{R} \sim 0 \cdot 9), n' \sim n \sim 1$, (105) gives $\varepsilon \sim 3°$. Thus the incident light need not be strictly collimated, and the large sources permitted are advantageous for observation of the fringes under high magnification.

When the light is quasi-monochromatic, with wavelength components covering a wavelength range $\Delta\lambda_0$ about mean wavelength $\bar{\lambda}_0$, the maxima of given order spread over a distance which corresponds to $\Delta m$ orders, where from (99), if we neglect wavelength dependence of $n'$ and $\phi, \Delta m = 2n'h\Delta\lambda_0/\overline{\lambda_0{}^2}$. The fringes are not significantly broadened by a departure from strict monochromatism if this quantity is small compared with $1/\mathscr{F}$, i.e. if

$$\Delta\lambda_0 \ll \frac{\left(\bar{\lambda}_0\right)^2}{2n'h\mathscr{F}}. \tag{106}$$

Again, when $n'h$ is only a few wavelengths, this is not a severe condition; for example, with $n'h \sim 5\lambda_0, \mathscr{F} = 30, \lambda_0 = 5500\text{Å}$, (106) gives $\Delta\lambda_0 \ll 18\text{Å}$. In consequence the source is not restricted to the type familiar in interferometry with large path differences, but instead high pressure arc discharges may be employed, which are very bright and are further advantageous for observations with high magnification.

We have been concerned above with light transmitted through the film, but a multiple-beam interference pattern may also be observed with reflected light. When there is no absorption at the reflecting surfaces, this pattern is complementary to the transmitted pattern in the sense that the sum of the intensities in the two is at each point equal to the incident intensity. The effect of absorption, which is always

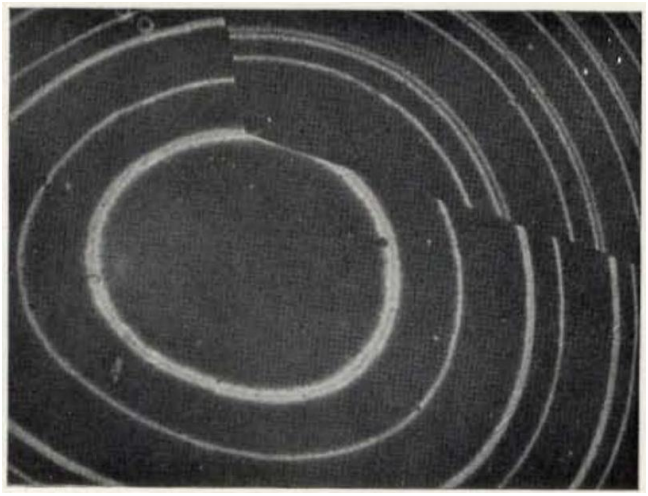


Fig. 7.76. Multiple-beam Fizeau fringes over a cleavage surface of mica (transmitted light). The fringes are given by the green ( 5461Å ) and yellow ( 5770Å, 5790Å ) lines of mercury. (After W. L. Wilcock.)

0

- C. Dufour and R. Picca, Rev. d'Opt., 24 (1945), 19.

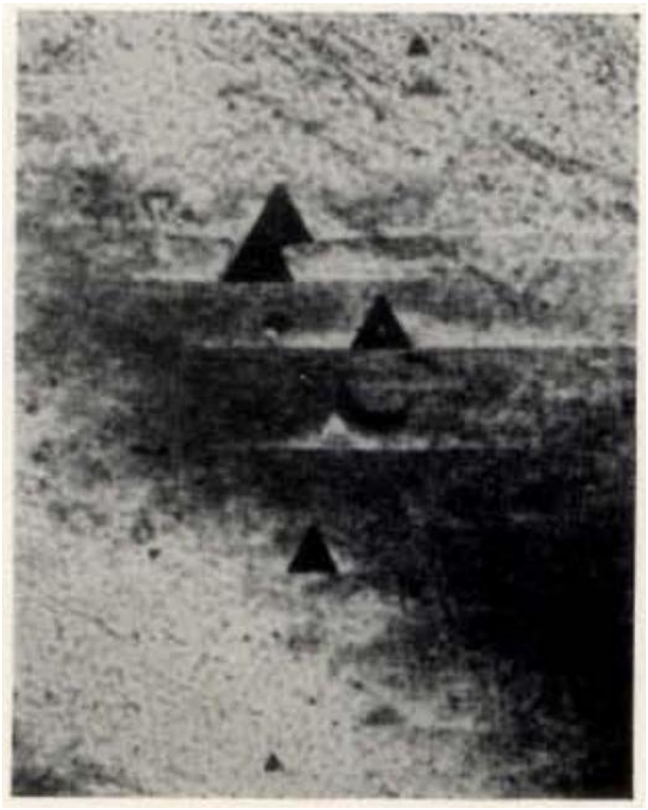


Fig. 7.77. Multiple-beam Fizeau fringe over a crystal surface of diamond.
(After W. L. Wilcock.)

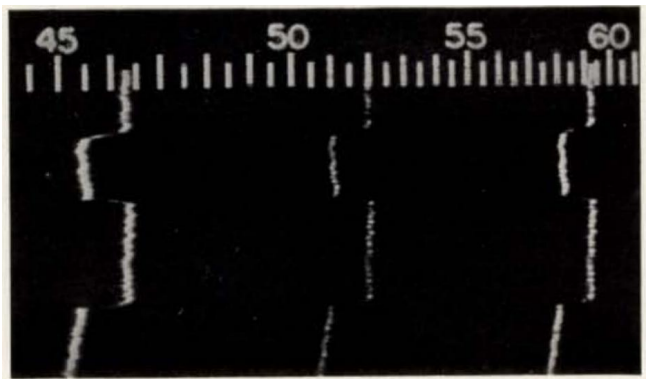



Fig. 7.79. Fringes of equal chromatic order given by a section of a diamond crystal surface. The scale is of wavelength in hundreds of Angstroms.
(After S. Tolansky and W. L. Wilcock, Proc. Roy. Soc., A, 191 (1947), 192.)

associated with practical reflecting coatings, has been considered by Holden.* He has shown that, with silver coatings of high reflectivity and low absorption, the pattern in reflected light is of dark fringes on a nearly uniform background; that these fringes are narrower than the corresponding bright fringes in transmitted light; but that the intensity at the minima of the reflected pattern is critically dependent on the absorption in the coating on the entrant surface of the film, and if this absorption is too high the fringes are invisible.

Multiple-beam Fizeau fringes are used in optical workshops to test the figure of high grade optical flats such as are required for the Fabry-Perot interferometer, and they have been employed extensively by Tolansky and his co-workers to study the topography of crystal and

metal surfaces which are nearly plane. † For this, the surface to be examined and the surface of an optical flat-each with a reflecting coating of silver-are held as close together as possible, and fringes in the air film between them are observed with a microscope of sufficient angular aperture to collect all the significant beams. If $\lambda = \lambda_0/n'$ is the wavelength in air, the fringes are contours of the unknown surface defined by planes parallel to the optical flat at intervals of $\lambda/2$. When the mutual inclination of the surfaces is sufficient to give a number of fringes in the field of view (see, for example, Fig. 7.76), the topography of the unknown surface can be quantitatively evaluated by measuring the deviation of the fringes from straight lines. In crossing a surface feature of height $\Delta h$, a fringe suffers a lateral displacement corresponding to $\Delta m$ orders, where

$$\Delta h = \frac{\lambda}{2}\Delta m; \tag{107}$$

so that $\Delta h$ can be determined if $\Delta m$ is known, and $\Delta m$ is, at least approximately, the ratio of the fringe displacement and the distance between adjacent fringes. When the reflectivity is high, surface features of very small height can be measured in this way. A displacement equal to the fringe half-width, which is easily measurable, corresponds to $\Delta h = \lambda/2\mathscr{F}$, and with $\lambda = 5500\text{Å}, \mathscr{F} = 40$, this is only about 70Å.‡ Whether a surface feature is an elevation or a depression can often be decided by observing the direction of motion of the fringes when the separation of the surfaces is altered, but in general it is necessary to use more than one wavelength. With fringes in transmitted light the several wavelengths can be used simultaneously, and the yellow lines of mercury $\lambda = 5770\text{Å}, 5790\text{Å}$ are convenient. The fringes then occur in close pairs whose separation increases with increasing order of interference; the direction of the local wedge angle is thus immediately visible, and if there is a discontinuity in the fringes, as when they cross a crystal cleavage step, it is possible to decide which are corresponding orders on opposite sides.

If the unknown surface is nearly plane, its inclination to the optical flat may be reduced until the whole field of view is occupied by only one fringe (see, for example, Fig. 7.77). Under these conditions extremely small variations of height can lead to detectable variations of intensity. Let $I, I + \Delta I$ be the intensities over adjacent areas with a small difference of height $\Delta h$, which corresponds to a change $\Delta\delta$ in $\delta$.

The contrast, defined by $|\Delta I|/I$, is maximum when $\delta$ is adjusted to make $|dI/d\delta|/I$ maximum. Considering the pattern in transmitted light, we have from (27)

$$\frac{I^{(t)}}{I^{(t)}_{\max}} = \frac{1}{1 + F\sin^2\frac{\delta}{2}} \tag{108}$$

whence

$$\frac{d}{d\delta}\left(\frac{I^{(t)}}{I^{(t)}_{\max}}\right) = \frac{-F\sin\delta}{2\left(1 + F\sin^2\frac{\delta}{2}\right)^2} \tag{109}$$

Hence

$$I^1\left|\frac{dI^{(t)}}{d\delta}\right| = \frac{F\sin\delta}{2\left(1 + F\sin^2\frac{\delta}{2}\right)} \tag{110}$$

0

• J. Holden, Proc. Phys. Soc. B., 62 (1949), 405.
† S. Tolansky, Multiple-beam Interferometry of Surfaces and Films (Oxford, Clarendon Press, 1948).
‡ Although the silver coating is about 500Å thick, it has been established that it does not alter the surface topography to an extent which is detectable by the technique.

so that for maximum contrast,

$$2\left(1+F\sin^2\frac{\delta}{2}\right)\cos\delta - F\sin^2\delta = 0 \tag{111}$$

When the reflectivity is high, $F$ is large and the values of $\delta$ which satisfy (111) are close to integral multiples of $2\pi$, so that we may write $\delta = 2m\pi \pm \varepsilon$, where $m$ is an integer and $\varepsilon$ is small compared with $\pi/2$. We then obtain from (111), neglecting powers of $\varepsilon$ higher than the second

$$\varepsilon = \frac{2}{\sqrt{F+2}} \sim \frac{2}{\sqrt{F}} \tag{112}$$

and (110) then gives

$$\frac{1}{I^{(t)}}\left|\frac{dI^{(t)}}{d\delta}\right|_{\max} \sim \frac{F\varepsilon}{2\left(1+\frac{F\varepsilon^2}{4}\right)} \sim \frac{\sqrt{F}}{2} \tag{113}$$

Hence to first order the maximum contrast is given by

$$\left|\frac{\Delta I^{(t)}}{I^{(t)}}\right|_{\max} = \frac{1}{I^{(t)}}\left|\frac{dI^{(t)}}{d\delta}\right|_{\max}|\Delta\delta| \sim \frac{\sqrt{F}}{2}|\Delta\delta|. \tag{114}$$

Since, from (99), $\Delta\delta$ and $\Delta h$ are related by $|\Delta\delta| = 4\pi|\Delta h|/\lambda$, and if the relation (22) is used, the maximum contrast may be expressed in the form

$$\left|\frac{\Delta I^{(t)}}{I^{(t)}}\right|_{\max} \sim \frac{4\mathscr{F}}{\lambda}|\Delta h|. \tag{115}$$

With $\lambda = 5500\text{Å}, \mathscr{F} = 40$, (115) gives $\left|\Delta I^{(t)}/I^{(t)}\right|_{\max} \sim 0.1$ when $\Delta h$ is only $\sim 3.5\text{Å}$, which is of the order of molecular dimensions.

We have considered so far the effects with quasi-monochromatic light. Suppose now that the film is illuminated at normal incidence with white light, and that by means of an achromatic lens $L$ the transmitted light is used to form an image of the film in the plane of the slit of a spectrograph (Fig. 7.78). For each wavelength component, the phase relationship of the beams reaching a point $P'$ of the slit is the same as at the point $P$ of the film conjugate to $P'$. Hence if the film is sufficiently
thin to satisfy condition (100), there are according to (99) maxima of intensity at $P'$ for those wavelengths ${\lambda_0}^{(m)}$ which satisfy the relation

$${\lambda_0}^{(m)} = \frac{2n'h}{m-\frac{\phi}{\pi}}, \quad m = 1, 2, \ldots \tag{116}$$

where $n'h$ is the optical thickness of the film at $P$; and the intensity at wavelengths between these maxima is given by the Arry formula. Thus when the reflectivity is high, the spectrum in the focal plane of the spectrograph is channelled with narrow bright fringes separated by much broader dark intervals, the wavelength separation of adjacent fringes being greater the thinner the film.

In the special case when $n'h$ is constant, ${\lambda_0}^{(m)}$ is constant and the fringes are straight lines parallel to the slit. These fringes are sometimes called Edser-Butler fringes. If $\phi$ is independent of wavelength they are spaced at equal intervals $\Delta\kappa_0 = 1/2n'h$

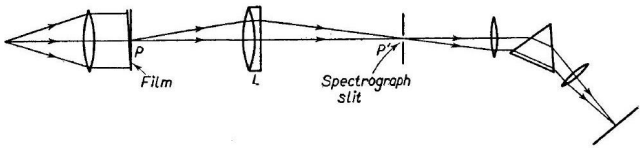



Fig. 7.78. Arrangement for the observation of white light fringes from a thin film.

of spectroscopic wave-number $\kappa_0 = 1/\lambda_0$, and because of this property they are useful for the calibration of spectrographs, particularly in the infra-red where standard wavelengths are not numerous. More generally, ${\lambda_0}^{(m)}$ varies along a fringe in a way which depends on the variation of optical thickness along the section of the film conjugate to the slit. In particular, if the film is an air film formed between an unknown surface and an optical flat, the fringes provide a powerful means of measuring the profile of a chosen section of the unknown surface. Such fringes, first described by Tolansky,* are sometimes called fringes of equal chromatic order (Fig. 7.79, f. p. 357).

Suppose the thicknesses of the film at two points $P_1, P_2$ are $h_1, h_2$, and let the maximum of order $m$ at the conjugate points $P_1', P_2'$ on the slit be at wavelengths in air ${\lambda_1}^{(m)}, {\lambda_2}^{(m)}$ respectively. Then using (116) we obtain, since $\lambda^{(m)} = {\lambda_0}^{(m)}/n'$,

$$\Delta h = h_2 - h_1 = \left(m - \frac{\phi_1}{\pi}\right) \frac{\lambda_2^{(m)} - \lambda_1^{(m)}}{2} + \frac{a}{2}\lambda_2^{(m)}, \tag{117a}$$

where

$$a = \frac{\phi_1 - \phi_2}{\pi}, \tag{117b}$$

and $\phi_1, \phi_2$ are the phase changes at reflection for wavelengths ${\lambda_1}^{(m)}, {\lambda_2}^{(m)}$. The quantity $m - \phi_1/\pi$ may be determined by measuring the wavelength separation of adjacent fringes. Thus if, for the point $P_1'$ of the slit, the maximum of order $m+1$ is at wavelength ${\lambda_1}^{(m+1)}$ we have also from (116)

$$\left(m - \frac{\phi_1}{\pi}\right) \lambda_1^{(m)} = \left(m + 1 - \frac{\phi_1'}{\pi}\right) \lambda_1^{(m+1)},$$

i.e.

$$m - \frac{\phi_1}{\pi} = (1+b) \frac{\lambda_1^{(m+1)}}{\lambda_1^{(m)} - \lambda_1^{(m+1)}} \tag{118a}$$

where

$$b = \frac{\phi_1 - \phi_1'}{\pi} \tag{118b}$$

and $\phi_1'$ is the phase change at reflection for wavelength ${\lambda_1}^{(m+1)}$. If the fringes are discontinuous, the wavelength separation of adjacent fringes on each side of the discontinuity must be measured in order to identify corresponding orders. From (117) and (118),

$$\Delta h = (1+b) \frac{\lambda_1^{(m+1)}}{\lambda_1^{(m)} - \lambda_1^{(m+1)}} \frac{\lambda_2^{(m)} - \lambda_1^{(m)}}{2} + \frac{a}{2}\lambda_2^{(m)}. \tag{119}$$

0

* S. Tolansky, Phil. Mag. (7), 36 (1945), 225.

With reflecting coatings of silver, the quantities $a$ and $b$, which arise from variation of phase change with wavelength, are negligible when the wavelength interval involved is not too large. Measurement of ${\lambda_1}^{(m)}$, ${\lambda_1}^{(m+1)}$ and ${\lambda_2}^{(m)}$ ) then gives $\Delta h$; and since by (119) $\Delta h$ is proportional to ${\lambda_2}^{(m)} - {\lambda_1}^{(m)}$, the profile of the chosen section of the unknown surface is obtained by plotting a single fringe on a scale proportional to wavelength.

## 9.10.8 7.6.8 Multiple-beam fringes with two plane parallel plates

(a) Fringes with monochromatic and quasi-monochromatic light Consider two plane parallel plates, with surfaces of high reflectivity, placed one behind the other and illuminated with plane waves of monochromatic light (Fig. 7.80). If we use the notation of §7.6.1, and distinguish the two plates by subscripts 1 and 2 , the intensity of the light transmitted through the first plate is, according to equations (27) and (26),

$$I_1^{(t)} = \frac{\mathscr{T}_1^2}{(1-\mathscr{R}_1)^2} \frac{1}{1+F_1 \sin^2 \frac{\delta_1}{2}} I^{(i)} \tag{120}$$

where $I^{(i)}$ is the intensity of the incident light. Similarly, the intensity of the light transmitted through the second plate, corresponding to incident intensity ${I_1}^{(t)}$, is

$$I_2^{(t)} = \frac{\mathscr{T}_2^2}{(1-\mathscr{R}_2)^2} \frac{1}{1+F_2 \sin^2 \frac{\delta_2}{2}} I_1^{(t)} \tag{121}$$

We will suppose that light reflected back from the second plate, and subsequently reflected forward from the first plate, is excluded; this implies that the distance between the plates is sufficiently large compared with their apertures, and if the plates are mutually parallel, that the incidence is not too close to normal. The intensity $I^{(t)}$ of the total transmitted light is then equal to ${I_2}^{(t)}$, and from (120) and (121)

$$\frac{I^{(t)}}{I^{(i)}} = \frac{\mathscr{T}_1^2}{(1-\mathscr{R}_1)^2} \frac{\mathscr{T}_2^2}{(1-\mathscr{R}_2)^2} \frac{1}{\left(1+F_1 \sin^2 \frac{\delta_1}{2}\right)} \frac{1}{\left(1+F_2 \sin^2 \frac{\delta_2}{2}\right)}. \tag{122}$$

There are thus fringes in the focal plane of the lens $L$, with a relative intensity distribution which is the product of the relative intensity distributions of the fringes given by each plate alone.

The phase differences $\delta_1$ and $\delta_2$ are, according to (25),

$$\delta_1 = \frac{4\pi}{\lambda_0} n_1' h_1 \cos\theta_1' + 2\phi_1, \quad \delta_2 = \frac{4\pi}{\lambda_0} n_2' h_2 \cos\theta_2' + 2\phi_2. \tag{123}$$

Let us consider the case when the plates are mutually parallel, so that the ring systems which would be given by each plate alone are concentric. When the angle of incidence

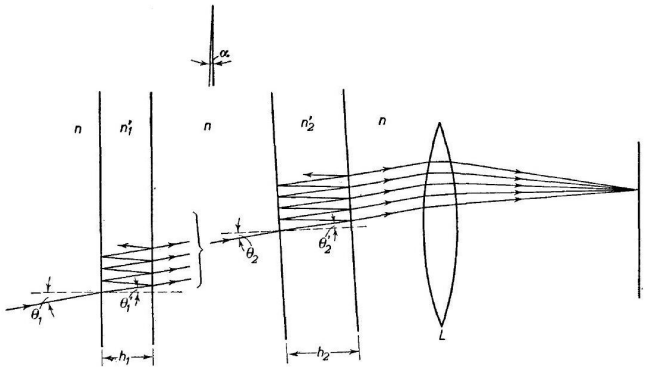



Fig. 7.80. Illustrating formation of fringes at infinity after multiple reflections in two plane parallel plates.

$\theta_1 = \theta_2 = \theta$ is small, $\cos\theta_1' \sim 1 - \theta_1'^2/2, \cos\theta_2' \sim 1 - \theta_2'^2/2$, and if $n$ is the refractive index of the surrounding medium, we have from the law of refraction $\theta_1' \sim n\theta/n_1'$, $\theta_2' \sim n\theta/n_2'$. Hence from (123), to the second order in $\theta$,

$$\delta_1 = \frac{4\pi}{\lambda_0} h_1 \left( n_1' - \frac{n^2\theta^2}{2n_1'} \right) + 2\phi_1, \quad \delta_2 = \frac{4\pi}{\lambda_0} h_2 \left( n_2' - \frac{n^2\theta^2}{2n_2'} \right) + 2\phi_2. \tag{124}$$

From (124) it follows that for a change of $\theta$ the corresponding changes of the $\delta$ 's are in the ratio

$$\frac{\Delta\delta_2}{\Delta\delta_1} = \frac{n_1' h_2}{n_2' h_1}, \tag{125}$$

where $\phi$ has been taken to be independent of $\theta$. If the refractive indices and thicknesses of the plates are such that in some direction $\theta = \theta_0$, the order of interference is integral for each plate, i.e. if

$$\theta = \theta_0, \quad \delta_1 = 2m_1\pi, \quad \delta_2 = 2m_2\pi, \quad (m_1 \text{ and } m_2 \text{ integers}), \tag{126}$$

the terms of (122) in $F_1$ and $F_2$ both have their maximum value of unity, and there is an absolute maximum of the intensity of the transmitted light, for $\theta = \theta_0$. If also

$$\frac{n_1' h_2}{n_2' h_1} = a, \tag{127}$$

where $a$ is a positive integer, we see from (125) that the integral orders ( $m_2 - a$ ), $(m_2 - 2a), \ldots$ of the second plate correspond to the same values of $\theta$ as the integral orders ( $m_1 - 1$ ), ( $m_1 - 2$ ), . . of the first plate, so that there are absolute maxima of $I^{(t)}/I^{(i)}$ in these directions also.*

The form of the intensity curve of the transmitted light is shown in Fig. 7.81. Between successive principal maxima there are ( $a - 1$ ) secondary maxima, corresponding to integral orders ( $m_2 - 1$ ), ( $m_2 - 2$ ), ... ( $m_2 - a + 1$ ), . . . . of the second plate; but if $F_1$ is large, and $a$ is not too large, the corresponding values of the term in

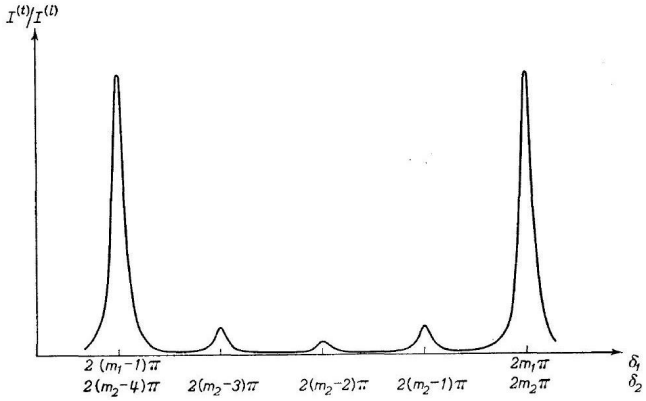



Fig. 7.81. Multiple-beam fringes of equal inclination from two plane parallel plates: ratio of transmitted and incident intensities as a function of phase differences $\delta_1$ and $\delta_2$. ( $\mathscr{R}_1 = 0.64, \mathscr{R}_2 = 0.64, a = 4$ ).

$F_1$ are small compared with unity, so that these secondary maxima are much weaker than the principal maxima.

Fringes of this kind from two Fabry-Perot interferometers have been used to examine the fine structure of spectral lines. The spacers are worked so that $h_2/h_1$ is approximately equal to the desired integer $a$, and fine adjustments are conveniently

- In practice only a finite number $p$ of coincidences is required, and then $n_1' h_2/n_2' h_1$ is sufficiently close to an integer $a$ if the mutual displacement of the maxima of the terms in $F_1, F_2$ at $\delta_1 = 2(m_1 - p)\pi, \delta_2 = 2(m_2 - pa)\pi$ is small compared with their half-widths. Using (125) and equation (21), these half-widths, in terms of $\delta_2$, are $4n_1' h_2/n_2' h_1 \sqrt{F_1}, 4/\sqrt{F_2}$. Hence the required condition is

$$\left| \frac{n_1' h_2}{n_2' h_1} 2p\pi - 2pa\pi \right| \ll \frac{n_1' h_2}{n_2' h_1} \frac{4}{\sqrt{F_1}} + \frac{4}{\sqrt{F_2}},$$

which we may write, for $(n_1' h_2/n_2' h_1) \sim a$,

$$\left| \frac{n_1' h_2}{n_2' h_1} - a \right| \ll \frac{1}{p} \left( \frac{a}{\mathscr{F}_1} + \frac{1}{\mathscr{F}_2} \right), \tag{126a}$$

where $\mathscr{F}_1, \mathscr{F}_2$ are the values of finesse for the fringes given by each plate alone. So long as $p$ is not too large and the order of interference $m_1$ for the first plate is not too small, relations (126) and (126a) may in practice be satisfied simultaneously. Thus with $\mathscr{F}_1 = \mathscr{F}_2 = 30$ and $a = 1$, (126a) requires $|(n_1' h_2/n_2' h_1) - a| \ll 1/15p$; whilst (126) can always be satisfied by a change of the order of $\lambda_0/4$ in $n_2' h_2$, and the corresponding change of $n_1' h_2/n_2' h_1$ is about $n_1' \lambda_0/4 n_2'^2 h_1 \sim n_1^2/2n_2'^2 m_1$, since $m_1 \sim 2n_1' h_1/\lambda_0$.

made by enclosing one of the interferometers in an air-tight container and varying the pressure-and hence the refractive index-of the air inside.*

When $a$ is greater than unity, the combination has the advantage of greater spectral range for given resolving power than can be obtained with a single interferometer. † The change of wavelength required to displace the pattern by an amount equal to the separation of adjacent principal maxima corresponds to a change of unity in the order of interference in the first plate, and a change of $a$ in the order of interference in the second plate. The spectral range without overlap of principal maxima is thus equal to the spectral range of the thinner interferometer, and is $a$ times the spectral range of the thicker interferometer (cf. (46)). The half-width of the

principal maxima corresponds to $\delta_1 = 2m_1\pi \pm \varepsilon_1/2$ or $\delta_2 = 2m_2\pi \pm \varepsilon_2/2$, ( $m_1, m_2$ integers), where from (122),

$$\left(1 + F_1 \sin^2 \frac{\varepsilon_1}{4}\right)\left(1 + F_2 \sin^2 \frac{\varepsilon_2}{4}\right) = 2$$

or by (125) and (127),

$$\left(1 + F_1 \sin^2 \frac{\varepsilon_2}{4a}\right)\left(1 + F_2 \sin^2 \frac{\varepsilon_2}{4}\right) = 2. \tag{128}$$

For $F_2$ sufficiently large, $\varepsilon_2$ is much less than $\pi/2$ and we may take $\sin \varepsilon_2 = \varepsilon_2$. Equation (128) then reduces to

$$\left(\frac{4}{\varepsilon_2}\right)^4 - \left(F_2 + \frac{F_1}{a^2}\right)\left(\frac{4}{\varepsilon_2}\right)^2 - \frac{F_1 F_2}{a^2} = 0 \tag{129}$$

whence

$$\varepsilon_2 = \frac{4\sqrt{2}}{\sqrt{\left(F_2 + \frac{F_1}{a^2}\right) + \sqrt{F_2^2 + \frac{6F_1F_2}{a^2} + \frac{F_1^2}{a^4}}}} \tag{130}$$

When the thinner interferometer is removed, $F_1 = 0$ and (130) reduces to $\varepsilon_2 = 4/\sqrt{F_2}$, in agreement with (21). Otherwise $\varepsilon_2$ is less than $4/\sqrt{F_2}$, so that the half-width of the principal maxima is less than the half-width of the fringes from the thicker interferometer. In practice $a$ is made equal to 3 or more, and we see from (130) that an increase of $F_2$ results in a much greater reduction of $\varepsilon_2$ than the same increase of $F_1$. We must also remember that the peak-transmission of the combination is equal to the product $\tau_1\tau_2$ of the peak-transmissions of the two interferometers, as may be seen from (122) and (33); and we have seen in §7.6.2 that, with available reflecting coatings, increases of $\mathscr{R}_1$ and $\mathscr{R}_2$ (and hence of $F_1$ and $F_2$ ) are accompanied by decreases of $\tau_1$ and $\tau_2$. For these reasons it is advantageous to make $F_1$ no larger than is necessary to suppress the secondary maxima to a satisfactory degree, and to make $F_2$ as large as intensity considerations permit. In these circumstances, $F_1/a^2$ is so small compared with $F_2$ that $\varepsilon_2 \sim 4/\sqrt{F_2}$ and the resolving power of the combination is nearly enough that of the thicker interferometer. The presence of secondary maxima is inconvenient if the source to be examined has spectral components of widely different intensities, since the principal maxima from a weak component

may be confused with the secondary maxima from a strong component. Such doubts may be removed by making observations with different thickness combinations.

When $a = 1$ there are no secondary maxima, and from (122) and (34) the contrast factor of the combination is equal to the product $\mathscr{C}_1\mathscr{C}_2$ of the contrast factors of the two interferometers. For given peak-transmission it is possible to obtain a much higher contrast factor than with a single interferometer, and the arrangement is of value for observations of weak satellites of spectral lines.*

(b) Fringes of superposition Fringes of great practical importance may be observed when two plane parallel plates, mutually inclined at angle $\alpha$, are illuminated near normal incidence with light which is so far from monochromatic that no fringes can be observed with either plate

0

- It is also possible, though less satisfactory, to make fine adjustments by tilting one interferometer relative to the other.

† The arrangement was first described by W. V. Houston, Phys. Rev. (2), 29 (1927), 478, and by E. Gehrcke and E. Lau, Zeits. f. tech. Physik, 8 (1927), 157.

alone. To understand this, † consider first the transmission of a monochromatic wave of wave number $k_0 = 2\pi/\lambda_0$ through one plate. Let $A^{(i)}(k_0)$ and $A^{(t)}(k_0)$ be the complex amplitude of the incident and of the transmitted wave respectively. Taking all the reflections into account, we have, according to (10), when (2) and (4) are also used,

$$A^{(t)}(k_0) = A^{(i)}(k_0)\,\mathscr{T}\sum_{p=0}^{\infty}\mathscr{R}^p e^{ip\delta} \tag{131}$$

so that the intensity $I^{(t)} = A^{(t)}A^{(t)\star}$ of the transmitted light is

$$I^{(t)}(k_0) = I^{(i)}(k_0)\,\mathscr{T}^2\sum_{p=0}^{\infty}\sum_{p'=0}^{\infty}\mathscr{R}^p\mathscr{R}^{p'}e^{i(p-p')\delta} \tag{132}$$

where $I^{(i)} = A^{(i)}A^{(i)\star}$ is the intensity of the incident light. If we set $|p-p'| = q$, (132) may be re-written in the form

$$\begin{aligned} I^{(t)}(k_0) &= I^{(i)}(k_0)\,\mathscr{T}^2\sum_{p=0}^{\infty}\mathscr{R}^{2p}\left\{1+\sum_{q=1}^{\infty}\mathscr{R}^q\left(e^{iq\delta}+e^{-iq\delta}\right)\right\} \\ &= I^{(i)}(k_0)\,\frac{\mathscr{T}^2}{1-\mathscr{R}^2}\left(1+2\sum_{q=1}^{\infty}\mathscr{R}^q\cos q\delta\right). \end{aligned} \tag{133}$$

With two plates in series, if we neglect the light reflected backwards and forwards between them, the intensity of the monochromatic light transmitted by both plates therefore is

$$I^{(t)}(k_0) = I^{(i)}(k_0)\,\frac{\mathscr{T}_1^2\,\mathscr{T}_2^2}{(1-\mathscr{R}_1^2)(1-\mathscr{R}_2^2)}\left(1+2\sum_{r=1}^{\infty}\mathscr{R}_1^r\cos r\delta_1\right)\left(1+2\sum_{s=1}^{\infty}\mathscr{R}_2^s\cos s\delta_2\right), \tag{134}$$

where the suffixes 1 and 2 refer to the first and the second plate respectively. The phase differences $\delta_1$ and $\delta_2$ are given by the expressions

$$\left.\begin{aligned} \delta_1 &= k_0\Delta\mathscr{S}_1 + 2\phi_1 = k_0 2n_1' h_1\cos\theta_1' + 2\phi_1, \\ \delta_2 &= k_0\Delta\mathscr{S}_2 + 2\phi_2 = k_0 2n_2' h_2\cos\theta_2' + 2\phi_2, \end{aligned}\right\} \tag{135}$$

where the various symbols have the same meaning as before.

Now if the light is not monochromatic we may regard it as superposition of monochromatic components of different frequencies. According to §7.5.8 the different components add incoherently, so that the total intensity is given by the sum (integral) of the intensities of the individual components. Thus from (134) we obtain the following expression for the total intensity of the light transmitted by both plates:
$I^{(t)} = \int \frac{\mathscr{T}_1{}^2\mathscr{T}_2{}^2}{(1-\mathscr{R}_1{}^2)(1-\mathscr{R}_2{}^2)} i^{(i)}(k_0)\,(1+2\sum_{r=1}^{\infty}\mathscr{R}_1{}^r\cos r\delta_1)\,(1+2\sum_{s=1}^{\infty}\mathscr{R}_2{}^s\cos s\delta_2)\,dk_0,$
where $i^{(i)}(k_0)$ represents the spectral intensity distribution of the incident light. The quantities $\mathscr{R}$ and $\mathscr{T}$ are in general functions of $k_0$, but we will assume that their variation with $k_0$ is

0

- L. C. Bradley and H. Kuhn, Nature, London, 162 (1948), 412.
  C. Dufour, Ann. de Physique, Paris (12), 6 (1951), 5.
  † The analysis given here follows essentially that of J. R. Benoit, C. Fabry and A. Perot, Trav. et Mem. Bur. Int. Poids et Mes., 15 (1913), 1.

negligible in the range over which $i^{(i)}(k_0)$ is appreciable. The first factors in (136) may then be taken outside the integral, and we may re-write the resulting expression in the form

$$I^{(t)} = \frac{\mathscr{T}_1^2\,\mathscr{T}_2^2}{(1-\mathscr{R}_1{}^2)\,(1-\mathscr{R}_2{}^2)} \int i^{(i)}(k_0)\left\{1 + 2\left[\sigma_1 + \sigma_2 + \sigma_{12}{}^{+} + \sigma_{12}{}^{-}\right)\right\} dk_0 \tag{137a}$$

where

$$\left.\begin{array}{rl} \sigma_1 & = \sum_{r=1}^{\infty} \mathscr{R}_1^r \cos r\delta_1 \\ \sigma_2 & = \sum_{s=1}^{\infty} \mathscr{R}_2^s \cos s\delta_2 \\ \sigma_{12}{}^{+} & = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mathscr{R}_1^r \mathscr{R}_2^s \cos\left(r\delta_1 + s\delta_2\right) \\ \sigma_{12}{}^{-} & = \sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mathscr{R}_1^r \mathscr{R}_2^s \cos\left(r\delta_1 - s\delta_2\right) \end{array}\right\} \tag{137b}$$

If the monochromatic components cover a wavelength range $\Delta\lambda_0$ about a mean wavelength $\bar{\lambda}_0$, the corresponding range of $k_0$ is $2\pi\Delta\lambda_0/\bar{\lambda}_0{}^2$ and when $\Delta\mathscr{S}_1$ and $\Delta\mathscr{S}_2$ are sufficiently large compared with the coherence length $\bar{\lambda}_0{}^2/\Delta\lambda$, the ranges of $\delta_1$ and $\delta_2$ are large compared with $2\pi$*. In these circumstances $\sigma_1, \sigma_2$ and $\sigma_{12}{}^{+}$will vary rapidly over the domain of integration and each will change sign many times. In consequence these terms will not contribute appreciably to $I^{(t)}$, and (137) reduces to †
$I^{(t)} \sim \frac{\mathscr{T}_1{}^2\mathscr{T}_2{}^2}{(1-\mathscr{R}_1{}^2)(1-\mathscr{R}_2{}^2)} \int i^{(i)}(k_0)\,(1 + 2\sum_{r=1}^{\infty}\sum_{s=1}^{\infty} \mathscr{R}_1{}^r \mathscr{R}_2{}^s \cos\left(r\delta_1 - s\delta_2\right))\, dk_0$.

In general the quantities $r\delta_1 - s\delta_2$ will also be at least of the order of $\delta_1$ or $\delta_2$, so that the integrals of the cosine terms in (138) will also be negligible, and $I^{(t)}$ is then effectively independent of $\delta_1$ and $\delta_2$. There is, however, an exception when

$$a\left(\Delta\mathscr{S}_1 + \frac{2\phi_1}{k_0}\right) - b\left(\Delta\mathscr{S}_2 + \frac{2\phi_2}{k_0}\right) = \varepsilon \tag{139}$$

where $a$ and $b$ are small integers without a common factor, and $|\varepsilon|$ is not large compared with $\bar{\lambda}_0{}^2/\Delta\lambda_0$. In this case, for

$$r = qa, \quad s = qb, \quad (q = 1, 2, 3, \ldots) \tag{140}$$

we have

$$r\delta_1 - s\delta_2 = qk_0\varepsilon \tag{141}$$

and the range of values of $|r\delta_1 - s\delta_2|$ is $|q\varepsilon\Delta k_0|$, which is not large compared with $2\pi q$; for these values of $r$ and $s$ the integrals of the cosine terms in (138) are not necessarily negligible, and we obtain

$$I^{(t)}(\varepsilon) \sim \frac{\mathscr{T}_1^2\,\mathscr{T}_2^2}{(1-\mathscr{R}_1^2)\,(1-\mathscr{R}_2^2)} \int i^{(i)}(k_0)\left(1 + 2\sum_{q=1}^{\infty}\left(\mathscr{R}_1{}^a\mathscr{R}_2{}^b\right)^q \cos qk_0\varepsilon\right) dk_0 \tag{142}$$

The series under the integral sum of (142) is identical with the series entering the expression (133) for the intensity of a monochromatic wave transmitted by a single plate, and in § 7.6.1 it

0

- As is evident on integrating (133) over all spectral components, this implies that with either plate alone the transmitted intensity is effectively independent of $\delta$, i.e. there are no fringes.
† A rigorous justification of the transition from (137a) to (142) would require more refined considerations than those given here.

has been evaluated in a closed form [eq. (13)]. Hence the sum of the series may immediately be written down by comparison,

$$1+2\sum_{q=1}^{\infty}\left(\mathscr{R}_1{}^a\mathscr{R}_2{}^b\right)^q\cos qk_0\varepsilon=\frac{1-\left(\mathscr{R}_1{}^a\mathscr{R}_2{}^b\right)^2}{\left(1-\mathscr{R}_1{}^a\mathscr{R}_2{}^b\right)^2+4\mathscr{R}_1{}^a\mathscr{R}_2{}^b\sin^2\frac{k_0\varepsilon}{2}}. \tag{143}$$

From (142) and (143),

$$I^{(t)}(\varepsilon)=\frac{\mathscr{T}_1^2\mathscr{T}_2^2\left\{1-\left(\mathscr{R}_1^a\mathscr{R}_2^b\right)^2\right\}}{\left(1-\mathscr{R}_1^2\right)\left(1-\mathscr{R}_2^2\right)\left(1-\mathscr{R}_1^a\mathscr{R}_2^b\right)^2}\int\frac{i^{(i)}\left(k_0\right)}{1+B\sin^2\left(\frac{k_0\varepsilon}{2}\right)}dk_0 \tag{144}$$

where

$$B=\frac{4\mathscr{R}_1{}^a\mathscr{R}_2{}^b}{\left(1-\mathscr{R}_1{}^a\mathscr{R}_2{}^b\right)^2} \tag{145}$$

By comparison with (15b) we see that the intensity distribution (144) is equivalent to a superposition of monochromatic intensity distributions, each of the form shown in Fig. 7.58. The intensity maxima of these distributions occur when $k_0\varepsilon$ is an integral multiple of $2\pi$, i.e. when

$$\varepsilon=m\lambda_0,\quad |m|=0,1,2,\ldots \tag{146}$$

If we take $n_1', n_2', \phi_1/k_0$ and $\phi_2/k_0$ to be constant over the spectral range involved in (144), the zero order maxima ( $\varepsilon=0$ ) of the monochromatic intensity distributions coincide in a central fringe. By (139) and (135), the position of this fringe in the focal plane of the lens $L$ (Fig. 7.80) is defined by

$$\frac{n_2'h_2\cos\theta_2'+c_2}{n_1'h_1\cos\theta_1'+c_1}=\frac{a}{b} \tag{147}$$

where $c_1, c_2$ are the values of $\phi_1/k_0, \phi_2/k_0$ respectively. On either side of this fringe, as $|\varepsilon|$ increases, the component patterns are mutually displaced because the scale of each is proportional to wavelength, and the fringes become less distinct. With white light there is a central white fringe in the position given by (147), with coloured maxima and minima on either side, and further away what appears to the eye to be uniform illumination. These fringes are called fringes of superposition. They are the multiple-beam form of Brewster's fringes (§ 7.5.6), and like them are straight lines,
parallel to the apex of the wedge formed between the plates, and with spacing inversely proportional to the wedge angle $\alpha$.

Fringes of superposition can be used, in a way suggested by Fabry and Buisson,* to determine the difference of optical thickness of two Fabry-Perot etalons whose optical thicknesses are very nearly in integral ratio $a$. For this purpose, one etalon is fixed and the other is tilted until the central white fringe passes through $O$, the focal point of the lens $L$ for light transmitted normally through the fixed etalon (Fig. 7.80). If the first etalon is fixed we then have $\theta_1=\theta_1'=0$, and $\theta_2=\alpha$, where $\alpha$ is the angle between the etalons. For $\alpha$ small we have, from the law of refraction, $\theta_2'\sim n\alpha/n_2'$, where $n$ is the refractive index of the air surrounding the etalons; $\cos\theta_2'\sim 1-\theta_2'^2/2\sim 1-n^2\alpha^2/2n_2'^2$; and from (147), putting $b=1$, we have, to the second power of $\alpha$,

$$\left(n_2'h_2+c_2\right)-a\left(n_1'h_1+c_1\right)=\frac{n^2h_2}{2n_2'}\alpha^2 \tag{148a}$$

Alternatively, if the second etalon is fixed the white fringe is at $O$ when

$$a\,(n_1'h_1 + c_1) - (n_2'h_2 + c_2) = \frac{n^2 a h_1}{2n_1'}\alpha^2 \tag{148b}$$

Even when the etalons are evacuated, as is sometimes the case, we may take $n_1' = n_2' = n$ if the difference $|a\,(n_1'h_1 + c_1) - (n_2'h_2 + c_2)|$ is sufficiently small, and measurement of $\alpha$ gives this difference in terms of the optical thickness of one of the etalons. Measurements with values of $a$ as high as ten have been made in this way.

A similar arrangement, but with etalons of approximately equal thickness, has been used to measure the refractive index and dispersion of air. † The etalons are first evacuated, and one is tilted to give fringes of convenient spacing with nearly monochromatic light of mean wavelength $\bar{\lambda}_0$. Air is then slowly admitted to the fixed etalon, of thickness $h$, causing a change $2(n-1)h$ of $\varepsilon$, where $n$ is the refractive index of the air. The corresponding number $\Delta m$ of fringes passing $O$ is counted, and from (146),

$$\Delta m = \frac{\Delta\varepsilon}{\bar{\lambda}_0} = \frac{2(n-1)h}{\bar{\lambda}_0} \tag{149}$$

giving ( $n-1$ ) if $h$ and $\bar{\lambda}_0$ are known.

## 9.11 Elements of the Theory of Diffraction

### 9.11.1 8.1. Introduction

In carrying out the transition from the general electromagnetic field to the optical field, which is characterized by very high frequencies (short wavelengths), we found that in certain regions the simple geometrical model of energy propagation was inadequate. In particular, we saw that deviations from this model must be expected in the immediate neighbourhood of the boundaries of shadows and in regions where a large number of rays meet. These deviations are manifested by the appearance of dark and bright bands, the diffraction fringes. Diffraction theory is mainly concerned with the field in these special regions; such regions are of great practical interest as they include the part of the image space in which the optical image is situated (region of focus).

The first reference to diffraction phenomena appears in the work of Leonardo da Vinci (1452-1519). Such phenomena were, however, first accurately described by Grimaldi in a book, published in 1665, two years after his death. The corpuscular theory, which, at the time, was widely believed to describe correctly the propagation of light, could not explain diffraction. Huygens, the first proponent of the wave theory, seems to have been unaware of Grimaldi's discoveries; otherwise he would have undoubtedly quoted them in support of his views. The possibility of explaining diffraction effects on the basis of a wave theory was not noticed until about 1818. In that year there appeared the celebrated memoir of Fresnel (cf. Historical Introduction) in which he showed that diffraction can be explained by the application of Huygens' construction (cf. §3.3.3) together with the principle of interference. Fresnel's analysis was later put on a sound mathematical basis by Kirchhoff (1882), and the subject has since then been extensively discussed by many writers.*

Diffraction problems are amongst the most difficult ones encountered in optics. Solutions which, in some sense, can be regarded as rigorous are very rare in diffraction theory. The first such solution was given as late as 1896 by A. Sommerfeld when, in an important paper, he discussed the diffraction of a plane wave by a perfectly conducting semi-infinite plane screen. Since then rigorous solutions of a small number of other diffraction problems (mainly two-dimensional) have also been found (cf. Chapter XI), but, because of mathematical difficulties, approximate methods must be used in most cases of practical interest. Of these the theory

of Huygens and Fresnel is by far the most powerful and is adequate for the treatment of the majority of problems encountered in instrumental optics. This theory and some of its applications form the main subject matter of this chapter.

## 9.11.2 8.2. The Huygens-Fresnel Principle

According to Huygens' construction (§ 3.3.3), every point of a wave-front may be considered as a centre of a secondary disturbance which gives rise to spherical wavelets,

and the wave-front at any later instant may be regarded as the envelope of these wavelets. Fresnel was able to account for diffraction by supplementing Huygens' construction with the postulate that the secondary wavelets mutually interfere. This combination of Huygens' construction with the principle of interference is called the Huygens-Fresnel Principle. Before applying it to the study of diffraction effects we shall verify that (with certain simple additional assumptions) the principle correctly describes the propagation of light in free space.

Let $S$ (Fig. 8.1) be the instantaneous position of a spherical monochromatic wavefront of radius $r_0$ which proceeds from a point source $P_0$, and let $P$ be a point at

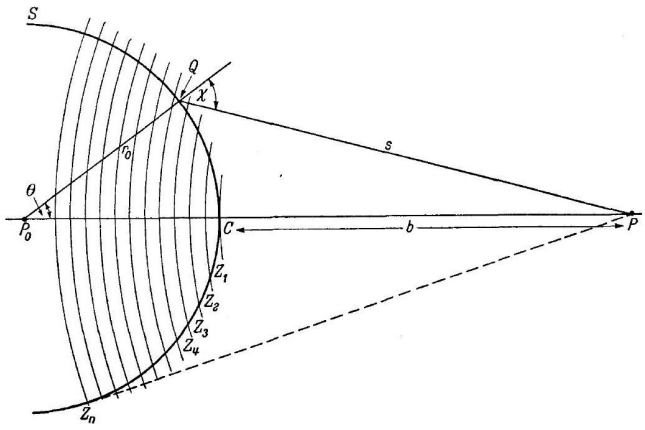



Fig. 8.1. Fresnel's zone construction.

which the light disturbance is to be determined. The time periodic factor $e^{-i\omega t}$ being omitted, the disturbance at a point $Q$ on the wave-front may be represented by $Ae^{ikr_0}/r_0$, where $A$ is the amplitude at unit distance from the source. In accordance with the Huygens-Fresnel principle we regard each element of the wave-front as the centre of a secondary disturbance which is propagated in the form of spherical wavelets, and obtain for the contribution $dU(P)$ due to the element $dS$ at $Q$ the expression

$$dU(P) = K(\chi)\frac{Ae^{ikr_0}}{r_0}\frac{e^{iks}}{s}dS,$$

where $s = QP$ and $K(\chi)$ is an inclination factor which describes the variation with direction of the amplitude of the secondary waves, $\chi$ being the angle (often called the angle of diffraction) between the normal at $Q$ and the direction $QP$. Following Fresnel we assume that $K$ is maximum in the original direction of propagation, i.e. for $\chi = 0$, and that it rapidly decreases with increasing $\chi$, being zero when $QP$ is tangential to the wave-front, i.e. when $\chi = \pi/2$; and

0

- For a fuller historical account of the development of the subject see C. F. Meyer, The Diffraction of Light, X-rays, and Material Particles (Chicago, The University Press, 1934).

finally, that only that part $S'$ of the primary wave contributes to the effect at $P$, which is not obstructed by obstacles which may be situated between $P_0$ and $P$. Hence the total disturbance at $P$ is given by

$$U(P) = \frac{Ae^{ikr_0}}{r_0} \iint_S \frac{e^{iks}}{s} K(\chi) dS. \tag{1}$$

To evaluate (1) we shall use the so-called zone construction of Fresnel. With centre at $P$, we construct spheres of radii

$$b, \quad b + \frac{\lambda}{2}, \quad b + \frac{2\lambda}{2}, \quad b + \frac{3\lambda}{2}, \ldots\ldots, b + \frac{j\lambda}{2}, \ldots$$

where $b = CP, C$ being the point of intersection of $P_0P$ with the wave-front $S$ (see Fig. 8.1). The spheres divide $S$ into a number of zones $Z_1, Z_2, Z_3 \ldots Z_j \ldots\ldots$

We assume that both $r_0$ and $b$ are large compared to the wavelength; then $K$ may be assumed to have the same value, $\boldsymbol{K_j}$, for points on one and the same zone. From the figure

$$s^2 = r_0^2 + (r_0 + b)^2 - 2r_0 (r_0 + b) \cos\theta$$

so that

$$sds = r_0 (r_0 + b) \sin\theta d\theta \tag{2}$$

and therefore

$$dS = r_0^2 \sin\theta d\theta d\phi = \frac{r_0}{r_0 + b} sds d\phi$$

$\phi$ being the azimuthal angle. Hence the contribution of the $j$ th zone to $U(P)$ is

$$\begin{aligned} U_i(P) &= 2\pi \frac{Ae^{ikr_0}}{r_0 + b} K_j \int_{b+(j-1)\lambda/2}^{b+j\lambda/2} e^{iks} ds \\ &= -\frac{2\pi i}{k} K_j \frac{Ae^{ik(r_0+b)}}{r_0 + b} e^{ikj\lambda/2} \left(1 - e^{-ik\lambda/2}\right) \end{aligned}$$

Since $k\lambda = 2\pi$, the last two factors reduce to

$$e^{ikj\lambda/2} \left(1 - e^{-ik\lambda/2}\right) = e^{i\pi j} \left(1 - e^{-i\pi}\right) = (-1)^j 2$$

so that

$$U_j(P) = 2i\lambda(-1)^{j+1} K_j \frac{Ae^{ik(r_0+b)}}{r_0 + b} \tag{3}$$

We note that the contributions of the successive zones are alternately positive and negative. The total effect at $P$ is obtained by summing all the contributions:

$$U(P) = 2i\lambda \frac{Ae^{ik(r_0+b)}}{r_0 + b} \sum_{j=1}^{n} (-1)^{j+1} K_j \tag{4}$$

The series

$$\Sigma = \sum_{j=1}^{n} (-1)^{j+1} K_j = K_1 - K_2 + K_3 - \ldots + (-1)^{n+1} K_n \tag{5}$$

can now be approximately summed by a method due to Schuster*.
First we write (5) in the form

$$\Sigma = \frac{K_1}{2} + \left(\frac{K_1}{2} - K_2 + \frac{K_3}{2}\right) + \left(\frac{K_3}{2} - K_4 + \frac{K_5}{2}\right) + \ldots\ldots \tag{6}$$

the last term being $\frac{1}{2}K_n$ or $\frac{1}{2}K_{n-1} - K_n$ according to $n$ being odd or even. Let us assume for the moment that the law which specifies the directional variation is such

that $K_j$ is greater than the arithmetic mean of its two neighbours $K_{j-1}$ and $K_{j+1}$. Then each of the bracketed terms in (6) is negative and it follows that
and

$$\left.\begin{array}{ll} \Sigma < \frac{K_1}{2} + \frac{K_n}{2} & \text{when } n \text{ is odd} \\ \Sigma < \frac{K_1}{2} + \frac{K_{n-1}}{2} - K_n & \text{when } n \text{ is even.} \end{array}\right\} \tag{7}$$

We can also write (5) in the form

$$\Sigma = K_1 - \frac{K_2}{2} - \left(\frac{K_2}{2} - K_3 + \frac{K_4}{2}\right) - \left(\frac{K_4}{2} - K_5 + \frac{K_6}{2}\right) - \ldots\ldots \tag{8}$$

the last term now being $-\frac{1}{2}K_{n-1} + K_n$ when $n$ is odd and $-\frac{1}{2}K_n$ when $n$ is even. Hence and

$$\left.\begin{array}{ll} \Sigma > K_1 - \frac{K_2}{2} - \frac{K_{n-1}}{2} + K_n & (n \text{ odd }) \\ \Sigma > K_1 - \frac{K_2}{2} - \frac{K_n}{2} & (n \text{ even }). \end{array}\right\} \tag{9}$$

Now each $K_j$ differs only slightly from its neighbouring values $K_{j-1}$ and $K_{j+1}$ so that the right-hand sides of the corresponding relations in (7) and (9) are practically equal, and, therefore, approximately,
and

$$\left.\begin{array}{ll} \Sigma = \frac{K_1}{2} + \frac{K_n}{2} & (n \text{ odd }) \\ \Sigma = \frac{K_1}{2} - \frac{K_n}{2} & (n \text{ even }). \end{array}\right\} \tag{10}$$

It may easily be verified that (10) remains valid when each $K_j$ is smaller than the arithmetic mean of its two neighbours, each of the bracketed terms in (6) and (8) then being positive. Moreover, (10) may be expected to remain valid even when only some of the bracketed terms are negative whilst the others are positive, for the series may then be divided into two parts according to the signs of the bracketed terms and a similar argument may be applied to each part. We may, therefore, conclude that the sum of the series is given by (10) unless the bracketed terms in (6) and (8) change sign so frequently that the error terms add up to an appreciable amount. If we exclude the later case it follows from (10) and (4) that

$$U(P) = i\lambda\,(K_1 \pm K_n)\,\frac{Ae^{ik(r_0+b)}}{r_0+b}, \tag{11}$$

the upper or lower sign being taken according as $n$ is odd or even. Using (3), equation (11) may be also written in the form

$$U(P) = \frac{1}{2}\left[U_1(P) + U_n(P)\right]. \tag{12}$$

0

- A. Schuster, Phil. Mag. (5), 31 (1891), p. 77.

For the last zone ($Z_n$) that can be seen from $P, QP$ is a tangent to the wave, i.e. $\chi = \pi/2$, and for this value of $\chi$, as already mentioned, $K$ was assumed to be zero Hence $K_n = 0$ and (11) reduces to

$$U(P) = i\lambda K_1 \frac{Ae^{ik(r_0+b)}}{r_0+b} = \frac{1}{2}U_1(P), \tag{13}$$

showing that the total disturbance at $P$ is equal to half of the disturbance due to the first zone.

Equation (13) is in agreement with the expression for the effect of the spherical wave if

$$i\lambda K_1 = 1,$$

i.e. if

$$K_1 = -\frac{i}{\lambda} = \frac{e^{-i\pi/2}}{\lambda} \tag{14}$$

The factor $e^{-i\pi/2}$ may be accounted for by assuming that the secondary waves oscillate a quarter of a period out of phase with the primary wave; the other factor can be explained by assuming that the amplitudes of the secondary vibrations are to the amplitudes of the primary vibrations in the ratio $1 : \lambda$. We can therefore conclude that, with these assumptions about the amplitude and phase of the secondary waves, the Huygens-Fresnel principle leads to the correct expression for the propagation of a spherical wave in free space. The additional assumptions must, however, be regarded as purely a convenient way of interpreting the mathematical expressions and as being devoid of any physical significance; the real justification of the factor (14) will become evident later (§ 8.3).

Still following Fresnel, let us consider the effect at $P$ when some of the zones are obstructed by a plane screen with a circular opening, perpendicular to $P_0P$ and with its centre on this line. The total disturbance at $P$ must now be regarded as due to wavelets from only those zones that are not obstructed by the screen. When the screen covers all but half of the first zone, (3) gives, on setting $j = 1$, and multiplying by $\frac{1}{2}$,

$$U(P) = i\lambda K_1 \frac{Ae^{ik(r_0+b)}}{r_0+b} = \frac{Ae^{ik(r_0+b)}}{r_0+b} \tag{15}$$

hence the disturbance at $P$ is now the same as would be obtained if no screen were present. When all the zones are covered except the first, (3) gives

$$U(P) = 2i\lambda K_1 \frac{Ae^{ik(r_0+b)}}{r_0+b} = 2\frac{Ae^{ik(r_0+b)}}{r_0+b} \tag{16}$$

so that the intensity $I(P) = |U(P)|^2$ is four times larger than if the screen were absent. When the opening is increased still further the intensity will decrease, since the first two terms in (4) have different signs. Moreover, since $K_1$ and $K_2$ are nearly equal, it follows that there will be almost complete darkness at $P$ when the opening is approximately equal to the first two zones. Thus, when the size of the opening is varied, there is a periodic fluctuation in the intensity at $P$. Similar results are obtained when the size of the opening and the source are fixed, but the position of the point $P$ of observation is varied along the axis; for then, as $P$ gradually approaches the screen, an increasingly larger number of zones is required to fill the opening completely.

All these results were found to be in good agreement with experiment. One prediction of Fresnel's theory made a strong impression on his contemporaries, and was, in fact, one of the decisive factors which temporarily ended the long battle between the corpuscular and the wave

theory of light in favour of the latter. It concerns the effect which arises when the first zone is obstructed by a small circular disc placed at right angles to $P_0P$. According to (5) the complex amplitude at $P$ is then given by

$$U(P) = 2i\lambda \frac{Ae^{ik(r_0+b)}}{r_0+b} [-K_2 + K_3 - K_4 + \ldots] \tag{17}$$

and, by a similar argument as before, the sum of the series in the brackets is $-K_2/2$. Since $K_2$ is assumed to differ only slightly from $K_1 = 1/i\lambda$, it follows that there is light in the geometrical shadow of the disc, and, moreover, that the intensity there is the same as if no disc were present.*

## 9.11.3 8.3. Kirchhoff's Diffraction Theory

### 8.3.1 The integral theorem of Kirchhoff

The basic idea of the Huygens-Fresnel theory is that the light disturbance at a point $P$ arises from the superposition of secondary waves that proceed from a surface situated between this point and the light source. This idea was put on a sounder mathematical basis by Kirchhoff †, who showed that the Huygens-Fresnel principle may be regarded as an approximate form of a certain integral theorem ‡ which expresses the solution of the homogeneous wave equation, at an arbitrary point in the field, in terms of the values of the solution and its first derivatives at all points on an arbitrary closed surface surrounding $P$.

We consider first a strictly monochromatic scalar wave

$$V(x, y, z, t) = U(x, y, z)e^{-i\omega t}. \tag{1}$$

In vacuum the space-dependent part then satisfies the time-independent wave equation

$$\left(\nabla^2 + k^2\right) U = 0, \tag{2}$$

where $k = \omega/c$. Equation (2) is also known as the Helmholtz equation.

Let $v$ be a volume bounded by a closed surface $S$, and let $P$ be any point within it; we assume that $U$ possesses continuous first- and second-order partial derivatives within and on this surface. If $U'$ is any other function which satisfies the same continuity requirements as $U$, we have by Green's theorem

$$\iiint_v \left(U\nabla^2 U' - U'\nabla^2 U\right) dv = -\iint_S \left(U\frac{\partial U'}{\partial n} - U'\frac{\partial U}{\partial n}\right) dS \tag{3}$$

0

- That a bright spot should appear at the centre of the shadow of a small disc was deduced from Fresnel's theory by S. D. Poisson in 1818. Poisson, who was a member of the committee of the French Academy which reviewed Fresnel's prize memoir, appears to have considered this conclusion contrary to experiment and so refuting Fresnel's theory. However, Arago, another member of the committee, performed the experiment and found that the surprising prediction was correct. A similar observation had been made a century earlier by Maraldi but had been forgotten.

† G. Kirchнoff, Berl. Ber. (1882), 641; Ann. d. Physik. (2) 18 (1883), 663; Ges. Abh. Nachtr., 22.

Kirchhoff's theory applies to the diffraction of scalar waves. As will be shown in § 8.4 a scalar theory is usually quite adequate for the treatment of the majority of problems of instrumental optics.

Vectorial generalizations of the Huygens-Fresnel principle have been proposed by many authors. The first satisfactory generalization is due to F. Kottler, Ann. d. Physik, 71 (1923), 457; 72 (1923), 320. [Cf. B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle, (Oxford, Clarendon Press, 2nd edition, 1950), p. 114].

‡ For monochromatic waves this theorem was derived earlier in acoustics by H. von Helmholtz, J. f. Math., 57 (1859), 7.

where $\partial/\partial n$ denotes differentiation along the inward* normal to $S$. In particular, if $U'$ also satisfies the time-independent wave equation, i.e. if

$$\left(\nabla^2 + k^2\right) U' = 0, \tag{4}$$

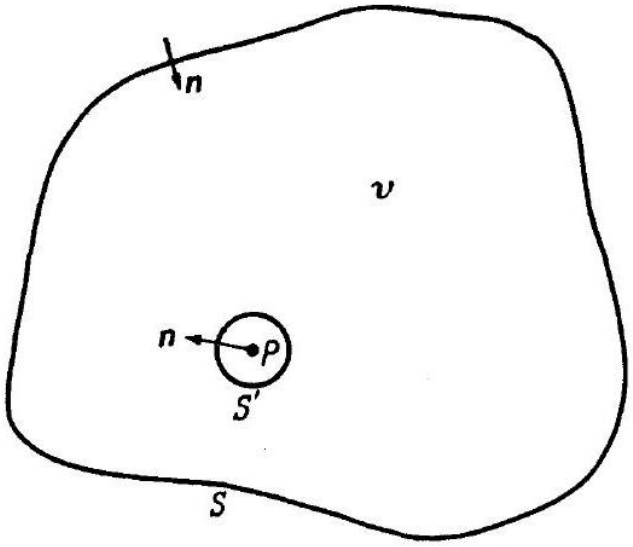



Fig. 8.2. Derivation of the Helmholtz-Kirchhoff integral theorem: region of integration.

then it follows at once from (2) and (4) that the integrand on the left of (3) vanishes at every point of $v$, and consequently

$$\iint_S \left( U \frac{\partial U'}{\partial n} - U' \frac{\partial U}{\partial n} \right) dS = 0 \tag{5}$$

Suppose we take $U'(x, y, z) = e^{iks}/s$, where $s$ denotes the distance from $P$ to the point $(x, y, z)$. This function has a singularity for $s = 0$, and since $U'$ was assumed to be continuous and differentiable, $P$ must be excluded from the domain of integration. We shall therefore surround $P$ by a small sphere of radius $\varepsilon$ and extend the integration throughout the volume between $S$ and the surface $S'$ of this sphere (Fig. 8.2). In place of (5), we then have

$$\iint_S + \iint_{S'} \left\{ U \frac{\partial}{\partial n} \left( \frac{e^{iks}}{s} \right) - \frac{e^{iks}}{s} \frac{\partial U}{\partial n} \right\} dS = 0$$

whence

$$\begin{aligned} \iint_S \left\{ U \frac{\partial}{\partial n} \left( \frac{e^{iks}}{s} \right) - \frac{e^{iks}}{s} \frac{\partial U}{\partial n} \right\} dS &= - \iint_{S'} \left\{ U \frac{e^{iks}}{s} \left( ik - \frac{1}{s} \right) - \frac{e^{iks}}{s} \frac{\partial U}{\partial n} \right\} dS' \\ &= - \iint_{\Omega} \left\{ U \frac{e^{ik\varepsilon}}{\varepsilon} \left( ik - \frac{1}{\varepsilon} \right) - \frac{e^{ik\varepsilon}}{\varepsilon} \frac{\partial U}{\partial s} \right\} \varepsilon^2 d\Omega \end{aligned} \tag{6}$$

where $d\Omega$ denotes an element of the solid angle. Since the integral over $S$ is independent of $\varepsilon$, we may replace the integral on the right-hand side by its limiting value

0

* Green's theorem is usually expressed in terms of the outward normal, but the inward normal is more convenient in the present application.

as $\varepsilon \to 0$; the first and third terms in this integral give no contribution in the limit, and the total contribution of the second term is $4\pi U(P)$. Hence

$$U(P) = \frac{1}{4\pi} \iint_S \left\{ U \frac{\partial}{\partial n} \left( \frac{e^{iks}}{s} \right) - \frac{e^{iks}}{s} \frac{\partial U}{\partial n} \right\} dS \tag{7}$$

This is one form of the integral theorem of Helmholtz and Kirchhoff.*
We note, that as $k \to 0$, the time-independent wave equation (2) reduces to Laplace's equation $\nabla^2 U = 0$, and (7) then goes over into the well-known formula of potential theory

$$U(P) = \frac{1}{4\pi} \iint_S \left\{ U \frac{\partial}{\partial n} \left( \frac{1}{s} \right) - \frac{1}{s} \frac{\partial U}{\partial n} \right\} dS \tag{8}$$

If $P$ lies outside the surface $S$, but $U$ is still assumed to be continuous and differentiable up to the second order within $S$, and if as before we take $U' = e^{iks}/s$, equation (3) remains valid throughout the whole volume within $S$. According to (5) the surface integral then has the value zero.

There is a complementary form of the Helmholtz-Kirchhoff theorem for the case when $U$ is continuous and differentiable up to the second order outside and on a closed surface $S$ (sources inside). In this case, however, as in other problems of propagation in an infinite medium, the boundary values on $S$ are no longer sufficient to specify the solution uniquely and additional assumptions must be made about the behaviour of the solution as $s \to \infty$. For a discussion of this case we must, however, refer elsewhere. †

So far we have considered strictly monochromatic waves. We now derive the general form of Kirchioff's theorem which applies to waves that are not necessarily monochromatic.

Let $V(x, y, z, t)$ be a solution of the wave equation

$$\nabla^2 V = \frac{1}{c^2} \frac{\partial^2 V}{\partial t^2}, \tag{9}$$

and assume that $V$ can be represented in the form of a Fourier integral

$$V(x, y, z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} U_\omega(x, y, z) e^{-i\omega t} d\omega \tag{10}$$

Then, by the Fourier inversion formula

$$U_\omega(x, y, z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} V(x, y, z, t) e^{i\omega t} dt \tag{11}$$

Since $V(x, y, z, t)$ is assumed to satisfy the wave equation (9), $U_\omega(x, y, z)$ will satisfy the time independent wave equation (2). If moreover $V$ obeys the appropriate regularity

conditions within and on a closed surface $S$, we may apply the Kirchhoff formula separately to each Fourier component $U_\omega(x, y, z) = U_\omega(P)$ :

---

0

- This theorem expresses $U(P)$ in terms of the values of both $U$ and $\partial U/\partial n$ on $S$. It may, however, be shown from the theory of Green's functions that the values of either $U$ or $\partial U/\partial n$ on $S$ are sufficient to specify $U$ at every point $P$ within $S$. (See for example F. Pockels: ber die Partielle Differentialgleichung $\left(\nabla^2 + k^2\right) U = 0$ (Leipzig, Teubner, 1891.) However, only in the simplest cases, e.g. when $S$ is a plane, is it possible to determine the appropriate Green's function [cf. A. Sommerfeld, Optics (New York, Academic Press, 1954), p. 199].

  † See for example B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 2nd ed., 1950), pp. 24-25.

$$U_\omega(P) = \frac{1}{4\pi}\iint_S \left\{ U_\omega \frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right) - \frac{e^{iks}}{s}\frac{\partial U_\omega}{\partial n} \right\} dS \tag{12}$$

When we change the order of integration and set $k = \omega/c$, (10) becomes,

$$\begin{aligned} V(P,t) &= \frac{1}{4\pi}\iint_S dS \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left\{ U_\omega \frac{\partial}{\partial n}\left(\frac{e^{-i\omega(t-s/c)}}{s}\right) - \frac{e^{-i\omega(t-s/c)}}{s}\frac{\partial U_\omega}{\partial n} \right\} d\omega \\ &= \frac{1}{4\pi}\iint_S dS \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \left\{ U_\omega \left\{ \frac{\partial}{\partial n}\left(\frac{1}{s}\right) + \frac{i\omega}{sc}\frac{\partial s}{\partial n} \right\} e^{-i\omega(t-s/c)} - \frac{e^{-i\omega(t-s/c)}}{s}\frac{\partial U_\omega}{\partial n} \right\} d\omega \end{aligned}$$

or using (10),

$$V(P,t) = \frac{1}{4\pi}\iint_S \left\{ [V]\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\frac{\partial s}{\partial n}\left[\frac{\partial V}{\partial t}\right] - \frac{1}{s}\left[\frac{\partial V}{\partial n}\right] \right\} dS \tag{13}$$

The square brackets denote "retarded values", i.e. values of the functions taken at the time $t - s/c$. The formula (13) is the general form of Kirchhoff's theorem.

It can also be seen by analogy with the previous case, that the value of the integra in (13) is zero when $P$ is outside $S$.

The last term in (13) represents the contribution of a distribution of sources of strength $-\frac{1}{4\pi}\frac{\partial V}{\partial n}$ per unit area, whilst the first two terms may be shown to represent a contribution of doublets of strength $V/4\pi$ per unit area, directed normally to the surface. Naturally these sources and doublets are fictitious, there being no deep physical significance behind such an interpretation.

### 8.3.2 Kirchhoff's diffraction theory

Whilst the integral theorem of Kirchhoff embodies the basic idea of the HuygensFresnel principle, the laws governing the contributions from different elements of the surface are more complicated than Fresnel assumed. Kirchhoff showed, however, that in many cases the theorem may be reduced to an approximate but much simpler form, which is essentially equivalent to the formulation of Fresnel, but which in addition gives an explicit formula for the inclination factor that remained undetermined in Fresnel's theory.

Consider a monochromatic wave, from a point source $P_0$, propagated through an opening in a plane opaque screen, and let $P$ as before be the point at which the light disturbance is to be determined. We assume that the linear dimensions of the opening, although large compared to the wavelength, are small compared to the distances of both $P_0$ and $P$ from the screen.

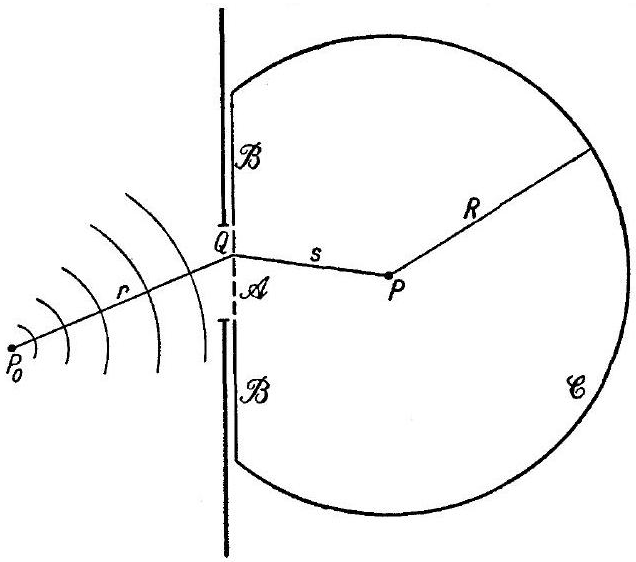
To find the disturbance at $P$ we take Kirchinoff's integral over a surface $S$ formed by (see Fig. 8.3(a)): (1) the opening $\mathscr{A}$, (2) a portion $\mathscr{B}$ of the non-illuminated side of the screen, and (3) a portion $\mathscr{C}$ of a large sphere of radius $R$, centred at $P$ which, together with $\mathscr{A}$ and $\mathscr{B}$, forms a closed surface.

Kirchhoff's theorem, expressed by equation (7), then gives

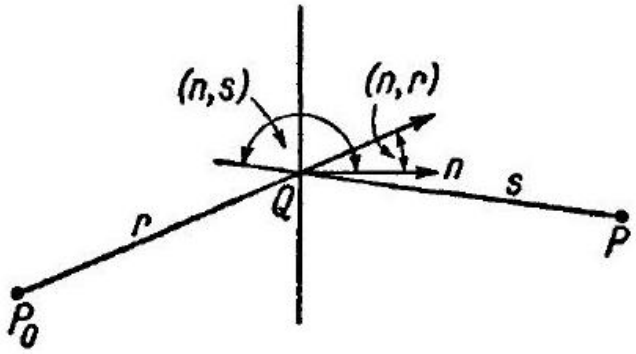
$$U(P) = \frac{1}{4\pi}\left[\iint_{\mathscr{A}} + \iint_{\mathscr{A}} + \iint_{\mathscr{C}}\right] \left\{ U\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right) - \left(\frac{e^{iks}}{s}\right)\frac{\partial U}{\partial n} \right\} dS \tag{14}$$

where, as before, $s$ is the distance of the element $dS$ from $P$ and $\partial/\partial n$ denotes differentiation along the inward normal to the surface of integration.

The difficulty is encountered that the values of $U$ and $\partial U/\partial n$ on $\mathscr{A}, \mathscr{B}$, and $\mathscr{C}$ which should be substituted into (14) are never known exactly. However, it is reasonable to suppose that everywhere on $\mathscr{A}$, except in the immediate neighbourhood of the rim of the opening, $U$ and $\partial U/\partial n$ will not appreciably differ from the values




(a) Fig. 8.3. Illustrating the derivation of the Fresnel-Kirchhoff diffraction formula.




(b)

obtained in the absence of the screen, and that on $\mathscr{B}$ these quantities will be approximately zero. Kirchhoff accordingly set

$$\left.\begin{array}{lll} \text{on } \mathscr{A}: & U = U^{(i)}, & \frac{\partial U}{\partial n} = \frac{\partial U^{(i)}}{\partial n}, \\ \text{on } \mathscr{B}: & U = 0 & \frac{\partial U}{\partial n} = 0, \end{array}\right\} \tag{15}$$

where

$$U^{(i)} = \frac{Ae^{ikr}}{r}, \quad \frac{\partial U^{(i)}}{\partial n} = \frac{Ae^{ikr}}{r}\left[ik - \frac{1}{r}\right]\cos(n, r) \tag{16}$$

are the values relating to the incident field (see Fig. 8.3 (b)) and $A$ is a constant. The approximations (15) are called Kirchhoff's boundary conditions and are the basis of Kirchhoff's diffraction theory.

It remains to consider the contribution from the spherical portion $\mathscr{C}$. Now it is evident that by taking the radius $R$ sufficiently large, the values of $U$ and $\partial U/\partial n$ on $\mathscr{C}$ may be made arbitrarily small, which suggests that the contribution from $\mathscr{C}$ may be neglected. However, by letting $R$ increase indefinitely, the area of $\mathscr{C}$ also increases beyond all limits, so that the condition $U \to 0$ and $\partial U/\partial n \to 0$ as $R \to \infty$ is not sufficient to make the integral vanish. A more precise assumption about the behaviour of the wave function at a large distance from

the screen must therefore be made, a point which we have already touched upon on p. 377 in connection with the uniqueness of solutions in problems involving an infinite medium. For our purposes it is sufficient to make the physically obvious assumption that the radiation field does not exist at all times but that it is produced by a source that begins to radiate at some particular instant of time $t = t_0$.* (This, of course, implies, that we now depart from strict monochromacy, since a perfectly monochromatic field would exist for all

times.) Then at any time $t > t_0$, the field fills a region of space the outer boundary of which is at distance not greater than $c\,(t - t_0)$ from $P_0, c$ being the velocity of light. Hence if the radius $R$ is chosen so large that at the time when the disturbance at $P$ is considered no contributions from $\mathscr{C}$ could have reached $P$ because at the appropriate earlier time the field has not reached these distant regions, the integral over $\mathscr{C}$ will vanish. Thus finally, on substituting into (14), and neglecting in the normal derivatives the terms $1/r$ and $1/s$ in comparison to $k$, we obtain

$$U(P) = -\frac{iA}{2\lambda}\iint_{\mathscr{A}} \frac{e^{ik(r+s)}}{rs}[\cos(n,r) - \cos(n,s)]dS, \tag{17}$$

which is known as the Fresnel-Kirchhoff diffraction formula.

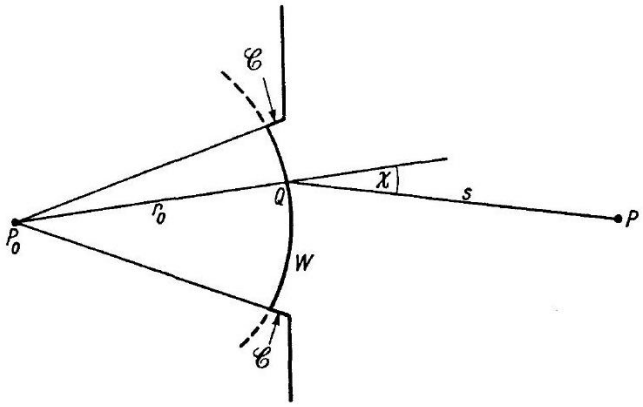



Fig. 8.4. Illustrating the diffraction formula (18).

It is evident that in place of $\mathscr{A}$ any other open surface, the rim of which coincides with the edge of the aperture, could have been chosen. In particular we may choose instead of $\mathscr{A}$ a portion $W$ of an incident wave front which approximately fills the aperture, together with a portion $\mathscr{C}$ of a cone with vertex at $P_0$ and with generators through the rim of the aperture (Fig. 8.4). If the radius of curvature of the wave is sufficiently large, the contribution from $\mathscr{C}$ may obviously be neglected. Also, on $W$, $\cos(n, r_0) = 1$. If further we set $\chi = \pi - (r_0, s)$ we obtain, in place of (17),

$$U(P) = -\frac{i}{2\lambda}\frac{Ae^{ikr_0}}{r_0}\iint_W \frac{e^{iks}}{s}(1 + \cos\chi)dS, \tag{18}$$

where $r_0$ is the radius of the wave-front $W$. This result is in agreement with Fresnel's formulation of Huygens' principle if, as the contribution from the element $dW$ of the wave-front we take

0

- This assumption is not essential but shortens the discussion. For a more formal argument see M. Born, Optik (Berlin, Springer, 1933), p. 149.

$$-\frac{i}{2\lambda}\frac{\mathrm{A}e^{ikr_0}}{r_0}\frac{e^{iks}}{s}(1+\cos\chi)dS. \tag{19}$$

Comparison of (18) with § 8.2 (1) gives for the inclination factor of Fresnel's theory the expression*

$$K(\chi) = -\frac{i}{2\lambda}(1+\cos\chi) \tag{20}$$

For the central zone $\chi = 0$, and (20) gives $K_1 = K(0) = -i/\lambda$ in agreement with § 8.2. (14). It is however not true, as Fresnel assumed, that $K(\pi/2) = 0$.

Returning now to the Fresnel-Kirchhoff diffraction formula (17), we note that it is symmetrical with respect to the source and the point of observation. This implies that a point source at $P_0$ will produce at $P$ the same effect as a point source of equal intensity placed at $P$ will produce at $P_0$. This result is sometimes referred to as the reciprocity theorem (or the reversion theorem) of Helmholtz.

So far we have assumed that the light on its passage from the source to $P$ does not encounter any other surface than the diffracting screen; the incident waves are then spherical. The analysis can be easily extended to cover more complicated cases, where the waves are no longer of such simple form. It is again found that, provided the radii of curvature at each point of the wave-front are large compared to the wave-length, and provided that the angles involved are sufficiently small, the results of Kirchfoff's theory are substantially equivalent to predictions based on the Huygens-Fresnel principle.

From the preceding discussion we can also immediately draw a conclusion which concerns the distribution of light diffracted by complementary screens, i.e. by screens which are such that the openings in one correspond exactly to the opaque portions of the others and vice versa. Let $U_1(P)$ and $U_2(P)$ denote respectively the values of the complex displacement when the first or the second screen alone is placed between the source and the point $P$ of observation, and let $U(P)$ be the value when no screen is present. Then, since $U_1$ and $U_2$ can be expressed as integrals over the openings, and since the openings in the two screens just add up to fill the whole plane,

$$U_1 + U_2 = U. \tag{21}$$

This result is known as Babinet's principle.*

From Babinet's principle two conclusions follow at once: If $U_1 = 0$, then $U_2 = U$; hence at points at which the intensity is zero in the presence of one of the screens, the intensity in the presence of the other is the same as if no screen was present. Further if $U = 0$, then $U_1 = -U_2$; this implies that, at points where $U$ is zero, the phases of $U_1$ and $U_2$ differ by $\pi$ and the intensities $I_1 = |U_1|^2, I_2 = |U_2|^2$ are equal. If for example a point source is imaged by an error-free lens, the light distribution $U$ in the image plane will be zero except in the immediate neighbourhood of the image $O$ of the source. If then complementary screens are placed between the object and the image one has $I_1 = I_2$ except in the neighbourhood of $O$.

The consequences of the basic approximation (15) of Kirchhoff's theory have been subject to many critical discussions, which showed, for example, that Kirchhoff's solution does not reproduce the assumed values in the plane of the aperture †. However, more recently it was shown by Wolf and Marchand ‡ that Kirchhoff's theory

0

- Expression (20) for the inclination factor was first derived by G. G. Stokes, Trans. Camb. Phil. Soc., 9 (1849), 1; reprinted in his Math. and Phys. Papers (Cambridge University Press, 2 (1883), 243).

0

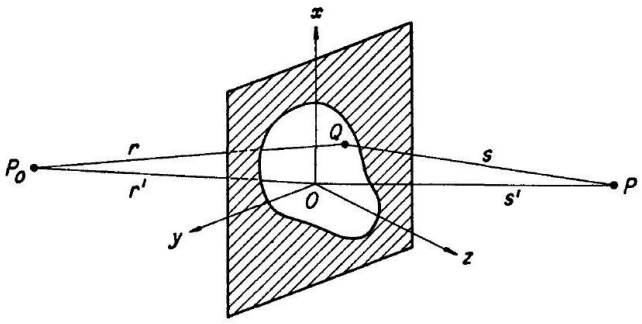
may be interpreted in a mathematically consistent way, as providing an exact solution to a somewhat different boundary value problem than that specified by eqs. (15) and (16). It turns out that Kirchhoff's theory is entirely adequate for the treatment of the majority of problems encountered in instrumental optics. This is mainly due to the smallness of the optical wavelengths in comparison with the dimensions of the diffracting obstacles.* In other problems, such as those relating to the behaviour of the field in the immediate neighbourhood of screens and obstacles, more refined methods have to be used; they must then be considered as boundary-value problems of electromagnetic theory, with the sources as appropriate singularities of the wave functions. Only in a very limited number of cases have such solutions been found; some of them will be discussed in Chapter XI.

### 8.3.3 Fraunhofer and Fresnel diffraction

We now examine more closely the Fresnel-Kirchhoff diffraction integral (17),

$$U(P) = -\frac{Ai}{2\lambda}\iint_{\mathscr{A}} \frac{e^{ik(r+s)}}{rs}[\cos(n,r) - \cos(n,s)]dS. \tag{22}$$

Fig. 8.5. Diffraction at an aperture in a plane screen.




As the element $dS$ explores the domain of integration, $r + s$ will in general change by very many wavelengths, so that, the factor $e^{ik(r+s)}$ will oscillate rapidly. On the other hand, if the distances of the points $P_0$ and $P$ from the screen are large compared to the linear dimensions of the aperture, the factor $[\cos(n,r) - \cos(n,s)]$ will not vary appreciably over the aperture. Further, we assume that if $O$ is any point in the aperture, the angles which the lines $P_0O$ and $OP$ make with $P_0P$ are not too large. We may then replace this factor by $2\cos\delta$, where $\delta$ is the angle between the line $P_0P$ and the normal to the screen. Finally the factor $1/rs$ may be

• A. Babinet, Compt. Rend., 4 (1837), 638. An analogous theorem of this type, which involves the electromagnetic field vectors rather than the single scalar $U$ and which may be considered as rigorous formulation of Babinet's principle, is given in §11.3.

† H. Poincar, Thorie mathmatique de la lumire (Paris, George Carr, II (1892)), pp. 187-8. See also B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 2nd ed., 1950), pp. 71-72 and G. Toraldo di Francia, Atti Fond. Giorgio Ronchi, XI (1956), § 6.

‡ E. Wolf and E. W. Marchand, J. Opt. Soc. Amer., 56 (1966), 1712. Also, it has been shown by F. Kottler, Ann. der Physik, 70 (1923), 405 that Kirchhoff's theory may be regarded as providing a rigorous solution to a certain saltus problem (problem with prescribed discontinuities rather than prescribed boundary values). This interpretation is of particular interest in connection with the problem of diffraction at a black (completely absorbing) screen. [See also F. Kottler, Progress in Optics, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1964), p. 281 and B. B. Baker and E. T. Copson, loc. cit., p. 98.]

An article by C. J. Bouwkamp, Rep. Progr. Phys. (London, Physical Society), 17 (1954), 35, contains references to numerous papers concerned with various modifications of Kirchnoff's theory.

replaced by $1/r's'$, where $r'$ and $s'$ are the distance of $P_0$ and $P$ from the origin and (22) then reduces to

$$U(P) \sim -\frac{Ai}{\lambda}\frac{\cos\delta}{r's'}\iint_{\mathscr{A}} e^{ik(r+s)}dS \tag{23}$$

We take a Cartesian reference system with origin in the aperture and with the $x$ - and $y$-axes in the plane of the aperture and choose the positive $z$ direction to point into the half-space that contains the point $P$ of observation (Fig. 8.5).

If $(x_0, y_0, z_0)$ and $(x, y, z)$ are the coordinates of $P_0$ and of $P$ respectively, and $(\xi, \eta)$ the coordinates of a point $Q$ in the aperture, we have

$$\left.\begin{aligned} r^2 &= (x_0-\xi)^2 + (y_0-\eta)^2 + z_0^2, \\ s^2 &= (x-\xi)^2 + (y-\eta)^2 + z^2, \end{aligned}\right\} \tag{24}$$

$$\left.\begin{aligned} r'^2 &= x_0^2 + y_0^2 + z_0^2 \\ s'^2 &= x^2 + y^2 + z^2 \end{aligned}\right\} \tag{25}$$

Hence

$$\left.\begin{aligned} r^2 &= r'^2 - 2\,(x_0\xi + y_0\eta) + \xi^2 + \eta^2, \\ s^2 &= s'^2 - 2(x\xi + y\eta) + \xi^2 + \eta^2. \end{aligned}\right\} \tag{26}$$

Since we assumed that the linear dimensions of the aperture are small compared to both $r'$ and $s'$ we may expand $r$ and $s$ as power series in $\xi/r', \eta/r', \xi/s'$ and $\eta/s'$. We then obtain

$$\left.\begin{aligned} r &\sim r' - \tfrac{x_0\xi+y_0\eta}{r'} + \tfrac{\xi^2+\eta^2}{2r'} - \tfrac{(x_0\xi+y_0\eta)^2}{2r'^3} - \dots \\ s &\sim s' - \tfrac{x\xi+y\eta}{s'} + \tfrac{\xi^2+\eta^2}{2s'} - \tfrac{(x\xi+y\eta)^2}{2s'^3} - \dots \end{aligned}\right\} \tag{27}$$

Substitution from (27) into (23) gives

$$U(P) = -\frac{i\cos\delta}{\lambda}\frac{Ae^{ik(r'+s')}}{r's'}\iint_{\mathscr{A}} e^{ikf(\xi,\eta)}d\xi d\eta \tag{28}$$

where

$$f(\xi,\eta) = -\frac{x_0\xi+y_0\eta}{r'} - \frac{x\xi+y\eta}{s'} + \frac{\xi^2+\eta^2}{2r'} + \frac{\xi^2+\eta^2}{2s'} - \frac{(x_0\xi+y_0\eta)^2}{2r'^3} - \frac{(x\xi+y\eta)^2}{2s'^3}\dots \tag{29}$$

If we denote by $(l_0, m_0)$ and $(l, m)$ the first two direction cosines

$$\left.\begin{aligned} l_0 &= -\tfrac{x_0}{r'}, \quad l = \tfrac{x}{s'} \\ m_0 &= -\tfrac{y_0}{r'}, \quad m = \tfrac{y}{s'} \end{aligned}\right\} \tag{30}$$

(29) may be written in the form

$$f(\xi,\eta) = (l_0-l)\,\xi + (m_0-m)\,\eta + \frac{1}{2}\Big\{\left(\frac{1}{r'}+\frac{1}{s'}\right)(\xi^2+\eta^2) - \frac{(l_0\xi+m_0\eta)^2}{r'} - \frac{(l\xi+m\eta)^2}{s'}\Big\}\dots \tag{31}$$

0

- Cf. S. Silver, J. Opt. Soc. Amer., 52 (1962), 131.

We have reduced the problem of determining the light disturbance at $P$ to the evaluation of the integral (28). Naturally the evaluation is simpler to carry out when the quadratic and higher order terms in $\xi$ and $\eta$ may be neglected in $f$. In this case one speaks of Fraunhofer diffraction; when the quadratic terms cannot be neglected, one speaks of Fresnel diffraction. Fortunately the simpler case of Fraunhofer diffraction is of much greater importance in optics.

Strictly speaking, the second and higher order terms disappear only in the limiting case $r' \to \infty, s' \to \infty$, i.e. when both the source and the point of observation are at infinity; (the factor $A$ outside the integral must then be assumed to tend to infinity
like $r's'$ ). It is, however, evident that the second order terms do not appreciably contribute to the integral if

$$\frac{1}{2}k\left|\left(\frac{1}{r'}+\frac{1}{s'}\right)\left(\xi^2+\eta^2\right)-\frac{\left(l_0\xi+m_0\eta\right)^2}{r'}-\frac{\left(l\xi+m\eta\right)^2}{s'}\right| \ll 2\pi. \tag{32}$$

We can immediately recognize certain conditions under which (32) will be satisfied. If we make use of inequalities of the form $\left(l_0\xi+m_0\eta\right)^2 \leqslant \left({l_0}^2+{m_0}^2\right)\left(\xi^2+\eta^2\right)$ and remember that ${l_0}^2, {m_0}^2, l^2$ and $m^2$ cannot exceed unity, we find that (32) will be satisfied if

$$|r'| \gg \frac{\left(\xi^2+\eta^2\right)_{\max}}{\lambda} \quad \text{and} \quad |s'| \gg \frac{\left(\xi^2+\eta^2\right)_{\max}}{\lambda}, \tag{33}$$

or if

$$\frac{1}{r'}+\frac{1}{s'}=0 \quad \text{and} \quad l_0^2, m_0^2, l^2, m^2 \ll \frac{|r'|\,\lambda}{\left(\xi^2+\eta^2\right)_{\max}}. \tag{34}$$

The conditions (33) give an estimate of the distances $r'$ and $s'$ for which the Fraunhofer representation may be used. Conditions (34) imply that Fraunhofer diffraction also occurs when the point of observation is situated in a plane parallel to that of the aperture, provided that both the point of observation and the source are sufficiently close to the $z$ axis. Here two cases may be distinguished: When $r'$ is negative, the wave-fronts incident upon the aperture are concave to the direction of the propagation, i.e. $P_0$ is a centre of convergence and not of divergence of the incident wave. This case is of great practical importance, as it arises in the image space of a well-corrected centred system that images a point source which is not far from the axis. A Fraunhofer pattern is then formed in the Gaussian image plane and may be considered as arising from the diffraction of the image-forming wave on the exit pupil. When $r'$ is positive, the wave-fronts are convex to the direction of propagation. The diffraction phenomena are virtual, being apparently formed on a screen through the source $P_0$. This case arises, for example, when an aperture is held in front of the eye, or the object glass of a telescope adjusted for distant vision of the light source.

To understand in more physical terms why Fraunhofer phenomena are observed in the focal plane of a well-corrected lens, let us compare first the two situations illustrated in Fig. 8.6. In (a) a pencil of rays from an infinitely distant point is incident upon the aperture in the direction specified by the direction cosines $l_0, m_0, n_0$. The effect observed at a very distant point $P$ in the direction $l, m, n$ may be regarded as arising from the superposition of plane waves originating at each point of the aperture and propagated in this direction. These waves (which have no existence in the domain of geometrical optics) may be called diffracted waves and the corresponding wavenormals the diffracted rays.

If now a well-corrected lens is placed behind the screen (Fig. 8.6 (b)) all the light diffracted in the ( $l, m, n$ ) direction will be brought to a focus $P'$ in the focal plane of the lens. Since the optical path from a wave-front of the diffracted pencil to $P'$ is the same for all the rays, one obtains substantially the same interference effects as in the first case; it being assumed, of course, that the lens is so large that it introduces no additional diffraction. More generally

still, the restriction that the wave incident upon the aperture is plane may also be removed, provided that the path lengths from the source to $P'$ are substantially the same for all the rays.

In the case of Fraunhofer diffraction, the four quantities $l_0, m_0, l, m$ enter (31) only in the combinations

$$p = l - l_0, \quad q = m - m_0 \tag{35}$$

Hence, within the range of validity of the above approximation, the effect is unchanged when the aperture is displaced in its own plane.

We shall write the integral governing Fraunhofer diffraction in the form

$$U(P) = C \iint_{\mathscr{A}} e^{-ik(p\xi+q\eta)} d\xi d\eta, \tag{36}$$

$C$ being the constant appearing in front of the integral (28). $C$ is defined in terms of quantities depending on the position of the source and of the point of observation, but in practice it is often more convenient to express it in terms of other quantities. Let

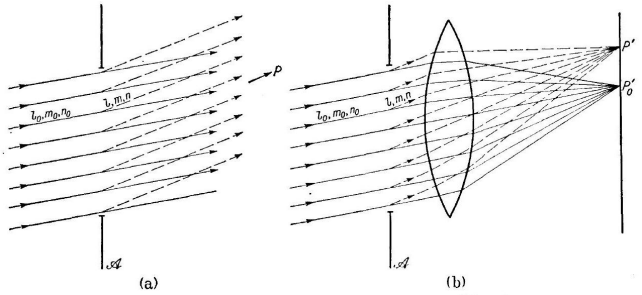



Fig. 8.6. Comparison of two cases of Fraunhofer diffraction.

$E$ be the total energy incident upon the aperture. By the law of conservation of energy the total energy that reaches the plane of observation must also be equal to $E$, so that we have the normalizing condition

$$\iint |U(p,q)|^2 dpdq = E \tag{37}$$

the integration extending over all possible values of $p$ and $q$. Equation (36) may be re-written as a Fourier integral

$$U(p,q) = \iint G(\xi,\eta) e^{-\frac{2\pi i}{\lambda}(p\xi+q\eta)} d\xi d\eta \tag{38}$$

where the pupil function* $G$ is given by

$$\begin{aligned} G(\xi,\eta) &= \text{ constant } (C) \text{ at points in the opening} \\ &= 0 \end{aligned} \tag{39}$$

at points outside opening

and the integral extends over the whole $(\xi,\eta)$ plane.

By Parseval's theorem for Fourier transforms †

$$\iint |G(\xi,\eta)|^2 d\xi d\eta = \left(\frac{1}{\lambda}\right)^2 \iint |U(p,q)|^2 dpdq \tag{40}$$

or, substituting from (37) and (39) and denoting by $D$ the area of the opening,

$$\frac{1}{\lambda^2}E = |C|^2 D \tag{41}$$

whence*

$$C = \frac{1}{\lambda}\sqrt{\frac{E}{D}} \tag{42}$$

The basic integral for Fraunhofer diffraction then takes the form

$$U(p,q) = \frac{1}{\lambda}\sqrt{\frac{E}{D}} \iint_{\mathscr{A}} e^{-ik(p\xi+q\eta)} d\xi d\eta \tag{43}$$

We note that the intensity $I_0 = |U(0,0)|^2$ at the centre of the pattern $p = q = 0$ is given by

$$I_0 = \left(\frac{1}{\lambda}\right)^2 \frac{E}{D} \left(\iint_{\mathscr{A}} d\xi d\eta\right)^2 = \frac{ED}{\lambda^2} = C^2 D^2 \tag{44}$$

In deriving (43) we have disregarded the fact that (36) was obtained subject to certain restrictions on the range of $p$ and $q$. The errors introduced by extending the integration in (40) over all $p$ and $q$ values is, however, negligible, since $U(p,q)$ is very small except in the neighbourhood of $p = q = 0$.

Let us now return to the basic diffraction integral (28). As the point $(\xi, \eta)$ explores the domain of integration, the function $f(\xi, \eta)$ will change by very many wavelengths, so that both the real and imaginary parts of the integrand will change sign many times. In consequence the contributions from the various elements will in general virtually cancel each other out (destructive interference). The situation is, however, different for an element which surrounds a point (called critical point or pole) where $f(\xi, \eta)$ is stationary. Here the integrand varies much more slowly and may be expected to give a significant contribution. Hence, when the wavelength is sufficiently small, the value of the integral is determined substantially by the behaviour of $f$ in the neighbourhood of points where $f$ is stationary. This is the principle of the method of stationary phase for determining the asymptotic behaviour of a certain class of integrals, and is discussed more fully in Appendix III. Here we only note the bearing of this result on the classification of diffraction phenomena:

On comparing (22) and (28) we see that $r + s = r' + s' + f$, so that (see Fig. 8.5)

$$f = P_0 Q + QP + \text{ constant} \tag{45}$$

Obviously $f$ considered as function of $Q$ will be stationary when $Q$ is collinear with $P_0$ and $P$. Hence the main contribution to the disturbance at $P$ comes from the immediate neighbourhood of the point $\bar{Q}$ where the line joining the source to the point of observation intersects the plane of the aperture. Now in the special case of Fraunhofer diffraction, $P_0$ and $P$ are effectively at infinity, so that there is no preferential point $\bar{Q}$. In this case the behaviour of the diffraction integral must, therefore, be expected to be somewhat exceptional.

In § 8.5-§ 8.8 we shall study the most important cases of Fraunhofer and Fresnel diffraction. But first we must justify the use of the single scalar wave function $U$ in calculations of the light intensity.

---

0

• More general pupil functions will be considered in § 8.6 and § 9.5.
† Cf. I. N. Sneddon, Fourier Transforms (New York, McGraw-Hill, 1951), pp. 25 and 44.

0

## 9.11.4 8.4. Transition to a Scalar Theory*

The only property of the $U$ function that we used in the derivation of the Kirchioff integral theorem was that it satisfies the homogeneous scalar wave equation. It therefore follows that this theorem and the conclusion of the preceding section apply to each of the Cartesian components of the field vectors, the vector potential, the Hertz vectors, etc., in regions where there are no currents and charges. To obtain a complete description of the field, the theorem must be applied separately to each of the Cartesian components. Fortunately it turns out that in the majority of problems encountered in optics an approximate description in terms of a single complex scalar wave function is adequate.

A complete description of an electromagnetic field requires the specification of the magnitude of the field vectors as well as their direction (polarization), both as functions of position and time. However, because of the very high frequencies of optical fields (of the order of $10^{14}$/sec.), one cannot measure the instantaneous values of any of these quantities, but only certain time averages over intervals that are large compared to the optical periods. Moreover, one usually deals with natural light, so that there is no preferential polarization direction of the observable (macroscopic) field. The quantity which is then of primary interest is the intensity $I$ defined in § 1.1.4 as the time average of the energy which crosses a unit area containing the electric and magnetic vector in unit time

$$I = \frac{c}{4\pi}|\langle E \wedge H\rangle|.$$

We shall show that the electromagnetic field which is associated with the passage of natural light through an optical instrument of moderate aperture and of conventional design is such that the intensity may approximately be represented in terms of a single complex scalar wave function by means of the formula †

$$I = |U|^2$$

and that the function $U$ may be calculated from the knowledge of the eikonal function of the system.

### 8.4.1 The image field due to a monochromatic oscillator

We consider a symmetrical optical system with a point source at $P_0$ (Fig. 8.7) emitting natural, quasi-monochromatic light of frequency $\omega_0$. We assume that the inclination to the axis of the rays which pass through the system is not large, say not more than 10° or 15°. At $P_0$ we choose a set of Cartesian axes ( $x_1, x_2, x_3$ ) with the $x_3$ direction along the principal ray. The source may be regarded as a dipole of moment $Q(t)$ which varies both in magnitude and direction with time $t$. The components of $Q(t)$ in the three directions will be written in the form of Fourier integrals,

$$Q_j(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} q_j(\omega) e^{-i\omega t} d\omega \quad (j = 1, 2, 3) \tag{1}$$

- We omit here a constant phase factor as it contributes nothing to the intensity $I = |U|^2$.

0

- We follow here substantially the analysis of O. Theimer, G. D. Wassermann and E. Wolf, Proc. Roy. Soc., A., 212 (1952), 426.
  † More generally it was shown by E. Wolf, Proc. Phys. Soc., 74 (1959), 269, that both the (timeaveraged) energy density and energy flow in an unpolarized quasi-monochromatic field may always be derived from one complex time-harmonic scalar wave function.

Since $Q_j(t)$ is real it follows that the complex quantities $q_j(\omega)$ satisfy the relations

$$q_j(-\omega) = q_j^\star(\omega) \tag{2}$$

where the asterisk denotes the complex conjugate. Consequently (1) may be written as

$$Q_j(t) = \mathscr{R}\left\{\sqrt{\frac{2}{\pi}}\int_0^\infty q_j(\omega)e^{-i\omega t}d\omega\right\} \quad (j=1,2,3) \tag{3}$$

$\mathscr{R}$ denoting the real part. Each Fourier component of (3) represents a monochromatic Hertzian oscillator with its axis along the $x_j$ direction.

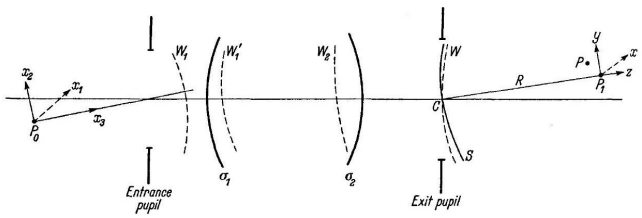



Fig. 8.7. Propagation of an electromagnetic wave through an optical system.

Let $|q_j(\omega)|$ and $\delta_j(\omega)$ be the amplitude and the phase of $q_j(\omega)$,

$$q_j(\omega) = |q_j(\omega)|\, e^{i\delta_j(\omega)}. \tag{4}$$

Since the source is assumed to emit quasi-monochromatic light, the modulus $|q_j(\omega)|$ will, for each $j$, differ appreciably from zero only within a narrow interval $\left(\omega_0 - \frac{1}{2}\Delta\omega,\ \omega_0 + \frac{1}{2}\Delta\omega\right)$. The assumption that the light is natural implies that $\delta_i(\omega)$ are rapidly and irregularly varying functions over the frequency range.*

Since the field may be regarded as a superposition of strictly monochromatic fields, it will be convenient to examine first the contributions from a single monochromatic Hertzian oscillator at $P_0$. As the field of such an oscillator is weak in the neighbourhood of its axis, and as we assume that the angles which the diameters of the entrance pupil subtend at $P_0$ are small, it follows that only the components $Q_1(t)$ and $Q_2(t)$ of $Q(t)$ will substantially contribute to the field. We shall therefore take as our typical oscillator one which has its axis in the $x_1x_2$-plane.

Let

$$\mathscr{R}\left\{q(\omega)\rho_0(\omega)e^{-i\omega t}\right\} \tag{5}$$

be the moment of this typical dipole, $\rho_0(\omega)$ being a unit vector in the direction of its axis. Such a dipole will produce at a point $T$ in vacuum, whose distance from $P_0$ is large compared to the wavelength $\lambda = (2\pi c/\omega)$, a field given by [see § 2.2 (64)]:

$$\left.\begin{aligned}\boldsymbol{E}_\omega &= \mathscr{R}\left\{\tfrac{\omega^2}{c^2 r}|q(\omega)|\boldsymbol{r}_0 \wedge (\boldsymbol{\rho}_0(\omega)\wedge\boldsymbol{r}_0)\, e^{i[\delta(\omega)-\omega(t-r/c)]}\right\},\\ \boldsymbol{H}_\omega &= \mathscr{R}\left\{\tfrac{\omega^2}{c^2 r}|q(\omega)|\boldsymbol{r}_0 \wedge \boldsymbol{\rho}_0(\omega)e^{i[\delta(\omega)-\omega(t-r/c)]}\right\},\end{aligned}\right\} \tag{6}$$

where $r_0$ denotes the unit radial vector.

Let $W_1$ be a typical geometrical wave-front in the object space at a distance from $P_0$ which is large compared with the wavelength. Since we assume that the angles which the rays make

0

• For a detailed discussion of this point see M. Planck, Ann. d. Physik, (4), 1 (1900), 61.

with the axis of the system are small, it follows immediately from (6) that at any particular instant of time the vectors $\boldsymbol{E}_{\boldsymbol{\omega}}$ and $\boldsymbol{H}_{\boldsymbol{\omega}}$ do not vary appreciably in magnitude and direction over $W_1$.

The effect of the first surface* $\sigma_1$ on the incident field is twofold. First, the amplitudes of the field vectors are diminished on account of reflection losses; secondly, the directions of vibrations are changed. Fresnel's formulae show that both these effects depend mainly on the magnitudes of the angle of incidence. If this angle is small (say 10° or so), reflection losses are also small (approx. 5 per cent) and the rotations of the planes of vibration do not exceed a few degrees (cf. § 1.5). Moreover, these effects are practically uniform over $\sigma_1$. Since the time-independent parts of $\boldsymbol{E}_{\omega}$ and $\boldsymbol{H}_{\omega}$ do not vary appreciably with position over the wave-front $W_1$, they will also not vary appreciably over the refracted wave-front $W_1'$ which follows the surface $\sigma_1$ (see Fig. 8.7). The same applies to the behaviour of the two fields over any other wave-front in the space between $\sigma_1$ and the second surface $\sigma_2$. For, as we showed in §3.1.3, in a homogeneous medium the direction of vibration along each ray remains constant, and also, since the wave-fronts are nearly spherical (centred on the Gaussian image of $P_0$ by the first surface), the amplitudes will be diminished almost in the ratio of their paraxial radii of curvature.

Repeating these arguments we finally arrive at a wave-front $W$ which passes through the centre $C$ of the exit pupil and find again that the time-independent parts of $\boldsymbol{E}_{\boldsymbol{\omega}}$ and $\boldsymbol{H}_{\boldsymbol{\omega}}$ do not vary appreciably over this wave-front. This result makes it immediately possible to write down an approximate mathematical representation for the field vectors in the region of the image.

We take rectangular Cartesian axes $(x, y, z)$ with origin at the Gaussian image $P_1$ of $P_0$, with the $z$ direction along $CP_1$. The field at all points in the region of the aperture except those in the immediate neighbourhood of the edge of the aperture can be approximately expressed in the form (cf. Chapter III)

$$\left.\begin{aligned}\boldsymbol{E}_{\omega}(x,y,z,t) &= \mathscr{R}\left\{\tfrac{\omega^2}{c^2}\boldsymbol{e}_{\omega}(x,y,z)e^{i\left\{\delta(\omega)-\omega\left[t-\frac{1}{c}\mathscr{S}_{\omega}(x,y,z)\right]\right]}\right\},\\ \boldsymbol{H}_{\omega}(x,y,z,t) &= \mathscr{R}\left\{\tfrac{\omega^2}{c^2}\boldsymbol{h}_{\omega}(x,y,z)e^{i\left\{\delta(\omega)-\omega\left[t-\frac{1}{c}\mathscr{S}_{\omega}(x,y,z)\right]\right\}}\right\},\end{aligned}\right\} \tag{7}$$

which may be regarded as generalization of (6). Here $\mathscr{S}_{\omega}(x, y, z)$ is the optical length from the object point to the point $(\ x, y, z\ )$, and $\boldsymbol{e}_{\omega}(x, y, z)$ and $\boldsymbol{h}_{\omega}(x, y, z)$ are mutually orthogonal real vectors. † In a homogeneous non-magnetic medium of refractive index $n$, these vectors satisfy the relation (see eq. (19) and (20) of §3.1)

$$|\boldsymbol{h}_{\omega}| = n\,|\boldsymbol{e}_{\omega}|\,. \tag{8}$$

We take a reference sphere $S$, centred on $P_1$, which passes through the centre $C$ of the exit pupil, and denote by $R$ its radius $CP_1$. In practice the distance between

$S$ and $W$ will nowhere exceed a few dozen wavelengths. Consequently on $S$ just as on $W$ the amplitude vectors $\boldsymbol{e}_{\omega}$ and $\boldsymbol{h}_{\omega}$ will be practically constant in magnitude and direction.

Let $P(X, Y, Z)$ be a point in the region of the image where the intensity is to be determined. If the angles which the diameters of the exit pupil subtend at $P$ are small, we may apply Kirchhoff's formula with the same approximation as in the previous section, and we find on

0

- We assume here that $\sigma_1$ is a refracting surface. If $\sigma_1$ is a mirror, no essential modifications of our argument are necessary, as is seen by inspection of Fresnel's formulae.
  † That $\boldsymbol{e}_{\omega}$ and $\boldsymbol{h}_{\omega}$ are real follows from the fact that the corresponding vectors in (6) are real (linear polarization) and that the state of polarization remains linear on each refraction [cf. p. 40]. Moreover, between any two consecutive surfaces, the state of polarization is constant along each ray, as shown in § 3.1.3.

integrating the expressions (7) over that part $S'$ of $S$ which approximately fills the exit pupil, if in addition we also neglect the variation of the inclination factor over $S'$, that

$$\left.\begin{aligned}\boldsymbol{E}_\omega(X,Y,Z,t) &= \mathscr{R}\tfrac{\omega^3}{2\pi i c^3}e^{i[\delta(\omega)-\omega t]}\iint_{S'}\tfrac{1}{s}\boldsymbol{e}_\omega\,(x',y',z')\,e^{i\frac{\omega}{c}[\mathscr{S}_\omega(x',y',z')+s]}dS,\\ \boldsymbol{H}_\omega(X,Y,Z,t) &= \mathscr{R}\tfrac{\omega^3}{2\pi i c^3}e^{i[\delta(\omega)-\omega t]}\iint_{S'}\tfrac{1}{s}\boldsymbol{h}_\omega\,(x',y',z')\,e^{i\frac{\omega}{c}[\mathscr{S}_\omega(x',y',z')+s]}dS,\end{aligned}\right\} \tag{9}$$

where $s$ is the distance from a typical point $(\,x',y',z'\,)$ on the reference sphere to $P$. Since the vectors $\boldsymbol{e}_\omega\,(\boldsymbol{x}',\boldsymbol{y}',\boldsymbol{z}')$ and $\boldsymbol{h}_\omega\,(x',y',z')$ do not vary appreciably over the surface of integration we may replace them by the values $\boldsymbol{e}_\omega(0,0,-R)$ and $\boldsymbol{h}_\omega(0,0,-R)$ which they take at the centre $C$ of the exit pupil. Now these vectors are orthogonal and satisfy (8), so that we may set, if in addition we take $n=1$,

$$\left.\begin{aligned}\boldsymbol{e}_\omega(0,0,-R) &= a(\omega)\boldsymbol{\alpha}(\omega),\\ \boldsymbol{h}_\omega(0,0,-R) &= a(\omega)\boldsymbol{\beta}(\omega),\end{aligned}\right\} \tag{10}$$

where $\alpha(\omega)$ and $\beta(\omega)$ are orthogonal unit vectors in the plane perpendicular to the $z$ direction. The relations (9) then become

$$\left.\begin{aligned}\boldsymbol{E}_\omega(X,Y,Z,t) &= \mathscr{R}\left\{\tfrac{\omega^2}{c^2}U_\omega(X,Y,Z)a(\omega)\alpha(\omega)e^{i[\delta(\omega)-\omega t]}\right\},\\ \boldsymbol{H}_\omega(X,Y,Z,t) &= \mathscr{R}\left\{\tfrac{\omega^2}{c^2}U_\omega(X,Y,Z)a(\omega)\boldsymbol{\beta}(\omega)e^{i[\delta(\omega)-\omega t]}\right\},\end{aligned}\right\} \tag{11}$$

where $U_\omega$ is the scalar wave function

$$U_\omega(X,Y,Z) = \frac{\omega}{2\pi i c}\iint_{S'}\frac{1}{s}e^{i\frac{\omega}{c}[\mathscr{S}_\omega(x',y',z')+s]}dS \tag{12}$$

From (11) we can immediately deduce by calculating the Poynting vector $S_\omega = c\,[\boldsymbol{E}_\omega \wedge \boldsymbol{H}_\omega]\,/4\pi$ and taking the time average, that the intensity at the point $P(X,Y,Z)$ due to the single dipole (represented by (5)) at $P_0$ is proportional to the square of the modulus of the scalar wave function $U_\omega(X,Y,Z)$. However, to justify the use of a single scalar wave function in calculating the intensity we must carry out the time averaging not for the monochromatic component but for the total field.

### 8.4.2 The total image field

We saw that the contributions of each frequency component to the total field may be regarded as arising essentially from two dipoles at $P_0$ with their axes along the $x_1$ and $x_2$ directions. Hence it follows from (1) and (11), if we also define contributions from negative frequencies by relations of the form (2), that the total field in the image region may be expressed approximately in the form

$$\left.\begin{aligned}E(X,Y,Z,t) &= \tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-\infty}^{\infty}\tfrac{\omega^2}{c^2}U_\omega(X,Y,Z)\left[a_1(\omega)\alpha_1(\omega)e^{i\delta_1(\omega)}\right.\\ &\qquad\left.+a_2(\omega)\alpha_2(\omega)e^{i\delta_2(\omega)}\right]e^{-i\omega t}d\omega\\ H(X,Y,Z,t) &= \tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-\infty}^{\infty}\tfrac{\omega^2}{c^2}\dot{U}_\omega(X,Y,Z)\left[a_1(\omega)\beta_1(\omega)e^{i\delta_1(\omega)}\right.\\ &\qquad\left.+a_2(\omega)\beta_2(\omega)e^{i\delta_2(\omega)}\right]e^{-i\omega t}d\omega\end{aligned}\right\} \tag{13}$$

Here suffixes 1 and 2 refer to the contributions from oscillators which have their axes along the $x_1$ and $x_2$ directions.

In order to determine the intensity in the image region it will be convenient to write down separate expressions for each of the Cartesian components of $\boldsymbol{E}$ and $\boldsymbol{H}$. Let $\theta_1(\omega)$ and $\theta_2(\omega)$ denote the angles which the unit vectors $\alpha_1(\omega)$ and $\alpha_2(\omega)$ make with the $x$ direction in the image space. Since $\alpha_1(\omega)$ and $\beta_1(\omega)$ and $\alpha_2(\omega)$ and $\beta_2(\omega)$ are real, mutually orthogonal vectors which lie in a plane perpendicular to the $z$ direction, it follows from (13) that the components of $\boldsymbol{E}$ and $\boldsymbol{H}$ are approximately given by*

$$\left.\begin{aligned} E_x(X,Y,Z,t) &= H_y(X,Y,Z,t) = \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega(X,Y,Z) f(\omega) e^{-i\omega t} d\omega \\ E_y(X,Y,Z,t) = -H_x(X,Y,Z,t) &= \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega(X,Y,Z) g(\omega) e^{-i\omega t} d\omega \\ E_z(X,Y,Z,t) = H_z(X,Y,Z,t) &= 0 \end{aligned}\right\} \tag{14}$$

where

$$\left.\begin{aligned} f(\omega) &= \tfrac{\omega^2}{c^2}\left[a_1(\omega)\cos\theta_1(\omega) e^{i\delta_1(\omega)} + a_2(\omega)\cos\theta_2(\omega) e^{i\delta_2(\omega)}\right] \\ g(\omega) &= \tfrac{\omega^2}{c^2}\left[a_1(\omega)\sin\theta_1(\omega) e^{i\delta_1(\omega)} + a_2(\omega)\sin\theta_2(\omega) e^{i\delta_2(\omega)}\right] \end{aligned}\right\} \tag{15}$$

It follows from (14) that the magnitude of the Poynting vector $S = c[E \wedge H]/4\pi$ can be expressed approximately in the form

$$|S| = \frac{c}{4\pi}\left[E_x^2 + E_y^2\right] = \frac{c}{4\pi}\left[H_x^2 + H_y^2\right] \tag{16}$$

We must now determine the time average of this quantity. †
For reasons of convergence we assume that the radiation field exists only between the instants $t = -T$ and $t = T$, where $T \gg 2\pi/\omega_0$; it is easy to pass to the limit $T \to \infty$ subsequently. It follows from (14), by the Fourier inversion theorem, that

$$U_\omega(X,Y,Z) f(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-T}^{T} E_x(X,Y,Z,t) e^{i\omega t} dt \tag{17}$$

with similar expressions involving $E_y, H_x$, and $H_y$. Now we have by (14)

$$\left\langle E_x^2 \right\rangle = \frac{1}{2T}\int_{-T}^{T} E_x^2 dt = \frac{1}{2T}\int_{-T}^{T} E_x dt \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} U_\omega f(\omega) e^{-i\omega t} d\omega \tag{18}$$

or, inverting the order of integration,

$$\begin{aligned} \left\langle E_x^2 \right\rangle &= \frac{1}{2T}\int_{-\infty}^{+\infty} U_\omega f(\omega) d\omega \frac{1}{\sqrt{2\pi}}\int_{-T}^{T} E_x e^{-i\omega t} dt \\ &= \frac{1}{2T}\int_{-\infty}^{+\infty} U_\omega f_\omega U_\omega^\star f_\omega^\star d\omega \quad \text{(by 17)} \\ &= \frac{1}{T}\int_0^{\infty} |U_\omega|^2 |f(\omega)|^2 d\omega \end{aligned} \tag{19}$$

since $U_{-\omega} f(-\omega) = U_\omega^\star f^\star(\omega)$. Similarly

$$\left\langle E_y^2 \right\rangle = \frac{1}{T}\int_0^{\infty} |U_\omega|^2 |g(\omega)|^2 d\omega. \tag{20}$$

Hence, the intensity $I(X,Y,Z)$, defined as the time average of the magnitude of the Poynting vector, is, according to (16), (19), and (20)

---

0

- It would be incorrect to conclude from (14) that the direction of the energy flow in the image region is necessarily everywhere parallel to $z$. For the relative errors in (14) may substantially affect the calculations of direction in regions where the intensity is small, e.g. in the neighbourhood of the dark rings in the Airy pattern.

† Our method of evaluating the average is due to M. Born and P. Jordan, Z. f. Phys., 33 (1925), 479.

$$I(X,Y,Z) = \frac{c}{4\pi T}\int_0^\infty |U_\omega(X,Y,Z)|^2 \left[|f(\omega)|^2 + |g(\omega)|^2\right] d\omega \tag{21}$$

Now if $|\Delta\omega|$ is sufficiently small $|U_\omega|$ will be practically independent of $\omega$ over the effective frequency range, so that $|U_\omega|$ may then be replaced by $|U_{\omega_0}|$ and taken outside the integral. The remaining term

$$\frac{c}{4\pi T}\int_0^\infty \left\{|f(\omega)|^2 + |g(\omega)|^2\right\} d\omega, \tag{22}$$

which is independent of $X, Y$, and $Z$ must also be independent of $T$ (implicitly contained in $f$ and $g$ on account of (17)), if a stationary phenomenon is observed. Hence (22) must be a constant ( $C$ say) and the intensity may therefore be finally written in the form

$$I(X,Y,Z) = C\,|U_{\omega_0}(X,Y,Z)|^2\,. \tag{23}$$

The constant $C$ depends in a complicated manner on the source and on the optical instrument; however one is usually only interested in the relative distribution of the intensity and not in its absolute value. The intensity may then simply be measured by the quantity $|U_{\omega_0}|^2$. Thus the complex scalar function (12) is adequate for calculating the intensity distribution in the image formed with a source of natural light by an optical system of moderate numerical aperture.

## 9.11.5 8.5. Fraunhofer Diffraction at Apertures of Various Forms

We shall now investigate the Fraunhofer diffraction pattern for apertures of various forms.

### 8.5.1 The rectangular aperture and the slit

Consider first a rectangular aperture of sides $2a$ and $2b$. With origin $O$ at the centre of the rectangle and with $O\xi$ and $O\eta$ axes parallel to the sides (Fig. 8.8), the Fraunhofer diffraction integral § 8.3 (36) becomes

$$U(P) = C\int_{-a}^{a}\int_{-b}^{b} e^{-ik(p\xi+q\eta)} d\xi d\eta = C\int_{-a}^{a} e^{-ikp\xi} d\xi \int_{-b}^{b} e^{-ikq\eta} d\eta$$

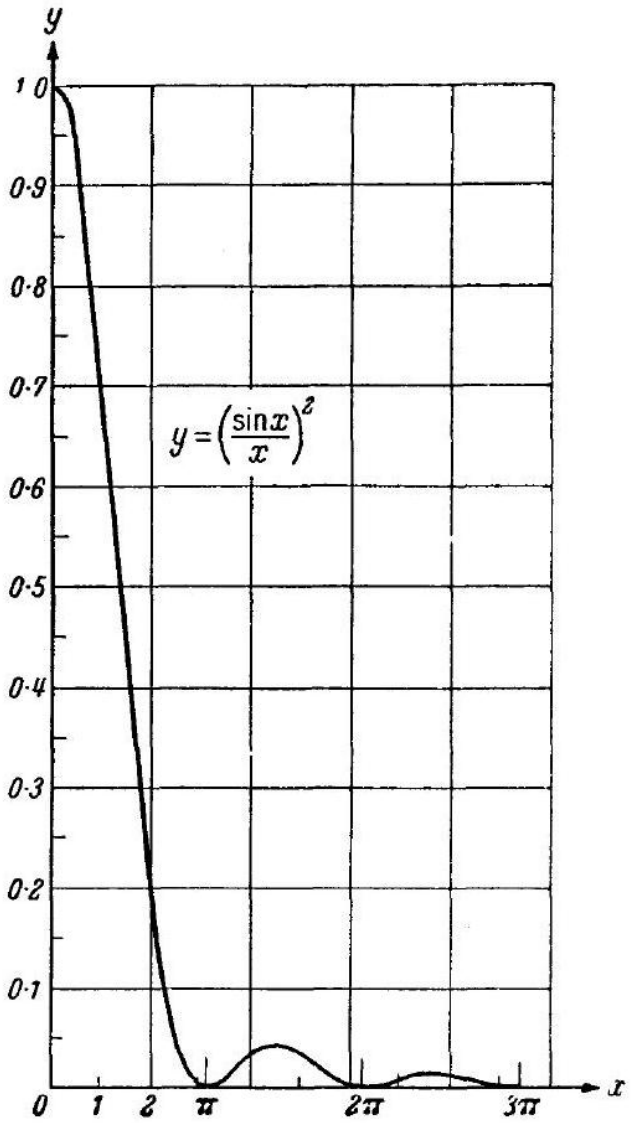



Fig. 8.9. Fraunhofer diffraction at a rectangular aperture. The function

$$y = \left(\frac{\sin x}{x}\right)^2$$

Now

$$\int_{-a}^{a} e^{-ikp\xi} d\xi = -\frac{1}{ikp}\left[e^{-ikpa} - e^{ikpa}\right] = 2\frac{\sin kpa}{kp}$$

with a similar expression for the other integral. Hence the intensity is given by

$$I(P) = |U(P)|^2 = \left(\frac{\sin kpa}{kpa}\right)^2 \left(\frac{\sin kqb}{kqb}\right)^2 I_0 \tag{1}$$

where by §8.3 (44) $I_0 = C^2D^2 = ED/\lambda^2$ is the intensity at the centre of the pattern, $E$ being the total energy incident upon the aperture and $D = 4ab$ the area of the rectangle.

The function $y = (\sin x/x)^2$ is displayed in Fig. 8.9. It has a principal maximum $y = 1$ at $x = 0$ and zero minima at $x = \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$. The minima separate the secondary maxima whose positions are given by the roots of the equation $\tan x - x = 0$

(see Table XIX). The roots asymptotically approach the values $x = (2m + 1)\pi/2, m$ being an integer.

$$y = \left(\frac{\sin x}{x}\right)^2$$

Table XIX
The first five maxima of the function

| $x$ | $y = \left(\frac{\sin x}{x}\right)^2$ |
|---|---|
| 0 | 1 |
| $1 \cdot 430\pi = 4 \cdot 493$ | 0.04718 |
| $2 \cdot 459\pi = 7 \cdot 725$ | $0 \cdot 01694$ |
| $3 \cdot 470\pi = 10 \cdot 90$ | $0 \cdot 00834$ |
| $4 \cdot 479\pi = 14 \cdot 07$ | 0.00503 |

We see that the intensity $I(P)$ is zero along two sets of lines parallel to the sides of the rectangle, given by

$$kpa = \pm u\pi, \quad kqb = \pm v\pi, \quad (u, v = 1, 2, 3, \ldots) \tag{2}$$

or, since $p = l - l_0, q = m - m_0, k = 2\pi/\lambda$,

$$l - l_0 = \pm\frac{u\lambda}{2a}, \quad m - m_0 = \pm\frac{v\lambda}{2b}. \tag{3}$$

Within each rectangle formed by pairs of consecutive dark lines the intensity rises to a maximum; all these maxima are, however, only a small fraction of the central maximum, and decrease rapidly with increasing distance from the centre (Fig. 8.10). The larger the opening, the smaller is the effective size of the diffraction pattern.

From the elementary diffraction pattern formed by coherent light from a point source, the diffraction pattern due to light from an extended source may be found by integration. If the source is coherent, it is the complex amplitude, and if it is incoherent, it is the intensity that must be integrated. The pattern due to a partially coherent source may also be determined from this elementary solution by a process of integration, taking into account the correlation which exists between the light from the different elements of the source (cf. Chapter X). A case of particular importance is that of a very long incoherent line source (e.g. a luminous wire), the light from which is diffracted by a narrow slit, parallel to the source. For simplicity of calculations we assume that the luminous wire as well as the slit are effectively infinitely long, and take the $y$-axis in the direction of the source. Since $q = m - m_0$ where $m_0$ specifies the position of a point source, it follows that the intensity $I'$ due to the line source is obtained by integrating (1) with respect to $q$ :

$$I' = \int_{-\infty}^{+\infty} I(P)dq = \frac{1}{kb}\left(\frac{\sin kpa}{kpa}\right)^2 I_0 \int_{-\infty}^{+\infty}\left(\frac{\sin t}{t}\right)^2 dt$$

Now*

$$\int_{-\infty}^{+\infty}\left(\frac{\sin t}{t}\right)^2 dt = \pi$$

so that

$$I' = \left(\frac{\sin kpa}{kpa}\right)^2 I_0' \tag{4}$$

where

$$I_0' = \frac{\lambda}{2b} I_0 = \frac{2aE}{\lambda} \tag{5}$$

The pattern is again characterized by the function $(\sin x/x)^2$, and consists of a succession of bright and dark fringes parallel to the line source and the slit. The constant $I_0'$ is the intensity at the central position $p = 0$.

### 8.5.2 The circular aperture

In a similar way we may investigate Fraunhofer diffraction at a circular aperture. It is now appropriate to use polar instead of rectangular coordinates. Let ( $\rho, \theta$ ) be the polar coordinates of a typical point in the aperture:

$$\rho \cos\theta = \xi, \quad \rho \sin\theta = \eta \tag{6}$$

and let $(w, \psi)$ be the coordinates of a point $P$ in the diffraction pattern, referred to the geometrical image of the source:

$$w \cos\psi = p, \quad w \sin\psi = q. \tag{7}$$

From the definition of $p$ and $q$ it follows that $w = \sqrt{p^2 + q^2}$ is the sine of the angle which the direction ( $p, q$ ) makes with the central direction $p = q = 0$. The diffraction integral § 8.3 (36) now becomes, if $a$ is the radius of the circular aperture,

$$U(P) = C \int_0^a \int_0^{2\pi} e^{-ik\rho w \cos(\theta - \psi)} \rho d\rho d\theta \tag{8}$$

Now we have the well-known integral representation of the Bessel functions †$J_n(z)$ :

$$\frac{i^{-n}}{2\pi} \int_0^{2\pi} e^{ix\cos\alpha} e^{in\alpha} d\alpha = J_n(x) \tag{9}$$

Equation (8) therefore reduces to

$$U(P) = 2\pi C \int_0^a J_0(k\rho w) \rho d\rho \tag{10}$$

Also, there is the well-known recurrence relation ‡

$$\frac{d}{dx}\left\{x^{n+1} J_{n+1}(x)\right\} = x^{n+1} J_n(x) \tag{11}$$

giving, for $n = 0$, on integration

0

- See, for example, W. Gröbner and N. Hofreiter, Integraltafel, Vol. II (Wien, Springer, 1950), p. 333.
† See, for example, E. Jahnke and F. Emde, Tables of Functions with Formulae and Curves (Leipzig, Teubner, 1933; reprinted by Dover Publications, New York, 4th ed., 1945), p. 149 ; or G. N. Watson, A Treatise on the Theory of Bessel Functions (Cambridge University Press, 1922), p. 20, equation 5 (with an obvious substitution).
‡ See for example E. Jahnke and F. Emde, loc. cit., p. 145 or E. T. Whittaker and G. N. Watson, A Course of Modern Analysis (Cambridge University Press, 4th ed., 1952), pp. 360-361.

$$\int_0^x x' J_0(x')\,dx' = xJ_1(x) \tag{12}$$

From (10) and (12) it follows that

$$U(P) = CD\left[\frac{2J_1(kaw)}{kaw}\right], \tag{13}$$

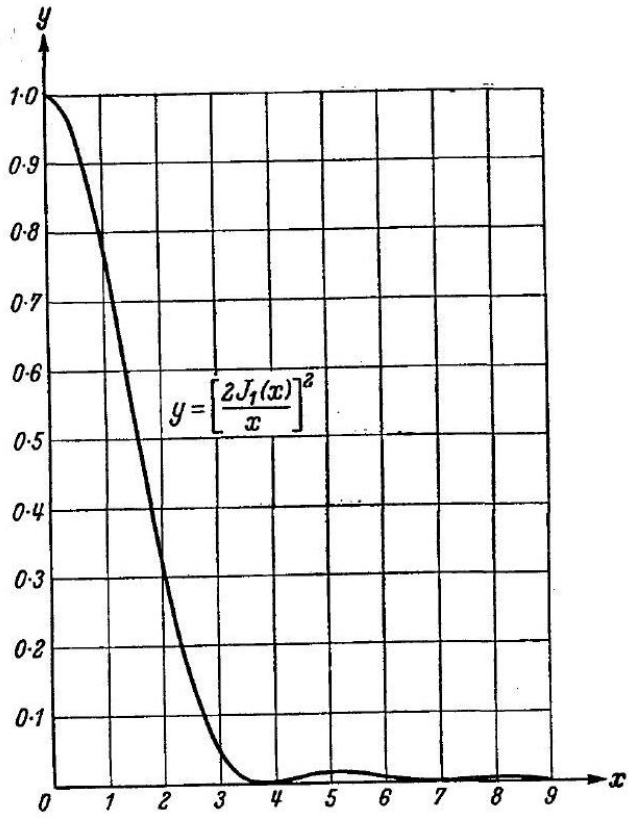



Fig. 8.11. Fraunhofer diffraction at a circular aperture. The function

$$y = \left(\frac{2J_1(x)}{x}\right)^2$$

where $D = \pi a^2$. Hence the intensity is given by

$$I(P) = |U(P)|^2 = \left[\frac{2J_1(kaw)}{kaw}\right]^2 I_0 \tag{14}$$

where by §8.3(44)$I_0 = C^2D^2 = ED/\lambda^2$. This is a celebrated formula first derived in a somewhat different form by Airy.*

The intensity distriburion in the neighbourhood of the geometrical image is characterized by the function $y = (2J_1(x)/x)^2$ shown in Fig. 8.11. It has its principal maximum $y = 1$ at $x = 0$, and with increasing $x$ it oscillates with gradually diminishing amplitude, in a similar way to the function $(\sin x/x)^2$ which we discussed in § 8.5.1. The intensity is zero (minimum) for values of $x$ given by $J_1(x) = 0$. The minima are

0

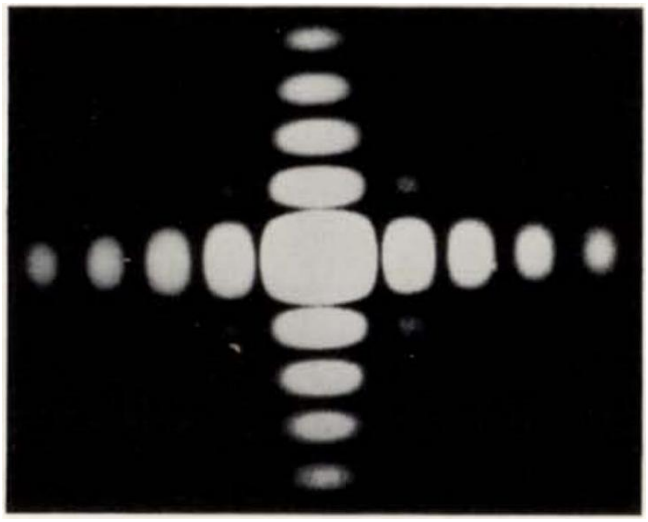


Fig. 8.10. Fraunhofer diffraction pattern of a rectangular aperture \$8 \mathrm{˜mm} \times 7 \mathrm{˜mm}\$, magnification 50×, mercury yellow light $\lambda = 5790$Å. To show the existence of the weak secondary maxima the central portion was overexposed.

(After H. Lipson, C. A. Taylor, and B. J. Thompson.)



Fig. 8.12. Fraunhofer diffraction pattern of a circular aperture (the Airy pattern) 6 mm in diameter, magnification 50×, mercury yellow light $\lambda = 5790$Å. To show the existence of the weak subsidiary maxima, the central portion was overexposed.
(After H. Lipson, C. A. Taylor, and B. J. Thompson.)

- G. B. Airy, Trans. Camb. Phil. Soc., 5 (1835), 283. Almost at the same time as Airy, Schwerd obtained an approximate solution by replacing the circle by a regular polygon with 180 sides.

Vectorial treatments of diffraction of a convergent spherical wave at a circular aperture, which take into account polarization properties of the field were published by W. Innatowski, Trans. Opt. Inst. Petrograd., 1 (1919) IV; V. A. Fock, ibid., 3 (1924), 24; H. H. Hopkins, Proc. Phys. Soc. 55 (1943), 116; R. Burtin, Optica Acta, 3 (1956), 104; B. Richards and E. Wolf, Proc. Roy. Soc., A, 253 (1959), 358; A. Boivin and E. Wolf, Phys. Rev., 138 (1965), B 1561; A. Boivin, J. Dow and E. Wolf, J. Opt. Soc. Amer., 57 (1967), 1171,

$$y = \left(\frac{2J_1(x)}{x}\right)^2.$$

Table XX
The first few maxima and minima of the function

| $x$ | $\left(\frac{2J_1(x)}{x}\right)^2$ | |
|---|---|---|
| 0 | 1 | Max. |
| $1 \cdot 220\pi = 3.833$ | 0 | Min. |
| $1.635\pi = 5.136$ | 0.0175 | Max. |
| $2 \cdot 233\pi = 7 \cdot 016$ | 0 | Min. |
| $2 \cdot 679\pi = 8 \cdot 417$ | 0.0042 | Max. |
| $3 \cdot 238\pi = 10 \cdot 174$ | 0 | Min. |
| $3.699\pi = 11.620$ | 0.0016 | Max. |

no longer strictly equidistant [see Table XX]. The positions of the secondary maxima are given by the values of $x$ that satisfy the equation

$$\frac{d}{dx}\left[J_1(x)/x\right] = 0,$$

or using the formula* (analogous to (11))

$$\frac{d}{dx}\left[x^{-n}J_n(x)\right] = -x^{-n}J_{n+1}(x), \tag{15}$$

by the roots of the equations $J_2(x) = 0$. With increasing $x$ the separation between two successive minima or two successive maxima approaches the value $\pi$, as in the previous case. The results show that the pattern consists of a bright disc, centred on the geometrical image $p = q = 0$ of the source, surrounded by concentric bright and dark rings (see Figs. 8.11 and 8.12). The intensity of the bright rings decreases rapidly with their radius and normally only the first one or two rings being bright enough to be visible to the naked eye. From Table XX it follows, since $x = 2\pi aw/\lambda$, that the radii of the dark rings are

$$w = \sqrt{p^2 + q^2} = 0.610\frac{\lambda}{a}, \quad 1.116\frac{\lambda}{a}, \quad 1.619\frac{\lambda}{a}, \ldots \tag{16}$$

The separation between two neighbouring rings approaches asymptotically the value $\lambda/2a$. The effective size of the diffraction pattern is again seen to be inversely proportional to the linear dimensions of the aperture.

It is also of interest to examine what fraction of the total incident energy is contained within the central core of the diffraction pattern. Denoting by $L(w_0)$ the fraction of the total energy contained within a circle of radius $w_0$ in the image plane, centred on the geometrical image, we have

0

- See for example E. Jahnke and F. Emde, loc. cit., p. 145, or E. T. Whittaker and G. N. Watson, loc. cit., p. 361.

$$
\begin{aligned}
L(w_0) &= \frac{1}{E}\int^{w_0}\int^{2\pi} I(w)wdwd\psi \\
&= \frac{D}{\lambda^2}\int_0^{w_0}\int_0^{2\pi}\left[\frac{2J_1(kaw)}{kaw}\right]^2 wdwd\psi \\
&= 2\int_0^{kaw_0}\frac{J_1^2(x)}{x}dx
\end{aligned}
\tag{17}
$$

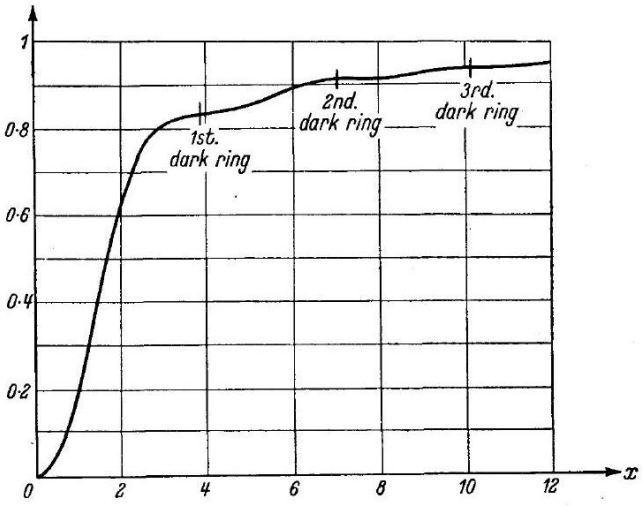



Fig. 8.13. The function $1-J_{0}{ }\hat{}{}{2}(x)-J_{1}{ }\hat{}{}{2}(x)$ representing the fraction of the total energy contained within circles of prescribed radii in the Fraunhofer diffraction pattern of a circular aperture.

Now from (11) for $n = 0$, we have, on multiplying by $J_1(x)$ and using (15) with $n = 0$,

$$
\begin{aligned}
\frac{J_1^2(x)}{x} &= J_0(x)J_1(x) - \frac{dJ_1(x)}{dx}J_1(x) \\
&= -\frac{1}{2}\frac{d}{dx}\left[J_0^2(x) + J_1^2(x)\right]
\end{aligned}
$$

The expression (17) now becomes, remembering that $J_0(0) = 1, J_1(0) = 0$,

$$
L(w_0) = 1 - J_0^2(kaw_0) - J_1^2(kaw_0), \tag{18}
$$

a formula due to Rayleigh.* This function is shown in Fig. 8.13. For the dark rings $J_1(kaw_0) = 0$, so that the fraction of the total energy outside any dark ring is simply ${J_0}^2(kaw_0)$. For the first, second, and third dark ring, ${J_0}^2(kaw_0)$ is equal to $0.162, 0.090$, and $0.062$ respectively. Thus more than 90 per cent of the light is contained within the circle bounded by the second dark ring.

### 8.5.3 Other forms of aperture

Fraunhofer diffraction at apertures of other forms may be studied in a similar manner, the calculations being particularly simple when curvilinear coordinates can be chosen so that one of the coordinate lines coincides with the boundary of the aperture.

0

We cannot discuss other cases in detail here,* but we shall derive a useful theorem concerning the modification of the pattern when the aperture is uniformly extended (or contracted) in one direction, and also consider Fraunhofer diffraction at a screen containing a large number of openings of the same size and shape.

Let $\mathscr{A}_1$ and $\mathscr{A}_2$ be two apertures such that the extension of $\mathscr{A}_2$ in a particular direction $(O\xi)$ is $\mu$ times that of $\mathscr{A}_1$. For Fraunhofer diffraction at $\mathscr{A}_1$, we have

$$U_1(p,q) = C \iint_{\mathscr{A}_1} e^{-ik(p\xi+q\eta)} d\xi d\eta \tag{19}$$

Similarly for Fraunhofer diffraction at $\mathscr{A}_2$,

$$U_2(p,q) = C \iint_{\mathscr{A}_2} e^{-ik(p\xi+q\eta)} d\xi d\eta \tag{20}$$

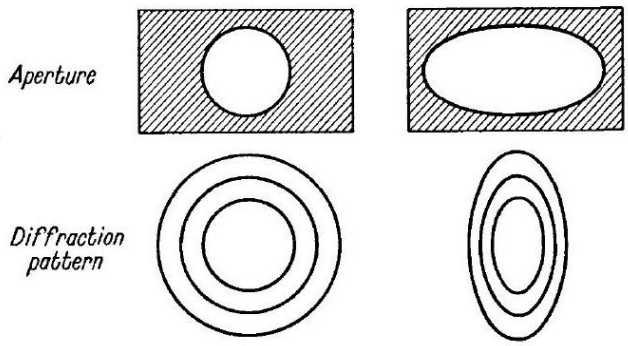



Fig. 8.14. Comparison of Fraunhofer diffraction at a circular and an elliptical aperture.

If in (20) we change the variables of integration from $(\xi, \eta)$ to $(\xi', \eta')$ where

$$\xi' = \frac{1}{\mu}\xi, \quad \eta' = \eta \tag{21}$$

we obtain

$$U_2(p,q) = \mu C \iint_{\mathscr{A}_1} e^{-ik(\mu p\xi'+q\eta')} d\xi' d\eta' = \mu U_1(\mu p, q) \tag{22}$$

This shows that when the aperture is uniformly extended in the ratio $\mu : 1$ in a particular direction, the Fraunhofer pattern contracts in the same direction in the ratio $1 : \mu$; and the intensity in the new pattern is $\mu^2$ times the intensity at the corresponding point of the original pattern. Using this result we may, for example, immediately determine the Fraunhofer pattern of an aperture which has the form of an ellipse or a parallelogram from that of a circle or rectangle respectively. Fig. 8.14 illustrates the case of an elliptical aperture.

We now consider the important case of a screen that contains a large number of identical and similarly oriented apertures. (According to Babinet's principle the results will also apply to the complementary distribution of obstacles.) Let $O_1, O_2$,

* Lord Rayleigh, Phil. Mag. (5), 11 (1881), 214. Also his Scientific Papers, 1, 513.

0

* Fraunhofer diffraction at an annular aperture is briefly considered in connection with resolving power in § 8.6.2.

$\ldots O_N$ be a set of similarly situated points, one in each aperture, and let the coordinates of those points referred to a fixed set of axes in the plane of the apertures be $(\xi_1, \eta_1), (\xi_2, \eta_2), \ldots (\xi_N, \eta_N)$. The light distribution in the Fraunhofer diffraction pattern is then given by

$$\begin{aligned} U(p,q) &= C\sum_n \iint_{\mathscr{A}} e^{-ik[(\xi_n+\xi')p+(\eta_n+\eta')q]}d\xi' d\eta' \\ &= C\sum_n e^{-ik[p\xi_n+q\eta_n]} \iint_{\mathscr{A}} e^{-ik(p\xi'+q\eta')}d\xi' d\eta' \end{aligned} \tag{23}$$

where the integration extends over any one opening $\mathscr{A}$ of the set. The integral expresses the effect of a single aperture, whilst the sum represents the superposition of the coherent diffraction patterns. If $I^{(0)}(p,q)$ is the intensity distribution arising from a single aperture, then, according to (23), the total intensity is given by

$$\begin{aligned} I(p,q) &= I^{(0)}(p,q) \mid \sum_n e^{-\,ik(p\xi_n+q\eta_n)}|^2 \\ &= I^{(0)}(p,q)\sum_n\sum_m e^{-ik[p(\xi_n-\xi_m)+q(\eta_n-\eta_m)]}. \end{aligned} \tag{24}$$

The simplest case, that of two openings, was considered earlier in §7.2, in connection with the theory of interference. However, we neglected there the dependence of $I^{(0)}$ on $p$ and $q$ (i.e. the effect of diffraction at each opening) and only studied the effect of superposition. It is easily seen that the earlier result (§7.2(17)) is in agreement with (24). For if $N = 2$, (24) reduces to

$$\begin{aligned} I &= I^{(0)}\left\{2 + e^{-ik[p(\xi_1-\xi_2)+q(\eta_1-\eta_2)]} + e^{-ik[p(\xi_2-\xi_1)+q(\eta_2-\eta_1)]}\right\} \\ &= 4I^{(0)}\cos^2\frac{1}{2}\delta, \end{aligned}$$

with

$$\delta = k\left[p\left(\xi_2-\xi_1\right)+q\left(\eta_2-\eta_1\right)\right].$$

Let us now consider the effect of a large number of apertures. We shall see that quite different results are obtained, depending on whether the apertures are distributed regularly or irregularly over the screen.

When the apertures are distributed irregularly over the screen, terms with different values of $m$ and $n$ in the double sum will fluctuate rapidly between +1 and -1 as $m$ and $n$ take on different values, and in consequence the sum of such terms will have zero mean value. Each remaining term ( $m = n$ ) has the value unity. Hence it follows that except for local fluctuations* the total intensity is $N$ times the intensity of the light diffracted by a single aperture:

$$I(p,q) \sim NI^{(0)}(p,q). \tag{25}$$

Diffraction effects of this type or, more often still, complementary effects (in the sense of Babinet's principle) may be easily observed, for example when a glass plate, dusted with lycopodium powder or covered with other particles of equal size and shape, is held in the path of light from a distant source. A piece of tin-foil pierced indiscriminately by a pin will also act as a diffracting screen of the type just considered.

Photographs of Fraunhofer diffraction patterns for apertures of various forms can be found in a paper by J. Scheiner and S. Hirayama, Abh. d. Königl. Akad. Wissensch., Berlin (1894), Anhang I. Photographs of Fresnel patterns were published by Y. V. Kathavate, Proc. Ind. Acad. Sci., 21 (1945), 177-210.

0

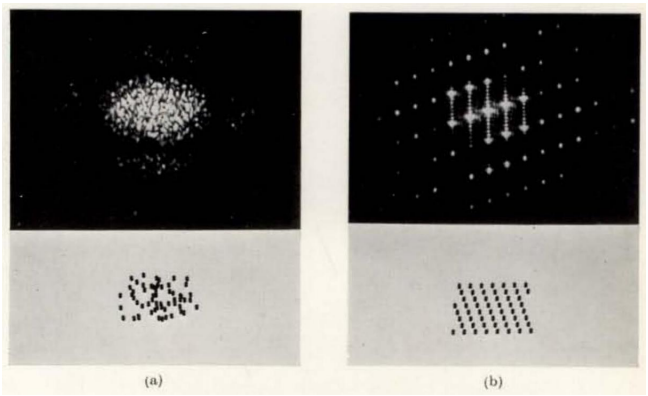



Fig. 8.15. Fraunhofer diffraction pattern from an irregular distribution (a), and a regular distribution (b), of 56 identical and similarly situated apertures in a plane screen. The form and distribution of the apertures is shown in the lower portions of the figures. Light: Mercury yellow, $\lambda = 5790$Å. (After H. Lipson, C. A. Taylor, and B. J. Thompson.)

The results are quite different when the openings are distributed regularly, for the terms with $m \neq n$ may now give appreciable contributions for certain values of $p$ and $q$. For example, if the points $O_n$ are so situated that for certain values of $p$ and $q$ the phases of all the terms for which $m \neq n$ are exact multiple of $2\pi$ their sum will be equal to $N(N-1)$ and so for large $N$ will be of the order of $N^2$. This enormous increase in intensity for particular directions, clearly illustrated in Fig. 8.15, is, as we shall see in the next section, of great importance in practice.

## 9.11.6 8.6. Fraunhofer Diffraction in Optical Instruments

### 8.6.1 Diffraction gratings

(a) The principle of the diffraction grating A diffraction grating may be defined as any arrangement which imposes on an incident wave a periodic variation of amplitude or phase, or both. We may characterize any particular grating by its transmission function, defined as follows:

Let a transparent or semi-transparent object (not necessarily periodic) cover a portion of a fictitious reference plane $\xi\eta$, and let it be illuminated by a plane monochromatic wave incident in a direction specified by the direction cosines $l_0, m_0$. Fig. 8.16 illustrates the arrangement, the $\eta$-axis being perpendicular to the plane of the

• Fluctuations of a somewhat different type arise when the apertures are not of the same form, but are distributed regularly or according to some statistical law (cf. M. v. Laue, Berl. Ber., (1914), 1144). Similar effects are observed in connection with diffraction of X -rays by liquids (cf. J. A. Prins, Naturwiss, 19 (1931), 435).

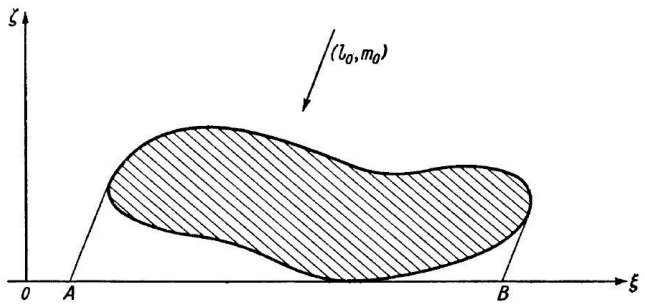



Fig. 8.16. Definition of the transmission function.

drawing. If no object were present, the disturbance in the $\xi\eta$-plane would be represented by the function $V_0(\xi, \eta) = A \exp\{ik\,(l_0\xi + m_0\eta)\}$, the factor $\exp(-i\omega t)$ being, as usual, omitted. Because of the presence of the object the disturbance will be modified and may be represented by some other function, which we denote by $V(\xi, \eta)$. The transmission function of the object is then defined as

$$F(\xi, \eta) = \frac{V(\xi, \eta)}{V_0(\xi, \eta)} \tag{1}$$

In general $F$ depends, of course, not only on $\xi$ and $\eta$ but also on the direction ( $l_0, m_0$ ) of illumination. The transmission function is in general complex, since both the amplitude and the phase of the light may be altered on passing through the object. In the special case when the object alters the amplitude but not the phase of the incident wave (i.e. if $\arg F \equiv 0$ ), we speak of an amplitude object; if it alters the phase but not the amplitude (i.e. $|F| = 1$ ) we speak of a phase object.

If we are concerned with reflected light rather than with light that is transmitted by an object, it is more appropriate to speak of a reflection function, defined in a
similar way, the only difference being that the reference plane is on the same side of the object as the incident light.

The ratio $|V/V_0|$ is practically unity for points outside the geometrical shadow (whose boundary is represented by points $A$ and $B$ in Fig. 8.16) cast by the object. If the portion outside the shadow region is covered by an opaque screen, the arrangements act as a diffracting aperture $\mathscr{A}$ with a non-uniform pupil function (cf. § 8.3 (39)). If the linear dimensions of $\mathscr{A}$ are large compared to the wavelength and if $F$ remains sensibly constant over regions whose dimensions are of the same order as the wavelength, the diffraction formula § 8.3 (23) remains valid under the same conditions as before, provided that the integrand of the diffraction integral is multiplied by $F$.

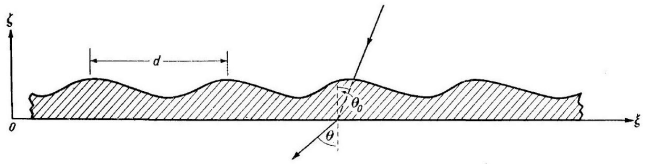



Fig. 8.17. Illustrating the theory of the diffraction grating.

Let us now consider a one-dimensional grating consisting of $N$ parallel grooves of arbitrary profile, ruled on one surface of a plane-parallel glass plate. Let the $\xi\eta$-plane coincide with the plane face of the plate, $\eta$ being the direction of the grooves and let $d$ be the period in the $\xi$-direction (see Fig. 8.17).

Assume that the direction of propagation of the wave incident upon the grating is in the plane of the figure, making an angle $\theta_0$ with $O\zeta$, and let $\theta$ denote the angle which $O\zeta$ makes with the line joining a very distant point of observation $P$ with the grating. As before we set $l_0 = \sin\theta_0, l = \sin\theta, p = l - l_0 = \sin\theta - \sin\theta_0$. The complex amplitude at $P$ is then immediately obtained from §8.5 (23), where the integrand must be multiplied by the transmission function $F$ of one periodic element. We may set $q = 0$ and

$$\xi_n = nd, \quad \eta_n = 0, \quad (n = 0, 1, \ldots N-1). \tag{2}$$

We then obtain

$$U(p) = U^{(0)}(p) \sum_{n=0}^{N-1} e^{-iknd p} = U^{(0)}(p) \frac{1 - e^{-iNkdp}}{1 - e^{-ikdp}}, \tag{3}$$

where*

$$U^{(0)}(p) = C \int_{\mathscr{A}} F(\xi) e^{-ikp\xi} d\xi \tag{4}$$

Hence

$$\begin{aligned} I(p) = |U(p)|^2 &= \frac{\left(1 - e^{-iNkdp}\right)}{\left(1 - e^{-ikdp}\right)} \cdot \frac{\left(1 - e^{iNkdp}\right)}{\left(1 - e^{ikdp}\right)} \left|U^{(0)}(p)\right|^2 \\ &= \frac{1 - \cos Nkdp}{1 - \cos kdp} I^{(0)}(p) \end{aligned} \tag{5}$$

where $I^{(0)}(p) = \left|U^{(0)}(p)\right|^2$. If we introduce the function

$$H(N, x) = \left(\frac{\sin Nx}{\sin x}\right)^2, \tag{6}$$

the formula (5) for the intensity may be written as

$$I(p) = H\left(N, \frac{kdp}{2}\right) I^{(0)}(p). \tag{5a}$$

Before discussing the implications of this basic formula we note that according to (3) the light distribution is the same as that due to a set of coherent secondary sources each characterized by the same amplitude function $\left|U^{(0)}(p)\right|$ and with phases that

0

- Since $F$ depends on $l_0$, the quantities $U^{(0)}$ and $I^{(0)}$ now depend both on $l$ and $l_0$ and not on the difference $l - l_0$ only. As we are only interested in effects for a fixed direction of incidence, we may regard $l_0$ as a constant and retain the previous notation.

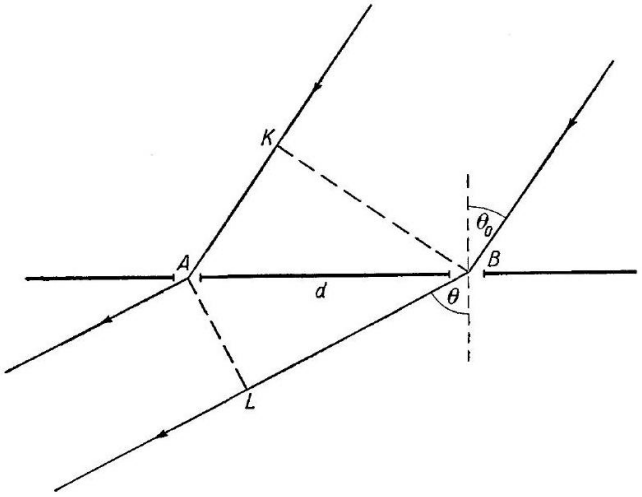



Fig. 8.18. Illustrating the theory of the diffraction grating.

differ from each other by integral multiples of $kdp$. To see the significance of this phase difference consider two corresponding points $A$ and $B$ in neighbouring grooves of the grating (Fig. 8.18). Since the effect of the grating is to impress a periodic variation on to the incident wave, it follows that the path difference between the light arriving at $A$ and at $B$ is the same as in the absence of the grating, i.e. it is equal to $AK = d\sin\theta_0, K$ denoting the foot of the perpendicular from $B$ on to the ray incident at $A$. Further, the light path from $B$ in the direction $\theta$ exceeds the light path from $A$ by $BL = d\sin\theta, L$ being the foot of the perpendicular from $A$ on to the ray diffracted at $B$ in the direction $\theta$. Hence the total path difference between light arriving at the distant point of observation from corresponding points in two neighbouring grooves is

$$BL - AK = d\,(\sin\theta - \sin\theta_0) = dp, \tag{7}$$

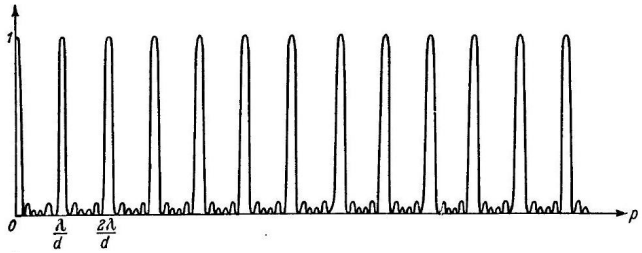
and the corresponding phase difference is $2\pi dp/\lambda = kdp$.

The formula (5a) expresses $I(p)$ as the product of two functions: one of them, $I^{(0)}$, represents the effect of a single period of the grating; the other, $H$, represents the effect of interference of light from different periods. The function $H(N, x)$ has maxima, each of height $N^2$, at all points where the denominator $\sin^2 x$ vanishes, i.e. where $x$ is zero or an integral multiple of $\pi$. Hence $H(N, kdp/2)$ has maxima of height $N^2$ when

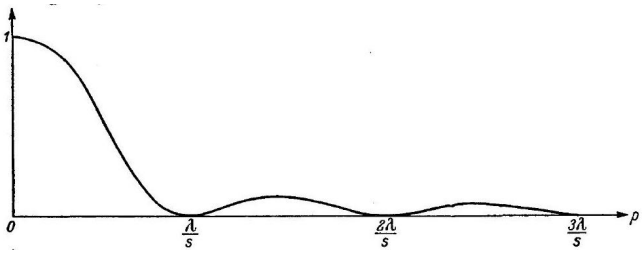
$$p \equiv \sin\theta - \sin\theta_0 = \frac{m\lambda}{d}. \quad (m = 0, \pm1, \pm2, \ldots). \tag{8}$$

The integer $m$ represents, according to (7), the path difference in wavelengths between light diffracted in the direction of the maximum, from corresponding points in two neighbouring grooves. In agreement with our earlier definition (§7.3.1), we call $m$ the order of interference. Between these principal maxima there are weak secondary maxima (see Fig. 8.19 (a)), the first secondary maximum being only a few per cent of

(a)



(b)




(c)

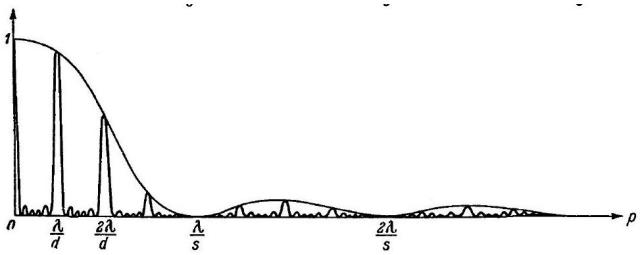



Fig. 8.19.

(a) The normalized interference function

$$\frac{1}{N^2}H(N, kdp/2) = \left[\frac{\sin(Nkdp/2)}{N\sin(kdp/2)}\right]^2.$$

(b) The normalized intensity function of a slit

$$I^{(0)}(p) = \left[\frac{\sin ksp/2}{ksp/2}\right]^2.$$

(c) The normalized intensity function of a grating consisting of $N$ similar equidistant parallel slits

$$\frac{1}{N^2}I(p) = \left[\frac{\sin(Nkdp/2)}{N\sin(kdp/2)}\right]^2 \left[\frac{\sin ksp/2}{ksp/2}\right]^2$$

Only the range $p \geqslant 0$ is shown, all the curves being symmetrical about the vertical axis $p = 0$.
the principal maximum when $N$ is large. The maxima are separated by points of zero intensity at $x = kdp/2 = \pm n\pi/N$, i.e. in directions given by

$$p \equiv \sin\theta - \sin\theta_0 = \frac{n\lambda}{Nd}, \quad (n = \pm 1, \pm 2, \ldots) \tag{9}$$

the case where $n/N$ is an integer being excluded.
The function $I^{(0)}(p)$ depends on the form of the grooves. Suppose that it has a principal maximum for some direction $p = p'$ and that on both sides of the maximum it falls off slowly in comparison with $H$. Then $I(p)$ will have the general form of the interference function $H$, but will be "modulated" by $I^{(0)}$. Thus $I(p)$ will still have fairly sharp maxima near the directions $p = m\lambda/d$. Since these directions (except for $m = 0$ ) depend on the wavelength, we see that the grating will decompose a beam of non-monochromatic light into spectral orders.

To illustrate these remarks let us consider a grating consisting of a succession of long equidistant slits (Fig. 8.20), each of width $s$ and length $L$, in an opaque screen.

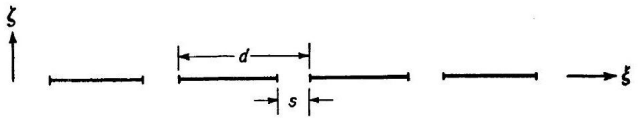



Fig. 8.20. Profile of a simple line grating.

If the grating is illuminated from a very distant line source parallel to the slits, the intensity $I^{(0)}$ is given by the expression § 8.5 (4) [with $2a = s, 2b = L$ ] and we obtain

$$I(p) = \frac{sE}{\lambda}\left(\frac{\sin\frac{Nkdp}{2}}{\sin\frac{kdp}{2}}\right)^2\left(\frac{\sin\frac{ksp}{2}}{\frac{ksp}{2}}\right)^2 \tag{10}$$

Curves representing the two factors in (10) and their product are shown in Fig. 8.19. The last factor in (10), which represents the effect of a single slit, has a principal maximum at $p = 0$ and minima given by $ksp/2 = n\pi$, i.e. at

$$p = \frac{n\lambda}{s}, \quad (n = \pm 1, \pm 2, \ldots) \tag{11}$$

separated by weak secondary maxima. We see that if $\lambda/s \gg \lambda/d$, i.e. if the width of each slit is small compared to $d$, the intensity $I(p)$ has in addition to a principal maximum at $p = 0$ a series of sharp, but progressively decreasing, maxima on either side of it, near directions given by (8).

Returning to the general case, it is evident that if the width of each groove is very small, of the order of a wavelength (as is often the case in practice) the formula (4), derived on the basis of Kirchhoff's approximation, can evidently no longer be expected to hold. In such cases more refined considerations must be made to determine the detailed distribution of the intensity. We may, however, expect that the main qualitative features indicated by our elementary theory, namely the existence of sharp maxima whose positions are substantially determined by the interference function $H$, remain even when the grooves are very narrow, provided, of course, that the intensity function of a single period varies slowly in an interval of the order $\Delta p = \lambda/d$.

Let us now consider the resolution that may be attained with a grating. The separation between a primary maximum of order $m$ and a neighbouring minimum is, according to (9), given by

$$\Delta p = \frac{\lambda}{Nd} \tag{12}$$

If the wavelength is changed by an amount $\Delta\lambda$, the $m$ th order maximum is, according to (8), displaced by an amount

$$\Delta' p = \frac{|m|}{d}\Delta\lambda. \tag{13}$$

Assuming that the lines of wavelength $\lambda \pm \frac{1}{2}\Delta\lambda$ will just be resolved when the maximum of the one wavelength coincides with the first minimum of the other (cf. p. 334) we have on the limit of resolution in the $m$ th order, $\Delta p = \Delta' p$, i.e.

$$\frac{\lambda}{\Delta\lambda} = |m|N. \tag{14}$$

Thus, the resolving power is equal to the product of the order number $m$ and the number $N$ of the grooves. For the $m$ th order we have, according to ( 8 ), that $d(\sin\theta - \sin\theta_0) = m\lambda$, so that we may also express the resolving power in the form $^-$

$$\frac{\lambda}{\Delta\lambda} = \frac{Nd\,|\sin\theta - \sin\theta_0|}{\lambda}. \tag{14a}$$

Because of (7) this implies that the resolving power is equal to the number of wavelengths in the path difference between rays that are diffracted in the direction $\theta$ from the two extreme ends (separated by distance Nd) of the grating. It is to be noted that since $|\sin\theta - \sin\theta_0|$ cannot exceed 2 , the resolving power that can be attained with a grating of overall width $w$ can never exceed the value $2w/\lambda$.

Let us illustrate the formula (14) by determining the number of grooves that a grating must have, in order to separate two lines which are a tenth of an ngström unit apart, near the centre of the visible region of the spectrum. In this case $\lambda \sim 5500\text{Å}, \Delta\lambda = 10^{-1}\text{Å}$, and if we observe in the second order ( $m = 2$ ), we must have, according to (14) $N \gtrsim 5 \cdot 5 \cdot 10^3/2 \cdot 10^{-1} = 27,500$, i.e. the grating must have at least 27,500 grooves.

For comparison let us consider the resolving power of a prism, in the position of minimum deviation, with a line source that is parallel to the edge $A$ of the prism (slit of the spectrograph). A pencil of parallel rays will be incident upon the prism and will be diffracted at a rectangle of width $l_1 = l_2$ (see Fig. 4.28). According to § 8.5 (2), the first minimum of the intensity is at an angular distance (assumed to be small)

$$p = \frac{\lambda}{l_1} \tag{15}$$

from the geometrical image of the slit. The change in the angular dispersion corresponding to a change of wavelength by amount $\Delta\lambda$, is, according to § 4.7 (36),

$$\Delta\varepsilon = \frac{t}{l_1}\frac{dn}{d\lambda}\Delta\lambda \tag{16}$$

where $t$ is the greatest thickness of the glass through which one of the extreme rays has passed, and $n$ is the refractive index of the material of the prism. Since at the limit of resolution $p \sim \Delta\varepsilon$ the resolving power is given by

$$\frac{\lambda}{\Delta\lambda} = t\left|\frac{dn}{d\lambda}\right| \tag{17}$$

Equation (17) shows, that, with given glass, the resolving power of a prism depends only on the greatest thickness of the glass traversed by the rays; in particular the resolving power is independent of the angle of the prism. As an example, suppose that the length of the base of the prism is equal to 5 cm and that it is made of heavy flint glass, for which $dn/d\lambda \sim 1000\ \text{cm}^{-1}$ at wavelength $\lambda = 5500\text{Å}$. If the prism is completely filled with light, then according to (17) it will resolve lines near the centre of the visible region which are not less than $\Delta\lambda$ apart, where $\Delta\lambda \sim 5 \cdot 5 \cdot 10^{-5}\ \text{cm}/5 \cdot 10^3 = 1 \cdot 1\text{Å}$. Thus a prism of this considerable size has a resolving power 10 times smaller than the grating of 27,500 grooves discussed before.

We have so far considered one-dimensional gratings only, but the analysis may easily be extended to two- and three-dimensional periodic arrangements of diffracting bodies. Two-dimensional gratings (called cross-gratings) find no practical applications, though their effects can often be observed, for example when looking at a bright source through a finely woven material (e.g. a handkerchief). The theory of three-dimensional gratings is, on the other hand, of great importance, such gratings being formed by a regular arrangement of atoms in a crystal. The lattice distances (distances between neighbouring atoms) are of the order of an Angström unit ( $10^{-8}$ cm ), this being also the order of magnitude of the wavelengths of X-rays. Hence, by sending a beam of X-rays through a crystal, diffraction patterns are produced, and from their analysis information about the structure of the crystal may be deduced.* Diffraction of

X-rays by crystals was predicted by von Laue † in 1912 and first observed by Friedrich and Knipping.

Another example of a grating-like structure is presented by ultrasonic waves in liquids. These are elastic waves produced by a piezo-electric oscillator, differing from ordinary sound waves only in having a frequency well above the upper limit of audibility. Such waves give rise to rarefactions and condensations in the liquid which then act on the incident light like a grating. The theory of this phenomenon is discussed in Chapter XII. For the rest of this section we shall restrict our attention to one-dimensional gratings as used in spectroscopic work.

(b) Types of grating ‡ The principle of the diffraction grating was discovered by Rittenhouse in 1785, § but this discovery attracted practically no attention. The principle was rediscovered by Fraunhofer* in 1819. Fraunhofer's first gratings were made by winding very fine wire round two parallel screws. Because of the relative ease with which wire gratings may be constructed these are occasionally used even to-day, particularly in the long-wavelength (infra-red) range. Later Fraunhofer made gratings with the help of a machine, by ruling through gold films deposited on a glass plate; also, using a diamond as a ruling point, he ruled the grooves directly on to the surface of glass.

Great advances in the technique of production of gratings were made by Rowland † who constructed several excellent ruling machines and also invented the so-called concave grating (discussed on pp. 412-414). Rowland's machine was able to rule gratings with grooves over 4 in . long over a length of 6 in ., and his first machine ruled about 14,000 grooves per inch, giving a resolving power in excess of 150,000 . Later Michelson ruled gratings considerably wider than 6 in ., with a resolving power approaching 400,000.

Most of the early gratings were ruled on speculum metal and glass, but the more recent practice is to rule the grooves on evaporated layers of aluminium. Since aluminium is a soft metal it causes less wear on the ruling point (diamond) and it also reflects better in the ultra-violet.

A perfect grating would have all the grooves strictly parallel and of identical form, but in practice errors will naturally occur. Quite irregular errors lead to a blurring of the spectrum and are not so serious as systematic errors, such as periodic errors of spacing. These errors give rise to spurious lines in the spectrum, known as ghosts. Often they can be distinguished from true lines only with difficulty.

High resolving power is not always the only important requirement in spectroscopic applications. When little energy is available, as for example in the study of spectra of faint stars or nebulae, or for work in the infra-red region of the spectrum, it is essential that as much light as possible should be diffracted into one particular order. Moreover, for precise wavelength measurements, a grating that gives high dispersion must be used. According to (8), the angular dispersion (with a fixed angle of incidence) is given by

$$\frac{d\theta}{d\lambda} = \frac{1}{\cos\theta}\frac{m}{d}, \tag{18}$$

0

- See for example M. von Laue, Röntgenstrahl-Interferenzen, Leipzig, Akademische Verlagsgesellschaft 2 Aufl. (1948); or The Crystalline State, ed. Sir L. Bragg (London, Bell and Sons), Vol. I, W. L. Bragg, A General Survey (1933); Vol. II, R. W. James, The Optical Principles of the Diffraction of X-rays (1948); Vol. III, H. Lipson and W. Cochran, The Determination of Crystal Structures (1953).
  † W. Friedrich, P. Knipping and M. von Laue, Mnchener Sitzungsber., 1912, p. 303; Ann.d. Physik, (4), 61 (1913), 971.
  ‡ A fuller account of methods of production of gratings and their development, see G. R. Harrison, J. Opt. Soc. Amer., 39 (1949), 413.
  § D. Rittenhouse, Trans. Amer. Phil. Soc., 2 (1786), 201. See also article by T. D. Cope in Journ. Franklin Inst., 214 (1932), 99.

so that to obtain high dispersion the spacing $d$ should be small or the observations must be made in high orders ( $m$ large). If, however, the grating is formed by a succession of opaque and transparent (or reflecting) strips, only a small fraction of the incident light is thrown into any one order. This drawback is overcome in modern practice by ruling the grooves to controlled shape. With a grating which consists of grooves of the form shown in Fig. 8.21, most of the light may be directed into one or two orders on one side of the central image. Gratings of this type, with fairly coarse grooves, are called echelette gratings, because they may be regarded as being intermediate between the older types of grating and the so-called echelon gratings which will be described later. Echelette gratings were first ruled by Wood ‡ on copper plates, using the natural edge of a selected carborundum crystal as a ruling point. Later they were ruled with diamond edges ground to the desired shape. They

had 2000-3000 grooves per inch, and when used with visible light they sent the greater part of the light into a group of spectra near the 15 th or 30 th order. Echelette gratings have considerable value in infra-red spectroscopy.

More recently methods have been developed for controlling the groove shape for gratings with much smaller groove spacing.* These blazed gratings, as they are called, have grooves of similar form as the echelettes, but form the most intense spectra in much lower orders (usually the first or second).

It appears that the resolving power of gratings of the type described is limited by practical considerations of manufacture to about 400,000 . For some applications (e.g. for the study of Zeeman effects and the hyperfine and isotope structure patterns), a resolving power that exceeds this value is required. For the attainment of such a high resolving power, Harrison † proposed the so-called echelle grating, which has wide, shallow grooves and is designed for use at an angle of incidence greater than 45°, the direction of incidence being normal to the narrow side of the step. These gratings

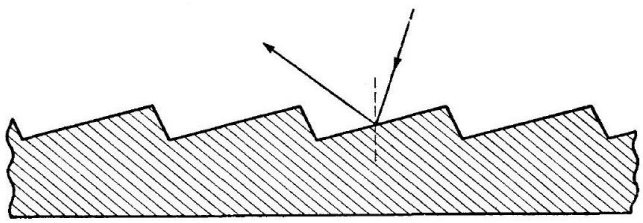


Fig. 8.21. Reflection grating with controlled groove form.

operate with relatively high orders ( $m \sim 1000$ ). A 10 in. echelle with 100 grooves per inch, designed for observation in the 1000 th order, has a resolving power of $1,000,000$‡.

Because a grating of good quality is very difficult to produce, replicas of original rulings are often used.§ These are obtained by moulding from an original ruled master grating.

Finally we must mention a "grating" of an entirely different construction, the echelon invented by Michelson. || It consists of a series of strictly similar planeparallel glass plates arranged in the form of a flight of steps (hence the name), as shown in Fig. 8.22. Each step retards the beam of light which passes through it by the same amount with respect to its

0

- J. Fraunhofer, Denkschr. Akad. Wiss. Mnchen, 8 (1821-1822), I. Ann. d. Physik., 74 (1823), 337. Reprinted in his collected works (Munich, 1888), 51, 117.
  † H. A. Rowland, Phil. Mag. (5), 13 (1882), 469. Nature, 26 (1882), 211. Phil. Mag., 16 (1883), 297.
  ‡ R. W. Wood, Phil. Mag., 20 (1910), 770; Ibid. 23 (1912), 310; A. Trowbridge and R. W. Wood, Ibid., 20 (1910), 886, 898.

neighbour. Because the breadth of each step is large compared to the wavelength, the effect of diffraction is confined to small angles, so that most of the light is concentrated in one or two spectra near the direction $\theta = 0$, and these correspond to very high orders, since the retardation introduced between successive beams is very many wavelengths.

The resolving power of the echelon depends not only on the path difference between the rays from the extreme ends of the grating but also (though to a much lesser extent) on the dispersion of the glass. If $n$ is the refractive index, $t$ the thickness of

each step, and $d$ its breadth (see Fig. 8.22), the path difference between rays diffracted from neighbouring steps is evidently $pd + (n-1)t$, it being assumed that $p$ is small. Hence the position of the principal maxima are given by

$$pd + (n-1)t = m\lambda, \quad (m = 0, 1, 2, \ldots) \tag{19}$$

If the wavelength is changed by an amount $\Delta\lambda$, the $m$ th order maximum is displaced by

$$\Delta' p = \left| m - t\frac{dn}{d\lambda} \right| \frac{\Delta\lambda}{d}. \tag{20}$$

The separation $\Delta p$ between a principal maximum of order $m$ and a neighbouring minimum is again given by (12), so that the condition $\Delta p = \Delta' p$ for the limit of resolution gives

$$\frac{\lambda}{\Delta\lambda} = N \left| m - \frac{dn}{d\lambda} t \right|. \tag{21}$$

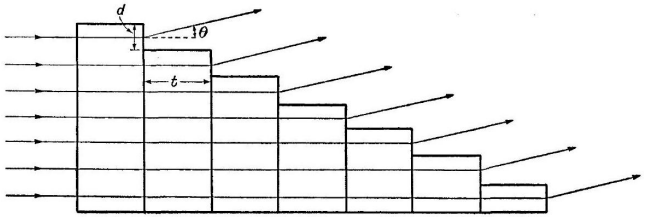



Fig. 8.22. Michelson's echelon.

Here we may substitute for $m$ the value $(n-1)t/\lambda$ obtained from (19) by neglecting the term $pd$, for the $p$ values for which the intensity is appreciable are of the order of $\lambda/d$, i.e. $pd$ is of the order of a wavelength, whilst $(n-1)t$ is of the order of many thousand wavelengths. We thus obtain the following expression for the resolving power of the echelon:

$$\frac{\lambda}{\Delta\lambda} \sim N \left| \frac{n-1}{\lambda} - \frac{dn}{d\lambda} \right| t. \tag{22}$$

0

- R. W. Wood, Nature, 140 (1937), 723; J. Opt. Soc. Amer., 34 (1944), 509; H. Babcock, 34 (1944), 1.
† G. R. Harrison, J. Opt. Soc. Amer., 39 (1949), 522.
‡ For a review of the theory and production of high resolution gratings, see G. W. Stroke, Progress in Optics, Vol. 2, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1963), 1.
§ First replicas were made by T. Thorp, British Patent No. 11,460 (1899); and later by R. J. Wallace, Astrophys. J., 22 (1905), 123; Ibid., 23 (1906), 96. Improved methods have been described by T. Merton, Proc. Roy. Soc. A, 201 (1950), 187.
‖ A. A. Michelson, Astrophys. J., 8 (1898), 37; Proc. Amer. Acad. Arts Sci., 35 (1899), 109.

The ratio $\frac{dn}{d\lambda}/\frac{n-1}{\lambda}$ is small. For flint glass near the centre of the visible region it has a value near -0.05 to -0.1 . Hence, under these circumstances, the resolving power of an echelon exceeds, by about 5 to 10 per cent, the resolving power of a line grating with $N$ grooves, when observation is made in the order $m = (n-1)t/\lambda$. One of Michelson's echelons consisted of twenty plates, each having a thickness $t = 18$ mm, and the breadth $d$ of each step was about 1 mm . Taking $n = 1.5$, the retardation between two successive beams measured in wavelengths of green light $\lambda = 5.10^{-5}$ cm was $m \sim 0 \cdot 5 \cdot 1 \cdot 8/5 \cdot 10^{-5} \sim 20,000$. Assuming $\frac{dn}{d\lambda}/\frac{n-1}{\lambda} = -0 \cdot 1$, this gives a resolving power of about $20(20,000 + 0 \cdot 1 \times 20,000) = 440,000$.

More important is the reflection echelon. Here each step is made highly reflecting by means of metallic coating, and the spectra formed by reflected light are observed. With a reflection echelon the resolving power is three to four times as large than with a transmission echelon of corresponding dimensions, since each step introduces a
retardation between successive beams of amount $2t$ instead of $(n-1)t \sim t/2$. Like the echelle grating the reflection echelon is capable of giving resolving power of over one million. Another advantage of the reflection echelon over the transmission echelon is that it may be used in the ultra-violet region of the spectrum, where glass absorbs. Although Michelson realized that advantages would be gained by using the instrument with reflected rather than transmitted light, technical difficulties prevented the production of a satisfactory reflection echelon for nearly thirty years until they were overcome by Williams.* Because of difficulties in assembling a large number of plates of equal thickness within the narrow permissible tolerance, the number of steps is limited in practice to about forty.

Finally, a few remarks must be made about overlapping of orders. Restricting ourselves to the visible region, i.e. considering wavelengths in the range $\lambda_1 = 0 \cdot 4\mu$ to

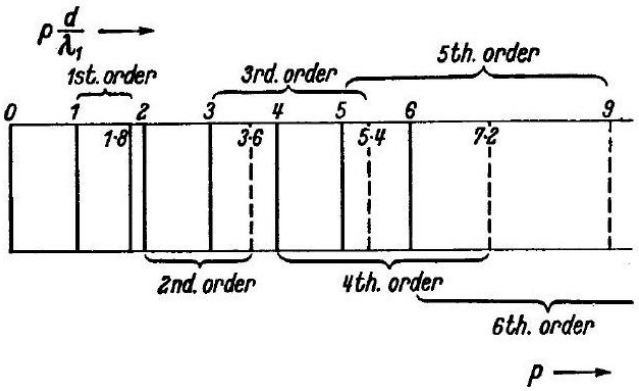



Fig. 8.23. The overlapping of grating spectra.

$\lambda_2 = 0.75\mu$, we see that the first-order spectrum does not quite reach to the spectrum of the second order: for the first-order spectrum covers the range from $p = \lambda_1/d$ to $p = \lambda_2/d = 0 \cdot 75\lambda_1/0 \cdot 4d = 1 \cdot 8\lambda_1/d$, whilst the second order begins at $p = 2\lambda_1/d$. On the other hand the spectrum of the second order extends across a part of the thirdorder spectrum, namely from $p = 2\lambda_1/d$ to $p = 2\lambda_2/d$, whilst the third order begins already at $2 \times 1 \cdot 8\lambda_1/d$. As the order increases the successive spectra overlap more and more (see Fig. 8.23). If the lines of wavelength $\lambda$ and $\lambda + \delta\lambda$ coincide in two successive orders ( $m+1$ )th and $m$ th, then

$$(m+1)\lambda = m(\lambda + \delta\lambda)$$

i.e.

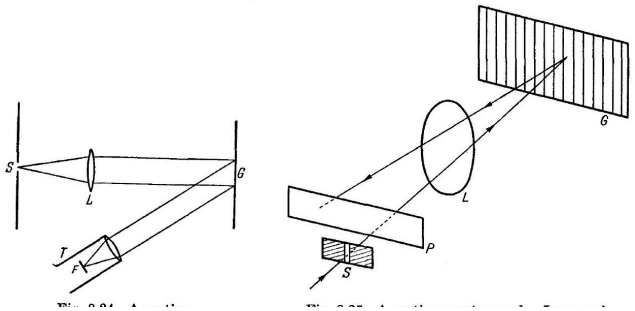
$$\frac{\delta\lambda}{\lambda} = \frac{1}{m}. \tag{23}$$

Thus the "free spectral range" is inversely proportional to the order.
The overlapping of orders was formerly used to compare wavelengths, in the socalled method of coincidences (cf. p. 340); this method has been superseded by simple interpolation between standard wavelengths determined interferometrically.

In conclusion let us summarize the main distinguishing features of the different types of gratings. We recall that, according to (14), high resolving power may be attained with either a large number of periods and relatively low orders, or with a moderate number of periods and large orders. Ordinary ruled gratings represent the low-order extreme ( $m \sim 1$ to 5 ), whilst the echelons represent the extreme of high orders ( $m \sim 20,000$ ). In between are the echellettes ( $m \sim 15$ to 30 ) and the echelles

( $m \sim 1000$ ). For particular applications one must bear in mind that the angular dispersion is directly proportional to the spectral order and inversely proportional to the period whilst the free spectral range is inversely proportional to the order.

(c) Grating spectrographs In a grating spectrograph coloured images of a slit source are produced in the various orders into which the grating separates the incident light. A simple arrangement is shown in Fig. 8.24. Collimated light from a slit source $S$ in the focal plane of a lens $L$ is incident on a reflection grating $G$, and the images of $S$ formed by the diffracted rays are observed in the focal plane $F$ of a telescope $T$. A modification of this arrangement,




known as Littrow's mounting, which has the advantage of compactness, is shown in Fig. 8.25. This is an autocollimation device, which needs only one lens. The slit is just below the plate $P$ and the lens is near the grating, which can be turned through a prescribed angle with respect to the direction of the incident beam.

In order to avoid losses of light which necessarily arise when the diffracted rays are focused by means of lenses, Rowland introduced the concave grating. Here the grooves are ruled on a concave highly reflecting metal surface, i.e. on a concave mirror, in such a way that their projections on a chord of the mirror surface are equidistant. A simple geometrical theorem indicates the possible positions of the slit and the plane of observation relative to the grating:

Let $Q$ be the midpoint of the surface of the grating and $C$ its centre of curvature, and describe a circle $K$ with centre at the midpoint $O$ of $QC$ and with radius $r = OQ = OC$ (Fig. 8.26). We shall prove that light from any point $S$ of the circle $K$ will be approximately reflected to a point $P$ and diffracted to points $P', P'', \ldots$ on the circle, each of these points being a focus for diffracted rays of a particular order. To show this, construct the reflected ray $QP$ corresponding to the incident ray $SQ$. If $\alpha = SQC$ is the angle of incidence, the angle $CQP$ of reflection is also equal to $\alpha$ and, moreover, the arc $SC$ is equal to the arc $CP$. Consider now another ray from $S$ incident upon the grating at a different point $R$. If the diameter of the circle is sufficiently large (in practice it is usually several feet), then no appreciable error is introduced by assuming $R$ to lie on the circle $K$. Hence, since $C$ is the centre of curvature of the grating, the angle of incidence $SRC$ and consequently also the angle of reflection are

0

- W. E. Williams, British Patent, 312534 (1926); Proc. Opt. Conv., 2 (1926), 982; Proc. Phys. Soc., 45 (1933), 699.

again equal to $\alpha$. Moreover, since the arc $CP$ is equal to the arc $SC$ it follows that the ray reflected at $R$ again passes through $P$.

Similar considerations apply to the diffracted rays. Let $\beta$ be the angle which a ray diffracted at $Q$ makes with $QP$. The corresponding ray of the same order, diffracted at

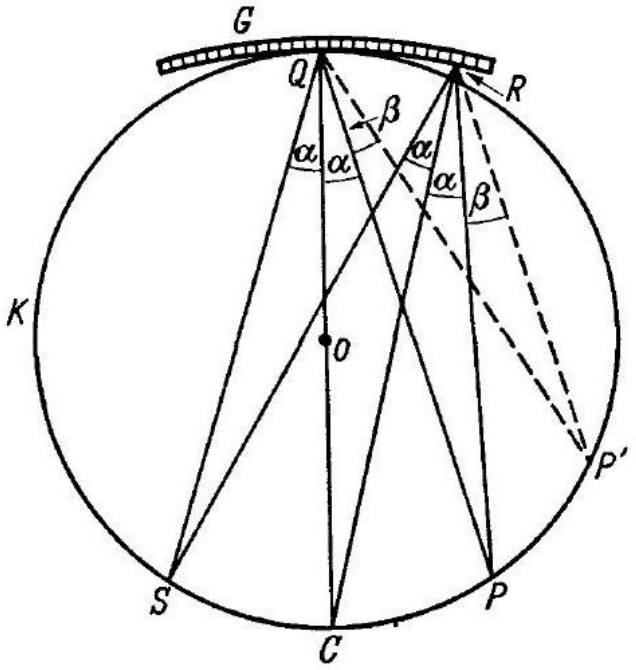



Fig. 8.26. Focusing with a concave grating (Rowland's circle).

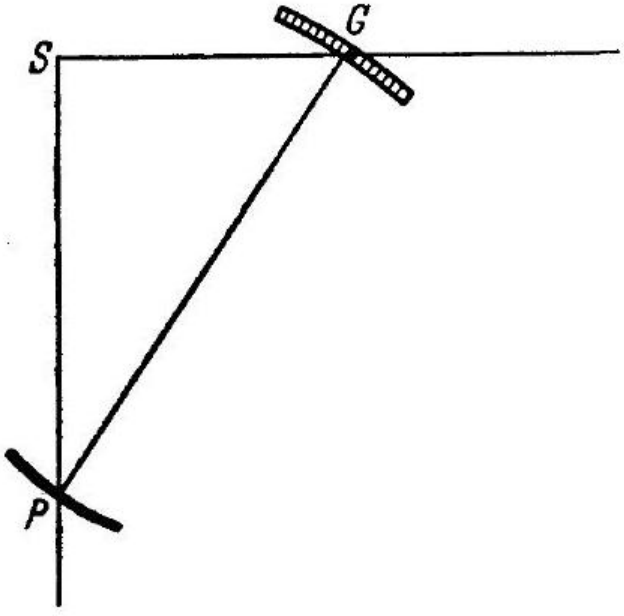



Fig. 8.27. Rowland's mounting for a concave grating.

$R$, will make the same angle ($\beta$) with $RP$. Hence the ray diffracted at $Q$ makes the same angle with $SQ$ as the ray diffracted at $R$ makes with $SR$, namely $2\alpha + \beta$. The two diffracted rays, therefore, meet in a point $P'$ of the circle $K$. Thus, to obtain sharp lines, the slit, the grating, and the plane of observation (photographic plate) should be

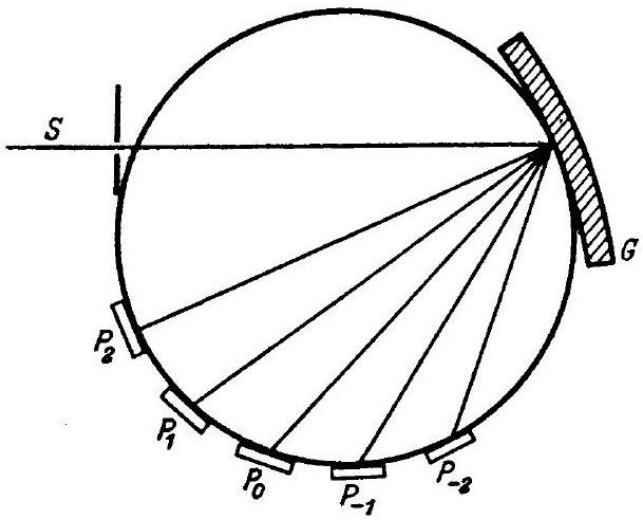



Fig. 8.28. Paschen's mounting for a concave grating.

situated on a circle, whose diameter is equal to the radius of curvature of the concave grating.

There are several mountings based on this principle. Rowland himself used the arrangement shown in Fig. 8.27. Here the grating $G$ and the plate holder $P$ are fixed to opposite ends of a movable girder, whose length is equal to the radius of curvature of the grating. The two ends of the girder are free to move along fixed tracks which are at right angles to each other. The slit $S$ is mounted immediately above their intersection in such a way that light falling normally on the slit proceeds along $SG$.

The slit is thus situated on a Rowland circle with diameter $PG$ and the order of the spectrum appearing on the plate depends on the position of the girder.

A different arrangement, shown in Fig. 8.28, avoids the use of mobile parts. Here a circular steel rail to which the slit $S$ and the grating $G$ are permanently attached plays the part of Rowland's circle. Round the rail a series of plate holders $P_0, P_1, P_{-1}, \ldots$ is set up, so that spectra of several orders can be photographed simultaneously. This arrangement, called Paschen's mounting, also has the advantage of great stability.

Another arrangement due to Eagle has, like the Littrow mounting for a plane grating, the advantage of compactness. Here the slit is immediately above or below the centre of the plate holder (Fig. 8.29), or it may be mounted at the side and the light may be thrown in the required direction by means of a small reflecting surface. To observe different portions of the spectrum, both the plate and the grating have to be rotated through the same amounts and in opposite senses, and their separation has to be changed, so that they are always tangential to the Rowland circle. One observes that part of the spectrum which is diffracted back at angles nearly equal to




Fig. 8.29. Eagle's mounting for a concave grating.

the angle of incidence. For this mounting to be strictly autocollimating, the slit $S$ should be at the centre of the plate.

Spectral lines obtained with a concave grating show the same aberrations as images obtained with a concave mirror, chiefly astigmatism. If, however, the concave grating is used in parallel light, the astigmatism may be made to vanish on the grating normal and to be very small over the whole usable spectrum.*

### 8.6.2 Resolving power of image-forming systems

The Fraunhofer diffraction formula § 8.3 (36) finds important applications in the calculation of the resolving power of optical systems. We have already introduced the concept of resolving power in connexion with interference spectroscopes in §7.6.3, and in the preceding section we have estimated the resolving power that can be attained with gratings and prisms. We shall now extend this concept to imageforming systems.

In a spectral apparatus (e.g. a line grating or the Fabry-Perot interferometer), the resolving power is a measure of the ability of the instrument to separate two neighbouring spectral lines of slightly different wavelengths. In an image-forming system, it is a measure of the ability to separate images of two neighbouring object points. In the absence of aberrations each point object would, according to geometrical optics, give rise to a sharp point image. Because of diffraction the actual image will nevertheless always be a finite patch of light. And if two such image patches (diffraction patterns) overlap, it will be more and more difficult to detect the presence of two objects, the closer the central intensity maxima are to each other. The limit down to

which the eye can detect the two objects is, of course, to some extent a matter of practical experience. With a photographic plate the contrast may be enhanced and so the limit of resolution decreased by suitable development. Nevertheless it is desirable to have some simple criterion which permits a rough comparison of the relative efficiency of different systems, and for this purpose Rayleigh's criterion discussed in § 7.6.3 may again be employed. According to this criterion two images are regarded as just resolved when the principal maximum of one coincides with the first minimum of the other. For a spectral apparatus, where the limit of resolution is a certain wavelength difference $\Delta\lambda$, the resolving power was defined as the quantity $\lambda/\Delta\lambda$. For an image-forming system the limit of resolution is some distance $\delta x$ or angle $\delta\theta$ and the resolving power is defined as the reciprocal (i.e. $1/\delta x$ or $1/\delta\theta$ ) of this quantity.

Let us consider first the limit of resolution of a telescope. For a distant object, the edge of the entrance pupil coincides with the circular boundary of the objective, and acts as the diffracting aperture. If $a$ is the radius of the objective aperture then, according to § 8.5 (16), the position of the first minimum of intensity referred to the central maximum is given by*

$$w = 0.61\frac{\lambda}{a}. \tag{24}$$

Now $w = \sqrt{p^2 + q^2}$ represents the sine of the angle $\phi$ which the direction ( $p, q$ ) makes with the central direction $p = q = 0$. This angle is usually so small that its sine may be replaced by the angle itself, and it then follows that (on the basis of Rayleigh's criterion) the angular separation of two stars that can just be resolved is $0.61\lambda/a$.

With a given objective, the angular size of the image as seen by the eye depends on the magnification of the eyepiece. It is impossible, however, to bring out detail not present in the primary image by increasing the power of the eyepiece, for each element of the primary image is

0

- Such an astigmatic-free mounting was described by F. L. O. Wadsworth, Astrophys. J., 3 (1896), 54. For a discussion of the aberration theory of gratings and grating mountings see an article by W. T. Welford in Progress in Optics, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 241.

a small diffraction pattern, and the actual image, as seen by the eyepiece, is only the ensemble of the magnified images of these patterns.

The largest telescope in existence (at Mount Palomar) has a diameter $2a \sim 5$ m. Neglecting for the moment the effect of the central obstruction in the telescope, the theoretical limit of resolution for light near the centre of the visible range ( $\lambda \sim 5 \cdot 6 \cdot 10^{-5}$ cm ) is seen to be

$$\phi \sim 0.61 \frac{5 \cdot 6 \cdot 10^{-5} \text{ cm}}{2 \cdot 5 \cdot 10^{2} \text{ cm}} \sim 1 \cdot 4 \cdot 10^{-7}$$

or, in seconds of arc,

$$\phi \sim 0.028''.$$

In §6.1 we quoted the value of 1 min . of arc for the limit of resolution of the eye. We can now give a more precise estimate. Since the diameter of the pupil of the eye varies from about 1.5 mm to about 6 mm (depending on the intensity of the light), it follows that the limit of resolution lies in the range (again taking $\lambda = 5 \cdot 6 \cdot 10^{-5}$ cm )

$$0 \cdot 61 \frac{5 \cdot 6 \cdot 10^{-5}}{0 \cdot 75 \cdot 10^{-1}} > \phi > 0 \cdot 61 \frac{5 \cdot 6 \cdot 10^{-5}}{3 \cdot 10^{-1}}$$

i.e.

$$4 \cdot 55 \cdot 10^{-4} > \phi > 1 \cdot 14 \cdot 10^{-4}$$

or, in minutes and seconds:

$$1'34'' > \phi > 0'24''.$$

So far we have assumed the aperture to be circular. Of considerable interest is also the case of an annular aperture, since in many telescopes, for example, the central portion of the circular aperture is obstructed by the presence of a secondary mirror. Suppose that the annular aperture is bounded by two concentric circles of radii $a$ and $\varepsilon a$, where $\varepsilon$ is some positive number less than unity. The light distribution in the Fraunhofer pattern is then represented by an integral of the form §8.5 (8), but with the $\rho$ integration extending only over the domain $\varepsilon a \leqslant \rho \leqslant a$. In place of §8.5 (13) we then obtain

$$U(P) = C\pi a^2 \left[\frac{2J_1(kaw)}{kaw}\right] - C\pi\varepsilon^2 a^2 \left[\frac{2J_1(k\varepsilon aw)}{k\varepsilon aw}\right] \tag{25}$$

so that the intensity is given by

$$I(P) = \frac{1}{(1-\varepsilon^2)^2} \left[\left(\frac{2J_1(kaw)}{kaw}\right) - \varepsilon^2 \left(\frac{2J_1(k\varepsilon aw)}{k\varepsilon aw}\right)\right]^2 I_0 \tag{26}$$

where $I_0 = |C|^2\pi^2 a^4 (1-\varepsilon^2)^2$ is the intensity at the centre $w = 0$ of the pattern. The position of the minima (zeros) of intensity are now given by the roots of the equation

$$J_1(kaw) - \varepsilon J_1(k\varepsilon aw) = 0, \quad (w \neq 0) \tag{27}$$

whilst the maxima are given by the roots of

$$J_2(kaw) - \varepsilon^2 J_2(k\varepsilon aw) = 0 \tag{28}$$

0

- According to § 7.6.3 the saddle-to-peak intensity ratio at the limit of resolution for diffraction at a slit aperture is $8/\pi^2 = 0.811$. The corresponding value for the present case (circular aperture) is 0.735 .

In deriving (28), the relation § 8.5 (15) was used, as in the case of circular aperture. For the unobstructed aperture ($\varepsilon = 0$) the first root of (27) is, of course, the value given by (24), namely $w = 3 \cdot 83/ka = 0 \cdot 61\lambda/a$. As $\varepsilon$ is increased, the first root of (27) decreases,* and with $\varepsilon = \frac{1}{2}$, for example, it is slightly less than $3 \cdot 15/ka = 0 \cdot 50\lambda/a$. As the principal maximum remains at $w = 0$ independently of $\varepsilon$, we see that on obstructing the central portion of the aperture the resolving power is increased. This improvement is, however, accompanied by a decrease in the brightness of the image. Also the secondary maxima become more pronounced so that the contrast is reduced. With $\varepsilon = \frac{1}{2}$, the first secondary maximum (at $w = 4.8/ka$ ) is 0.10 of the principal maximum, as compared with the value 0.018 (at $w = 5.14/ka$ ) for a circular aperture (see Fig. 8.30).

The obstruction of the central part of the circular aperture corresponds to the replacement of the pupil function (cf. §8.3(39)) $G(\xi, \eta) = C$ or 0 according as $0 \leqslant \rho \leqslant a$ or $\rho > a$ by $G(\xi, \eta) = C$ or 0 according as $\varepsilon a \leqslant \rho \leqslant a$ or $\rho < \varepsilon a, \rho > a$. Naturally it is possible to alter the pupil function in other ways. A general method for modifying a pupil function consists in depositing on one or more surfaces of the system a thin, partially transmitting film of suitable substance. The same effect may be achieved by means of a specially constructed "filter", for example, a hollow lens of appropriate form filled with an absorbing liquid. The problem arises of determining the form of the pupil function which, in some agreed sense, would give the best possible image.

This problem has been investigated by a number of workers.* Of particular interest is a result due to Toraldo di Francia, † that the pupil function may be so chosen as to make the radius of the first dark ring arbitrarily small and at the same time the dark zone surrounding the central ring arbitrarily large. However, by gradually decreasing the radius of the first dark ring, less and less light is utilized in the central

0

- From (26) it follows by a simple calculation that as $\varepsilon \to 1, I/I_0 \to {J_0}^2(kaw)$. Since the first zero of the equation $J_0(x) = 0$ is at $x = 2 \cdot 40$, it follows that with increasing $\varepsilon$ the radius of the first dark ring approaches the value given by $w = 2 \cdot 40/ka = 0 \cdot 38\lambda/a$.

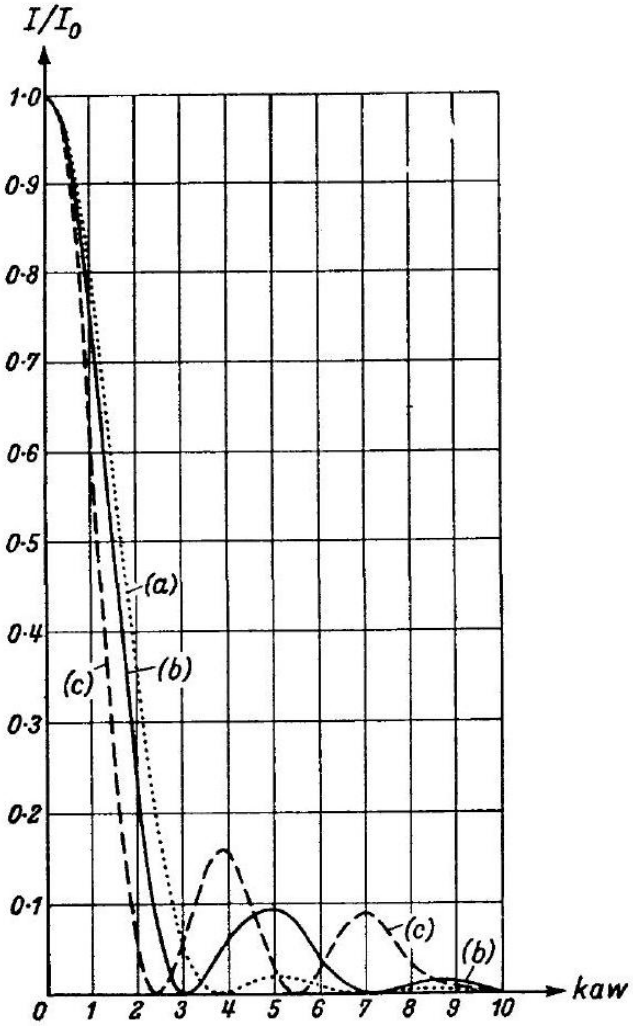



Fig. 8.30. Illustrating the effect of central obstruction on the resolution. Normalized intensity curves for Fraunhofer patterns of (a) Circular aperture, (b) Annular aperture with $\varepsilon=\frac{1}{2}$, and (c) Annular aperture with $\varepsilon \to 1$.

[After G. C. Steward, The Symmetrical Optical System (Cambridge University Press, 1928), p. 89.]

disc, so that the smallest practicable size of the disc and hence the resolving power is limited by the amount of light available.

A partial suppression of the secondary maxima by an appropriate modification of the pupil function is known as apodization. ‡ In spectroscopic analysis it facilitates the detection of satellites of spectral lines, whilst in astronomical applications it facilitates the resolution of double stars of appreciably different apparent brightness.

0

• See, for example, R. Straubel, P. Zeeman Verh. (Den Haag, Martinus Nijhoff, 1935), p. 302; R. K. Luneburg, Mathematical Theory of Optics (Mimeographed Lecture Notes, Brown University, Providence, R.I. (1944); printed version published by University of California Press, Berkeley and Los Angeles, 1964), §50); H. Osterberg and J. E. Wilkins, J. Opt. Soc. Amer., 39 (1949), 553; G. Lansraux, Rev. d'Optique 32 (1953), 475.

A brief review of investigations in this field is given in an article by E. Wolf, Rep. Progr. Phys. (London, Physical Society), 14 (1951), 109.

† G. Toraldo di Francia, Suppl. Nuovo Cimento, 9 (1952), 426.

‡ A thorough review of investigations on this subject was given by P. Jacquinot and B. RoizenDossier in Progress

The conventional theory of resolving power, as outlined in this section, is particularly appropriate to direct visual observations. With other methods of detection (e.g. photometric) the presence of two objects of much smaller angular separation than indicated by Rayleigh's criterion may often be revealed. In this connexion it is also of interest to compare the relative resolving efficiency of a telescope and a Michelson Stellar Interferometer (§7.3.6). If the presence of two stars is judged by means of the first vanishing of the fringes formed by the interferometer, and if $d$ is the maximum distance by which its outer mirrors may be separated, then according to § 7.3 (34) double stars down to angular separation $\phi \sim \frac{1}{2}\lambda/d$ may be detected with this instrument. Comparison of this value with (24) shows that, to detect double stars of this separation with a telescope used visually, the diameter $2a$ of its objective would have to be about $2 \cdot 4d$.

### 8.6.3 Image formation in the microscope

In the elementary theory of resolving power which we have just outlined, light from the two object points was assumed to be incoherent. This assumption is justified when the two objects are self-luminous, e.g. with stars viewed by a telescope. The intensity observed at any point in the image plane is then equal to the sum of the intensities due to each of the object points.

In a microscope the situation is, as a rule, much more complicated. The object is usually non-luminous and must, therefore, be illuminated with the help of an auxiliary system. Owing to diffraction on the aperture of the illuminating system (condenser), each element of the source gives rise to a diffraction pattern in the object plane of the microscope. The diffraction patterns which have centres on points that are sufficiently close to each other partly overlap, and in consequence the light vibrations at neighbouring points of the object plane are in general partially correlated. Some of this light is transmitted through the object with or without a change of phase, whilst the rest is scattered, reflected, or absorbed. In consequence, it is in general impossible to obtain, by means of a single observation, or even by the use of one particular arrangement, a faithful enlarged picture showing all the small-scale structural variations of the object. Various methods of observation have, therefore, been developed, each suitable for the study of certain types of objects, or designed to bring out particular features.

We shall briefly outline the theory of image formation in a microscope, confining our attention first of all to the two extreme cases of completely incoherent and perfectly coherent illumination. Partially coherent illumination will be discussed in § 10.5.2.

(a) Incoherent illumination We first consider a self-luminous object (e.g. an incandescent filament of an electric bulb). Let $P$ be the axial point of the object and $Q$ a neighbouring point in the object plane, at a distance $Y$ from $P$, and let $P'$ and $Q'$ be the images of these points (Fig. 8.31). Further let $\theta$ and $\theta'$ be the angles which the marginal rays of the axial pencils make with the axis.

If $a'$ is the radius of the region (assumed to be circular) in which the beam of light converging on $P'$ intersects the back focal plane $\mathscr{F}'$ and if $D'$ is the distance between the back focal plane and the image plane, then, since $\theta'$ is small,

$$\theta' = \frac{a'}{D'}. \tag{29}$$

Further, if $w = \sqrt{p^2 + q^2}$ is the separation of $Q'$ from $P'$ measured in "diffraction units" [cf. § 8.3 (35) and § 8.5 (7)], i.e. the sine of the angle which the two points subtend at the centre of the diffracting aperture, then we have to a good approximation,

in Optics, Vol. 3, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1964), p. 29.

$$Y' = w\mathrm{D}'. \tag{30}$$

Let $n$ and $n'$ be the refractive indices, $\lambda$ and $\lambda'$ the wavelengths in the object and image spaces, and $\lambda_0$ the wavelength in vacuum. Then, since according to § 8.5 (16) the first minimum of the diffraction pattern of $P$ is given by $w = 0.61\lambda'/a'$, we have, at the limit of resolution

$$Y' = 0.61\lambda'\frac{\mathrm{D}'}{a'} = 0.61\frac{\lambda'}{\theta'} = 0.61\frac{\lambda_0}{n'\theta'}. \tag{31}$$

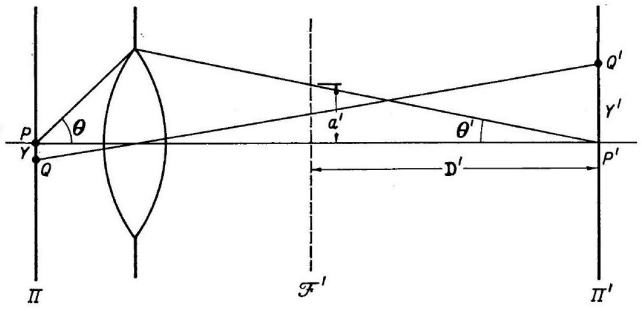



Fig. 8.31. Illustrating the theory of resolving power of the microscope.

A microscope must, of course, be so designed that it gives a sharp image not only of an axial point but also of neighbouring points of the object plane. According to § 4.5.1 the sine condition must therefore be satisfied, i.e.*

$$nY\sin\theta = -n'Y'\sin\theta'$$

Since $\theta'$ is small we may replace $\sin\theta'$ by $\theta'$. On substituting for $Y'$ into (31), we finally obtain

$$|Y| \sim 0.61\frac{\lambda_0}{n\sin\theta} \tag{32}$$

This formula gives the distance between two object points which a microscope can just resolve when the illumination is incoherent and the aperture is circular.

The quantity $n\sin\theta$ which enters into (32) is the numerical aperture, [cf. §4.8(13)] and must be large if a high resolving power is to be achieved. Means for obtaining a large numerical aperture were discussed in § 6.6.

(b) Coherent illumination-Abbe's theory We now consider the other extreme case, namely when the light emerging from the object may be treated as strictly coherent. This situation is approximately realized when a thin object of relatively simple structure is illuminated by light from a sufficiently small source via a condenser of low aperture (cf. § 10.5.2).

The first satisfactory theory of resolution with coherent illumination was formulated and also illustrated with beautiful experiments, by E. Abbe.* According to Abbe, the object acts as a diffraction grating, so that not only every element of the aperture of the objective, but also every element of the object must be taken into account in determining the complex disturbance at any particular point in the image plane. Expressed mathematically, the transition from the object to the image involves two integrations, one extending over the object plane, the other

0

• The minus sign appears here because $\theta'$ corresponds to $-\gamma_1$ of § 4.5.

extending over the aperture. In Abbe's theory, diffraction by the object is first considered and the effect of the aperture is taken into account in the second stage. An alternative procedure, in which the order is reversed, is also permissible and leads naturally to the same result. †

To illustrate Abbe's theory we consider first the imaging of a grating-like object which is illuminated by a plane wave incident normally on to the object plane

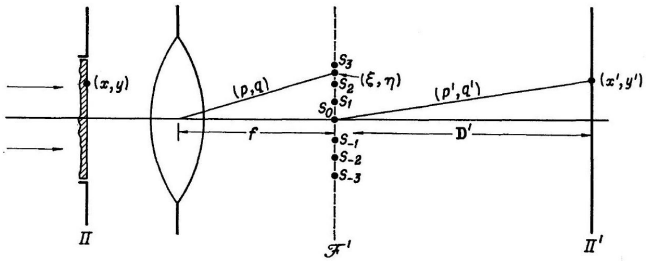



Fig. 8.32. Illustrating the Abbe theory of image formation in a microscope, with coherent illumination.

(Köhler's central illumination). The wave is diffracted by the object and gives rise to a Fraunhofer diffraction pattern of the grating (cf. §8.6.1), in the back focal plane $\mathscr{F}'$ of the objective. In Fig. 8.32 the maxima (spectra of successive orders) of this pattern are denoted by $\ldots S_{-2}, S_{-1}, S_0, S_1, S_2, \ldots$. Every point in the focal plane may be considered to be a centre of a coherent secondary disturbance, whose strength is proportional to the amplitude at that point. The light waves that proceed from these secondary sources will then interfere with each other and will give rise to the image of the object in the image plane $\Pi'$ of the objective. To obtain a faithful image it is necessary that all the spectra contribute to the formation of the image. Strictly this is never possible because of the finite aperture of the objective. We shall see later that the exclusion of some of the spectra may result in completely false detail appearing in the image. For practical purposes it is evidently sufficient that the aperture shall be large enough to admit all those spectra that carry an appreciable amount of energy.

Let us express these considerations in more precise terms without restricting ourselves to a grating-like object. If $x, y$ are the coordinates of a typical point in the object plane and $f$ is the distance of the focal plane $\mathscr{F}'$ from the lens objective, the disturbance at a point

$$\xi = pf, \quad \eta = qf \tag{33}$$

of the $\mathscr{F}'$ plane (see Fig. 8.32) is given by the Fraunhofer formula

$$U(\xi, \eta) = C_1 \iint_{\mathscr{A}} F(x, y) e^{-ik\left[\frac{\xi}{f}x + \frac{\eta}{f}y\right]} dxdy \tag{34}$$

where $F$ is the transmission function of the object, $C_1$ is a constant, and the integration is taken over the area $\mathscr{A}$ of the object plane II covered by the object.

0

- Ernst Abbe, Archiv. f. Mikroskopische Anat., 9 (1873), 413. Also his Gesammelte Abhandlungen, 1 (Jena: G. Fischer, 1904), 45 and O. Lummer and F. Reiche, Die Lehre von der Bildtstehung im Mikroskop von Ernst Abbe, (Braunschweig, Vieweg, 1910). A good account of Abbe's theory was also given in A. B. Porter, Phil. Mag. (6), 11 (1906), 154.
  † A theory of image formation in the microscope which is equivalent to this alternative approach was formulated by Lord Rayleigh in Phil. Mag. (5), 42 (1896), 167; also his Scientific Papers, 4 (Cambridge University Press, 1903), p. 235.

Next consider the transition from the back focal plane $\mathscr{F}'$ to the image plane II'. If, as before, $D'$ denotes the distance between $\mathscr{F}'$ and $\Pi'$, and $V(x', y')$ is the disturbance at a typical point

$$x' = p'\mathbf{D}', \quad y' = q'\mathbf{D}', \tag{35}$$

of the image plane, we have for Fraunhofer diffraction on the aperture $\mathscr{B}$ in $\mathscr{F}'$

$$V(x', y') = C_2 \iint_{\mathscr{A}} U(\xi, \eta) e^{-ik\left[\frac{x'}{\mathbf{D}'}\xi + \frac{y'}{\mathbf{D}'}\eta\right]} d\xi d\eta \tag{36}$$

it being assumed that $a'/D' \ll 1$ (see figure 8.31). Substitution from (34) into (36) gives

$$V(x', y') = C_1 C_2 \iint_{\mathscr{A}} \iint_{\mathscr{B}} F(x, y) e^{-i\frac{k}{f}\left[\left(x + \frac{f}{D'}x'\right)\xi + \left(v + \frac{f}{D'}y'\right)\eta\right]} dx dy d\xi d\eta \tag{37}$$

Now if $F(x, y)$ is defined as zero for all points of the object plane that lie outside $\mathscr{A}$, the integration with respect to $x$ and $y$ may formally be extended from $-\infty$ to $+\infty$. Also, if the aperture $\mathscr{B}$ is so large that $|U(\xi, \eta)|$ is negligible for points of the $\mathscr{F}'$-plane that lie outside $\mathscr{B}$, the integrations with respect to $\xi$ and $\eta$ may likewise each be extended over the range from $-\infty$ to $+\infty$. Noting also that [cf. §4.3 (10) where $f'$ and $Z'$ correspond to our $f$ and $-\mathrm{D}'$ respectively]

$$\frac{f}{\bar{D}'} = -\frac{1}{M} \tag{38}$$

where $M(< 0)$ is the magnification between $\Pi$ and $\Pi'$, we obtain by the application of the Fourier integral theorem*

$$V(x', y') = CF\left(\frac{x'}{M}, \frac{y'}{M}\right) = CF(x, y) \tag{39}$$

where ( $x, y$ ) is the object point whose image is at ( $x', y'$ ), and

$$C = C_1 C_2 \lambda^2 f^2$$

is a constant. Hence to the accuracy here implied † the image is strictly similar to the object (but inverted), provided the aperture is large enough.

To show that completely false detail may appear in the image if some of the spectra that carry appreciable energy are excluded, we consider a one-dimensional gratinglike object consisting of $N$ equidistant congruent slits of width $s$, separated by

opaque regions, with period $d$. For simplicity the aperture will be assumed to be rectangular with two of its sides parallel to the strips.

According to § 8.6 (3)

$$U(\xi) = C_1' \left(\frac{\sin\frac{k\xi s}{2f}}{\frac{k\xi s}{2f}}\right) \frac{1 - e^{-iNkd\xi lf}}{1 - e^{-ikd\xi lf}} \tag{40}$$

where for $U^{(0)}$ there has been substituted the expression relating to diffraction on a rectangular aperture and $C_1'$ is a constant (cf. §8.5.1). If the rectangular aperture extends in the $\xi$ direction throughout the range

0

- See for example R. Courant and D. Hilbert, Methods of Mathematical Physics (New York. Interscience Publishers, 1953), Vol. 1, p. 79.
  † As pointed out on p. 384, the Fraunhofer approximation used here is restricted to the case when the object points as well as the image points are sufficiently close to the axis.

$$-a \leqslant \xi \leqslant a,$$

the disturbance in the image plane is, by (36) and (40), given by ( $C'$ denoting a constant)

$$V(x') = C' \int_{-a}^{a} \frac{\sin \frac{ks\xi}{2f}}{\frac{ks\xi}{2f}} \frac{1 - e^{-iNkd\xi/f}}{1 - e^{-ikd\xi/f}} e^{-ikx'\xi/\mathrm{D}'} d\xi \tag{41}$$

The position of principal maxima of the integrand are given by the roots of the equation $1 - \exp[-ikd\xi/f] = 0$, i.e. by $\xi = mf\lambda/d$, where $m$ is an integer. Between these principal maxima there are weak secondary maxima. If $N$ is large, the principal maxima are very sharp and the secondary maxima negligible in comparison. To a good approximation we may then replace the integral by a sum of integrals, each extending from the midpoint $Q_m$ of the interval between two successive principal maxima to the next midpoint $Q_{m+1}$. In each interval we may replace the argument by the central value $\xi = mf\lambda/d = 2\pi mf/kd$, and obtain for $V$ the following expression:

$$V(x') \sim V_0 \sum_{-\bar{m}<m<\bar{m}} \frac{\sin \frac{m\pi s}{d}}{\frac{m\pi s}{d}} e^{\frac{2\pi i m x'}{d'}} \tag{42}$$

Here

$$\bar{m} = \frac{ad}{\lambda f}, \quad d' = Md = -\frac{\mathrm{D}'}{f} d \tag{43}$$

and $V_0$ is the integral

$$V_0 = C' \int_{Q_m}^{Q_{m+1}} \frac{1 - e^{-iNkd\xi/f}}{1 - e^{-ikd\xi/f}} d\xi \tag{44}$$

which, apart from small correction terms in the high orders, is practically independent of $m$. The series (42) may be re-written in real form as

$$\frac{V(x')}{V_0} = 1 + 2 \sum_{1<m<\bar{m}} \frac{\sin \frac{m\pi s}{d}}{\frac{m\pi s}{d}} \cos \frac{2\pi m x'}{d'} \tag{45}$$

Suppose first that the length $a$ of the aperture is very large. The summation may then formally be extended over the whole infinite range ( $\bar{m} = \infty$ ), and we can easily
verify that the image is then strictly similar to the object. For this purpose we expand the transmission function $F$ of the grating-like object (see Fig. 8.33),

$$\left.\begin{array}{lll} F(x) = F_0 & \text{when} & 0 < |x| < s/2 \\ = 0 & \text{when} & s/2 < |x| < d/2 \end{array}\right\} \tag{46}$$

into a Fourier series

$$F(x) = c_0 + 2 \sum_{m=1}^{\infty} c_m \cos \frac{2\pi m x}{d}. \tag{47}$$

Then

$$c_0 = \frac{F_0 s}{d}, \quad c_m = F_0 \frac{\sin \frac{\pi m s}{d}}{\pi m} \quad (m = 1, 2, 3 \ldots). \tag{48}$$

We see that apart from a constant factor this series is the same as (45).

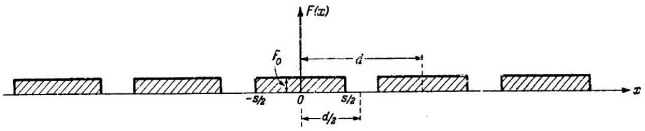



Fig. 8.33. A grating-like object.

Suppose now that the length $a$ of the aperture is decreased. If $a$ is so small that only the zero-order spectrum contributes to the image, i.e. if $\bar{m} = ad/\lambda f$ is only a fraction of unity, then according to (45) $V(x') =$ constant, so that the image plane is uniformly illuminated. (This result is, of course, not strictly true, as we have neglected certain error terms; in reality there is a weak drop in intensity towards the edge.)

If in addition to the zero-order spectrum the two spectra of the first order ( $S_1, S_{-1}$ ) are also admitted by the aperture, i.e. if $\bar{m} = ad/\lambda f$ is slightly greater than unity, then we see from (45) that

$$\frac{V(x')}{V_0} = 1 + 2\frac{\sin\frac{\pi s}{d}}{\frac{\pi s}{d}}\cos\frac{2\pi x'}{d'}. \tag{49}$$

The image has now the correct periodicity $x' = d'$, but a considerably flattened intensity distribution. By increasing the aperture more and more the image is seen to resemble the object more and more closely.

A completely false image is obtained when the lower orders are excluded. If for example all orders except the second are excluded, then

$$\frac{V(x')}{V_0} = 2\frac{\sin\frac{2\pi s}{d}}{\frac{2\pi s}{d}}\cos\frac{4\pi x'}{d'}, \tag{50}$$

so that the image has the period $x' = d'/2$; the "image" shows twice the number of lines that are in fact present in the object,

Finally let us estimate the resolving power. Consider again the situation illustrated in Fig. 8.31, but assume now that the light from $P$ and $Q$ is coherent. Then the distribution in the image plane arises essentially from the coherent superposition of the two Arry diffraction patterns, one centred on $P'$, the other on $Q'$. The complex amplitude at a point situated between $P'$ and $Q'$ at distance $w_1$ (measured in "diffraction units") from $P'$ is given by

$$U(w_1) = \left[\frac{2J_1(kaw_1)}{kaw_1} + \frac{2J_1[ka(w-w_1)]}{ka(w-w_1)}\right]U_0, \tag{51}$$

$w$ being the distance between $P'$ and $Q'$ and the other symbols having the same meaning as before. The intensity is, therefore, given by

$$I(w_1) = \left[\frac{2J_1(kaw_1)}{kaw_1} + \frac{2J_1(ka(w-w_1)}{ka(w-w_1)}\right]^2 I_0 \tag{52}$$

Now in the case of incoherent illumination, $P'$ and $Q'$ were considered as resolved when the principal intensity maximum of the one pattern coincided with the first minimum of the other. The intensity at the midpoint ( $kaw \sim 1.92$ ) between the two maxima is then equal to $2\,[2J_1(1.92)/1.92]^2 \sim 0.735$ of the maximum intensity of either, i.e. the combined intensity curve has a dip of about 26.5 per cent between the principal maxima. (This corresponds to the value 19 per cent for a slit aperturecf. Fig. 7.62.) If we consider a dip of this amount as again substantially determining the limit of resolution, the critical separation $w = 2w_1$ is obtained from the relation

$$\frac{I(w_1)}{I_0} = \left[2\frac{2J_1(kaw_1)}{kaw_1}\right]^2 = 0.735. \tag{53}$$

The first root of this transcendental equation is $w_1 \sim 2.41/ka$, so that the critical separation measured in ordinary units is

$$Y' = 2w_1 \mathrm{D}' \sim \frac{2 \cdot 41}{\pi}\frac{\mathrm{D}'\lambda'}{a} = \frac{0 \cdot 77\lambda'}{\theta'} = \frac{0 \cdot 77\lambda_0}{n'\theta'} \tag{54}$$

To relate $Y'$ to the corresponding separation $Y$ of the object points we use the sine condition (with the approximation $\sin\theta' \sim \theta'$ ), and finally obtain for the limit of resolution with coherent illumination the expression

$$|Y| = 0.77\frac{\lambda_0}{n\sin\theta} \tag{55}$$

Apart from a larger numerical factor (which in any case is somewhat arbitrary as it depends on the form of the object and aperture and on the sensitivity of the receptor), we obtain the same expression as in the case of incoherent illumination [eq. (32)]. Thus with light of a given wavelength the resolving power is again substantially determined by the numerical aperture of the objective.

(c) Coherent illumination-Zernike's phase contrast method of observation* We have defined a phase object as one which alters the phase but not the amplitude of the incident wave. An object of this type is of non-uniform optical thickness, but

does not absorb any of the incident light. Such objects are frequently encountered in biology, crystallography, and other fields. It is evident from the preceding discussion that with ordinary methods of observation little information about phase objects can be obtained. For the complex amplitude function that specifies the disturbance in the image plane is then similar to the transmission function of the object* and, as the eye (or any other observing instrument) only distinguishes changes in intensity, one can only draw conclusions about the amplitude changes but not about the phase changes introduced by the object.

To obtain information about phase objects, special methods of observation must be used, for example, the so-called central dark ground method of observation where the central order is excluded by a stop, or the Schlieren method, where all the spectra on one side of the central order are excluded. The most powerful method, which has the advantage that it produces an intensity distribution which is directly proportional to the phase changes introduced by the object, is due to Zernike † and was first described by him in 1935. It is known as the phase contrast method.

To explain the principle of the phase contrast method, consider first a transparent object in the form of a one-dimensional phase grating. The transmission function of such an object is by definition (see p. 401) of the form

$$F(x) = e^{i\phi(x)} \tag{56}$$

where $\phi(x)$ is a real periodic function, whose period ( $d$ say) is equal to the period of the grating. We assume that the magnitude of $\phi$ is small compared to unity, so that we may write

$$F(x) \sim 1 + i\phi(x). \tag{57}$$

0

- For fuller discussion of the phase contrast method see, for example, M. Franon, Le contraste de phase en optique et en microscopie (Paris, Revue d'Optique, 1950); and A. H. Bennett, H. Jupnik, H. Osterberg, and O. W. Richards, Phase Microscopy (New York, J. Wiley & Sons, 1952).

If we develop $F$ into a Fourier series

$$F(x) = \sum_{m=-\infty}^{\infty} c_m e^{\frac{2\pi i m x}{d}} \tag{58}$$

then, since $F$ is of the form (56) and $\phi$ is real and numerically small compared to unity,

$$c_0 = 1, \quad c_{-m} = -c_m^\star \quad (m \neq 0) \tag{59}$$

The intensity of the $m$ th order spectra is proportional to $|c_m|^2$.
In the phase contrast method of observation a thin plate of transparent material called the phase plate is placed in the back focal plane $\mathscr{F}'$ of the objective and by means of it the phase of the central order ( $S_0$ in Fig. 8.32) is retarded or advanced with respect to the diffraction spectra ( $S_1, S_{-1}, S_2, S_{-2}, \ldots$ ) by one-quarter of a period. This means that the complex amplitude distribution in the focal plane is altered from a distribution characterized by the coefficients $c_m$, to a distribution characterized by coefficients $c'_m$, where

$$c'_0 = c_0 e^{\pm i\pi/2} = \pm i, \quad c'_m = c_m \quad (m \neq 0) \tag{60}$$

the positive or negative sign being taken according as the phase of the central order is retarded or advanced. The resulting light distribution in the image plane will now no longer represent the phase grating (57), but rather a fictitious amplitude grating

$$G(x) = \pm i + i\phi(x). \tag{61}$$

Hence the intensity in the image plane will now be proportional to (neglecting $\phi^2$ in comparison to unity)

$$I(x') = |G(x)|^2 = 1 \pm 2\phi(x), \tag{62}$$

where as before $x' = Mx, M$ being the magnification. This relation shows that with the phase contrast method of observation, phase changes introduced by the object are transformed into changes in intensity, the intensity at any point of the image plane being (apart from an additive constant) directly proportional to the phase change due to the corresponding element of the object.* When the phase of the central order is retarded with respect to the diffraction spectra (upper sign in (61)), regions of the object which have greater optical thickness will appear brighter than the mean illumination, and one then speaks of a bright phase contrast; when the phase of the central order is advanced, regions of greater spectral thickness will appear darker and one then speaks of a dark phase contrast (Figs. 8.34 and 8.35).

To obtain good resolution, the aperture of the illuminating system is often of annular rather than circular form (cf. § 8.6.2). In this case the annular region of $\mathscr{F}'$ through which the direct (undiffracted) light passes plays the role of the central order $S_0$ of Fig. 8.32, and it is this light which must then be retarded or advanced by a quarter period.

The phase-changing plate may be produced by evaporating a thin layer of a suitable dielectric substance on to a glass substrate. If $n$ is the refractive index of the substance and

0

- Strict similarity would actually be attained only if the objective had an infinite aperture. Because the aperture is always finite, some details of the phase structures can be seen. In some cases the visibility of such "images" is enhanced, at the expense of resolution, by a slight defocusing of the instrument (cf. H. H. Hopkins, contribution in M. Franon, Le contraste de phase et le contraste par interfrences (Paris, Revue d'Optique, 1952), 142).
  † F. Zernike, Z. Tech. Phys., 16 (1935), 454; Phys. Z., 36 (1935), 848; Physica, 9 (1942), 686, 974.

$d$ the thickness of the layer, then for a retardation of a quarter of a period one must have $d = \lambda/4(n-1)$. A retardation of the central order by this amount is, of course, equivalent to an advance of the diffracted spectra by three-quarters of a period, and vice versa. It is possible to increase the sensitivity of the method by using slightly absorbing instead of a dielectric coating. We shall return to this point later.

It remains to show that the phase contrast method is not restricted to phase objects of periodic structure. For this purpose we divide the integral (34) into two parts:

$$U(\xi,\eta) = U_0(\xi,\eta) + U_1(\xi,\eta) \tag{63}$$

where

$$\left.\begin{aligned} U_0 &= C_1 \iint_{\mathscr{A}} e^{-\frac{ik}{f}[\xi x+\eta y]} dxdy \\ U_1 &= C_1 \iint_{\mathscr{A}} [F(x,y)-1] e^{-\frac{ik}{f}[\xi x+\eta y]} dxdy \end{aligned}\right\} \tag{64}$$

$U_0$ represents the light distribution that would be obtained in the plane $\mathscr{F}'$ if no object were present, whilst $U_1$ represents the effect of diffraction. Now the "direct light" $U_0$ (corresponding to the central order $S_0$ of Fig. 8.32), will be concentrated in only a small region $\mathscr{B}_0$ of the $\mathscr{F}'$-plane, around the axial point $\xi = \eta = 0$. On the

other hand a very small fraction of the diffracted light will, in general, reach this region, most of it being diffracted to other parts of this plane.*

Suppose that the region $\mathscr{B}_0$ through which the direct light passes is covered by a phase plate. The effect of the plate may be described by a transmission function

$$A = ae^{i\alpha} \tag{65}$$

For a plate that only retards or advances the light which is incident upon it, $a = 1$; for a plate that also absorbs light $a < 1$. The light emerging from the aperture will be represented by

$$U'(\xi,\eta) = AU_0(\xi,\eta) + U_1(\xi,\eta) \tag{66}$$

so that, according to (36), the distribution of the complex amplitude in the image is given by

$$V(x',y') = V_0(x',y') + V_1(x',y') \tag{67}$$

where

$$\left.\begin{aligned} V_0 &= AC_2 \iint_{\mathscr{A}} U_0(\xi,\eta) e^{-\frac{ik}{D'}[x'\xi+y'\eta]} d\xi d\eta \\ V_1 &= C_2 \iint_{\mathscr{A}} U_1(\xi,\eta) e^{-\frac{ik}{D'}[x'\xi+y'\eta]} d\xi d\eta \end{aligned}\right\} \tag{68}$$

Now the aperture $\mathscr{B}$ greatly exceeds in size the region $\mathscr{B}_0$, and since $U_0$ was seen to be practically zero outside $\mathscr{B}_0$, no appreciable error is introduced by extending the domain of integration in $V_0$ over the whole $\mathscr{F}'$-plane. Moreover, if $\mathscr{B}$ is assumed to be so large as to admit all the diffracted rays that carry any appreciable energy, the integral for $V_1$ may likewise be given infinite limits. Finally, if as before the transmission function $F(x,y)$ is defined as zero at points of the object plane outside the region covered by the object, the integrals (64) may also

0

- The approximation implicit in the elementary theory of the phase contrast method are discussed by J. Picht, Zeitschr. f. Instrkde, 56 (1936), 481; ibid. 58 (1938), 1 and by F. D. Kahn, Proc. Phys. Soc., B 68 (1955), 1073.

be taken with infinite limits. We then obtain, on substituting from (64) into (68), and using the Fourier integral theorem and the relation (38),

$$\left.\begin{aligned} V_0(x', y') &= CA \\ V_1(x', y') &= C\left[F\left(\tfrac{x'}{M}, \tfrac{y'}{M}\right) - 1\right] = C[F(x, y) - 1]. \end{aligned}\right\} \tag{69}$$

From (67) and (69) it follows that the intensity in the image plane is given by

$$I(x', y') = |V(x', y')|^2 = |C|^2 |A + F(x, y) - 1|^2. \tag{70}$$

With a phase object

$$F(x, y) = e^{i\phi(x,y)} \tag{71}$$

and (70) reduces to †

$$I(x', y') = |C|^2 \left[a^2 + 2\{1 - a\cos\alpha - \cos\phi(x, y) + a\cos(\alpha - \phi(x, y))\}\right] \tag{72}$$

$$I(x', y') = 2C^2[1 - \cos\phi(x, y)].$$

Since $\phi$ was assumed to be small, (72) may be written as

$$I(x', y') = |C|^2 \left[a^2 + 2a\phi(x, y)\sin\alpha\right] \tag{73}$$

and, if the phase difference introduced by the plate represents a retardation or advance by a quarter of a period, then $\alpha = \pm\pi/2$ and (73) reduces to

$$I(x', y') = |C|^2 \left[a^2 \pm 2a\phi(x, y)\right] \tag{74}$$

When the plate does not absorb any of the incident light ( $a = 1$ ) we have again the expression (62). The intensity changes are then directly proportional to the phase variations of the object. With a plate that absorbs a fraction $a^2$ of the direct light the ratio of the second term to the first term in (73) has the value $\pm\phi/a$, so that the contrast of the image is enhanced. For example, by weakening the direct light to one-ninth of its original value, the sensitivity of the method is increased three times.

## 9.11.7 8.7. Fresnel Diffraction at a Straight Edge

### 8.7.1 The diffraction integral

Having considered various cases of Fraunhofer diffraction, we now turn our attention to the more general case of Fresnel diffraction.

The basic diffraction integral § 8.3 (28) may be written in the form

$$U(P) = B(C + iS), \tag{1}$$

where

0

- This point was investigated in detail by J. Preht, Zeitschr. f. Instrkde, 58 (1938), 1. See also F. Zernike, Mon. Not. Roy. Astr. Soc., 94 (1934), 382-383, where it is discussed in a somewhat different connexion. † The special case when $a = 0$ corresponds to the dark-ground method of observation. According to (72), the intensity distribution is then given by

$$\left.\begin{aligned} B &= -A\tfrac{i}{\lambda}\cos\delta\tfrac{e^{ik(r'+s')}}{r's'} \\ C &= \iint_{\mathscr{A}}\cos\{kf(\xi,\eta)\}d\xi d\eta \\ S &= \iint_{\mathscr{A}}\sin\{kf(\xi,\eta)\}d\xi d\eta \end{aligned}\right\} \tag{3}$$

The intensity $I(P) = |U(P)|^2$ at the point $P$ of observation is then given by

$$I(P) = |B|^2\left(C^2 + S^2\right). \tag{4}$$

We must now retain in the expansion § 8.3 (31) for $f(\xi,\eta)$ terms in $\xi$ and $\eta$ at least up to the second order.

As before we take the plane of the aperture $\mathscr{A}$ as the $xy$-plane. To simplify the calculations we choose as the $x$ direction the projection of the line $P_0P$ on to the plane of the aperture (Fig. 8.36). Thus with a source in a prescribed position our reference system will in general be different for different points of observation.

According to § 8.3 (30), we now have $l = l_0, m = m_0$, so that the linear terms in $f(\xi,\eta)$ disappear. The direction cosines of the rays $P_0O$ and $OP$ are

$$\left.\begin{aligned} l &= l_0 = \sin\delta, \\ m &= m_0 = 0, \\ n &= n_0 = \cos\delta, \end{aligned}\right\} \tag{5}$$

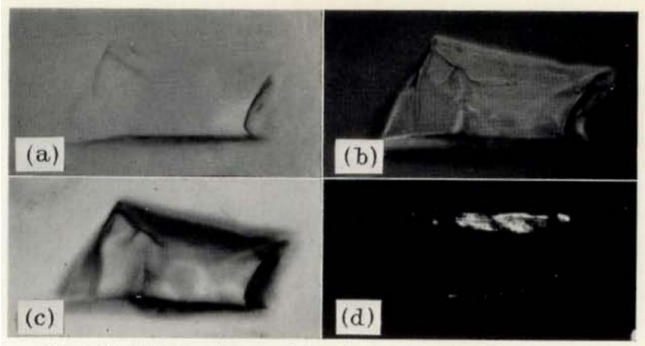



Fig. 8.34. Microscope images of glass fragments ( $n = 1.52$ ) mounted in clarite ($n = 1.54$), 100×.

(a), Bright field image; (b) and (c), Phase contrast images; (d), Dark field image.
[After A. H. Bennett, H. Jupnik, H. Osterberg, and O. W. Richards, Trans. Amer. Microscop. Soc., 65 (1946), 126.]

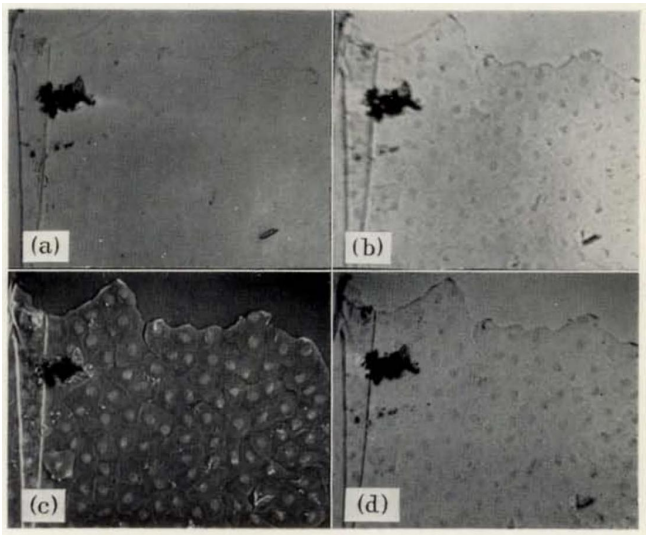



Fig. 8.35. Microscope images of epithelium from frog nictitating membrane, 100×.
(a), Bright field image at full aperture, N.A. = 0.25; (b), Bright field image with aperture half filled; (c), Phase-contrast image (bright contrast) at full aperture; (d), Phase-contrast image (dark contrast) at full aperture.

[After A. H. Bennett, H. Jupnik, H. Osterberg, and O. W. Richards, Trans. Amer. Microscop. Soc., 65 (1946), 119.]
where, as before, $\delta$ denotes the angle between the line $P_0P$ and the $z$-axis. The expression § 8.3 (31) for $f(\xi, \eta)$ reduces to

$$f(\xi, \eta) = \frac{1}{2}\left(\frac{1}{r'} + \frac{1}{s'}\right)\left(\xi^2 \cos^2 \delta + \eta^2\right) + \dots \tag{6}$$

If we neglect terms of third and higher order in $\xi$ and $\eta$, the integrals (3) become

$$\left.\begin{aligned} C &= \iint_{\mathscr{A}} \cos\left\{\tfrac{\pi}{\lambda}\left(\tfrac{1}{r'} + \tfrac{1}{s'}\right)\left(\xi^2 \cos^2 \delta + \eta^2\right)\right\} d\xi d\eta, \\ S &= \iint_{\mathscr{A}} \sin\left\{\tfrac{\pi}{\lambda}\left(\tfrac{1}{r'} + \tfrac{1}{s'}\right)\left(\xi^2 \cos^2 \delta + \eta^2\right)\right\} d\xi d\eta. \end{aligned}\right\} \tag{7}$$

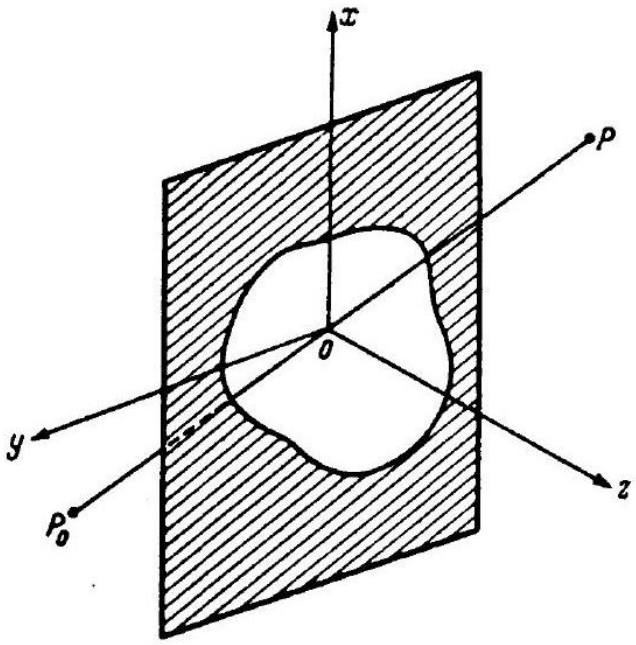



Fig. 8.36. Fresnel diffraction at an opening in a plane opaque screen.

It is convenient to introduce new variables of integration $u, v$, defined by

$$\left.\begin{array}{l} \frac{\pi}{\lambda}\left(\frac{1}{r'}+\frac{1}{s'}\right)\xi^2\cos^2\delta=\frac{\pi}{2}u^2, \\ \frac{\pi}{\lambda}\left(\frac{1}{r'}+\frac{1}{s'}\right)\eta^2=\frac{\pi}{2}v^2. \end{array}\right\} \tag{8}$$

Then

$$d\xi d\eta=\frac{\lambda}{2}\frac{dudv}{\left(\frac{1}{r'}+\frac{1}{s'}\right)\cos\delta}$$

and our integrals become

$$\left.\begin{array}{l} C=b\iint_{\mathscr{A}}\cos\left\{\frac{\pi}{2}\left(u^2+v^2\right)\right\}dudv \\ S=b\iint_{\mathscr{A}}\sin\left\{\frac{\pi}{2}\left(u^2+v^2\right)\right\}dudv. \end{array}\right\} \tag{9}$$

where

$$b=\frac{\lambda}{2\left(\frac{1}{r'}+\frac{1}{s'}\right)\cos\delta}. \tag{10}$$

The integration now extends over the region $\mathscr{A}'$ of the $(u, v)$ plane into which the region $\mathscr{A}$ of the aperture is transformed by means of (8).

### 8.7.2 Fresnel's integrals

If $\mathscr{A}'$ is a rectangle with sides parallel to the axes of $u$ and $v$, the integrals are simplified still further by means of the identities

$$\left.\begin{array}{l} \cos\left[\frac{\pi}{2}\left(u^2+v^2\right)\right]=\cos\left(\frac{\pi}{2}u^2\right)\cos\left(\frac{\pi}{2}v^2\right)-\sin\left(\frac{\pi}{2}u^2\right)\sin\left(\frac{\pi}{2}v^2\right), \\ \sin\left[\frac{\pi}{2}\left(u^2+v^2\right)\right]=\sin\left(\frac{\pi}{2}u^2\right)\cos\left(\frac{\pi}{2}v^2\right)+\cos\left(\frac{\pi}{2}u^2\right)\sin\left(\frac{\pi}{2}v^2\right). \end{array}\right\} \tag{11}$$

To evaluate (9) in this case we must consider the integrals

$$\left.\begin{array}{l} \mathscr{C}(w)=\int_0^w\cos\left(\frac{\pi}{2}\tau^2\right)d\tau \\ \mathscr{S}(w)=\int_0^w\sin\left(\frac{\pi}{2}\tau^2\right)d\tau \end{array}\right\} \tag{12}$$

$\mathscr{C}(w)$ and $\mathscr{S}(w)$ are known as Fresnel's integrals. They are of importance in connexion with many diffraction problems and have been extensively studied. We must briefly consider some of their properties.*

First we derive series expressions for $\mathscr{C}(w)$ and $\mathscr{S}(w)$. Expanding the cosine and sine under the integral signs into power series and integrating term by term we find that

$$\left.\begin{aligned}\mathscr{C}(w) &= w\left[1-\tfrac{1}{2!5}\left(\tfrac{\pi}{2}w^2\right)^2+\tfrac{1}{4!9}\left(\tfrac{\pi}{2}w^2\right)^4-\ldots\right]\\ \mathscr{S}(w) &= w\left[\tfrac{1}{1!3}\left(\tfrac{\pi}{2}w^2\right)-\tfrac{1}{3!7}\left(\tfrac{\pi}{2}w^2\right)^3+\tfrac{1}{5!11}\left(\tfrac{\pi}{2}w^2\right)^5-\ldots\right].\end{aligned}\right\} \tag{13}$$

The series (13) are convergent for all values of $w$ but are suitable for computations only when $w$ is small. When $w$ is large the integrals may be evaluated from series in inverse powers of $w$. We re-write (12) as

$$\mathscr{C}(w) = \mathscr{C}(\infty) - \int_w^\infty \frac{d}{d\tau}\left(\sin\frac{\pi}{2}\tau^2\right)\frac{d\tau}{\pi\tau} \tag{14}$$

Integration by parts gives

$$\mathscr{C}(w) = \mathscr{C}(\infty) + \frac{1}{\pi w}\sin\left(\frac{\pi}{2}w^2\right) + \int_w^\infty \frac{d}{d\tau}\left(\cos\frac{\pi}{2}\tau^2\right)\frac{d\tau}{\pi^2\tau^3}$$

Integrating again by parts and continuing this process we obtain

$$\mathscr{C}(w) = \mathscr{C}(\infty) - \frac{1}{\pi w}\left[P(w)\cos\left(\frac{\pi}{2}w^2\right) - Q(w)\sin\left(\frac{\pi}{2}w^2\right)\right],$$

and similarly

$$\mathscr{S}(w) = \mathscr{S}(\infty) - \frac{1}{\pi w}\left[P(w)\sin\left(\frac{\pi}{2}w^2\right) + Q(w)\cos\left(\frac{\pi}{2}w^2\right)\right],\Bigg\} \tag{15}$$

where

$$\left.\begin{aligned}Q(w) &= 1-\tfrac{1.3}{(\pi w^2)^2}+\tfrac{1.3.5.7}{(\pi w^2)^4}-\ldots\\ P(w) &= \tfrac{1}{\pi w^2}-\tfrac{1.3.5}{(\pi w^2)^3}+\tfrac{1.3.5.7.9}{(\pi w^2)^5}\ldots\end{aligned}\right\} \tag{16}$$

To evaluate the integrals $\mathscr{C}(\infty)$ and $\mathscr{S}(\infty)$, we combine them into a complex integral

$$\mathscr{C}(\infty) + i\mathscr{S}(\infty) = \int_0^\infty e^{i\frac{\pi}{2}\tau^3}d\tau \tag{17}$$

and introduce a new variable of integration

$$\zeta = \tau\sqrt{-\frac{i\pi}{2}} = \tau\frac{i-1}{2}\sqrt{\pi}, \quad \tau = -\zeta\frac{i+1}{\sqrt{\pi}}.$$

The real path of integration $0 \leqslant \tau \leqslant \infty$ goes over into a path along a line through the origin and inclined at 45° to the real axis in the complex $\zeta$ plane. Now it is easy to see that if the integral is taken along a line parallel to the imaginary axis then, with increasing distance

0

- Of the numerous tables of Fresnel's integrals we may refer to the following: British Association Report (Oxford, 1926), 273-5. E. Jahnke and F. Emde, Tables of Functions with Formulae and Curves (Leipzig and Berlin, Teubner; reprinted by Dover Publications, New York, 4th ed., 1945), p. 35. G. N. Watson, A Treatise on the Theory of Bessel Functions (Cambridge University Press, 2nd ed., 1944), p. 744. T. Pearcey, Tables of Fresnel Integrals to Six Decimal Places (Cambridge University Press, 1956).

from the origin, the integral tends to zero. It then follows from Cauchy's residue theorem that the integral taken along any oblique line through the origin is equal to the value of the integral taken along the real axis. Hence

$$\mathscr{C}(\infty) + i\mathscr{S}(\infty) = \frac{i+1}{\sqrt{\pi}} \int_0^\infty e^{-\zeta^2} d\zeta = \frac{i+1}{2}.$$

(The real integral with respect to $\zeta$ is the well known Gaussian error integral,* and has the value $\sqrt{\pi}/2$.) Thus

$$\left.\begin{aligned} \mathscr{C}(\infty) &= \textstyle\int_0^\infty \cos\left(\frac{\pi}{2}\tau^2\right) d\tau = \frac{1}{2}, \\ \mathscr{S}(\infty) &= \textstyle\int_0^\infty \sin\left(\frac{\pi}{2}\tau^2\right) d\tau = \frac{1}{2}. \end{aligned}\right\} \tag{18}$$

The relations (15), together with (16) and (18), express Fresnel's integrals in series of inverse powers of $w$. These are divergent (asymptotic) series, which provide a good approximation to the integrals when $w$ is large by taking only a limited number of terms into account (cf. Appendix III).

The behaviour of Fresnel's integrals may be illustrated by means of an elegant geometrical representation due to Cornu. †$\mathscr{C}$ and $\mathscr{S}$ are regarded as rectangular

Cartesian coordinates of a point $P$. As $w$ takes on all the possible values, the point $P$ describes a curve. Since $\mathscr{C}(0) = \mathscr{S}(0) = 0$ the curve passes through the origin, and since

$$\mathscr{C}(-w) = -\mathscr{C}(w), \quad \mathscr{S}(-w) = -\mathscr{S}(w), \tag{19}$$

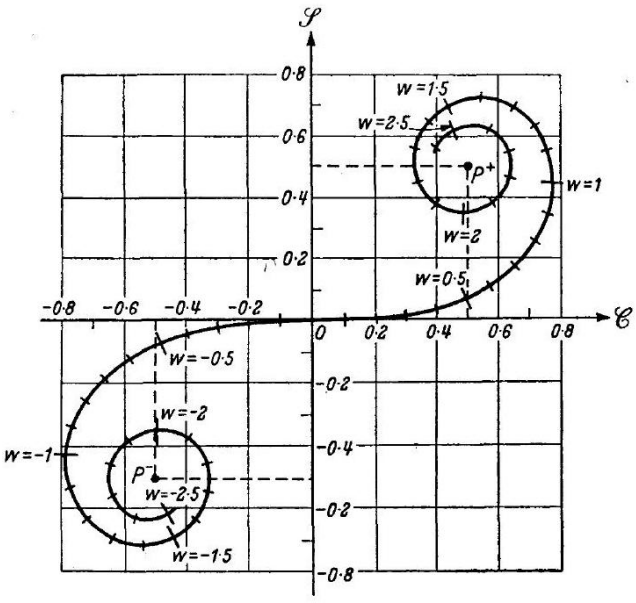



Fig. 8.37. The Corno spiral.

it is antisymmetric with respect to both axes. If $dl$ is an element of arc of the curve then

0

• See, for example, R. Courant, Differential and Integral Calculus, Vol. 1 (London and Glasgow. Blackie and Sons Ltd., 2nd ed., 1942), p. 496.
† A. Cornu, Journ. de Phys., 3 (1874), 5, 44.

$$dl^2 = d\mathscr{C}^2 + d\mathscr{S}^2 = \left[\left(\frac{d\mathscr{C}}{dw}\right)^2 + \left(\frac{d\mathscr{S}}{dw}\right)^2\right](dw)^2$$
$$= \left[\cos^2\left(\frac{\pi}{2}w^2\right) + \sin^2\left(\frac{\pi}{2}w^2\right)\right](dw)^2$$

i.e.

$$(dl)^2 = (dw)^2. \tag{20}$$

Hence if $l$ is measured in the sense of increasing $w$, the parameter $w$ represents the length of arc of the curve measured from the origin.

Let $\theta$ be the angle which the tangent to the curve makes with the $\mathscr{C}$ axis. Then

$$\tan\theta = \frac{d\mathscr{S}}{d\mathscr{C}} = \frac{\frac{d\mathscr{S}}{dw}}{\frac{d\mathscr{C}}{dw}} = \frac{\sin\left(\frac{\pi}{2}w^2\right)}{\cos\left(\frac{\pi}{2}w^2\right)} = \tan\left(\frac{\pi}{2}w^2\right)$$

i.e.

$$\theta = \frac{\pi}{2}w^2. \tag{21}$$

Thus $\theta$ increases monotonically with $|w|$. Since $\theta = 0$ when $w = 0$ the tangent touches the $\mathscr{C}$ axis at the origin. When $w^2 = 1$ then $\theta = \pi/2$, so that the tangent is then perpendicular to the $\mathscr{C}$ axis. When $w^2 = 2, \theta = \pi$ and the tangent is then parallel to the $\mathscr{C}$ axis again but is oriented in the negative direction. Since according
to (18) and (19)$\mathscr{C}(\infty) = -\mathscr{C}(-\infty) = \frac{1}{2}, \mathscr{S}(\infty) = -\mathscr{S}(-\infty) = \frac{1}{2}$, the two branches of the curve approach the points $P^+$and $P^-$with coordinates are $\left(\frac{1}{2}, \frac{1}{2}\right)$ and $\left(-\frac{1}{2}, -\frac{1}{2}\right)$ respectively. This curve is known as the Cornu spiral (see Fig. 8.37) and is useful in discussions of the general properties of Fresnel diffraction patterns.

### 8.7.3 Fresnel diffraction at a straight edge

We now consider Fresnel diffraction at a semi-infinite plane bounded by a sharp straight edge. This problem is of special interest with regard to the behaviour of the field near the boundaries of geometrical shadows. We restrict our attention to the case where the line $P_0P$ and also its projection (our $x$ axis) on to the half-plane is

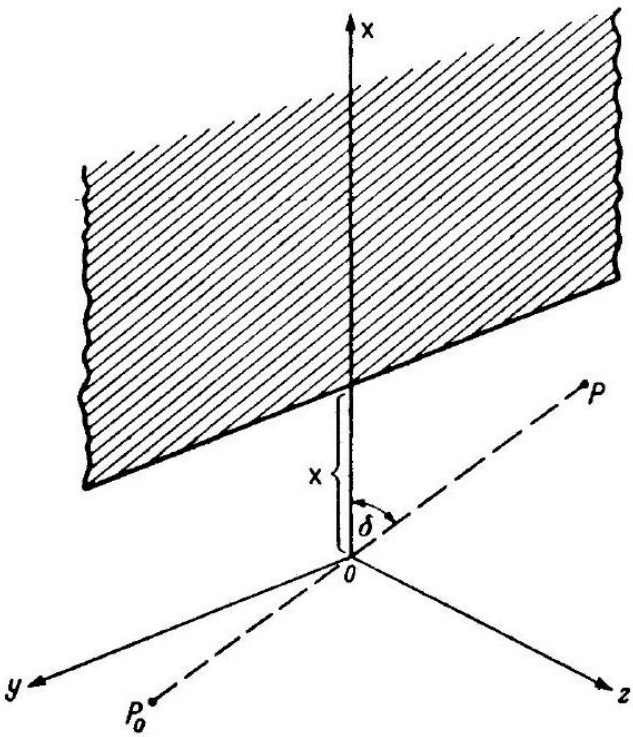



Fig. 8.38. Fresnel diffraction at a straight edge.

perpendicular to the edge (Fig. 8.38). If $x$ is the distance of the edge from the origin (which lies on the line $P_0P$ ), the integration extends throughout the region

$$-\infty < \xi < x, \quad -\infty < \eta < \infty,$$

or, in terms of $u$ and $v$,

$$-\infty < u < w, \quad -\infty < v < \infty, \tag{22}$$

where

$$w = \sqrt{\frac{2}{\lambda}\left(\frac{1}{r'} + \frac{1}{s'}\right)}\, x \cos\delta. \tag{23}$$

The point of observation $P$ lies in the illuminated region or in the geometrical shadow according as $x > 0$ or $x < 0$.

The diffraction integrals (9) become

$$\left.\begin{aligned} C &= b\int_{-\infty}^{w} du \int_{-\infty}^{+\infty} dv \left\{\cos\left(\tfrac{\pi}{2}u^2\right)\cos\left(\tfrac{\pi}{2}v^2\right) - \sin\left(\tfrac{\pi}{2}u^2\right)\sin\left(\tfrac{\pi}{2}v^2\right)\right\} \\ S &= b\int_{-\infty}^{w} du \int_{-\infty}^{\infty} dv \left\{\sin\left(\tfrac{\pi}{2}u^2\right)\cos\left(\tfrac{\pi}{2}v^2\right) + \cos\left(\tfrac{\pi}{2}u^2\right)\sin\left(\tfrac{\pi}{2}v^2\right)\right\} \end{aligned}\right\} \tag{24}$$

We have violated here a condition used in the derivation of (9), namely that the linear dimensions of the domain of integration shall be small compared to the distances $P_0O$ and $OP$. To justify the approximate validity of these formulae also in the present case, a more careful discussion of the error terms is necessary. We shall omit it here, since the diffraction by a half-plane will be treated again later by rigorous methods [§ 11.5].

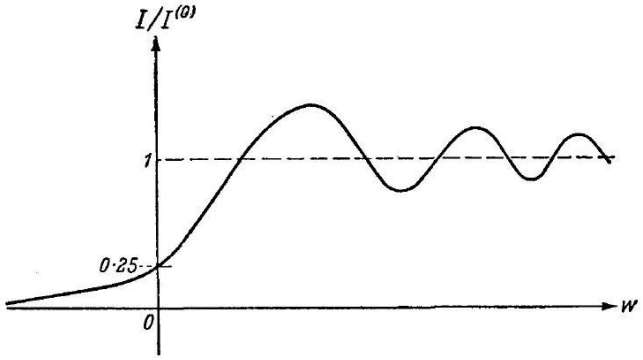



Fig. 8.39. Intensity distribution in the Fresnel diffraction pattern of a straight edge.

From the relations (18) and (19), we have

$$\left.\begin{array}{l}\int_{-\infty}^{w}\cos\left(\frac{\pi}{2}\tau^2\right)d\tau=\int_{-\infty}^{0}+\int_{0}^{w}=\mathscr{C}(\infty)+\mathscr{C}(w)=\frac{1}{2}+\mathscr{C}(w)\\ \int_{-\infty}^{+\infty}\cos\left(\frac{\pi}{2}\tau^2\right)d\tau=1\end{array}\right\}\tag{25}$$

and similarly

$$\left.\begin{array}{l}\int_{-\infty}^{w}\sin\left(\frac{\pi}{2}\tau^2\right)d\tau=\frac{1}{2}+\mathscr{S}(w),\\ \int_{-\infty}^{+\infty}\sin\left(\frac{\pi}{2}\tau^2\right)d\tau=1\end{array}\right\}\tag{26}$$

Hence (24) becomes

$$\left.\begin{array}{l}C=b\left\{\left[\frac{1}{2}+\mathscr{C}(w)\right]-\left[\frac{1}{2}+\mathscr{S}(w)\right]\right\},\\ S=b\left\{\left[\frac{1}{2}+\mathscr{C}(w)\right]+\left[\frac{1}{2}+\mathscr{S}(w)\right]\right\},\end{array}\right\}\tag{27}$$

and substitution into (4) gives finally an expression for the intensity:

$$I=\frac{1}{2}\left\{\left[\frac{1}{2}+\mathscr{C}(w)\right]^2+\left[\frac{1}{2}+\mathscr{S}(w)\right]^2\right\}I^{(0)},\tag{28}$$

where

$$I^{(0)}=4|B|^2b^2=\frac{|A|^2}{(r'+s')^2}\tag{29}$$

The behaviour of the intensity function (28) can be deduced from the Cornu spiral. The quantity $2I/I^{(0)}$ is seen to be equal to the square of the distance of the point $w$ of the Cornu Spiral from the "asymptotic point" $P^-\left(-\frac{1}{2},-\frac{1}{2}\right)$. Thus if the point of observation is in the illuminated region ( $w>0$ ), $I/I^{(0)}$ oscillates with diminishing amplitudes as the distance from the edge increases and approaches asymptotically the value unity, as may be expected on the basis of geometrical optics. The maximum value of the intensity is not at the edge of the geometrical shadow, but some distance away from it, in the directly illuminated region (Fig. 8.39). On the edge of the shadow
$(w=0), I/I^{(0)}=\frac{1}{4}$. In the shadow region, $I/I^{(0)}$ decreases monotonically towards zero. These predictions are found to be in good agreement with experimental results.

## 9.11.8 8.8. The Three-Dimensional Light Distribution near Focus

In § 8.3.3 it was shown that the light distribution in the focal plane of a well-corrected lens arises essentially from Fraunhofer diffraction on the aperture of the lens, and in § 8.5 the

Fraunhofer diffraction patterns for apertures of various forms were studied in detail. To obtain a fuller knowledge of the structure of an optical image, we must study the light distribution not only in the geometrical focal plane but also in the neighbourhood of this plane. The knowledge of the three-dimensional (Fresnel) distribution near focus is of particular importance in estimating the tolerance in the setting of the receiving plane in an image-forming system.

The properties of the out-of-focus monochromatic images of a point source by a circular aperture were first discussed in detail by E. Lommel in a classical memoir.* Starting from the Huygens-Fresnel integral, Lommel succeeded in expressing the complex disturbance in terms of convergent series of Bessel functions and also confirmed experimentally the phenomena predicted on the basis of these calculations. Almost at the same time as Lommel, H. Struve † published a similar though a less comprehensive analysis relating to the circular aperture. He did not work out the numerical consequences in such detail but gave useful approximations for the intensity near the edge of the geometrical shadow, where the series expansions are rather slowly convergent. Asymptotic approximations relating to points of observations at distances of many wavelengths from the focus were derived some years later by K. Sohwarzschild. ‡

The investigations of Lommel and Struve attracted relatively little attention, and in 1909 the problem was treated again by P. Debye, § whose discussion established certain general features of the diffracted field both near and far away from the focus. In more recent times the analysis of these authors was extended and diagrams were published which show in detail the structure of the field in this complex region; and the results were broadly confirmed by experiment, both with light and with microwaves (short radio waves).

In discussing the light distribution near focus, we shall take as our starting point the analysis of Lommel and Struve, but it will be convenient and instructive to begin from the integral representation of the field in the form employed by Debye.

### 8.8.1 Evaluation of the diffraction integral in terms of Lommel functions

Consider a spherical monochromatic wave emerging from a circular aperture and converging towards the axial focal point $O$. We shall consider the disturbance $U(P)$ at a typical point $P$ in the neighbourhood of $O$. The point $P$ will be specified by a position vector $\boldsymbol{R}$ relative to $O$, and it will be assumed that the distance $R = OP$ as well as the radius $a(\gg \lambda)$ of the aperture are small compared to the radius $f = CO$ of the wave-front $W$ that momentarily fills the aperture (Fig. 8.40).

If $s$ denotes the distance from the point of observation $P$ to a point $Q$ on $W$ and $A/f$ is the amplitude at $Q$ of the incident wave we have, by the application of the HuygensFresnel principle,

$$U(P) = -\frac{i}{\lambda}\frac{Ae^{-ikf}}{f}\iint_W \frac{e^{iks}}{s}dS, \tag{1}$$

where, since only small angles are involved, the variation of the inclination factor over the wave-front has been neglected. If $\boldsymbol{q}$ denotes the unit vector in the direction $OQ$, we have, to a good approximation

$$s - f = -\boldsymbol{q}\cdot\boldsymbol{R}. \tag{2}$$

0

- E. Lommel, Abh. Bayer. Akad., 15, Abth. 2, (1885), 233; a later paper (ibid. 15, Abth. 3, (1886), 531) deals with diffraction by a slit, by an opaque strip and by a straight edge.
  † H. Struve, Mm. de l'Acad. de St. Petersbourgh (7), 34 (1886), 1.
  ‡ K. Schwarzschild, Sitzb. Mnchen. Akad. Wiss., Math.-Phys. Kl., 28 (1898), 271.
  § P. Debye, Ann. d. Physik. (4), 30 (1909), 775.

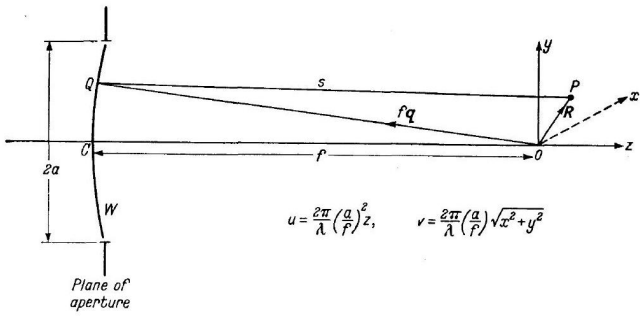



Fig. 8.40. Diffraction of a converging spherical wave at a circular aperture: Notation.

Also, the element $dS$ of the wave-front is given by

$$dS = f^2 d\Omega, \tag{3}$$

where $d\Omega$ is the element of the solid angle that $dS$ subtends at $O$. Moreover, we may replace $s$ by $f$ in the denominator of the integrand without introducing an appreciable error. Equation (1) then becomes

$$U(P) = -\frac{i}{\lambda} A \iint_{\Omega} e^{-ikq\cdot\boldsymbol{R}} d\Omega, \tag{4}$$

the integration now extending over the solid angle $\Omega$ which the aperture subtends at the focus. Equation (4) is the Debye integral and expresses the field as a superposition of plane waves of different directions of propagation (specified by the vectors $\boldsymbol{q}$ which fill $\Omega$ ).

Before discussing the evaluation of the Debye integral we note the interesting fact that, being a sum of elementary solutions (plane waves), it represents a rigorous solution of the wave equation which in the limit $f \to \infty$ (aperture at infinite distance) is valid throughout the whole of space. Of course (4) is not a rigorous solution of our original problem, since no account has been taken of the nature of the screen, the exact boundary conditions being approximated to by those of the Kirchhoff diffraction theory. The true solution to our problem would include not only contributions of plane waves that are propagated in the directions of the incident geometrical rays, but of waves propagated in all possible directions.* If, however, the above mentioned conditions are satisfied ( $f \gg a \gg \lambda, f \gg R$ ), only contributions of waves that are included in (4) are appreciable.

To evaluate (4) we first express the integrand in a more explicit form. We take Cartesian axes at $O$, with the $z$ direction along $Oz$. Let $(x, y, z)$ be the coordinates of $P$ and $(\xi, \eta, \zeta)$ those of $Q$. We set

$$\left.\begin{aligned} \xi &= a\rho\sin\theta, \quad x = r\sin\psi \\ \eta &= a\rho\cos\theta, \quad y = r\cos\psi \end{aligned}\right\} \tag{5}$$

Since $Q$ lies on the spherical wave-front $W$,

$$\zeta = -\sqrt{f^2 - a^2\rho^2} = -f\left[1 - \frac{1}{2}\frac{a^2\rho^2}{f^2} + \ldots\right] \tag{6}$$

Then

0

- This corresponds to the representation of the field in terms of a so-called angular spectrum of plane waves. Cf. § 11.4.2.

$$\begin{aligned}\boldsymbol{q}\cdot\boldsymbol{R} &= \frac{x\boldsymbol{\xi}+y\eta+z\zeta}{f}\\ &= \frac{a\rho r\cos(\theta-\psi)}{f} - z\left[1-\frac{1}{2}\frac{a^2\rho^2}{f^2}+\ldots\right].\end{aligned} \tag{7}$$

It is useful at this stage to introduce dimensionless variables $u$ and $v$, which together with $\psi$ specify the position of $P$ :

$$u=\frac{2\pi}{\lambda}\left(\frac{a}{f}\right)^2 z,\quad v=\frac{2\pi}{\lambda}\left(\frac{a}{f}\right)r=\frac{2\pi}{\lambda}\frac{a}{f}\sqrt{x^2+y^2} \tag{8}$$

We note that the point $P$ lies in the direct beam of light or in the geometrical shadow according as $|v/u|\leqq 1$.

From (7) and (8) it follows that if terms above the second power in $a\rho/f$ are neglected in comparison to unity,

$$k\boldsymbol{q}\cdot\boldsymbol{R}=v\rho\cos(\theta-\psi)-\left(\frac{f}{a}\right)^2 u+\frac{1}{2}u\rho^2 \tag{9}$$

Further, the element of the solid angle is

$$d\Omega=\frac{dS}{f^2}=\frac{a^2\rho d\rho d\theta}{f^2} \tag{10}$$

Hence (4) becomes

$$U(P)=-\frac{i}{\lambda}\frac{a^2A}{f^2}e^{i\left(\frac{f}{a}\right)^2 u}\int_0^1\int_0^{2\pi} e^{-i\left[v\rho\cos(\theta-\psi)+\frac{1}{2}u\rho^2\right]}\rho d\rho d\theta \tag{11}$$

The integral with respect to $\theta$ is the same as one that we encountered in connection with $\mathrm{F}_{\mathrm{radnhofer}}$ diffraction on a circular aperture (§ 8.5.2). It is equal to $2\pi J_0(v\rho)$, where $J_0(v\rho)$ is the Bessel function of zero order. Hence (11) becomes

$$U(P)=-\frac{2\pi i a^2A}{\lambda f^2}e^{i\left(\frac{f}{a}\right)^2 u}\int_0^1 J_0(v\rho)e^{-\frac{1}{2}iu\rho^2}\rho d\rho \tag{12}$$

It is convenient to consider separately the real and imaginery parts of the integral. We set

$$2\int_0^1 J_0(v\rho)e^{-\frac{1}{2}u\rho^2}\rho d\rho=C(u,v)-iS(u,v) \tag{13}$$

where

$$\left.\begin{aligned}C(u,v)&=2\textstyle\int_0^1 J_0(v\rho)\cos\left(\frac{1}{2}u\rho^2\right)\rho d\rho\\ S(u,v)&=2\textstyle\int_0^1 J_0(v\rho)\sin\left(\frac{1}{2}u\rho^2\right)\rho d\rho.\end{aligned}\right\} \tag{14}$$

These integrals may be evaluated in terms of the Lommel functions

$$\left.\begin{aligned}V_n(u,v)&=\textstyle\sum_{s=0}^{\infty}(-1)^s\left(\frac{u}{v}\right)^{n+2s}J_{n+2s}(v),\\ V_n(u,v)&=\textstyle\sum_{s=0}^{\infty}(-1)^s\left(\frac{v}{u}\right)^{n+2s}J_{n+2s}(v),\end{aligned}\right\} \tag{15}$$

introduced by Lommel for this purpose.* Using the relation § 8.5 (11)

$$\frac{d}{dx}\left[x^{n+1}J_{n+1}(x)\right]=x^{n+1}J_n(x)$$

$C(u,v)$ may be written as

$$C(u,v) = \frac{2}{v}\int_0^1 \frac{d}{d\rho}\left[\rho J_1(v\rho)\right]\cos\left(\frac{1}{2}u\rho^2\right)d\rho$$
$$= \frac{2}{v}\left[J_1(v)\cos\frac{1}{2}u + u\int_0^1 \rho^2 J_1(v\rho)\sin\left(\frac{1}{2}u\rho^2\right)d\rho\right] \tag{16}$$

on integrating by parts. Again using the relation § 8.5 (11), integrating by parts and continuing this process, we obtain

$$C(u,v) = \frac{\cos\frac{1}{2}u}{\frac{1}{2}u}\left[\left(\frac{u}{v}\right)J_1(v) - \left(\frac{u}{v}\right)^3 J_3(v) + \ldots\right]$$
$$+ \frac{\sin\frac{1}{2}u}{\frac{1}{2}u}\left[\left(\frac{u}{v}\right)^2 J_2(v) - \left(\frac{u}{v}\right)^4 J_4(v) + \ldots\right]$$
$$= \frac{\cos\frac{1}{2}u}{\frac{1}{2}u}U_1(u,v) + \frac{\sin\frac{1}{2}u}{\frac{1}{2}u}U_2(u,v) \tag{17a}$$

In a similar way we find

$$S(u,v) = \frac{\sin\frac{1}{2}u}{\frac{1}{2}u}U_1(u,v) - \frac{\cos\frac{1}{2}u}{\frac{1}{2}u}U_2(u,v) \tag{17b}$$

These formulae are valid at all points in the neighbourhood of the focus, but are only convenient for computations when $|u|v\ |< 1$, i.e. when the point of observation lies in the geometrical shadow. When $|u/v| > 1$, i.e. when the point of observation is in the illuminated region, it is more appropriate to use expansions involving positive

powers of $v/u$. These may be derived in a similar manner by integrating by parts with respect to the trigonometric term. The first step gives

$$C(u,v) = \frac{2}{u}\int_0^1 J_0(v\rho)\frac{d}{d\rho}\left[\sin\left(\frac{1}{2}u\rho^2\right)\right]d\rho$$
$$= \frac{2}{u}\left[J_0(v)\sin\frac{1}{2}u + v\int_0^1 J_1(v\rho)\sin\left(\frac{1}{2}u\rho^2\right)\rho d\rho\right] \tag{18}$$

where the relation § 8.5 (15)

$$\frac{d}{dx}\left[x^{-n}J_n(x)\right] = -x^{-n}J_{n+1}(x)$$

has been used. Integrating by parts again and using the last relation together with the well-known formula (which may be deduced from the series expansion for $J_n(x)$ )

$$\lim_{x\to 0}\frac{J_n(x)}{x^n} = \frac{1}{2^n n!}, \tag{19}$$

0

- For fuller discussions of these functions we refer to Lommex's memoirs (loc. cit.) and the following books: G. N. Watson, A Treatise on the Theory of Bessel Functions (Cambridge University Press, 1922), pp. 537-50; A. Gray, G. B. Mathews, and T. M. MacRobert, A Treatise on Bessel Functions (London, Macmillan, 2nd ed. 1922), Chapter XIV; and J. Walker, The Analytical Theory of Light (Cambridge University Press, 1904), p. 396.

we obtain

$$C(u,v) = \frac{\sin\frac{1}{2}u}{\frac{1}{2}u}\left[J_0(v) - \left(\frac{v}{u}\right)^2 J_2(v) + \ldots\right] - \frac{\cos\frac{1}{2}u}{\frac{1}{2}u}\left[\left(\frac{v}{u}\right)J_1(v) - \left(\frac{v}{u}\right)^3 J_3(v) + \ldots\right] + \frac{2}{u}\left[\frac{v^2}{2u} - \frac{1}{3!}\left(\frac{v^2}{2u}\right)^3 + \ldots\right].$$

The series in the first two lines are two of the Lommel $V_n$ functions and the series in the third line will be recognized as the expansion of $\sin(v^2/2u)$. Hence

$$C(u,v) = \frac{2}{u}\sin\frac{v^2}{2u} + \frac{\sin\frac{1}{2}u}{\frac{1}{2}u}V_0(u,v) - \frac{\cos\frac{1}{2}u}{\frac{1}{2}u}V_1(u,v). \tag{20a}$$

In a similar way we obtain for the other integral the expression

$$S(u,v) = \frac{2}{u}\cos\frac{v^2}{2u} - \frac{\cos\frac{1}{2}u}{\frac{1}{2}u}V_0(u,v) - \frac{\sin\frac{1}{2}u}{\frac{1}{2}u}V_1(u,v). \tag{20b}$$

This completes the formal solution of our problem. We shall now discuss some of the implications of these formulae.

### 8.8.2 The distribution of intensity

According to (12), (13), (17), and (20), the intensity $I = |U|^2$ in the neighbourhood of focus is given by the two equivalent expressions

$$I(u,v) = \left(\frac{2}{u}\right)^2 \left[U_1^2(u,v) + U_2^2(u,v)\right] I_0, \tag{21a}$$

and

$$I(u,v) = \left(\frac{2}{u}\right)^2 \left[1 + V_0^2(u,v) + V_1^2(u,v) - 2V_0(u,v)\cos\left\{\frac{1}{2}\left(u + \frac{v^2}{u}\right)\right\} - 2V_1(u,v)\sin\left\{\frac{1}{2}\left(u + \frac{v^2}{u}\right)\right\}\right] I_0, \tag{21b}$$

where

$$I_0 = \left(\frac{\pi a^2 |A|}{\lambda f^2}\right)^2 \tag{22}$$

is the intensity at the geometrical focus $u = v = 0$.

It follows from (15) that $U_1(-u,v) = -U_1(u,v), U_2(-u,v) = U_2(u,v), V_0(-u,v) = V_0(u,v), V_1(-u,v) = -V_0(u,v)$. Accordingly $I(u,v)$ remains unchanged when $u$ is replaced by $-u$. Hence, in the neighbourhood of the focus the intensity distribution is

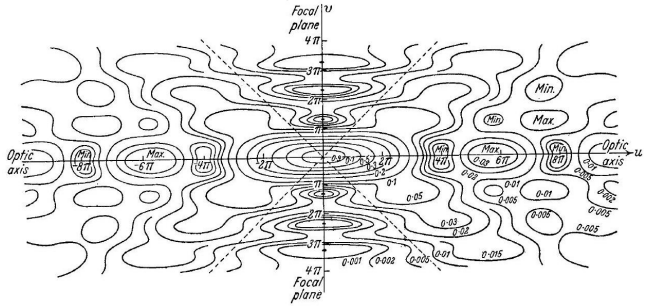



Fig. 8.41. Isophotes [contour lines of the intensity $I(u, v)$ ] in a meridional plane near focus of a converging spherical wave diffracted at a circular aperture. The intensity is normalized to unity at focus. The dotted lines represent the boundary of the geometrical shadow. When the figure is rotated about the $u$-axis, the minima on the $v$-axis generate the Airy dark rings. (Adapted from E. H. Linfoot and E. Wolf, Proc. Phys. Soc., B, 69 (1956), 823.)

symmetrical about the geometrical focal plane. Naturally, the distribution is also symmetrical about the axis $v = 0$.

From the formulae (21), Lommel computed the intensity distribution in a number of selected receiving planes near focus, and verified experimentally some of the predictions.* The lines of equal intensity (called isophotes) near focus, constructed from Lommel's data are shown in Fig. 8.41. †

Of particular interest is the tubular structure of the bright central portion of the diffraction image seen clearly in the figure and postulated on experimental grounds already in 1894 by Taylor.* It is this structure that is responsible for the tolerance in the setting of a receiving plane in an image-forming system.

We shall now consider several special cases of interest.

(a) Intensity in the geometrical focal plane. For points in the geometrical focal plane $u = 0$ and (21a) reduces to

$$I(0, v) = 4 \lim_{u \to 0} \left[ \frac{U_1^2(u, v) + U_2^2(u, v)}{u^2} \right] I_0 \tag{23}$$

From the defining equation for the $U_n$ functions it follows that

$$\lim_{u \to 0} \left[ \frac{U_1(u, v)}{u} \right] = \frac{J_1(v)}{v}, \quad \lim_{u \to 0} \left[ \frac{U_2(u, v)}{u} \right] = 0 \tag{24}$$

0

- Related experiments were also described by M. E. Hufford and H. T. Davis, Phys. Rev., 33 (1926), 589 and C. A. Taylor and B. J. Thompson, J. Opt. Soc. Amer., 48 (1958), 844. Similar experiments with microwaves were carried out by M. P. Bachynski and G. Bekefi, J. Opt. Soc. Amer., 47 (1957), 428, for the case of circular aperture, and by P. A. Mathews and A. L. Cullen, Proc. Inst. Elect. Engrs., Pt. C., 103 (1956), 449, for rectangular aperture.

  † A similar, but less detailed diagram was published by M. Berek, Z. Phys., 40 (1926), 421. It is reproduced in some books with errors (incorrect position of the geometrical shadow and interchange of axes). Another version of the diagram, substantially in agreement with Fig. 8.41, was given by F. Zernike and B. R. A. Nijboer in their contribution to La Thorie des Images Optiques, Paris, Revue d'Optique (1949), 227. It is based on a different, but equivalent expansion (§ 9.4, eq. (12)) of the diffraction integral. Corresponding diagrams, calculated on the basis of electromagnetic theory, showing contours of the electric energy density and of the energy flow, were published by A. Borvin and E. Wolf, Phys. Rev., 138 (1965), B 1561 and A. Boivin, J. Dow and E. Wolf, J. Opt. Soc. Amer., 57 (1967), 1171.

Similar diagrams for the annular aperture were published by E. H. Linfoot and E. Wolf, Proc. Phys. Soc., B, 66 (1953), 145. Some extensions of Lommel's analysis to diffraction by concentric arrays of ring-shaped apertures were discussed by A. Boivin, J. Opt. Soc. Amer., 42 (1952), 60.

so that

$$I(0,v) = \left[\frac{2J_1(v)}{v}\right]^2 I_0 \tag{25}$$

We thus obtain the Airy formula § 8.5 (14) for Fraunhofer diffraction at a circular aperture, as was to be expected.
(b) Intensity along the axis. For points on the axis, $v = 0$, and the two $V_n$ functions entering the expression (21b) reduce to

$$V_0(u,0) = 1, \quad V_1(u,0) = 0$$

Hence

$$\begin{aligned} I(u,0) &= \frac{4}{u^2}\left[2 - 2\cos\frac{1}{2}u\right] I_0 \\ &= \left(\frac{\sin u/4}{u/4}\right)^2 I_0 \end{aligned} \tag{26}$$

Thus the intensity along the axis is characterized by the function $(\sin x/x)^2$ which we discussed in §8.5.1 in connection with Fraunhofer diffraction at a rectangular aperture. The first zero of intensity on the axis is given by $u/4 \equiv \pi a^2 z/2\lambda f^2 = \pm\pi$, i.e. it is at a distance $z = \pm 2f^2\lambda/a^2$ from the focus.

It is usual to regard a loss of about 20 per cent in intensity at the centre of the image patch as permissible. Since $\left(\frac{\sin u/4}{u/4}\right)^2$ decreases by this amount when the receiving plane is displaced from the central position ( $u = 0$ ) to $u \sim 3 \cdot 2$, it follows that the focal tolerance $\Delta z$ is approximately

$$\Delta z = \pm 3 \cdot 2\frac{\lambda}{2\pi}\left(\frac{f}{a}\right)^2 \sim \pm\frac{1}{2}\left(\frac{f}{a}\right)^2 \lambda \tag{27}$$

With an $f/10$ pencil for example ( $f/a = 20$ ), and with light of wavelength $\lambda = 5.10^{-5}$ cm, the focal tolerance is about $\pm 0 \cdot 5 \cdot 20^2 \cdot 5 \cdot 10^{-5}$ cm $= \pm 0 \cdot 1$ mm.
(c) Intensity along the boundary of the geometrical shadow. For points on the boundary of the geometrical shadow $u = \pm v$. Since the distribution is symmetrical with

respect to the geometrical focal plane we may, without loss of generality, take $u = +v$. The $U_n$ functions then reduce to

$$U_1(u,u) = \sum_{s=0}^{\infty}(-1)^s J_{2s+1}(u), \quad U_2(u,u) = \sum_{s=0}^{\infty}(-1)^s J_{2s+2}(u) \tag{28}$$

We recall the well-known identities of Jacobi.*

$$\left.\begin{aligned} \cos(u\cos\theta) &= J_0(u) + 2\textstyle\sum_{s=1}^{\infty}(-1)^s J_{2s}(u)\cos 2s\theta \\ \sin(u\cos\theta) &= 2\textstyle\sum_{s=0}^{\infty}(-1)^s J_{2s+1}(u)\cos(2s+1)\theta \end{aligned}\right\} \tag{29}$$

0

• H. D. Taylor, Mon. Not. Roy. Astr. Soc., 54 (1894), 67.

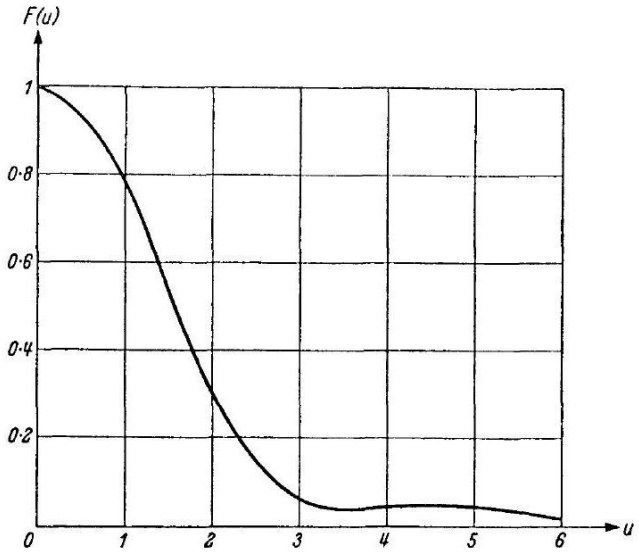



Fig. 8.42. The variation of intensity along the boundary of the geometrical shadow.

$$\text{The function } F(u) = \frac{1 - 2J_0(u)\cos u + {J_0}^2(u)}{u^2}.$$

Setting $\theta = 0$, it follows on comparison with (28), that

$$\left.\begin{aligned} U_1(u,u) &= \tfrac{1}{2}\sin u \\ U_2(u,u) &= \tfrac{1}{2}\left[J_0(u) - \cos u\right], \end{aligned}\right\} \tag{30}$$

and (21a) reduces to

$$I(u,u) = \frac{1 - 2J_0(u)\cos u + J_0^2(u)}{u^2} I_0. \tag{31}$$

This function is shown in Fig. 8.42.

### 8.8.3 The integrated intensity

It is also desirable to determine the fraction $L$ of the (time averaged) total energy that falls within a small circle of prescribed radius $r_0$ about the axial point in the receiving plane $u =$ constant. If

$$E = \pi a^2 \left(\frac{|A|}{f}\right)^2 \tag{32}$$

denotes the total energy incident on to the aperture in unit time, the required fraction of energy is given by

$$\begin{aligned} L\,(u, v_0) &= \frac{1}{E}\int_0^{r_0}\int_0^{2\pi} I(u,v) r dr d\psi \\ &= \frac{1}{2I_0}\int_0^{v_0} I(u,v) v dv \end{aligned} \tag{33}$$

0

- See for example G. N. Watson, Theory of the Bessel Functions (Cambridge University Press, 1922) p. 22.

where

$$v_0 = \frac{2\pi}{\lambda}\left(\frac{a}{f}\right) r_0. \tag{34}$$

If Lommel's expressions (21) for the intensity are substituted into (33), the integral may be developed into series involving Bessel functions. The derivation is lengthy and we shall therefore only give the final results, due to Wolf.*

Again two formally different expressions are obtained, one of which is convenient for computations when the boundary of the small circle is in the geometrical shadow, the other when it is in the direct beam of light. Suppressing the suffix zero, i.e. writing $v$ in place of $v_0$, one has in the first case ( $|v/u| \geqslant 1$ )

$$L(u,v) = 1 - \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1}\left(\frac{u}{v}\right)^{2s} Q_{2s}(v) \tag{35a}$$

where

$$Q_{2s}(v) = \sum_{p=0}^{2s}(-1)^p \left[J_p(v)J_{2s-p}(v) + J_{p+1}(v)J_{2s+1-p}(v)\right] \tag{36}$$

In the second case ($|v/u| \leqslant 1$)

$$\begin{aligned} L(u,v) = {} & \left(\frac{v}{u}\right)^2 \left[1 - \sum_{s=0}^{\infty} \frac{(-1)^s}{2s+1}\left(\frac{v}{u}\right)^{2s} Q_{2s}(v)\right] \\ & - \frac{4}{u}\left[Y_1(u,v)\cos\frac{1}{2}\left(u+\frac{v^2}{u}\right) + Y_2(u,v)\sin\frac{1}{2}\left(u+\frac{v^2}{u}\right)\right], \end{aligned} \tag{35b}$$

where the $Q$ 's are again given by (36) and $Y_1$ and $Y_2$ are two of the functions

$$\begin{aligned} Y_n(u,v) &= \sum_{s=0}^{\infty}(-1)^s(n+2s)\left(\frac{v}{u}\right)^{n+2s} J_{n+2s}(v) \\ &= \frac{1}{2}\left[\frac{v^2}{u}V_{n-1}(u,v) + uV_{n+1}(u,v)\right]. \end{aligned} \tag{37}$$

In Fig. 8.43, the contour lines of $L(u,v)$, computed from these formulae are displayed. They may be regarded as analogous in a certain sense to the rays of geometrical optics.

We note that in the special case when the receiving plane coincides with the focal plane ( $u = 0$ ), ( $35a$ ) reduces to

$$\begin{aligned} L(0,v) &= 1 - Q_0(v) \\ &= 1 - J_0^2(v) - J_1^2(v), \end{aligned} \tag{38}$$

in agreement with Rayleigh's formula § 8.5 (18).
Of special interest is the case when the circle over which the integral (33) is

0

- E. Wolf, Proc. Roy. Soc., A, 204 (1951), 533. Asymptotic approximations for $L(u,v)$ were derived by J. Focke, Optica Acta, 3 (1956), 161.

extended coincides with the cross-section of the geometrical cone of rays. Then $|v|u\ |= 1$ and the series in (35a) and (35b) can be summed, and give*

$$L(u,u) = 1 - J_0(u)\cos u - J_1(u)\sin u. \tag{39}$$

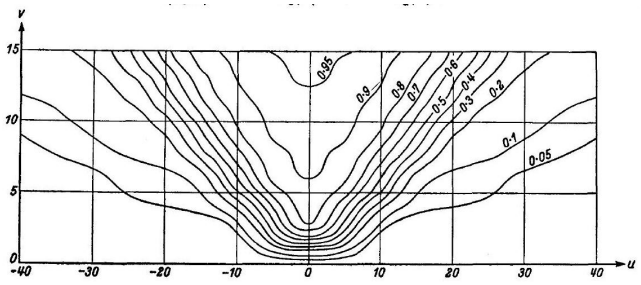



Fig. 8.43. Contour lines of $L(u,v)$, giving the fraction of the total energy which falls within small circles centred on the axis in selected receiving planes $u$ = constant.

(After E. Wolf, Proc. Roy. Soc., A, 204 (1951), 542.)

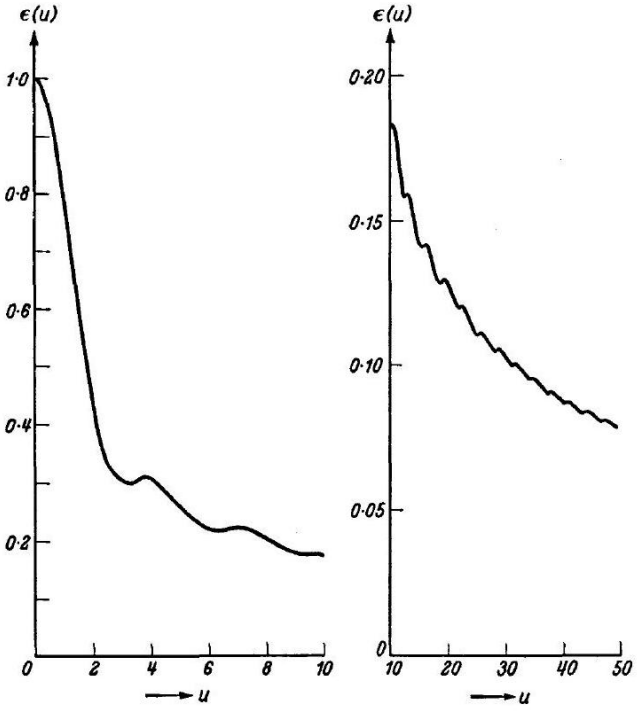



Fig. 8.44. The fraction $\varepsilon(u)$ of the total energy within the geometrical shadow. (After E. Wolf, Proc. Roy. Soc., A, 204 (1951), 544.)

Hence the expression

$$\varepsilon(u) = J_0(u)\cos u + J_1(u)\sin u \tag{40}$$

gives, for the receiving plane $u$ = constant, the fraction of the total energy in the geometrical shadow. The function $\varepsilon(u)$ is plotted in Fig. 8.44; it is not monotonically

0

• E. Wolf, loc. cit., p. 539.

decreasing but has maxima (apart from $u = 0$ ) when $J_1(u) = 0$ and minima when $\sin u = 0 (u \neq 0)$.

### 8.8.4 The phase behaviour

Finally we consider the behaviour of the phase of the disturbance in the neighbourhood of the focus. According to (12) and (13) the phase is, apart from the additive term ( $-\omega t$ ) given by*

$$\phi(u, v) = \left(\frac{f}{a}\right)^2 u - \chi(u, v) - \frac{\pi}{2}, \quad (\mathrm{mod} 2\pi), \tag{41}$$

where

$$\cos\chi = \frac{C}{\sqrt{C^2 + S^2}}, \quad \sin\chi = \frac{S}{\sqrt{C^2 + S^2}}, \tag{42}$$

the positive square root being taken in (42).
We note that, unlike the intensity distribution, the phase distribution cannot be expressed in terms of $u$ and $v$ alone, but has a structure which depends on the angular aperture of the geometrical pencil of rays. It also follows that each "branch" of the multivalued function $\phi(u, v)$ is continuous in $u$ and $v$ at all points where the intensity does not vanish, and that at points of zero intensity it is indeterminate. At the focus $u = v = 0$, one of its values is $-\pi/2$.

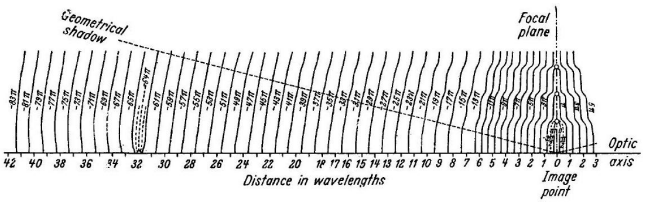



Fig. 8.45. Profiles of the co-phasal surfaces $\phi(u, v)$ = constant near the geometrical focus, calculated with $\lambda=5.10\hat{}\{-5\}$ \mathrm{˜cm}, a=2.5 \mathrm{˜cm}, f=10 \mathrm{˜cm}$.
(After G. W. Farnell, Canad. J. Phys., 35 (1957), 780.) †

The co-phasal surfaces (surfaces $\phi$ = constant) are, of course, surfaces of revolution about the $u$-axis. We shall now show that they possess a further symmetry, expressed by the relation

$$\phi(-u, v) + \phi(u, v) = -\pi \quad (\mathrm{mod} 2\pi). \tag{43}$$

From (14),

$$C(-u, v) = C(u, v), \quad S(-u, v) = -S(u, v). \tag{44}$$

From (42) it then follows that

$$\cos\chi(-u, v) = \cos\chi(u, v), \quad \sin\chi(-u, v) = -\sin\chi(u, v), \tag{45}$$

0

- The symbol mod$2\pi$ on the right of an equation denotes that the two sides of the equation are indeterminate to the extent of an additive constant $2m\pi$ where $m$ is any integer.
† The values associated with the surfaces of Fig. 8.45 differ by $\pi$ from those of Farnell's paper. This is so because, for the sake of consistency with our analysis, the phase at the focus is taken as $-\pi/2$in

so that

$$\chi(-u, v) = -\chi(u, v) \tag{46}$$

The relation (43) now follows from (46) and (41). Hence the reflection in the plane $u = 0$ of any co-phasal surface $\phi = \phi_0$ is a co-phasal surface $\phi = -\pi - \phi_0$.

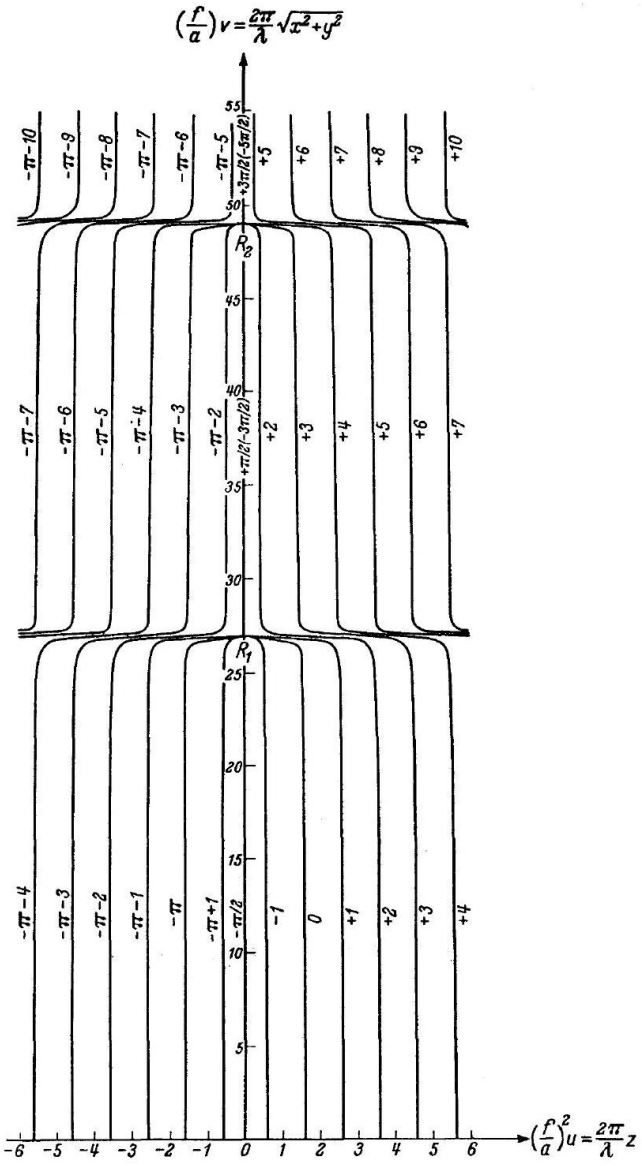



Fig. 8.46. Profiles of co-phasal surfaces in the immediate neighbourhood of the geometrical focal plane of an $f/3.5$ homocentric pencil. $O R_{1}$ and $OR_2$ are the radii of the first and second Airy dark rings.

(After E. H. Linfoot and E. Wolf, Proc. Phys. Soc., B, 69 (1956), 827.)

In Fig. 8.45 the profiles of the co-phasal surfaces of an $f/2$ homocentric pencil are shown. Far away from the focus these surfaces coincide with the spherical wavefronts of geometrical optics, but they become more and more deformed as the focal region is approached. The profiles in the immediate neighbourhood of the geometrical focal plane are shown in Fig. 8.46. It is seen that close to the focus, the surfaces are substantially plane; they are, however, spaced

accordance with (51), whereas in the paper by Farnell it is taken as $+\pi/2$.

only $1 - a^2/4f^2$ times as far apart as they would be in a parallel beam of light of the same wavelength. Moreover (see Fig. 8.41), the intensity is not uniform over each co-phasal surface. In the immediate neighbourhood of the Atry dark rings (denoted by $R_1$ and $R_2$ in Fig. 8.46) the co-phasal surfaces show a more complicated behaviour. Proceeding from the geometrical focus in the $v$ direction, the phase is seen to be constant between any two successive dark rings, but changes abruptly by $\pi$ on crossing each ring. That this must be so may be seen from the expression (derived easily with the help of (24), (41) and (46)) for the complex disturbance in the geometrical focal plane:

$$U(0,v) = \frac{1}{i}\frac{2J_1(v)}{v}\sqrt{I_0}. \tag{47}$$

Since this expression is purely imaginary for all values of $v$, the phase of $U(0,v)$ can only be equal to $-\pi/2$ or $+\pi/2(\mathrm{mod}2\pi)$, and since it changes sign on crossing each dark ring, the phase must then undergo a sudden jump by $\pi$. A discontinuity of this amount also occurs on crossing each axial point of zero intensity.

It is also of interest to study how the phase of the field changes as the point of observation moves along each ray through the focus. It is convenient to compare this variation with that of a spherical wave converging to the focus in the halfspace $z < 0$ and diverging from it in the half-space $z > 0$. If the phase $\tilde{\phi}$ of this comparison wave is taken as zero at the focus we have

$$\left.\begin{array}{lll} \tilde{\phi}(u,v) = & -kR \quad \text{when} \quad u \leqslant 0 \\ & +kR \quad \text{when} \quad u \geqslant 0 \end{array}\right\} \tag{48}$$

where as before $R = \sqrt{x^2+y^2+z^2} > 0$ is the distance of the point of observation from the focus. The difference

$$\delta(u,v) = \phi(u,v) - \tilde{\phi}(u,v) \tag{49}$$

is called the phase anomaly.
From (43), (48), and (49) it follows that

$$\delta(-u,v) + \delta(u,v) = -\pi \pmod{2\pi} \tag{50}$$

whilst at the focus itself we have

$$\delta(0,0) = \phi(0,0) = -\frac{\pi}{2} \pmod{2\pi} \tag{51}$$

The behaviour of the phase anomaly along selected rays through the focus of an $f/3.5$ homocentric pencil is shown in Fig. 8.47.

The figures show that on passing through the focus along any ray except the axial one, $\delta$ undergoes a rapid but continuous change in phase of $\pi$. This effect was observed
a long time ago by Gouy* and has been the subject of many investigations. † Along the axis however the phase anomaly has a singular behaviour: it fluctuates periodically between the values 0 and $-\pi$.

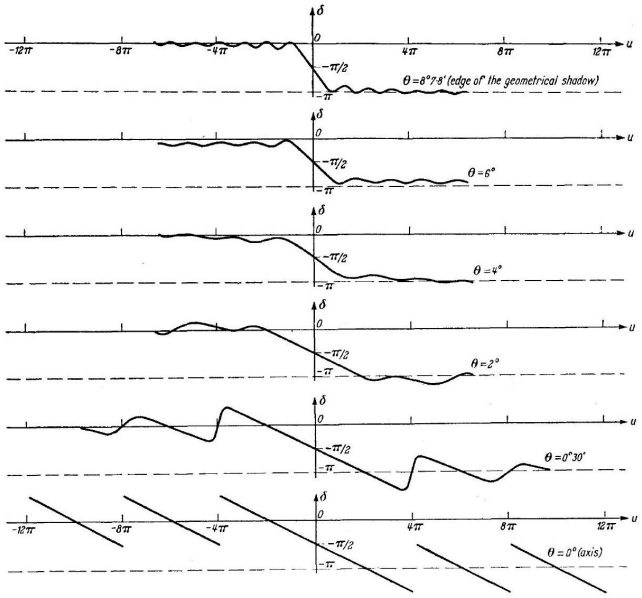



Fig. 8.47. Phase anomaly $\delta$ along geometrical rays through the focus of a homocentric $f/3.5$ pencil. The angle $\theta$ denotes the inclination of the ray to the axis.
(After E. H. Linfoot and E. Wolf, Proc. Phys. Soc., B, 69 (1956), 827.)

It can be shown by considering the asymptotic approximation to the HuygensFresnel integral for large values of $k$ (small wavelengths), that as light advances along a ray the phase changes suddenly by $\pi/2$ on passing through either of the two principal centres of curvature of the associated wave-fronts. ‡ The case which we have just considered corresponds to the special situation when the two centres of curvature coincide. Thus the change by half period is associated even with the geometrical

optics solution.* Fig. 8.47 shows how the discontinuity of geometrical optics goes over into a continuous transition when the finiteness of the wavelength is taken into account. Finally a reference may be made to a paper by Farnell, † which describes an experimental investigation into the structure of the phase distribution in the focal region of a microwave lens; very good agreement with the predictions of the theory was found.

## 9.11.9 8.9. The Boundary Diffraction Wave

If the edge of a diffracting aperture or obstacle is observed from points within the geometrical shadow, it appears luminous. This result was already known to Thomas Young ‡ who attempted, prior to Fresnel, to explain diffraction on a wave-theoretical basis. Young believed that the incident light undergoes a kind of reflection at the edge of the diffracting body, and he regarded the diffraction pattern as arising from the interference of the incident wave and the reflected "boundary wave". However, Young's views were expressed in a qualitative manner only and did not gain much recognition.

0

- L. G. Gouy, Compt. Rend. Acad. Sci. Paris, 110 (1890), 1251; Ann Chim. (Phys.) (6), 24 (1891), 145.
† For a survey of the literature see F. Reiche, Ann. d. Physik, (4), 29 (1909), 56, ibid., 401. References to more recent investigations are given in E. H. Linfoot and E. Wolf, Proc. Phys. Soc., B, 69 (1956), 827.
‡ H. Poincar, Thorie mathmatique de la lumire, II (Paris, Georges Carr, (1892)), pp. 168-74. See also J. Walker, The Analytical Theory of Light (Cambridge University Press, 1904), pp. 91-93.

That Young's theory contained an element of truth became evident after SommerFELD in 1894 obtained a rigorous solution for the diffraction of plane waves by a plane, semi-infinite reflecting screen (see § 11.5). This solution shows that in the geometrical shadow the light is propagated in the form of a cylindrical wave that appears to proceed from the edge of the screen, whilst in the illuminated region it is represented as superposition of the cylindrical wave and of the original incident wave.

The question arises whether also under more general conditions diffraction can be accounted for as the combined effect of an incident wave and a boundary wave. This problem had been investigated before the appearance of Sommerfeld's paper by Maggi, § but his results appear to have been forgotten. It was later investigated independently and much more fully by Rubinowicz.|| The Maggi-Rubinowicz theory was developed further by Miyamoto and Wolf. I

Consider a monochromatic light wave from a point source $P_0$ propagated through

an aperture in a plane opaque screen. As before we assume that the linear dimensions of the aperture are large compared with the wavelength but small compared to the distance of $P_0$ and of the point of observation $P$ from the screen, and that the angles of incidence and of diffraction are small. Then we have in the approximations of the Kirchhoff theory (§ 8.3.2)

$$U(P) = \frac{1}{4\pi} \iint_{\mathscr{A}} \left\{ \frac{e^{ikr}}{r} \frac{\partial}{\partial n} \left( \frac{e^{iks}}{s} \right) - \left( \frac{e^{iks}}{s} \right) \frac{\partial}{\partial n} \left( \frac{e^{ikr}}{r} \right) \right\} dS \tag{1}$$

where $\mathscr{A}$ denotes the diffracting aperture and the other symbols have the same meaning as before. We construct a closed surface bounded by (1) the opening $\mathscr{A}$, (2) the surface of a truncated cone $\mathscr{B}$ whose vertex is at $P_0$ and whose generators pass through the edge of the aperture, and (3) a portion $\mathscr{C}$ of a large sphere centred

0

- For discussions of this point see also A. Rubrnowicz, Phys. Rev., 54 (1938), 931; and C. J. Bouwkamp, Physica, 7 (1940), 485.

† G. W. Farnell, Canad. J. Phys., 36 (1958), 935.

‡ Thomas Young, Phil. Trans. Roy. Soc., 20 (1802), 26.

§ G. A. Maggi, Annali di Matem. (2), 16 (1888), 21. Maggi's analysis is also discussed in a paper by F. Kottler, Ann. d. Physik, (4), 70 (1923), 413; and in B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 1950, 2nd ed.), p. 74.

Experimental evidence for the "existence" of the boundary wave was found by W. Wien, Inaug. Diss., Berlin, 1886; E. Maey, Ann. d. Physik. (9), 49 (1893), 69; and A. Kalaschnikow, Journ. Russ. Phys. Chem. Ges., 44 (1912), Phys. Teil, 133. See also S. Banerji, Phil. Mag. (6), 37 (1919), 112; and S. K. Mitra, ibid. (6), 38 (1919), 289.

|| A. Rubinowicz, Ann. d. Physik. (4), 53 (1917), 257; ibid. (4), 73 (1924); ibid., 81 (1926), 153; Acta Phys. Polonica, 12 (1953), 225. See also G. N. Ramachandran, Proc. Indian Acad. Sci. A., 21 (1945), 165; L. C. Martin, Proc. Phys. Soc., 55 (1943), 104; ibid., 62B (1949), 713. Y. V. Kathavate, ibid. A., 21 (1945), 177; R. S. Ingarden, Acta Phys. Polon., 14 (1955), 77; O. Laporte and J. Meixner, Zeitschr. f. Phys., 153 (1958), 129.

A very comprehensive account of researches relating to the boundary wave is given in A. Rubinowicz' book Die Beugungswelle in der Kirchhoffschen Theorie der Beugung (Warszawa, Polska Akademia Nauk, 1957).

IT K. Miyamoto and. E. Wolf, J. Opt. Soc. Amer., 52 (1962) 615, 626; K. Miyamoto, Proc. Phys. Soc., 79 (1962), 617. See also E. W. Marchand and E. Wolf, J. Opt. Soc. Amer., 52 (1962), 761; A. Rubinowicz, ibid., 52 (1962) 717; Acta Phys. Polonica 21 (1962), 61,451; Progress in Optics, Vol. 4, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 199.

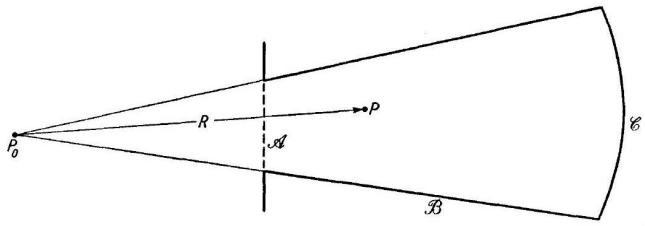



Fig. 8.48. Derivation of the boundary diffraction wave.

on $P$ (Fig. 8.48). If $R$ denotes the distance from $P_0$ to $P$, we have rigorously from Kirchhoff's integral theorem,

$$\frac{1}{4\pi}\iint_{\mathscr{A}+\mathscr{B}+\mathscr{C}}\left\{\frac{e^{ikr}}{r}\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right)-\frac{e^{iks}}{s}\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right\}dS=\frac{e^{ikR}}{R}\text{ or }0 \tag{2}$$

according to the point $P$ lying inside or outside the surface. Now in the same way as in §8.3.2, the contribution from $\mathscr{C}$ can be made negligible by taking the radius of the sphere sufficiently large. We then obtain from (1) and (2),

$$U(P)=U^{(g)}(P)+U^{(d)}(P) \tag{3}$$

where

$$\left.\begin{array}{rl} U^{(g)}(P) & =\frac{e^{ikR}}{R} \\ & =0 \\ \text{when } P \text{ is in the direct beam} & \\ & =0 \text{ is in the geometrical shadow} \end{array}\right\} \tag{4}$$

and

$$U^{(d)}(P)=-\frac{1}{4\pi}\iint_{\mathscr{B}}\left\{\frac{e^{ikr}}{r}\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right)-\left(\frac{e^{iks}}{s}\right)\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\right\}dS \tag{5}$$

$U^{(g)}$ represents the disturbance as predicted by geometrical optics, so that $U^{(d)}$ must represent the effect of diffraction. We shall now show that $U^{(d)}$ may be transformed into a line integral along the edge of the aperture.

We note first that the spheres $r=$ constant cut orthogonally the truncated cone $\mathscr{B}$. Hence on $\mathscr{B}$,

$$\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)=0. \tag{6}$$

Also

$$\frac{d}{dn}\left(\frac{e^{iks}}{s}\right)=\frac{d}{ds}\left(\frac{e^{iks}}{s}\right)\cos(n,s)=\left(\frac{ik}{s}-\frac{1}{s^2}\right)e^{iks}\cos(n,s). \tag{7}$$

Hence (5) reduces to

$$U^{(d)}(P)=-\frac{1}{4\pi}\iint_{\mathscr{S}}\frac{e^{ik(r+s)}}{rs}\left(ik-\frac{1}{s}\right)\cos(n,s)dS. \tag{8}$$

We can take as the element $dS$ the area $ABB'A'$ bounded by segments of two neighbouring generators and by the arcs of circles in which the spheres $r=$ constant

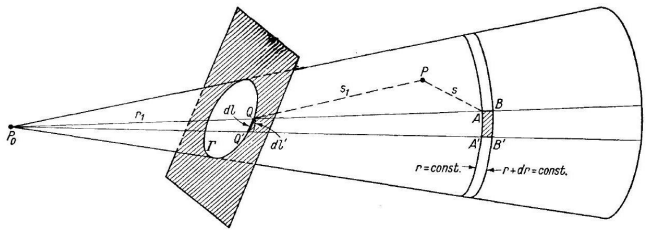



Fig. 8.49. Derivation of the boundary diffraction wave.

$$P_0A = r, \quad P_0Q = r_1, \quad PA = s, \quad PQ = s_1.$$

and $r + dr =$ constant intersect the cone (Fig. 8.49). If $d\phi$ is the angle between the two generators, then

$$dS = rdrd\phi. \tag{9}$$

Let $Q$ and $Q'$ be the points of intersection of the two generators with the edge $\Gamma$ of the aperture and let $dl$ be the length of the element of $\Gamma$ between these two points. If $dl'$ denotes the corresponding element of arc of the circle in which the sphere with radius $r_1 = P_0Q$ intersects the cone then

$$dl' = r_1 d\phi = dl \cos(dl, dl') = dl \sin(r_1, dl). \tag{10}$$

From (9) and (10)

$$dS = \frac{r}{r_1} \sin(r_1, dl)\, drdl \tag{11}$$

Also, by projection on the normal to the cone at $A$ and $Q$ (the normals at these points being parallel),

$$s \cos(n, s) = s_1 \cos(n, s_1). \tag{12}$$

On substituting from (11) and (12) into (8) it follows that

$$\begin{aligned} U^{(d)}(P) &= -\frac{1}{4\pi} \iint_{\mathscr{S}} \frac{e^{ik(r+s)}}{rs} \left(ik - \frac{1}{s}\right) \frac{s_1}{s} \cos(n, s_1) \frac{r}{r_1} \sin(r_1, dl)\, drdl \\ &= -\frac{1}{4\pi} \int_\Gamma dl \frac{s_1}{r_1} \cos(n, s_1) \sin(r_1, dl) \int_{r_1}^{\infty} e^{ik(r+s)} \left(\frac{ik}{s^2} - \frac{1}{s^3}\right) dr \end{aligned} \tag{13}$$

Next we shall show that the integrand of the second integral in (13) is a total differential, namely that

$$e^{ik(r+s)} \left[\frac{ik}{s^2} - \frac{1}{s^3}\right] = \frac{d}{dr} \left\{ \frac{e^{ik(r+s)}}{s\,[s + r - r_1 + s_1 \cos(s_1, r_1)]} \right\}. \tag{14}$$

We have, on carrying out the differentiation on the right of (14),

$$\frac{d}{dr}\left\{\frac{e^{ik(r+s)}}{s\left[s+r-r_1+s_1\cos\left(s_1,r_1\right)\right]}\right\}$$
$$=\frac{e^{ik(r+s)}}{s\left[s+r-r_1+s_1\cos\left(s_1,r_1\right)\right]}\left\{ik\left(1+\frac{ds}{dr}\right)-\frac{1}{s}\frac{ds}{dr}\right.$$
$$\left.-\frac{1}{\left(s+r-r_1+s_1\cos\left(s_1,r_1\right)\right.}\cdot\left[1+\frac{ds}{dr}\right]\right\} \tag{15}$$

Now from the triangle $APQ$

$$s^2=s_1^2+\left(r-r_1\right)^2+2s_1\left(r-r_1\right)\cos\left(s_1,r_1\right), \tag{16}$$

whence, differentiating with respect to $r$ whilst keeping $r_1$ and $s_1$ fixed,

$$s\frac{ds}{dr}=r-r_1+s_1\cos\left(s_1,r_1\right) \tag{17}$$

On substituting from this equation into (15), the identity (14) follows. Hence

$$\int_{r_1}^{\infty}e^{ik(r+s)}\left(\frac{ik}{s^2}-\frac{1}{s^3}\right)dr=\left[\frac{e^{ik(r+s)}}{s\left(s+r-r_1+s_1\cos\left(s_1,r_1\right)\right.}\right]_r^{\infty}$$
$$=-\frac{e^{ik(r_1+s_1)}}{s_1^2\left[1+\cos\left(s_1,r_1\right)\right]} \tag{18}$$

and (13) finally becomes

$$U^{(d)}(P)=\frac{1}{4\pi}\oint_{\Gamma}\frac{e^{ik(r_1+s_1)}}{r_1s_1}\cdot\frac{\cos\left(n,s_1\right)}{\left[1+\cos\left(s_1,r_1\right)\right]}\sin\left(r_1,dl\right)dl \tag{19}$$

This formula, together with (3) and (4), is the Rubinowicz representation of the Kirchhoff diffraction integral and may be regarded as the mathematical formulation of Young's theory. It expresses the effect of diffraction in terms of the incident wave that is propagated in accordance with the laws of geometrical optics and a boundary diffraction wave that may be thought of as arising from the scattering of the incident radiation by the boundary of the aperture.

Since $U$ is a continuous function of position, it follows from (4) that the boundary wave $U^{(d)}$ is discontinuous across the edge of the geometrical shadow so as to compensate for the discontinuity in the "geometrical wave" $U^{(g)}$. The discontinuity in $U^{(d)}$ arises from the factor $[1+\cos\left(s_1,r\right)]$ in the denominator in (19).

By the same argument which was given in connexion with the classification of diffraction phenomena on p. 386 and which illustrates the physical significance of the principle of the stationary phase [see Appendix III], it follows that only those points of the domain of integration contribute substantially to $U^{(d)}$ for which the phase of the integrand is stationary, i.e. for which

$$\frac{d}{dl}\left[k\left(r_1+s_1\right)\right]=0. \tag{20}$$

This relation may also be written in the form of a "reflection law",

$$\cos\left(r_1,dl\right)=-\cos\left(s_1dl\right) \tag{21}$$

## 9.11.10 8.10. Gabor's Method of Imaging by Reconstructed Wavefronts (Holography)

In an attempt to improve the resolving power of the electron microscope, Gabor* has proposed a two-step method of optical imagery. In the first step an object is illuminated with a coherent electron wave or a coherent light wave. The object is assumed to be such that a considerable part of the wave penetrates undisturbed through it. † A diffraction pattern, called hologram, which is formed by the interference of the secondary wave arising from the presence of the object with the strong background wave, is recorded on a photographic plate. If the plate, suitably processed, is replaced in the original position and is illuminated by the background wave alone, the wave that is transmitted by the plate contains information about the original object, and this can be extracted from the photograph by optical processes. In order to "reconstruct" the object from this "substitute" wave, it is only necessary to send it through a suitable image-forming system, and an image will appear in the plane conjugate to the plane in which the object was situated. We shall only be concerned with the optical principle involved ‡.

### 8.10.1 Producing the positive hologram

Consider a monochromatic wave from a small source $S$, impinging on a semi-transparent object $\sigma$ (Fig. 8.50(a)). Let $\mathscr{H}$ be a screen some distance behind the object and let $U = Ae^{i\varphi}$ represent the complex disturbance at a typical point of $\mathscr{H}$, $A$ being the (real) amplitude, and $\psi$ the phase of the disturbance. We may regard $U$ as the sum of two terms,

$$U = U^{(i)} + U^{(s)} = e^{i\psi_i}\left[A^{(i)} + A^{(s)}e^{i(\psi_s - \psi_i)}\right]. \tag{1}$$

Here $U^{(i)} = A^{(i)}e^{i\varphi_i}$ denotes the incident wave (or coherent background); it is the field which would be produced at $\mathscr{H}$ in the absence of the object. The other term, $U^{(s)} = A^{(s)}e^{iw_s}$ represents the secondary, or diffracted, wave and it is this wave that contains information about the object. In terms of the amplitudes and phases, the amplitude of the total disturbance $U$ may according to (1) be written as

$$A = \sqrt{UU^\star} = \sqrt{A^{(i)2} + A^{(s)2} + 2A^{(i)}A^{(s)}\cos(\psi_s - \psi_i)}. \tag{2}$$

As usual, we have suppressed the time harmonic factor $e^{-i\omega t}$, and thus have implicitly assumed that the secondary wave issuing from the object is of the same frequency as the incident wave. An object of this type is called a Rayleigh scatterer; to a good approximation, practically all non-fluorescent objects are of this type.

---

0

- D. Gabor, Nature, 161 (1948), 777; Proc. Roy. Soc., A, 197 (1949), 454; Proc. Phys. Soc., B, 64 (1951), 449.
  † This means that, if the wave is expressed as the sum of the incident wave and a diffracted secondary wave, the scattering of the secondary wave is neglected. This neglection represents what is usually known as Born's first approximation, and finds many applications in the theory of scattering of X-rays and electrons. For the limits of validity of this approximation in electron microscopy and interferometry see D. Gabor, Rev. Mod. Phys., 28 (1956), 260.
  ‡ For a detailed treatment of this subject and for an account of various modifications and applications of this technique, see for example, J. B. DeVelis and G. O. Reynolds, Theory and Applications of Holography (Reading, Mass., Addison-Wesley Publishing Company, 1967) and H. M. Smith, Principles of Holography (New York, J. Wiley and Sons, 1969).

Suppose now that a photographic plate is placed in the $\mathscr{H}$ plane. Let $\alpha$ be the transmission factor of the plate, defined by analogy with the transmission function $F$ of § 8.6.1 as the ratio of the complex amplitude of the wave emerging from the plate

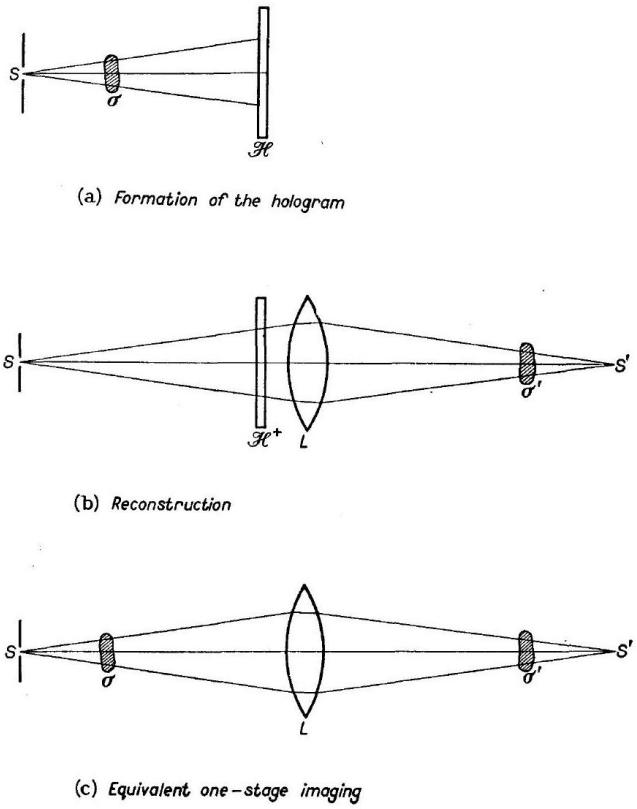



Fig. 8.50. Illustrating Gabor's method of imaging by reconstructed wave-fronts.

to that of the wave incident on the plate. The corresponding transmission factor for the intensity is $\tau = \alpha\alpha^\star$; and the quantity

$$D = -\log_{10}\tau = -\log\alpha\alpha^\star \tag{3}$$

is called the density of the plate. The product $E$ of the intensity $I = A^2$ of light that reaches the plate and the time $t$ of exposure,

$$E = It, \tag{4}$$

is called simply exposure (or light sum) and the curve giving $D$ against $\log_{10} E$ is known as the Hurter-Driffield curve; a typical form of this curve is shown in Fig. 8.51. In the range between the points $P$ and $Q$ the curve is practically a straight line, and if $\Gamma$ is its slope, the density of the negative is evidently given by

$$D := D_0 + \Gamma\log_{10}\frac{E}{E_0}, \tag{5}$$

where $D_0$ and $E_0$ are constants. Using (3) it follows that

$$\tau = \tau_0\left(\frac{E}{E_0}\right)^{-\Gamma}. \tag{6}$$

For pure absorption, without phase change, $\alpha$ is a real number, the square root $\sqrt{\tau}$ of the intensity transmission. Thus, in this case, the amplitude transmission factor $\alpha_n$ of the "negative hologram" is given by a relation of the form

$$\alpha_n = (K_n A)^{-\Gamma_n}\,, \tag{7}$$

where $K_n$ is proportional to the square root of the time of the exposure.
Suppose now that we take a positive print of the negative hologram. The amplitude transmission factor $\alpha_p$ of the positive is

$$\alpha_p = \left[K_p\,(K_n A)^{-\Gamma_n}\right]^{-\Gamma_p} = K A^{\Gamma}, \tag{8}$$

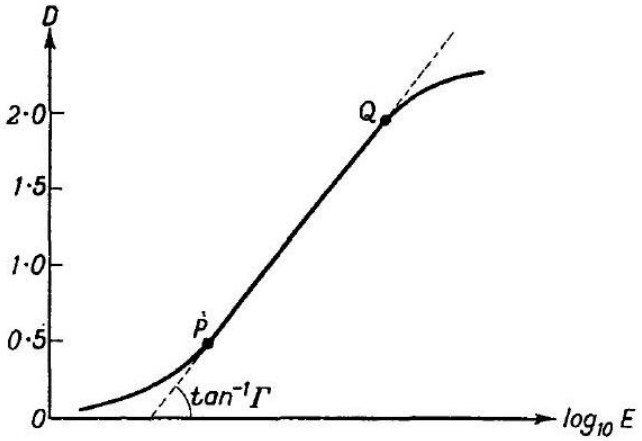



Fig. 8.51. The Hurter-Driffield curve (photographic response).

where $\Gamma = \Gamma_n\Gamma_p$ is the "overall gamma" of the negative-positive process and $K = K_p{}^{-\Gamma_p} K_n{}^{\Gamma}$.

### 8.10.2 The reconstruction

In the reconstruction process (Fig. 8.50 (b)) the positive hologram ( $\mathscr{H}^+$), whose amplitude transmission factor is given by (8), is illuminated by the coherent background $U^{(i)}$ alone. The background is obtained simply by removing the object, otherwise preserving the geometry of the original arrangement. A substitute wave $U'$ is transmitted by the plate, and this, according to ( 2 ) and ( 8 ), is represented by

$$U' = \alpha_p U^{(i)} = K A^{(i)} e^{i\psi_i} \left[A^{(i)^2} + A^{(s)^2} + 2A^{(i)}A^{(s)}\cos\left(\psi_s - \psi_i\right)\right]^{\frac{1}{2}\Gamma}. \tag{9}$$

If we chose $\Gamma = 2$, then

$$U' = K A^{(i)^2} e^{i\psi_i}\left[A^{(i)} + \frac{A^{(s)^2}}{A^{(i)}} + A^{(s)} e^{i(\psi_s - \psi_i)} + A^{(s)} e^{-i(\psi_s - \psi_i)}\right]. \tag{10}$$

On comparing (10) and (1) it is seen that if $A^{(i)}$ is constant, i.e. if the background is uniform, the substitute wave $U'$ contains a component, called the reconstructed wave, proportional to $U$ (the first and third term in (10)). The remainder of (10) consists of two terms. One has the same phase as the background and an amplitude $A^{(s)^i}/A^{(i)}$ times that of the background. This term can be made very small by making the

background sufficiently strong.* The other term has the same amplitude $\left(KA^{(i)^2}A^{(s)}\right)$ as the reconstructed wave, but has a phase shift of opposite sign relative to the background. We say that it represents a conjugate wave and we shall show that it may be regarded as being due to a fictitious object of a similar nature as the true object, but situated in a different plane.

To show this we return for the moment to the arrangement of Fig. 8.50(a) and denote by $O_1$ any point in the object $\sigma$, and by $P$ any point in the $\mathscr{H}$-plane (Fig. 8.52). When the object is illuminated from the point source $S$, the point $P$ may be assumed to receive light along two rays, namely along the direct ray $SP$ (associated with the coherent background) and along the "diffracted secondary ray" $O_1P$ (corresponding to the secondary wave). Suppose first that there is no change of phase on diffraction at $O_1$. Then the diffracted light at $P$ is delayed with respect to the direct light by the path difference $O_1P - O_1A$, where $A$ is the point on the intersection with $SO_1$ of the

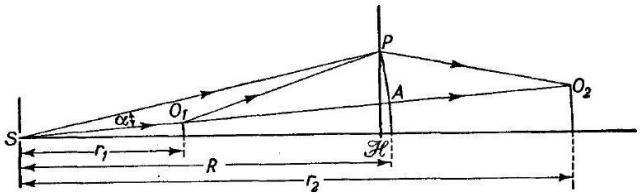



Fig. 8.52. Illustrating the position of the conjugate object.

sphere centered on $S$ and passing through $P$. Now the conjugate wave is advanced relative to the direct wave by the same amount as the diffracted secondary wave is retarded with respect to it, so that the conjugate wave catches up with the direct wave at that point $O_2$ on the line $SO_1$ where

$$O_1P - O_1A = PO_2 - AO_2 \tag{11}$$

If $r_1, R$, and $r_2$ are the distances of the points $O_1, A$, and $O_2$ from $S$, and $\alpha$ is the angle between $SP$ and the line $SO_1AO_2$ (see Fig. 8.52), equation (11) may be written as

$$\sqrt{(R\cos\alpha - r_1)^2 + R^2\sin^2\alpha} - (R - r_1) = \sqrt{(r_2 - R\cos\alpha)^2 + R^2\sin^2\alpha} - (r_2 - R)\,, \tag{12}$$

Expanding both sides in powers of $\alpha$ and retaining leading terms only, it follows that

$$\frac{1}{r_1} + \frac{1}{r_2} = \frac{2}{R}. \tag{13}$$

If the hologram were curved to a sphere of radius $R$, the relation (13) would hold with respect to any point on it, and it follows that the conjugate wave could then be regarded as being due to a fictitious object which is the image of the original object in a spherical mirror, of radius $R$ and centre $S$ (cf. §4.4 (16)). This result evidently also holds, as a good approximation, in the case of a plane hologram, provided that the off-axis angles of the rays are small enough. If, moreover, $r_1/R$ is small compared to unity, then according to (13) $r_2 \sim -r_1$, so that the conjugate object and the true object are situated symmetrically about the source $S$.

It has been assumed so far that there is no phase change on diffraction by the object. If there is a change of phase, the preceding considerations regarding the position of the conjugate object still apply, provided that the conjugate object is assumed to

0



Hologram

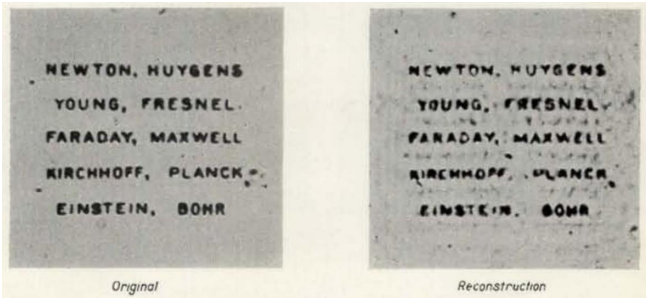



Fig. 8.53. Imaging by reconstructed wave-fronts.
The object was a microphotograph of 1.5 mm diam., illuminated with light of wavelength $\lambda = 4358$ A through a pinhole of diameter \$0 \cdot 2 \mathrm{˜mm}\$, reduced by a microscope objective to $5\mu$ nominal diameter at 50 mm from the object. Geometrical magnification 12. Effective aperture of lens used in reconstruction was $0 \cdot 025$. Noisy background chiefly due to imperfections of illuminating objective.
(After D. Gabor, Proc. Roy. Soc., A, 197 (1949), 454.)

- Strong coherent background is also utilized in a method due to F. Zernike (cf. Proc. Phys. Soc., 61 (1948), 158) for the display of weak secondary interference fringes.

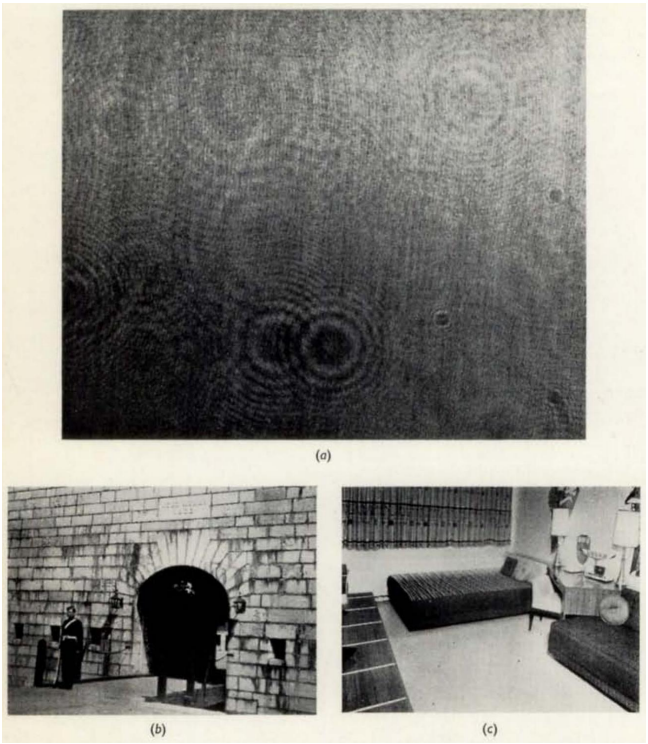



Fig. 8.54. Imaging by reconstructed wavefronts. Reconstruction of two objects from a single hologram.

Hologram (a) and reconstructions (b) and (c) of two transparencies which were illuminated with diffused coherent light. The transparencies were placed 14 and 24 in . from the hologram recording plate in such positions that neither obscured the other when viewed from the position where the hologram was recorded. The observable structures seen in the hologram are mainly diffraction patterns produced by dust particles.
(After E. N. Leith and J. Upatneks, J. Opt. Soc. Amer., 54 (1964), 1297.)

In Fig. 10.6 are shown photographs of the patterns observed with such an arrangement, for various separations $d$. The corresponding theoretical curves, computed from the formula (39) are also shown. The chain lines represent the envelopes

$$\left.\begin{aligned} I_{\max}(\phi, d) &= 2\left(\frac{2J_1(u)}{u}\right)^2 \left\{1 + \left|\frac{2J_1(v)}{v}\right|\right\}, \\ I_{\min}(\phi, d) &= 2\left(\frac{2J_1(u)}{u}\right)^2 \left\{1 - \left|\frac{2J_1(v)}{v}\right|\right\}. \end{aligned}\right\} \tag{41}$$

It is of interest to note that when $\beta = \pi$ (cases (D) and (E)), the intensity at the centre of each pattern has a relative minimum, not a maximum, in agreement with our general considerations. The variation of the degree of coherence with the separation

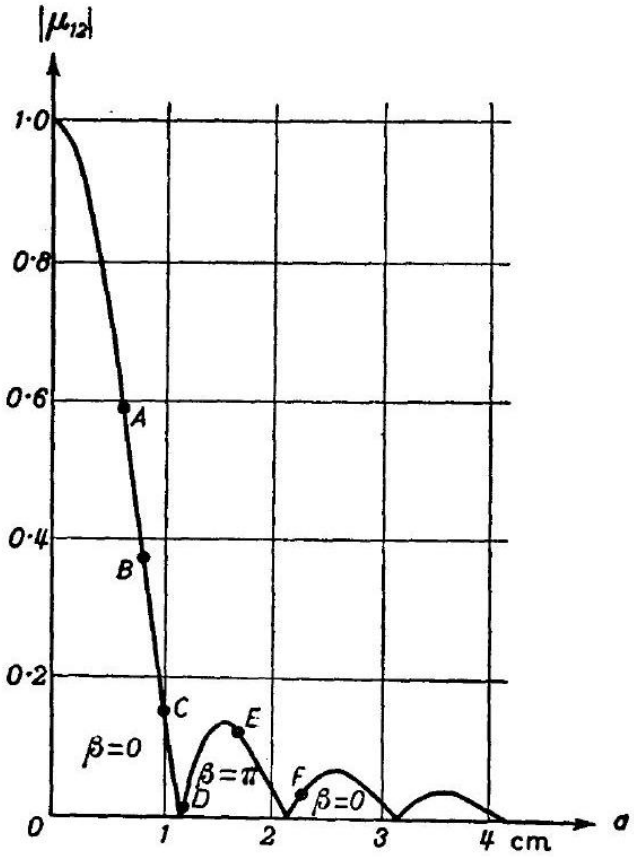



Fig. 10.7. Two-beam interference with partially coherent light. The degree of coherence as function of the separation $d$ of the two illuminated apertures in the Diffractometer. ( $\rho_{1}=0.45 \times 10\hat{}\{-2\} \mathrm{~cm}$, R=152 $\mathrm{~cm}$, $\bar{\lambda}=5790$ Å; incoherent illumination of $\sigma_1$ assumed.)

of the two apertures is shown in Fig. 10.7, where also the six values corresponding to the photographs of Fig. 10.6 are indicated by the appropriate letters.

## 9.12 The Diffraction Theory of Aberrations

In Chapter V we studied the effects of aberrations on the basis of geometrical optics. In that treatment the image was identified with the blurred figure formed by the points of intersection of the geometrical rays with the image plane. Since geometrical optics gives an approximate model valid in the limit of very short wavelengths, it is to be expected that the geometrical theory gradually loses its validity as the aberrations become small. For example, in the limiting case of a perfectly spherical convergent wave issuing from a circular aperture, geometrical optics predicts for the focal plane an infinite intensity at the focus and zero intensity elsewhere, whereas, as has been shown in §8.5.2, the real image consists of a bright central area surrounded by dark and bright rings (the Arry pattern). In the neighbourhood of the focal plane the light distribution has also been seen to be of a much more complicated nature (cf. Fig. 8.41) than geometrical optics suggests. We are thus led to the study of the effects of aberrations on the basis of diffraction theory.

The first investigations in this field are due to Rayleigh.* His main contribution was the formulation of a criterion (discussed in §9.3) which, in an extended form, has come to be widely used for determining the maximum amounts of aberrations that may be tolerated in optical instruments. The subject was carried further by the researches of many writers who

investigated the effects of various aberrations, † and we may mention, in particular, the more extensive treatments by Steward, Picht, and Born. ‡

A very extensive diffraction treatment of image formation in the presence of aberrations is due to Nisoer, § carried out partly in collaboration with Zernike. It is concerned with influence of small aberrations, when the departures of the wave-frouts from spherical form are only a fraction of the wavelength. The effects of large aberrations were studied on the basis of diffraction theory by van Kampen ‖ with the help of asymptotic approximations; this treatment is based on a formal extension to

functions of two variables of the principle of stationary phase, which has since been formulated rigorously, first by J. Focke (see Appendix III, p. 752).

In the main part of this chapter we shall give an account of the Number-Zernike theory and examine the structure of diffraction images affected by primary aberrations. In the final section (§ 9.5) we shall go over from point objects to extended objects and investigate imaging with coherent and incoherent illumination. Imaging with partially coherent illumination will be considered in Chapter X.

## 9.12.1 9.1. The Diffraction Integral in the Presence of Aberrations

### 9.1.1 The diffraction integral

Consider a centred optical system with a point source of monochromatic light $P_0$ (Fig. 9.1). We take a Cartesian system of axes, with origin at the Gaussian image

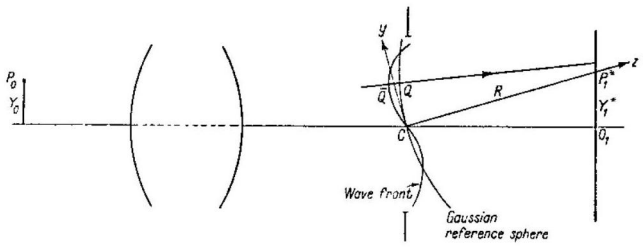



Fig. 9.1. Choice of the reference system and notation.

0

- Lord Rayleigh, Phil. Mag., (5), 8 (1879), 403. Reprinted in his Scientific Papers (Cambridge University Press, 1899), 1, 428.

† A historical survey of diffraction theory of aberrations was given by E. Worf in Rep. Progr. Phys. (London, Physical Society), 14 (1951), 95.

: G. C. Steward, Phil. Trans. Roy. Soc., A, 225 (1925), 131; also his book The Symmetrical Optical System (Cambridge University Press, 1928). J. Picht, Ann. d. Physik, (4), 77 (1925), 685. Ilid., 80 (1926), 491; also his Optische Abbildung (Braunschweig, Vieweg, 1931). M. Born, Naturwissenschaften, 20 (1932), 921; and his Optik (Berlin, Springer, 1933), p. 202.

§ B. R. A. Nuroeer, Thesis, University of Groningen, 1942. The main part was also published in Physica, 10 (1943), 679; ibid., 13 (1947), 605; F. Zernike and B. R. A. Nijboer, contribution to La Thorie des Images Optiques (Paris, Revue d'Optique, 1949), p. 227. An extension of the theory to somewhat larger aberrations was discussed by K. Nienhuis and B. R. A. Nijвoen, Physica, 14 (1948), 590. Mention must also be made of a thesis by K. Nienuus (University of Groningen, 1948), which is mainly concerned with the experimental study of the effects of aberrations. Some of the beautiful photographs obtained by Nienhurs aro reproduced in §9.4.

Experimental investigations at microwave frequencies, concerning the structure of the image region in the presence of aberrations were described by M. P. Bachynske and G. Bekeft in the paper referred to on p. 440, and also in Trans. Inst. Radio Eng., AP-4 (1956), 412.

in. G. van Kampen, Physica, 14 (1949), 575; ibid., 16 (1950), 817; ibid. 25 (1958), 437.

point $P_1^\star$ of $P_0$, with the $z$-axis along $CP_1^\star$, where $C$ is the centre of the exit pupil. The $y$-axis is taken in the meridional plane (the plane containing $P_0$ and the axis of the system). The off-axis distances of $P_0$ and $P_1^\star$ will be denoted by $Y_0$ and $Y_1^\star$ respectively.

As in Chapter $V$ the deformation of the wave-fronts in the region of the exit pupil will be described by the aberration function $\Phi$. Let $\bar{Q}$ and $Q$ be the points in which a ray in the image space intersects the wave-front through $C$ and the Gaussian reference sphere respectively. Assuming that the refractive index of the image space is unity, $\Phi$ (taken as positive in Fig. 9.1) represents the distance $\bar{Q}Q$ measured along the ray.

Let $R$ denote the radius $CP_1^\star$, of the Gaussian reference sphere and let $s$ be the distance between $Q$ and an arbitrary point $P$ in the region of the image. The disturbance at $Q$ is represented by $Ae^{ik(\Phi-R)}/R$, where $A/R$ is the amplitude at $Q$. According to the Huygens-Fresnel principle the disturbance at $P$ is given by

$$U(P) = -\frac{i}{\lambda}\frac{Ae^{-ikR}}{R}\iint \frac{e^{ik[\Phi+s]}}{s}dS \tag{1}$$

where the integration extends over the portion of the reference sphere that approximately fills the exit pupil. In (1) we have assumed that the angles involved are small, so that the variation of the inclination factor over the reference sphere may be neglected; we have also assumed that the amplitude of the wave is substantially constant over the wave-front, so that $A$ can be taken outside the integral.

Let $(\xi, \eta, \zeta)$ be the coordinates of $Q$ and $(x, y, z)$ the coordinates of $P$, and let $a$ be the radius of the exit pupil. As in § 8.8, which was concerned with the special case of an aberration-free wave ( $\Phi \equiv 0$ ), we set

$$\left.\begin{aligned} \xi &= a\rho\sin\theta, \quad x = r\sin\psi, \\ \eta &= a\rho\cos\theta, \quad y = r\cos\psi, \end{aligned}\right\} \tag{2}$$

and we have, as in § 8.8 (2) and § 8.8 (9)*

$$k(s-R) = -v\rho\cos(0-\psi) - \frac{1}{2}u\rho^2 + \left(\frac{R}{a}\right)^2 u \tag{3}$$

where $u$ and $v$ are the two "optical coordinates" of $P$,

$$u = \frac{2\pi}{\lambda}\left(\frac{a}{R}\right)^2 z, \quad v = \frac{2\pi}{\lambda}\left(\frac{a}{R}\right)\sqrt{x^2+y^2} \tag{4}$$

It will now be convenient to regard $\Phi$ as function of $Y_1^\star, \rho$, and $\theta$,

$$\Phi = \Phi\left(Y_1^\star, \rho, \theta\right) \tag{5}$$

The element of the Gaussian reference sphere is $dS = a^2\rho d\rho d\theta$, and if the angle which $CP_1^\star$ makes with the axis of the system is small, the range of integration may be taken as $0 \leqslant \rho \leqslant 1, 0 \leqslant \theta < 2\pi$. Moreover, for points of observations in the region of the image, $s$ may be replaced by $R$ in the denominator of the integrand. Hence (1) becomes, on substitution from (3),

$$U(P) = U(u, v, \psi) = -\frac{i}{\lambda}\frac{Aa^2}{R^2}e^{i\left(\frac{R}{a}\right)^3 u}\int_0^1\int_0^{2\pi} e^{i\left[k\oplus\left(Y_1^\star,\rho,\theta\right) - v\rho\cos(\theta-\varphi) - \frac{1}{4}u\rho^2\right]}\rho d\rho d\theta \tag{6}$$

so that the intensity at $P$ is

$$I(P) = |U(P)|^2 = \left(\frac{A\alpha^2}{\lambda R^2}\right)^2\left|\int_0^1\int_0^{2\pi} e^{i\left[k\Phi\left(Y_1^*,\rho,\theta\right) - v\rho\cos(\theta-\varphi) - \psi\rho^2\right]}\rho d\rho d\theta\right|^2 \tag{7}$$

It is convenient to express the intensity $I(P)$ as fraction of the intensity $I^\star$ which would be obtained at the Gaussian image point $P_1^\star$ if no aberrations were present. According to (7),

$$I^\star = \pi^2 \left(\frac{Aa^2}{\lambda R^2}\right)^2 \tag{8}$$

so that the normalized intensity is †

$$i(P) = \frac{I(P)}{I^\star} = \frac{1}{\pi^2}\left|\int_0^1 \int_0^{2\pi} e^{i\left[k\Phi\left(Y_1^\star,\rho_1,\theta\right) - v\rho\cos(\theta-\varphi) - \frac{1}{2}u\rho^2\right]}\rho d\rho d\theta\right|^2 \tag{9}$$

In the absence of aberrations the intensity is a maximum at the Gaussian image point. When aberrations are present, this will in general no longer be the case, and we may call the point of maximum intensity the diffraction focus. ‡ Often one is

- $R$ now corresponds to $f$ of § 8.8. It is of interest to note that if equations (2) and (3) of § 8.8 are used, the diffraction integral may again be expressed in the form of an angular spectrum of plane waves (cf. J. Focke, Optica Acta, 3 (1956), 110).
  i No confusion should arise between the symbol $i$ used for the normalized intensity, and the ane aymbol used for $\sqrt{-1}$, since the former always occurs with arguments, e.g. $i(P), i(u,v,\psi)$, ore.
  : In general there may be, of course, more than one diffraction focus, but the diffraction focus unique, if the aberrations are sufficiently small.
  interested only in the maximum intensity in a particular plane of observation; this value [when normalized as in (9)] will be called the Strehl intensity.*

From (9) some simple results, which will be needed later, may immediately be deduced.

### 9.1.2. The displacement theorem. Change of reference sphere

Let $\Phi$ and $\Phi'$ be two aberration functions such that

$$\Phi' = \Phi + H\rho^2 + K\rho\sin\theta + L\rho\cos\theta + M \tag{10}$$

where $H, K, L$, and $M$ are constants of order $\lambda$. Further let $i(u,v,\psi)$ and $i'(u,v,\psi)$ be the corresponding normalized intensities. Then by (9)

$$i(u,v,\psi) = \frac{1}{\pi^2}\left|\int_0^1 \int_0^{2\pi} e^{if(u,v,\psi;\rho,\theta)}\rho d\rho d\theta\right|^2 \tag{11}$$

where

$$f(u,v,\psi;\rho,\theta) = k\Phi - v\rho\cos(\theta-\psi) - \frac{1}{2}u\rho^2, \tag{12}$$

with a similar expression for $i'$. Now according to (10) the last expression may also be written in the form

$$\begin{aligned} f(u,v,\psi;\rho,\theta) &= k\Phi' - k\left[H\rho^2 + K\rho\sin\theta + L\rho\cos\theta + M\right] - v\rho\cos(\theta-\psi) - \frac{1}{2}u\rho^2 \\ &= k\Phi' - v'\rho\cos\left(\theta-\psi'\right) - \frac{1}{2}u'\rho^2 - kM \\ &= f'\left(u',v',\psi';\rho,\theta\right) - kM \end{aligned} \tag{13}$$

where

$$u' = u + 2kH, \quad v' \sin\psi' = v \sin\psi + kK, \quad v' \cos\psi' = v \cos\psi + kL \tag{14}$$

According to (2) and (4), equations (14) represent the transformation

$$z' = z + 2\left(\frac{R}{a}\right)^2 H, \quad x' = x + \left(\frac{R}{a}\right) K, \quad y' = y + \left(\frac{R}{a}\right) L. \tag{15}$$

From (11) and (13) it follows that

$$i(u, v, \psi) = i'(u', v', \psi') \tag{16}$$

We have thus established the following displacement theorem: The addition to an aberration function of a term $H\rho^2 + K\rho\sin\theta + L\rho\cos\theta + M$, where $H, K, L$, and $M$ are constants of order $\lambda$, results in no change in the three-dimensional intensity distribution near focus apart from a displacement of the distribution as a whole in accordance with the transformation (15); that is, a shift of amount $2(R/a)^2 H$ occurs along the principal direction $CP_1^\star$ away from the exit pupil and shifts of amounts $(R/a)K$ and $(R/a)L$ occur in the positive $x$ - and $y$-directions respectively.

The additive terms on the right-hand side of (10) may be interpreted as representing a change of reference sphere. Suppose that we choose a new reference sphere, centred on a point $P'(x', y', z')$ in the image region, and of radius $R'$ such that the new sphere is at most a few wavelengths away from the Gaussian sphere. Let a ray $\bar{Q}Q$ intersect the new reference sphere at a point $N$. Then the wave aberration $\Phi'$ referred to this new sphere is (see Fig. 9.2)

$$\Phi' = \bar{Q}N = \bar{Q}Q - NQ \sim \bar{Q}Q - NG, \tag{17}$$

where $G$ is the point in which the line $NP'$ intersects the Gaussian reference sphere, the refractive index of the image space being assumed to be unity as before. Now $\bar{Q}Q = \Phi$ is the wave aberration referred to the Gaussian sphere and $NG = NP' - GP' = R' - s$, where $s$ denotes the distance from $G$ to $P'$. Hence (17) may be written as

$$\Phi' \sim \Phi + s - R' = \Phi + \frac{\lambda}{2\pi}\left[-v\rho\cos(\theta - \psi) - \frac{1}{2}u\rho^2 + \left(\frac{R}{a}\right)^2 u\right] + (R - R'), \tag{18}$$

where (3) was used. Here $u, v$, and $\psi$ are given by (3) and (4) with $x', y', z'$ written in place of $x, y, z$. The relation (18) may be written in the form (10) with

$$H = -\frac{1}{2}\left(\frac{a}{R}\right)^2 z', \quad K = -\left(\frac{a}{R}\right) x', \quad L = -\left(\frac{a}{R}\right) y', \quad M = z' + R - R'. \tag{19}$$

### 9.1.3 A relation between the intensity and the average deformation of wave-fronts

When the aberrations are small it is possible to express the intensity at the centre of the reference sphere in terms of the mean square value of the wave aberration. Let $\Phi_P$ be the wave aberration referred to a reference sphere centred on a point $P$ in the image region. Then, according to equations (9) and (18), the normalized intensity at $P$ may be expressed in the form

0

- This concept is due to K. Strehl, Z.f. Instrumkde., 22 (1902), 213, who called it "Definitionsholligkeit". In the English literature the less appropriate term "definition" is often used.

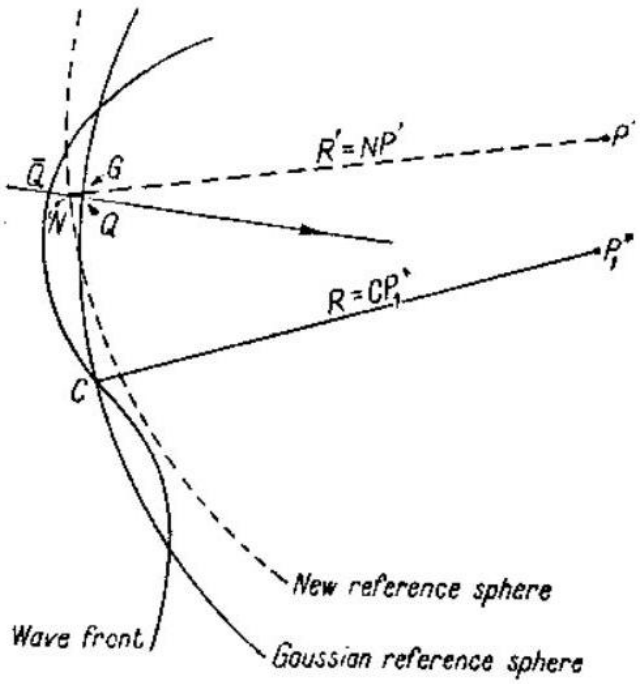



Fig. 9.2. Change of reference sphere.

$$i(P) = \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi} e^{ik\Phi_P}\rho d\rho d\theta\right|^2 = \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi}\left[1 + ik\Phi_P + \frac{1}{2}(ik\Phi_P)^2 + \ldots\right]\rho d\rho d\theta\right|^2 \quad (20)$$

Let $\overline{\Phi_P^n}$ denote the average value of the $n$th power of $\Phi_P$, i.e.

$$\overline{\Phi_P^n} = \frac{\int_0^1\int_0^{2\pi}\Phi_P^n\rho d\rho d\theta}{\int_0^1\int_0^{2\pi}\rho d\rho d\theta} = \frac{1}{\pi}\int_0^1\int_0^{2\pi}\Phi_P^n\rho d\rho d\theta \quad (21)$$

If we assume that the aberrations are so small that we may neglect third and higher powers of $k\Phi_P$ in (20), the intensity at $P$ may be written in the form

$$i(P) \sim \left|1 + ik\overline{\Phi_P} - \frac{1}{2}k^2\overline{\Phi_P^2}\right|^2 = 1 - \left(\frac{2\pi}{\lambda}\right)^2\left[\overline{\Phi_P^2} - \left(\overline{\Phi_P}\right)^2\right]. \quad (22)$$

The quantity in the brackets on the right is the "mean-square deformation" $(\Delta\Phi)^2$ of the wave-front,

$$(\Delta\Phi_P)^2 = \frac{\int_0^1\int_0^{2\pi}\left(\Phi_P - \overline{\Phi_P}\right)^2\rho d\rho d\theta}{\int_0^1\int_0^{2\pi}\rho d\rho d\theta} = \overline{\Phi_P^2} - \left(\overline{\Phi_P}\right)^2 \quad (23)$$

so that (22) may be written as

$$i(P) \sim 1 - \left(\frac{2\pi}{\lambda}\right)^2(\Delta\Phi_P)^2 \quad (24)$$

This formula implies that, when the aberrations are small, the normalized intensity at the centre of the reference sphere in the region of focus is independent of the nature of the aberration and is smaller than the ideal value unity by an amount proportional to the mean-square deformation of the wave-front.

## 9.12.2 9.2. Expansion of the Aberration Function

### 9.2.1 The circle polynomials of Zernike

In our discussion of the effects of aberrations on the basis of geometrical optics (Chapter V), we expanded the aberration function $\Phi$ in a power series. In the present treatment, where integrations over the unit circle must be carried out, it is more appropriate to expand $\Phi$ in terms of a complete set of polynomials that are orthogonal over the interior of the unit circle.* Many sets of polynomials with this property can be constructed; there is, however, one such set, introduced by Zernike, † which has certain simple properties of invariance. In Appendix VII it is shown how these circle polynomials of Zernike may be derived and some of their properties are discussed; here we shall only summarize the formulae needed in the present chapter.

The circle polynomials of Zernike are polynomials $V_n^l(X,Y)$ in two real variables $X, Y$, which, when expressed in polar coordinates ( $X = \rho\sin\theta, Y = \rho\cos\theta$ ), are of the form

$$V_n^l(\rho\sin\theta, \rho\cos\theta) = R_n^l(\rho)e^{il\theta}, \tag{1}$$

where $l \geqslant 0$, and $n \geqslant 0$ are integers, $n \geqslant |l|$, and $n - |l|$ is even. The orthogonality and normalizing properties are expressed by the formulae

$$\iint_{X^2+Y^2\leqslant 1} V_n^{l\star}(X,Y)V_{n'}^{l'}(X,Y)dXdY = \frac{\pi}{n+1}\delta_{ln'}\delta_{nn'} \tag{2}$$

where $\varepsilon_{ij}$ is the Kronecker symbol and where the asterisks denotes the complex conjugate. The radial functions $R_n^l(\rho)$ are polynomials in $\rho$, containing the powers $\rho^n, \rho^{n-2}, \ldots \rho^{|l|}$ and, as is shown in Appendix VII, are closely related to Jacobi's

polynomials (terminating hypergeometric series). As is seen from (1) and (2), the radial polynomials satisfy the relations

$$\int_0^1 R_n^l(\rho)R_{n'}^l(\rho)\rho d\rho = \frac{1}{2(n+1)}\delta_{nn'} \tag{3}$$

They are, given by the formulae ( $m = |l|$ )

$$R_n^{\pm m}(\rho) = \frac{1}{\left(\frac{n-m}{2}\right)!\rho^m}\left\{\frac{d}{d\left(\rho^2\right)}\right\}^{\frac{n-m}{2}}\left\{\left(\rho^2\right)^{\frac{n+m}{2}}\left(\rho^2-1\right)^{\frac{n-m}{2}}\right\} \tag{4}$$

$$= \sum_{s=0}^{\frac{n-m}{2}}(-1)^s\frac{(n-s)!}{s!\left(\frac{n+m}{2}-s\right)!\left(\frac{n-m}{2}-s\right)!} \tag{5}$$

The normalization has been chosen so that for all permissible values of $n$ and $m$,

$$R_n^{\pm m}(1) = 1. \tag{6}$$

The radial polynomials have the generating function

0

- The term "complete" implies that any reasonably well-behaved function can be expanded as a series of functions of the set. For a more precise definition of the term see, for example, R. Courant and D. Hilbert, Methods of Mathematical Physics, Vol. I (1st English edition, New York, Interscience Publishers, 1953), pp. 51-54.
  † F. Zernike, Physica, 1 (1934), 689.

$$\frac{\left[1+z-\sqrt{1-2z\left(1-2\rho^2\right)+z^2}\right]^m}{(2z\rho)^m\sqrt{1-2z\left(1-2\rho^2\right)+z^2}}=\sum_{s=0}^{\infty}z^s R_{m+2s}^{\pm m}(\rho). \tag{7}$$

When $m=0$, the left-hand side reduces to the generating function for the Legendre polynomials* of argument $2\rho^2-1$, so that

$$R_{2n}^0(\rho)=P_n\left(2\rho^2-1\right). \tag{8}$$

In Table XXI the explicit form of the polynomials for the first few values of the indices is given.

Table XXI
The radial polynomials \$R_{n}\hat{}{}{m}(\rho)\$ for $m \leqslant 8, n \leqslant 8$

| \n 70 | 0 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 0 | | $2\varphi^2-1$ | | $6p^4-8p^2+1$ | | $20\rho^6-30\rho^4+12\rho^2-1;$ | |
| 1 | | | $3\rho^3-2\rho$ | | $10\rho^3-12\rho^3+3\rho$ | | $35p^7-60p$ |
| 2 | | $\rho^2$ | | $4\rho^4-3\rho^2$ | | $15\rho^6-20\rho^4+6\rho^2$ | |
| 3 | | | $\rho^2$ | | $5\rho^5-4\rho^3$ | | $21\rho^7-30\rho$ |
| | | | | $\rho^4$ | | $6\rho^6-5\rho^4$ | |
| 5 | | | | | $\rho^s$ | | $7p^7-6p^6$ |
| 8 | – | | | | | $\rho^2$ | |
| 8 | □ | | | | | | $p^7$ |
| | | | | | | | |

The following relation (also proved in the Appendix) is of great importance in the Nijboer-Zernike theory:

$$\int_0^1 R_n^m(\rho)J_m(v\rho)\rho d\rho=(-1)^{\frac{n-m}{2}}\frac{J_{n+1}(v)}{v} \tag{9}$$

where $J$ is a Bessel function of the first kind.
Instead of the complex polynomials $V$, one may use the real polynomials*

$$\left.\begin{aligned} U_n^m &= \tfrac{1}{2}\left[V_n^m+V_n^{-m}\right]=R_n^m(\rho)\cos m\theta, \\ U_n^{-m} &= \tfrac{1}{2i}\left[V_n^m-V_n^{-m}\right]=R_n^m(\rho)\sin m\theta. \end{aligned}\right\} \tag{10}$$

In our applications we shall need the polynomials $U_n^m=R_n^m(\rho)\cos m0$ only. This is so because the wave distortions are symmetrical about the meridional plane $\theta=0$ and consequently the aberration function is an even function of $\theta$.

### 9.2.2 Expansion of the aberration function

Following Nijboer, we expand the aberration function $\Phi$ in terms of Zernike's circle polynomials. As in §5.1 it follows from symmetry that the variables enter the expansion only in the combination $Y_1^{\star 2}$, $\rho^2$ and $Y_1^{\star}\rho\cos\theta$, so that the expansion must be of the form

0

- See, for example, R. Courant and D. Hilbert, loc. cit., p. 85.

$$\Phi\left(Y_1^\star, \rho, \theta\right) = \sum_l \sum_n \sum_m a_{lnm} Y_1^{\star 2l+m} R_n^m(\rho) \cos m\theta \tag{11}$$

where $l, n$, and $m$ are non-negative integers, $n \geqslant m, n-m$ is even and the $a$ 's are constants.

As we shall mainly be concerned with the diffraction image of a fixed object point ( $Y_1^\star$ constant), it is convenient to supress the explicit dependence of $\Phi$ on $Y_1^\star$ and re-write (11) in the form

$$\Phi = A_{00} + \frac{1}{\sqrt{2}} \sum_{n=2}^{\infty} A_{n0} R_n^0(\rho) + \sum_{n=1}^{\infty} \sum_{m=1}^{n} A_{nm} R_n^m(\rho) \cos m\theta \tag{12}$$

The coefficients $A_{nm}$ are functions of $Y_1^\star$, and the factor $1/\sqrt{2}$ has been introduced in the second term to simplify the final formulae.

If the aberrations are sufficiently small, the normalized intensity at the Gaussian focus may be expressed in a simple form in terms of the coefficients $A$. We have, on substituting from (12) into §9.1 (21) and using the orthogonality relations (3),

$$\left.\begin{array}{ll} \bar{\Phi} & = A_{00} \\ \overline{\Phi^2} & = A_{00}^2 + \frac{1}{2} \sum_{n=1}^{\infty} \sum_{m=0}^{n} \frac{A_{nm}^2}{n+1}. \end{array}\right\} \tag{13}$$

The first relation implies that $A_{00}$ represents the mean retardation of the wave-front behind the Gaussian reference sphere. The second relation is "Parseval's formula" for the orthogonal set of functions $R_n^m(\rho) \cos m\theta$. On substituting from (13) into equation (22) of §9.1, it follows that the normalized intensity at the Gaussian focus is

$$i\left(P_1^\star\right) = 1 - \frac{2\pi^2}{\lambda^2} \sum_{n=1}^{\infty} \sum_{m=0}^{n} \frac{A_{nm}^2}{n+1} \tag{14}$$

A marked advantage of the expansion in terms of the circle polynomials arises in connection with the important problem of the "balancing" of aberrations of different orders against each other in such a way as to obtain maximum intensity. Suppose that the aberration is represented by a single "power series" term

$$\Phi = A'_{nm} \rho^n \cos^m \theta, \tag{15}$$

where $A'$ is a small constant, of the order of a wavelength or less. We enquire whether it is possible to increase the intensity $i\left(P_1^\star\right)$ by introducing aberrations of lower order. More precisely, we wish to choose constants $A'_{pq}$ in the expression

$$\Phi' = A'_{nm} \rho^n \cos^m \theta + \sum_{\mathcal{D}<n} \sum_{q \leqslant p} A'_{pq} \rho^p \cos^q \theta \tag{16}$$

so as to make the intensity at the Gaussian focus as large as possible.

With any particular choice of the constants $A'_{pq}$, the aberration function (16) may also be expressed in terms of the circle polynomials, in the form

$$\Phi' = \varepsilon_{nm} A_{nm} R_n^m(\rho) \cos m\theta + \sum_{p<n} \sum_{q \leqslant p} \varepsilon_{pq} A_{pq} R_p^q(\rho) \cos q\theta, \tag{17}$$

0

- A pictorial representation of some of these polynomials is given in an article by J. Y. Wang and D. E. Silva, Appl. Opt., 19 (1980), 1510.

where

$$\left.\begin{array}{lll} \varepsilon_{nm} & = \frac{1}{\sqrt{2}} & \text{when} \quad m = 0, n \neq 0, \\ & = 1 & \text{otherwise.} \end{array}\right\} \tag{18}$$

According to (5) the highest power of $\rho$ in $R_n^m(\rho)$ is $n$ and appears in the radial polynomial $R_n^m(\rho)$ with a coefficient $n!/\left[\frac{1}{2}(n+m)\right]!\left[\frac{1}{2}(n-m)\right]!$ Hence, on comparing the coefficients of $\rho^n \cos^m\theta$ (or of $\rho^n \cos m\theta$ ) in (16) and (17), it follows that

$$A_{nm} = \varepsilon_{nm} \frac{\left[\frac{1}{2}(n+m)\right]!\left[\frac{1}{2}(n-m)\right]!}{n!2^{m-1}} A'_{nm} \quad \text{when } n \neq 0. \tag{19a}$$

It is also evident that

$$A_{00} = A'_{00} \tag{19b}$$

Now with $A'_{nm}$ and consequently $A_{nm}$ fixed, it follows from (14) that the maximum intensity at $P_1^\star$ is obtained by making all the coefficients under the summation sign in (17) identically zero. The aberration function then becomes

$$\Phi' = \varepsilon_{nm} A_{nm} R_n^m(\rho) \cos m\theta \tag{20}$$

and the intensity at $P_1^\star$ now is

$$i\,(P_1^\star) = 1 - \frac{2\pi^2}{\lambda^2} \frac{A_{nm}^2}{n+1} \tag{21}$$

$A_{nm}$ being given, in terms of the coefficient $A'_{nm}$, by (19). It is now evident, that in the single aberration term $A_{nm} R_n^m(\rho) \cos m\theta$ of the expansion (12), a number of terms of the form $A'_{pq}\rho^p \cos^q\theta$ have been combined, with $p = n, n-2, \ldots m; q = m, m-2, \cdots 1$ or 0 in such a way, that for a given (sufficiently small) value of the coefficient of $0^n \cos^m 0$, the normalized intensity at the Gaussian focus is a maximum.

We illustrate this result by a simple example. Suppose that a system suffers from is small amount of sixth-order spherical aberration [ $\Phi = A'_{60}\rho^6$ ], and that we are in a position to introduce a controlled amount of fourth-order spherical aberration [$A'_{40}\rho^4$] and defocusing [ $A'_{20}\rho^2$ ]. We seek the values of the coefficients $A'_{40}$ and $A'_{20}$ which make the intensity at the diffraction focus as large as possible. Problems of
this type were first studied by Richter,* who showed that the maximum is obtained when the two coefficients are chosen so that

$$\frac{A'_{40}}{A'_{60}} = -\frac{3}{2}, \quad \frac{A'_{20}}{A'_{60}} = \frac{3}{5} \tag{22}$$

A glance at Table XXI, on p. 465 , shows that this is precisely the ratio of the corresponding coefficients in the polynomial $R_6^0(\rho)$ :

$$R_6^0(\rho) = 20\rho^6 - 30\rho^4 + 12\rho^2 - 1 \tag{23}$$

We see that, provided the aberrations are sufficiently small, the introduction of ZERNIKE's circle polynomials automatically solves the problem of balancing of aberrations, in the sense explained; moreover, with the help of the displacement theorem, it also enables the determination of the position of the diffraction focus.

## 9.12.3 9.3. Tolerance Conditions for Primary Aberrations

Before considering the difficult problem of determining the intensity distribution in the diffraction image in the presence of aberrations, we consider the much simpler problem of estimating the maximum amounts of aberrations that may be tolerated in an optical system.

It is evident from the discussion of the preceding section that, when aberrations are present, the maximum intensity in the diffraction image is smaller than the intensity at the Gaussian focus (centre of the Arry pattern) in an aberration free system of the same aperture and focal length. It was shown first by Rayleigh † that when a system suffers from primary spherical aberration of such an amount that the wave-front in the exit pupil departs from the Gaussian reference sphere by less than a quarter wavelength, the intensity at the Gaussian focus is diminished by less than 20 per cent -a loss of light that can usually be tolerated. Later workers found that in the presence of other commonly occurring aberrations the quality of the image is likewise not seriously affected when the wave-front deformation is less than a quarter of a wavelength. This result has become known as Rayleigh's quarter wavelength rule and is a useful criterion for the amount of aberration that can be tolerated in an image-forming system. This rule is, of course, only a rough guide as to the desirable state of correction of a system, since the light distribution in the image depends not only on the maximum deformation but also on the shape of the wave-fronts (type of aberration). Moreover, the loss of light that may be tolerated depends naturally on the particular use to which the instrument is put, and more stringent tolerances have to be imposed in certain cases.

When the condition $|\Phi_{\max}| = \lambda/4$ is applied to aberrations of different types, somewhat different values for the intensity at the diffraction focus are obtained. It seems more appropriate to formulate tolerance criteria which correspond to a prescribed value of the intensity at the diffraction focus. Criteria of this type were considered by Marchal, ‡who used the relation that exists between the intensity

at the centre of the reference sphere and the root-mean-square deviation of the wavefront from spherical form.

When the aberrations are sufficiently small, the intensity at a point $P$ in the region of the image may according to equation (24) of §9.1 be expressed in the form

$$i(P) \sim 1 - \left(\frac{2\pi}{\lambda}\right)^2 (\Delta\Phi_P)^2 \tag{1}$$

Following Marchal we shall regard a system to be well corrected when the normalized intensity at the diffraction focus $F$ is greater than or equal to 0.8 . Now from (1), $i(F) \geqslant 0.8$ when $|\Delta\Phi_F| \lesssim \lambda/14$, so that this condition is equivalent to the requirement that the root-mean-square departure of the wave-front from a reference sphere that is centred on the diffraction focus shall not exceed the value $\lambda/14$.*

Let us now determine the position of the diffraction focus and the tolerances for primary (Seidel) aberrations. In the notation of the present chapter, each primary aberration represents a wave-front deformation of the form †

$$\Phi = a'_{lnm} Y_1^{\star 2l+m} \rho^n \cos^m \theta \tag{2}$$

where $2l + m + n = 4$. It is convenient to set

0

- R. Richter, Z.f. Instrunkde., 45 (1025), 1.
† Lord Rayleigi, Phil. Mag., (5), 8 (1879), 403. Reprinted in his Scientific Papers (Cambridge University Press, 1899), 1, 432-435.
‡ A. Marchal, Rev. d'Optique, 26 (1947), 257.

$$A'_{lnm} = a'_{lnm} Y_1^{\star 2l+m}, \tag{3}$$

and (2) becomes

$$\Phi = A'_{lnm} \rho^n \cos^m \theta \tag{4}$$

The constant $A'$ can easily be expressed in terms of the Seidel coefficients $B, C, D$, $E$, and $F$, introduced in Chapter V. If we take the arbitrary constant $\lambda_0$ in equations (7) and (8) of § 5.2 equal to unity, then $\lambda_1$ denotes the magnification between the pupil planes, and if we remember that we now have $n_1 = 1$, the variables $\rho$ and $y_0$ of §5.3 (7) correspond to $a\rho/\lambda_1$ and $-\lambda_1 Y_1^\star/R$ of the present section; comparing (4) with § 5.3 (7) we obtain

- Marchal's condition is actually based on a somewhat different inequality. It follows from § 9.1 (20) that

$$\begin{aligned} i(\ \mathrm{F}) &= \frac{1}{\pi^2} \left| \int_0^1 \int_0^{2\pi} e^{ik\Phi_r} \rho d\rho d\theta \right|^3 \\ &> \frac{1}{\pi^2} \left| \int_0^1 \int_0^{2\pi} \cos(k\Phi_r)\, p d\rho d\theta \right|^2 \end{aligned}$$

Now if $k|\Phi_P| < \pi/2$, i.e. if $|\Phi_F| < \lambda/4, \cos(k\Phi_F)$ may evidently be replaced by $1 - \frac{1}{2}(k\Phi_F)^2$ in this inequality. If, moreover, the radius of the reference sphere is chosen so that $\bar{\Phi}_F = 0$, then $(\Delta\Phi_F)^2 = \bar{\Phi}_F^2$ and the inequality becomes

$$i(F) \gtrsim \left[1 - \frac{2\pi^2}{\lambda^2} (\Delta\Phi_F)^2\right]^3. \tag{Ia}$$

Marchal's criterion is based on the inequality (1a) but the difference is evidently of no great practical consequence if the aberrations aro small. For our purposes, the use of the relation (1) rather than (1a) has the advantage that it is more directly rclated to the oxtremal properties of Zernike's circlo polynomials.

† We attach a prime to the coefficients in the power series mpresentation, whilst unprimed coefficients refer to the representation in terms of the ZERNIKI circle polynomials.

$$\left.\begin{aligned} A'_{040} &= -\tfrac{1}{4} \left(\tfrac{a}{\lambda_1}\right)^4 B, \\ A'_{031} &= -\left(\tfrac{a}{\lambda_1}\right)^3 \left(\tfrac{\lambda_1 Y_1^\star}{R}\right) F, \\ A'_{022} &= -\left(\tfrac{a}{\lambda_1}\right)^2 \left(\tfrac{\lambda_1 Y_1^\star}{R}\right)^2 C, \\ A'_{120} &= -\tfrac{1}{2} \left(\tfrac{a}{\lambda_1}\right)^2 \left(\tfrac{\lambda_1 Y_1^\star}{R}\right)^2 D, \\ A'_{111} &= -\left(\tfrac{a}{\lambda_1}\right) \left(\tfrac{\lambda_1 Y_1^\star}{R}\right)^3 E. \end{aligned}\right\} \tag{5}$$

In the expansion in terms of the circle polynomials, a typical term represents the aberration*

$$\Phi = \varepsilon_{nm} A_{lnm} R_n^m(\rho) \cos m\theta. \tag{6}$$

The terms with indices $l, m$, and $n$ such that $2l+m+n = 4$ (primary aberrations) are shown in the last column of Table XXII. It is seen that some of the Seidel terms are now accompanied by terms of lower degrees, and these, according to the displacement theorem (§ 9.1.2), give rise to a bodily displacement of the intensity distribution.

Table XXII
Representation of the primary aberrations

| Type of aberration | $l$ | $n$ | $m$ | Representation in form (4) | Representation in form (6) |
|---|---|---|---|---|---|
| Spherical aberration | 0 | 4 | 0 | $A'_{040}\rho^4$ | $\frac{1}{\sqrt{2}}A_{040}R_4^0(\rho) = \frac{1}{\sqrt{2}}A_{040}\left(6\rho^4 - 6\rho^2 + 1\right)$ |
| Coma | 0 | 3 | 1 | $\mathrm{A}'_{091}\mathrm{o}^3\cos\theta$ | $\mathrm{A}_{031}R_3^2(\rho)\cos\theta = A_{081}\left(3\rho^3 - 2\rho\right)\cos\theta$ |
| Astigmatism | 0 | 2 | 2 | $A'_{082}\rho^2\cos^2\theta$ | $A_{022}R_2^2(\rho)\cos 2\theta = A_{022}\rho^2\left(2\cos^2\theta - 1\right)$ |
| Curvature of field | 1 | 2 | 0 | $A'_{120}\rho^2$ | $\frac{1}{\sqrt{2}}A_{120}R_2^0(\rho) = \frac{1}{\sqrt{2}}A_{120}\left(2\rho^3 - 1\right)$ |
| Distortion | 1 | 1 | 1 | $A'_{111}\rho\cos\theta$ | $A_{111}R_1^1(\rho)\cos\theta = A_{111}\rho\cos\theta$ |

Now according to the theorem of §9.2.2, the image affected by an aberration represented by (6) has maximum intensity at the Gaussian focus. Hence from a comparison of corresponding terms in the last two columns in Table XXII, we may immediately determine the coordinates of the diffraction focus of an image affected

by a primary aberration. We illustrate this by considering primary spherical aberration in detail. This aberration is represented by

$$\Phi = A'_{040}\rho^4. \tag{7}$$

The corresponding expression (6) is

$$\Phi = \frac{1}{\sqrt{2}}A_{040}R_4^0(\rho) = \frac{1}{\sqrt{2}}A_{040}\left(6\rho^4 - 6\rho^2 + 1\right). \tag{8}$$

Now if

$$A'_{040} = \frac{6}{\sqrt{2}}A_{040} \tag{9}$$

then, according to the displacement theorem, the intensity distribution will be the same in both cases; but the distribution corresponding to (7) will be displaced relative to that corresponding to (8) in accordance with equation (15) of §9.1, with

$$H = 6A_{040}/\sqrt{2} = A'_{040}, \quad K = L = 0, \quad M = -A_{040}/\sqrt{2} = A'_{040}/6,$$

i.e. according to the transformation

$$x' = x, \quad y' = y, \quad z' = z + 2\left(\frac{R}{a}\right)^2 A'_{040} \tag{10}$$

Since the diffraction image associated with (8) has maximum intensity at the origin $x = y = z = 0$, the diffraction focus $F$ for primary spherical aberration represented by (7) is at the point

$$x_F = y_F = 0, \quad z_F = 2\left(\frac{R}{a}\right)^2 A'_{040}. \tag{11}$$

The point $F$ specified by (11) has a simple geometrical interpretation. Let $\Delta Y$ and $\Delta Z$ be the lateral and longitudinal spherical aberrations, measured as positive when the ray intersects

0

- The factor $\varepsilon_{nm}$ which is equal to unity except for $m = 0, n \neq 0$ when it is equal to $1/\sqrt{2}$, is retained here for the sake of uniformity with the formulae of § 9.2.

the axes on the positive side of the Gaussian focus. By §5.1 (17), with $\Phi = A'_{040}\rho^4, \rho = Y/a, \mathrm{D}_1 \sim -R \sim -R', X_0 = X = 0, n_1 = 1,$

$$\Delta Y = Y_1 - Y_1^\star = 4\left(\frac{R}{a}\right)\left(\frac{Y}{a}\right)^3 A'_{040} \tag{12a}$$

and hence by elementary geometry and the use of the preceding relation

$$\Delta Z = Z - Z_1^\star \sim \frac{R}{Y}\Delta Y = 4\left(\frac{R}{a}\right)^2\left(\frac{Y}{a}\right)^2 A'_{040} \tag{12b}$$

For the marginal ray ($Y = a$) this gives $(\Delta Z)_{\max} = 4\left(\frac{R}{a}\right)^2 A'_{040}$, so that (11) is seen to imply that* the diffraction focus in the presence of a small amount of primary spherical aberration is situated midway between the paraxial and marginal foci.

Next let us determine the tolerance for primary spherical aberration. For any aberration characterized by (6) the normalized intensity at the Gaussian focus will be according to § 9.2 (14), greater than or equal to 0.8 if

$$1 - \frac{2\pi^2}{\lambda^2}\frac{A_{lnm}^2}{n+1} \geqslant 0 \cdot 8$$

ieprovided that

$$|A_{lmm}| \lesssim \frac{\lambda\sqrt{n+1}}{10}. \tag{13}$$

In particularfor primary spherical aberration this gives
orby9

$$|A_{040}| \lesssim 0 \cdot 22\lambda,$$

$$|A'_{040}| \lesssim 0.94\lambda. \tag{14}$$

This is the required tolerance condition for primary spherical aberration and implies that the maximum deviation of the wavefront from the Gaussian reference sphere must be less than 0.94 of the wavelength

In a strictly similar manner we may find the position of the diffraction focus and the tolerance for the other primary aberrationsIn particularthe diffraction focus in the presence of a small amount of primary astigmatism is found to be at the point whose coordinates are

$$x_F = y_F = 0, \quad z_F = \left(\frac{R}{a}\right)^2 A'_{022}. \tag{15}$$

This result has again a simple physical interpretationAccording to5and §5318 the radii $R_t$ and $R_s$ of the tangential and sagittal focal surfaces are given byassuming $n_1 = 1$ again

$$\frac{1}{R_t} = -\frac{4}{a^2}\left(\frac{R}{Y_1^\star}\right)^2 A'_{022}, \quad \frac{1}{R_3} = 0, \tag{16}$$

so that the abscissa $Z_t$ and $Z_s$ of the two focal lines are

0

- We neglect here the small effect arising from the different choice of the $z$-directions in § 5.1 and in the present discussion.

$$Z_t = -\frac{Y_1^{\star 2}}{2R_t} = 2\left(\frac{R}{a}\right)^2 A'_{022}, \quad Z_t = 0. \tag{17}$$

Table XXIII
Position of the diffraction focus and tolerance conditions for primary aberrations

| ~~Type of aberration~~ | Coordinates of diffraction focus $F$ | | | ~~Tolerance condition $[i(F) > 0.8]$~~ |
|---|---|---|---|---|
| | $x_F$ | $y_F$ | $z_F$ | |
| Spherical aberration | 0 | 0 | $2\left(\frac{R}{a}\right)^2 A'_{040}$ | $\lvert A'_{040}\rvert \lesssim 0.94\lambda$ |
| Coma | 0 | $\frac{2}{3}\left(\frac{R}{a}\right) A'_{b31}$ | 0 | $\mid A'_{031}.1.60\lambda$ |
| Astigmatism | 0 | 0 | $\left(\frac{R}{a}\right)^2 A'_{023}$ | $\lvert A'_{\text{ozs}}\mid \lesssim 0.35\lambda$ |
| Curvature of field | 0 | 0 | $2\left(\frac{R}{a}\right)^2 A'_{110}$ | |
| Distortion | 0 | $\left(\frac{R}{a}\right) A_{111}$ | 0 | |

Hence (15) implies that the diffraction focus in the presence of a small amount of primary astigmatism is situated midway between the tangential and sagittal focal lines.

- Since primary curvature of field and primary distortion are represented by terms of the second and first degree in $\rho$ respectively, it follows, in accordance with the displacement theorem, that the only effect of these aberrations is a "bodily shift" of the three-dimensional distribution associated with an aberration-free image. Thus in the presence of a small amount of primary curvature of field or primary distortion, the normalized intensity $i$ at the diffraction focus is unity, but the diffraction focus does not coincide with the Gaussian image point.

In Table XXIII the results relating to primary aberrations are summarized.

## 9.12.4 9.4. The Diffraction Pattern Associated with a Single Aberration

We now consider the diffraction image in the presence of an aberration represented by a single term of the expansion § 9.2 (11),

$$\Phi = a_{lnm} Y_1^{\star 2l+m} R_n^m(\rho) \cos m\theta \tag{1}$$

As before, we suppress the explicit dependence on $Y_1^\star$, and set

$$\alpha_{lnm} = \frac{2\pi}{\lambda} a_{lnm} Y_1^{\star 2l+m} = \frac{2\pi}{\lambda} \varepsilon_{nm} A_{lnm}. \tag{2}$$

We also set

$$C = -\frac{i\pi A}{\lambda}\left(\frac{a}{R}\right)^2 e^{i\left(\frac{R}{a}\right)^2 u}. \tag{3}$$

The diffraction integral § 9.1 (6) then becomes

$$U(u,v,\psi) = \frac{C}{\pi}\int_0^1 \int_0^{2\pi} e^{i\left[-v\rho\cos(\theta-\psi) - 1u\rho^2 + \alpha_{1nm} R_n^m(\rho)\cos m\theta\right]} \rho d\rho d\theta \tag{!}$$

The integral (4) may be developed in an infinite series, by expanding both the terms $e^{-iv\rho\cos(\theta-\psi)}$ and $e^{i\alpha_{2mm} R_n^m(\rho)\cos m\theta}$ with the help of the $\mathrm{J}_{\text{ACOBI}}$ identity [§ 8.8 (29)]

$$e^{iz\cos\phi} = J_0(z) + 2\sum_{s=1}^{\infty} i^s J_s(z)\cos s\phi. \tag{5}$$

Multiplying the two expansions together we find that

$$e^{i[-v\rho\cos(\theta-\varphi)+\alpha_{lnm}R_n^m(\rho)\cos m\theta]}$$
$$= 4\sum_{s=0}^{\prime}\sum_{s'=0}^{\infty} i^s(-i)^{s'} J_s\left[\alpha_{lnm}R_n^m(\rho)\right] J_{s'}(v\rho)\cos ms\theta\cos\left[s'(\theta-\psi)\right], \tag{6}$$

where the prime on the summation sign implies that the terms in $s = 0$ and $s' = 0$ are each to be taken with a factor $\frac{1}{2}$. We substitute this double series into (4) and integrate with respect to $\theta$ term by term. This gives

$$U(u,v,\psi) = 4C\sum_{s=0}^{\infty}(-i)^{(m-1)s}\cos ms\psi\int_0^1 e^{-jiu\rho^2} J_s\left[\alpha_{inm}R_n^m(\rho)\right] J_{ms}(v\rho)\rho d\rho, \tag{7}$$

the term $s = 0$ being again taken with the factor $\frac{1}{2}$.

As we are interested in small aberrations ( $\alpha$ small), we develop the term $J_J\left[\alpha_{\text{inm}}\, R_n^m(\rho)\right]$ under the integral sign in a power series, and re-arrange the resulting expression according to powers of $\alpha_{\text{inm}}$ . This gives

$$U(u,v;\psi) = C\left[U_0 + i\alpha_{lnm}U_1 + (i\alpha_{lnm})^2 U_3 + (i\alpha_{lnm})^3 U_3 + (i\alpha_{lnm})^4 U_4 + \ldots\right], \tag{8a}$$

where

$$U_3 = \frac{1}{2.3!}\left\{3(-i)^m\cos m\psi\int_0^1 e^{-\frac{1}{2}iu\rho^2}\left\{R_n^m(\rho)\right\}^3 J_m(v\rho)\rho d\rho \right.$$
$$\left. +(-i)^{3m}\cos 3m\psi\int_0^1 e^{-\frac{1}{2}iu\rho^2}\left\{R_n^m(v\rho)\right\}^3 J_{3m}(v\rho)\rho d\rho\right\}$$

$$U_1$$

$$+i^{2n}$$

$$\frac{1}{2^2.4!}\left\{3\int_0^1 e^{-\frac{1}{2}iu\rho^2}\left\{R_n^m(\rho)\right\}^4 J_0(v\rho)\rho d\rho \quad +4\right.$$

$$+i^{4}$$

Nijboer found that where both $u$ and $\alpha_{\text{lnm}}$ are of the order of unity, about four terms in the expansion (8a) suffice to give the intensity to within a few per cent.

To evaluate the integrals in (8b) we may proceed as follows. We express the factor $e^{-iup^2}$ in terms of the radial polynomials, with the help of the following well-known formula due to Bater*

$$e^{iz\cos\phi} = \left(\frac{\pi}{2z}\right)^{\dagger}\sum_{s=0}^{\infty} i^s(2s+1)J_{s+1}(z)P_s(\cos\phi), \tag{9}$$

where the $P$ 's are the Legendre polynomials. If we set $\cos\phi = 2\rho^2 - 1$ and use the relation $P_s\left(2\rho^2-1\right) = R_{2s}^0(\rho)$[9.2(8)], it follows that

$$e^{-\frac{1}{2}u\rho^2} = e^{-\frac{1}{2}iu_e - \frac{1}{2}iu\left(2\rho^2-1\right)} = e^{-\frac{1}{4}iu}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(-i)^s(2s+1)J_{s+\frac{1}{4}}\left(\frac{1}{4}u\right)R_{2s}^0(\rho). \tag{10}$$

On substitution from (10) into (8b) integrals are obtained each of which consists of a Bessel function multiplied by a product of the radial polynomials. Now these integrals can be evaluated by the use of the formula §9.2 (9)

$$\int_0^1 R_n^m(\rho) J_m(v\rho)\rho d\rho = (-1)^{\frac{n-m}{2}} \frac{J_{n+1}(v)}{v} \tag{11}$$

provided each of the products of the radial polynomials is expressed as a linear combination of the form $\sum_p A_p R_p^m(\rho)$, with $m$ equal to the order of the Bessel function by which the product is multiplied. It is not easy to obtain a general expression for the coefficients $A_p$, but if $m$ and $n$ are not too large such linear relations may easily be established with the help of Table XXI, p. 465, as will be seen later from some examples. For the discussion of methods relating to more general cases we refer the reader to Nuboer's thesis.

In the special case of an aberration-free wave we obtain, on substituting from (10) into (7),

$$\begin{aligned} U(u,v,v) &= 2Ce^{-\frac{1}{2}iu}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(-i)^s(2s+1)J_{s+\frac{1}{4}}\left(\frac{1}{4}u\right)\int_0^1 R_{2s}^0(\rho)\cdot J_0(v\rho)\rho d\rho \\ &= 2Ce^{-\frac{1}{2}iu}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(i)^s(2s+1)J_{s+\frac{1}{4}}\left(\frac{1}{4}u\right)\frac{J_{2s+1}(v)}{v} \end{aligned} \tag{12}$$

where (11) was used. Though formally different, this expansion is equivalent to the series developments of Lommel, given in § 8.8.1.

From the expansion (8) some general properties of the diffraction image may immediately be deduced. It is seen that $U$ remains unchanged when $\psi$ is replaced by $\psi+2\pi\mu/m(\mu=1,2,\ldots m)$; hence the $z$-axis is an $m$-fold axis of symmetry. Moreover the planes through the $z$-axis which makes angles $\pi\mu/m$ with the plane $x=0$ are planes of symmetry. When $m=0$, there is, of course, rotational symmetry.

Next consider the symmetry with respect to the plane $z=0$. We observe that when $u$ is replaced by $-u$, all integrals in (8) change into their conjugates. Now, if $m$ is odd, the coefficients by which these integrals are multiplied are all real. In this case, therefore, $U(u,v,\psi)$ is changed into its conjugate, and, in consequence, the intensity remains unaltered. Hence, if $m$ is odd, the intensity distribution is symmetrical with respect to the plane $z=0$. If, on the other hand $m$ is even (but $m\neq 0$ ), those coefficients which involve a factor $\cos 2\mu m\psi$ ( $\mu$ being an integer) are real, whereas the others, which involve a factor $\cos(2\mu+1)m\psi$, are pure imaginary. Therefore, if $u$ is replaced by $-u$ and at the same time $\psi$ is replaced by $\psi+\pi/m, U$ now changes into its conjugate. Hence if $m$ is even but different from zero, the intensity at any point in the plane $z=$ constant is the same as the intensity at the point resulting from reflection in the plane $z=0$ and an additional rotation through an angle $\pi/m$ about the $z$-axis. It follows that whereas, as has already been pointed out, the $z$-axis is in general an $m$-fold axis of symmetry, it is, when $m$ is even, a $2m$-fold axis of symmetry with respect to the pattern in the plane $z=0$. When $m=0$ (spherical aberration), the diffraction image is not symmetrical with respect to the plane $z=0$.

Finally, we observe that when the sign of the aberration constant $\alpha_{\text{inm}}$ is changed, the intensity distribution is not altered if $u$ is replaced by $-u$ when $m$ is even, and if $\psi$ is replaced by $\psi+\pi$ when $m$ is odd.

---

0

- See, for example, G. N. Watson, A Treatise on the Theory of Bessel Functions (Cambridge University Press, 2nd ed., 1944), p. 368.

We now briefly consider the structure of the image when the system suffers from a primary aberration of a small amount.

### 9.4.1 Primary spherical aberration

In this case $l = m = 0$ and $n = 4$. The aberration function is independent of 0 and the three-dimensional diffraction image has rotational symmetry about the principal direction $v = 0$.

According to (8) the expansion of the diffraction integral in powers of $\alpha$ is

$$U(u, v, \psi) = C\left[U_0 + i\left(\alpha_{040}\right)U_1 + \left(i\alpha_{040}\right)^2 U_2 + \ldots\right] \tag{13}$$

where $U_0$ represents the disturbance in the aberration-free image and $U_1, U_2$, are given by the other expressions in (8b), with $m = 0$ and $n = 4$. In particular,

$$\dot{U}_1 = 2\int_0^1 e^{-\frac{1}{2}u\rho^2} R_4^0(\rho) J_0(v\rho)\rho d\rho \tag{14}$$

A substitution for $e^{-\text{tiup}\,\rho^2}$ from (10) gives

$$U_1 = 2e^{-\frac{1}{2}u}\sqrt{\frac{2\pi}{u}}\sum_{s=0}^{\infty}(-i)^s(2s+1)J_{s+1}\left(\frac{1}{4}u\right)\int_0^1 R_{2s}^0(\rho)R_4^0(\rho)J_0(v\rho)\rho d\rho \tag{15}$$

To evaluate the integral on the right, the product of the radial polynomials will be replaced, as already explained, by a linear combination of radial polynomials whose upper index is equal to the order of the Bessel function (zero in this case). The linear combination is easily found by using the relation § 9.2 (8)

$$R_{2s}^0(\rho) = P_s\left(2\rho^2 - 1\right) \tag{16}$$

and certain well known relations involving the Legendre polynomials. We have

$$R_{2s}^0(\rho)R_4^0(\rho) = P_s\left(2\rho^2 - 1\right)P_2\left(2\rho^2 - 1\right). \tag{17}$$

Now $P_2(t) = \frac{1}{2}\left(3t^2 - 1\right)$ so that the right-hand side of (17) may be written as

$$P_3(t)P_2(t) = \frac{3}{2}t^2 P_3(t) - \frac{1}{2}P_3(t). \tag{18}$$

Applying several times the recurrence relation*

$$tP_s(t) = \frac{1}{2s+1}\left[(s+1)P_{s+1}(t) + sP_{s-1}(t)\right] \tag{19}$$

it follows that

$$P_2(t)P_s(t) = a_s P_{s+2}(t) + b_s P_s(t) + c_s P_{s-2}(t), \tag{20}$$

where

$$a_s = \frac{3}{2}\frac{(s+2)(s+1)}{(2s+3)(2s+1)}, \quad b_s = \frac{(s+1)s}{(2s+3)(2s-1)}, \quad c_s = \frac{3}{2}\frac{s(s-1)}{(2s+1)(2s-1)}. \tag{21}$$

Hence (17) becomes

$$R_{2s}^0(\rho)R_4^0(\rho) = a_s R_{2s+4}^0(\rho) + b_s R_{2s}^0(\rho) + c_s R_{2s-4}^0(\rho) \tag{22}$$

Finally, substituting from (22) into (15) and using (11), it follows that

$$U_1 = 2e^{-\frac{1}{2}iu}\sqrt{\frac{2\pi}{u}}\frac{1}{v}\sum_{\varepsilon=0}^{\infty}(i)^3(2s+1)J_{s+\frac{1}{4}}\left(\frac{1}{4}u\right)\left[a_3 J_{2s+5}(v) + b_3 J_{2s+1}(v) + c_3 J_{2s-3}(v)\right]. \qquad (23)$$

In a similar way † one may obtain series developments for $U_2, U_3, \ldots$ Using these expansions one may then calculate the intensity $I = |U|^2$ at a number of points

in the region of the image and construct the isophotes (lines of equal intensity). In any plane at right angle to the principal direction $v = 0$ the isophotes are, of course, circles.

Fig. 9.3 shows the isophotes in a meridional plane for primary spherical aberration $\Phi = -0.48\lambda\rho^4$, whilst in Figs. 9.4 and 9.5 photographs are reproduced which show the appearance of images in various receiving planes in the presence of spherical aberration of somewhat larger amounts.*

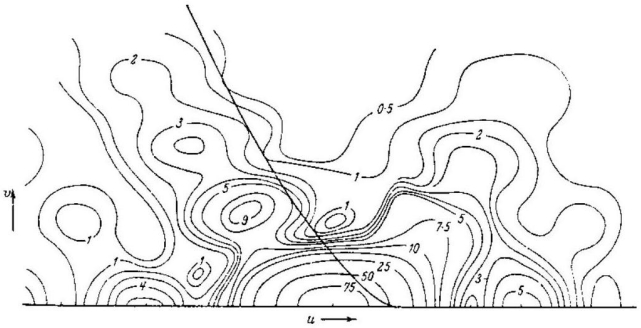



Fig. 9.3. Isophotes in a meridional plane, in presence of primary spherical aberration $\Phi=0.48 \lambda \rho\hat{}\{4\}$. The thick line indicates the geometrical ceustic. The intensity is normalized to 100 at the centre of the aberration-free image. Strehl intensity in best receiving plane: 0.95 .
[After F. Zerntke and B. R. A. Nijboer, contribution to La Thorie des Imajes Optiques (Paris, Revue d'Optique, 1949), p. 232.]

### 9.4.2 Primary coma

We now have $l = 0, n = 3, m = 1$. According to Table XXIII, p. 472, the diffraction focus is in the plane $z = 0$ and the disturbance in this plane is given by

$$U(0, v, \psi) = C\left[U_0(0, v, \psi) + (i\alpha_{031})\, U_1(0, v, \psi) + (i\alpha_{031})^2\, U_2(0, v, \psi) + \ldots\right] \qquad (24)$$

With $u = 0$, the integral $U_0$, defined in (8 b), represents the disturbance $2J_1(v)/v$ in the focal plane of aberration-free system (Airy pattern), and $U_1$ can immediately be evaluated by the use of (II). To evaluate $U_2, U_3$. . we must again express the products of the circle polynomials by the appropriate linear combination of such polynomials. In particular, it may be verified with the help of Table XXI, p. 465, that

$$\left(R_3^1\right)^2 = \frac{1}{4}R_0^0 + \frac{1}{20}R_2^0 + \frac{1}{4}R_4^0 + \frac{9}{20}R_5^0 = \frac{2}{5}R_2^2 + \frac{3}{5}R_6^2. \qquad (25)$$

(Fig. (a) after B. R. A. Nijboer, Thesis, University of Groningen, 1942, p. 62;
Fig. (b) after K. Nienhuis and B. R. A. Nijboer, Physica 14 (1949), 599.)

0

• See, for example, E. T. Whittaker and G. N. Watson, A Course of Modern Analysis (Cambridge University

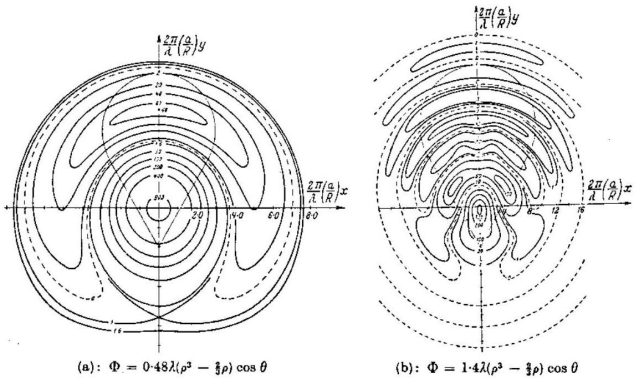



Fig. 9.6. Isophotes in the plane $z = 0$ in the presence of primary coma. The dotterd curves represent lines of zero intensity. The boundary of the geometrical confusion figure is also shown. The intensity is normalized to 1000 at the centre of the aberration-free image. Strehl intensity: (a) 0.879 ; (b) 0.306 .

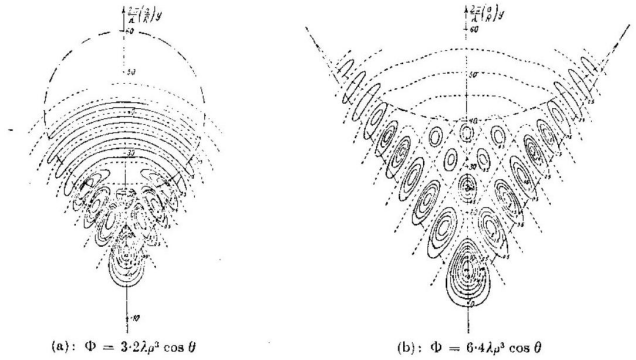



Fig. 9.7. Lsophotes in the plane $z = 0$ in the presence of primary coma. The intensity is normalized to 100 at the centre of the aberration-free image.

Using these relations in the expression for $U_2$ in (8b) and applying (11) the integrals in $U_2$ are immediately evaluated and we have in all

$$\left.\begin{aligned} U_0(0,v,\psi) &= \tfrac{2J_1(v)}{v} \\ U_1(0,v,\psi) &= i\cos\psi\tfrac{2J_4(v)}{v} \\ U_2(0,v,\psi) &= \tfrac{1}{2v}\left\{\tfrac{1}{4}J_1(v) - \tfrac{1}{20}J_3(v) + \tfrac{1}{4}J_5(v) - \tfrac{9}{20}J_7(v)\right\} \\ &\quad -\cos 2\psi\left(\tfrac{2}{5}J_3(v) + \tfrac{3}{5}J_7(v)\right) \end{aligned}\right\} \tag{26}$$

The isophotes for primary coma of various amounts are shown in Figs. 9.6 and 9.7. The data

Press, fourth ed. 1940), p. 308.

† For details see F. Zernike and B. R. A. Nijboer, contribution to La Thorie des Images Optiques (Paris, Revue d'Optique. 1949), p. 227.

0

- The expansion (13) is unsuitable for computing the intensity when the aberrations are not small compared to the wavelength. Isophotes in a meridional plane for primary spherical aberration of several wavelengths were determined with the help of a mechanical integrator by A. Marchal. and are published in E. H. Linfoot, Recent Advances in Optics (Oxford, Clarendon Press, 1055), pp. 60-61, and in M. Franon's contribution to Encyclopedia of Physics, XXIV (Ed. S. Flegge, Berlin, Springer, 1956), pp. 321-322; and by J. Focke, Optica Acta, 3 (1956), 110 , who und aqumptotic approximations. See also J. Picht, Ann. d. Physik, (4), 77 (1925), 685.

for Fig. 9.6 were computed from series expansions, those for Fig. 9.7 by numerical integrations. Photographs showing images affected by primary coma are given in Fig. 9.8. The figures show that when the aberration is of the order of a wavelength, the image resembles neither the Alry pattern nor the pattern predicted by geometrical optics. As the aberration is increased the true image soon becomes of the form specified by geometrical optics, but is broken up by a series of dark bands; these may be shown to arise from the interference of rays diffracted from diametrically opposite points of the aperture.

Fig. 9.6 also illustrates the general result established in §9.2, that when a small aberration is represented in terms of a circle polynomial, the intensity pattern is displaced so as to have its maximum at the origin of the coordinates.

### 9.4.3 Primary astigmatism

The effect of a small amount of primary astigmatism may be investigated in a similar manner. We now have $l = 0, n = m = 2$, and as shown in § 9.3 the diffraction focus is midway between the two focal lines. We consider the light distribution in the central plane, i.e. the plane through this point, at right angles to the principal direction. When the aberration is represented in terms of the appropriate circle polynomial $[A_{022}R_2^2(\rho)\cos 20]$ the central plane is the plane $u = 0$.

The disturbance in the central plane is given by

$$U(0, v, \psi) = C\left[U_0(0, v, \psi) + (i\alpha_{022})\, U_1(0, v, \psi) + (i\alpha_{022})^2\, U_2(0, v, \psi)\ldots\right] \tag{27}$$

where $U_0(0, v, \psi)$ represents as before the Airy pattern distribution and $U_1(0, v, \psi)$ may immediately be evaluated with the help of (11). To evaluate $U_2$ we use the identities

$$\left.\begin{aligned}(R_2^2)^2 &= \tfrac{1}{3}R_0^0 + \tfrac{1}{2}R_2^0 + \tfrac{1}{6}R_4^0,\\ &= R_4^4,\end{aligned}\right\} \tag{28}$$

and proceed as before. We obtain in all

$$\left.\begin{aligned}U_0(0, v, \psi) &= \tfrac{2J_1(v)}{v}\\ U_1(0, v, \psi) &= -2\cos 2\psi \tfrac{2J_3(v)}{v}\\ U_2(0, v, \psi) &= \tfrac{1}{2v}\left\{\tfrac{1}{3}J_1(v) - \tfrac{1}{2}J_3(v) + \tfrac{1}{6}J_5(v) + \cos 4\psi J_5(v)\right\}\end{aligned}\right\} \tag{29}$$

In Fig. 9.9 isophote-diagrams of astigmatic images are shown. Fig. 9.9 (a) was computed from the expansion (27), with terms up to and including the fourth power of $\alpha$ taken into account. Photographs of astigmatic images are given in Fig. 9.10 and 9.11. It is seen that when only a small amount of astigmatism is present, the isophotes in the central plane are circular near the centre but have a more complex form in the outer part of the image. When the astigmatism is increased, the image has a cushion-like appearance, and is crossed by interference fringes.

As regards the two remaining primary aberrations (curvature and distortion), we have already seen that they do not affect the structure of the three-dimensional

(Fig. (a) after B. R. A. Nijboer, Thesis, University of Groningen, 1942, p. 55;
Fig. (b) after K. Nienhuis, Thesis, University of Groningen, 1948, p. 13.)

image, but only the position of the diffraction focus. The isophote diagrams in the region of focus are therefore identical with that for an aberration-free image (Fig. 8.41), but are displaced relative to the Gaussian focus by the amounts indicated in Table XXIII, p. 472.

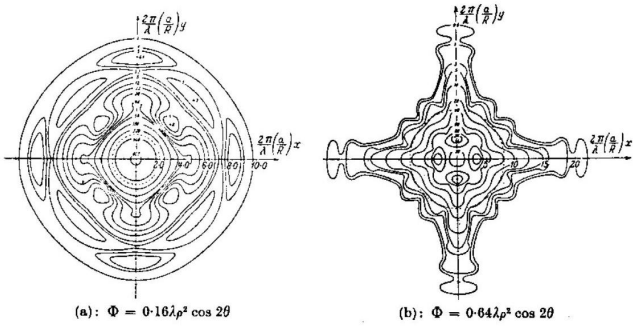



Fig. 9.9. Isophotes in the central plane in the presence of primary astigmatism. The dotted circles represent the boundary of the geometrical confusion figure. The intensity is normalized to 1000 at the centre of the aberration-free image.
$$
\text { Streml intensity: (a) } 0.84 ; (b) 0.066 .
$$

## 9.12.5 9.5. Imaging of Extended Objects

So far we have been concerned with images of point sources. We shall now describe some general methods, based on the techniques of Fourier transforms, relating to imaging of extended objects. These methods were developed chiefly by Duffieux,* partly in collaboration with lansraux, and were later extended and applied to particular problems by many writers. †

We consider separately imaging with coherent and imaging with incoherent light.

### 9.5.1 Coherent illumination

We shall specify points in the Gaussian image plane and in the plane of the exit pupil by Cartesian coordinates ( $x_1, y_1$ ) and ( $\xi, \eta$ ), respectively, referred to parallel axes with their origins at the axial points. Points in the object plane will conveniently be specified by scale normalized coordinates ( $x_0, y_0$ ) which are such that if ( $X_0, Y_0$ ) are the Cartesian coordinates of a typical object point and $M$ the lateral magnification, then

$$x_0 = MX_0, \quad y_0 = MY_0, \tag{1}$$

so that an object point and its Gaussian image point have now the same coordinate numbers.*

The imaging properties of the system may be characterized by means of a transmission function $K(x_0, y_0; x_1, y_1)$, defined as the complex amplitude, per unit area of the $x_0y_0$-plane, at the point ( $x_1, y_1$ ) in the Gaussian image plane, due to a disturbance of unit amplitude and zero phase at the object point ( $x_0, y_0$ ). The transmission function depends, of course, also on the wavelength $\lambda$ of the light, but as we shall only be concerned with monochromatic light we need not consider this dependence.

0

- P. M. Duffieux, L'Intgrale de Fourier et ses Applications l'Optique (Rennes, 1946); P. M. Duffieux and G. Lansraux, Rev. d'Optique, 24 (1945), 65, 151, 215.
  † See, for example, A. Blanc-Lapierre, Ann. de l'Inst. Henri Poincar, 13 (1953), 245; H. H. Hopkins, Proc. Roy. Soc., A, 217 (1953), 408; ibid., A, 231 (1955), 91; Proc. Phys. Soc., B, 69 (1956), 562; K. Miyamoto, Progress in Optics, Vol. I, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1961), 41.

Let $U_0(x_0, y_0)$ represent the complex disturbance in the plane of the object. The element at ( $x_0, y_0$ ) makes a contribution $dU_1(x_1, y_1) = U_0(x_0, y_0) K(x_0, y_0; x_1, y_1) dx_0 dy_0$ to the disturbance at the point $(x_1, y_1)$ in the image plane. Hence the total disturbance at ( $x_1, y_1$ ) is

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} U_0(x_0, y_0) K(x_0, y_0; x_1, y_1) dx_0 dy_0 \tag{2}$$

The integral extends only formally over an infinite domain, since $U_0 K$ is zero outside the area which does not send light into the image space of the system.

Now when we were dealing with point sources, we specified the properties of the system by the complex disturbance in the exit pupil, and this was characterized by the aberration function and by an amplitude factor, the latter being assumed to be constant in systems of moderate aperture. It is not difficult to find an expression for the transmission function in terms of these quantities. For this purpose we consider first the limiting form of (2) when the source reduces to a point source of unit strength and zero phase at the point $x_0 = x_0', y_0 = y_0'$, i.e. when

$$U_0(x_0, y_0) = \delta(x_0 - x_0') \delta(y_0 - y_0'), \tag{3}$$

where $\delta$ is the Dirac delta function [cf. Appendix IV]. Then (2) gives

$$U_1(x_1, y_1) = K(x_0', y_0'; x_1, y_1), \tag{4}$$

i.e the transmission function $K$ represents the disturbance due to the point source (3). Let us take the Gaussian reference sphere with centre at the Gaussian image point $x_1' = x_0', y_1' = y_0'$. Let $R$ be the radius of this reference sphere and let

$$H(x_0', y_0'; \xi, \eta) = \frac{i}{\lambda} G(x_0', y_0'; \xi, \eta) \frac{e^{-ikR}}{R} \tag{5}$$

be the disturbance at a typical point $(\xi, \eta)$ on this sphere, due to the point source (3). Apart from an additive factor $\pi/2$, the phase of $G$ is then the aberration function $\Phi$

- $(x_0, y_0)$ and $(x_1, y_1)$ may be regarded as the Seidel variables of § 5.2 with the choice $l_1 = C = 1$
  of the erbitrary constants.
  of the system, whilst the amplitude of $G$ is a measure of the non-uniformity in the amplitude of the image-forming wave. The factor $i/\lambda$ has been introduced on the right-hand side of (5) to simplify later formulae. Now by the Huygens-Fresnel principle, the disturbance in the image plane is related to the disturbance on the Gaussian reference sphere by the formula (small angles of diffraction assumed)

$$U_1(x_1, y_1) = -\frac{i}{\lambda} \iint H(x_0', y_0'; \xi, \eta) \frac{e^{iks}}{s} d\xi d\eta \tag{6}$$

where $s$ is the distance from the point ( $\xi, \eta$ ) of this sphere to the point ( $x_1, y_1$ ) in the Gaussian image plane, and the integral extends over the portion of the reference sphere that approximately fills the aperture. We also have according to equations (2) and (7) of §8.8, with $x = x_1 - x_0', y = y_1 - y_0', z = 0$ and with $R$ in place of $f$,

$$s \sim R - \frac{(x_1 - x_0')\xi + (y_1 - y_0')\eta}{R} \tag{7}$$

From the formulae (4)-(7) we obtain

$$K(x_0, y_0; x_1, y_1) = \frac{1}{(\lambda R)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G(x_0, y_0; \xi, \eta)\, e^{-\frac{2\pi i}{\lambda R}[(x_1 - x_0)\xi + (y_1 - y_0)\eta]} d\xi d\eta \tag{8}$$

$G$ being taken to be zero at points $(\xi, \eta)$ which are outside the opening. This is the required relation between the transmission function $K$ and the "pupil function" $G$ of the system.

Since $K$ may be regarded as the disturbance in the image of a point source, it has, when considered as function of $x_1, y_1$, a fairly sharp maximum at or near the Gaussian image point $x_1 = x_0, y_1 = y_0$ and falls off rapidly, though as a rule not monotonically, with increasing distance from this point. In a well corrected system $K$ will only be appreciable in an area whose size is of the order of the first dark ring of the Airy pattern. Considered as function of $(x_0, y_0)$, the transmission function varies slowly as this point explores the object surface. More precisely, the working field may be divided up into regions, each large compared to the finest detail that the system can resolve, with the property that in each such region $A, K$ is to a good approximation a function of the displacement vector from the Gaussian image point, but not of the position of the image point itself. For example, in a well corrected system, $K(x_0, y_0; x_1, y_1)$ represents, apart from a constant factor, the Arry pattern, centred on the Gaussian image point of $(x_0, y_0)$. In such cases we may write

$$K(x_0, y_0; x_1, y_1) = K_\Delta(x_1 - x_0, y_1 - y_0). \tag{9}$$

A region $A$ with this property is said to be an isoplanatic region of the system. We shall restrict our discussion to objects that are so small that they fall within such an isoplanatic region.* In this case the equations (2) and (8) may be replaced by

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} U_0(x_0, y_0)\, K(x_1 - x_0, y_1 - y_0)\, dx_0 dy_0 \tag{2a}$$

and

$$K(x_1 - x_0, y_1 - y_0) = \frac{1}{(\lambda R)^2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G(\xi, \eta) e^{-\frac{2\pi i}{\lambda R}\{(x_1 - x_0)\xi + (y_1 - y_0)\eta\}} d\xi d\eta \tag{8a}$$

the function $G$ now being independent of the object point.

Let us represent $U_0, U_1$, and $K$ as Fourier integrals:

$$U_0(x_0, y_0) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathscr{U}_0(f, g) e^{-2\pi i[f x_0 + g y_0]} df dg \tag{10a}$$

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathscr{U}_1(f, g) e^{-2\pi i[f x_1 + g y_1]} df dg \tag{10b}$$

$$K(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathscr{K}(f, g) e^{-2\pi i[f x + g v]} df dg \tag{10c}$$

Then, by the Fourier inversion formula

0

- A thorough discussion of the conditions under which (9) holds has been published by P. Dumontet, Optica Acta, 2 (1955), 53.

$$\mathscr{U}_0(f,g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_0(x_0,y_0)\, e^{2\pi i[fx_0+gy_0]}dx_0dy_0 \tag{11a}$$

$$\mathscr{U}_1(f,g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_1(x_1,y_1)\, e^{2\pi i[fx_1+gy_1]}dx_1dy_1 \tag{11b}$$

$$\mathscr{K}(f,g) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K(x,y)e^{2\pi i[fx+qy]}dxdy \tag{llc}$$

According to equation (2a) $U_1$ is a convolution (also called resultant or Faltung) of $U_0$ and $K$; and on Fourier inversion we obtain by the convolution theorem,* the simple relation

$$\mathscr{U}_1(f,g) = \mathscr{U}_0(f,g)\mathscr{K}(f,g) \tag{12}$$

This equation implies that if the disturbances in the object plane and in the image plane are each considered as a superposition of space-harmonic components of all possible "spatial frequencies" $f, g$, then each component of the image depends only on the corresponding component of the object, and the ratio of the components is $\mathscr{K}$. Thus the transition from the object to the image is equivalent to the action of a linear filter. Moreover, comparison of (10c) and (8a) shows that

$$\mathscr{K}\left(\frac{\xi}{\lambda R}, \frac{\eta}{\lambda R}\right) = G(\xi,\eta), \tag{13}$$

so that the frequency response function (also called transmission factor) $\mathscr{K}(f,g)$ for coherent illumination is equal to the value of the pupil function $G$ at the point

$$\xi = \lambda R f, \quad \eta = \lambda R g \tag{14}$$

of the Gaussian reference sphere.
Since $G$ is zero at points in the $\xi,\eta$-plane which are outside the boundary of the aperture, the spectral amplitudes belonging to frequencies above a certain value are not transmitted by the system. If the aperture is circular, and of radius $a$, then evidently frequency pairs such that

$$f^2 + g^2 > \left(\frac{a}{\lambda R}\right)^2 \tag{15}$$

are not transmitted. To illustrate this result consider a one-dimensional object the properties of which do not vary in the $x$ direction. Let $\Delta y_0$ be the period belonging to the frequency $g$. Then by (15) the system can only transmit information about spectral components for which

$$\Delta y_0 = \frac{1}{g} > \frac{\lambda}{\sin\theta_1} \tag{16}$$

where $\sin 0_1 \sim a/R$ is the angular semi-aperture on the image side, assumed to be small. Now $y_0 = MY_0$, where $M$ is the linear magnification, and if we assume that the system obeys the sine condition then (cf. §4.5.1) $n_0 \sin\theta_0/n_1 \sin\theta_1 = M$, and (16) may be written as

$$\Delta Y_0 > \frac{\lambda_0}{n_0 \sin\theta_0} \tag{17}$$

where $\lambda_0 = n_1\lambda$ is the vacuum wavelength and $n_0 \sin\theta_0$ is the numerical aperture of the system. Thus if the disturbance across the object plane varies sinusoidally with displacement, information about it can only be obtained in the image plane if the period exceeds the value given by the right-hand side of (17).

### 9.5.2 Incoherent illumination

We now consider the case where the light that emanates from the different elements of the object plane is incoherent, e.g. when the object is a primary source. Using the same coordinates as before, let $I_0(x_0, y_0)$ be the intensity at a typical point in the object plane. The intensity of the light that reaches the point $(x_1, y_1)$ in the plane of the image from the element $dx_0 dy_0$ centred on the object point $(x_0, y_0)$ is $dI_1(x_1, y_1) = I_0(x_0, y_0) |K(x_0, y_0; x_1, y_1)|^2 dx_0 dy_0$, where $K$ is again the transmission function of the system. Since the object is assumed to be incoherent, the intensities from the different elements of the object plane are additive, so that the total intensity at $(x_1, y_1)$ is given by

$$I_1(x_1, y_1) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I_0(x_0, y_0) |K(x_0, y_0; x_1, y_1)|^2 dx_0 dy_0 \tag{18}$$

If we again restrict ourselves to sufficiently small objects, we may replace (18) by*

$$I_1(x_1, y_1) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I_0(x_0, y_0) |K(x_1 - x_0, y_1 - y_0)|^2 dx_0 dy_0 \tag{19}$$

Equation (19) shows that for imaging with incoherent illumination the intensity distribution in the image is a convolution of the intensity distribution in the object with the squared modulus of the transmission function. We represent these functions as Fourier integrals of the form (10), and denote by $\mathscr{I}_0(f, g)$, $\mathscr{I}_1(f, g)$, and $\mathscr{L}(f, g)$

- We note that to replace (18) by (19) it is not necessary that $K$ should satisfy the full isoplanatic condition (9); it is sufficient that the modulus of $K$ alone satisfies it, i.e. that throughout the region $A$ occupied by the object, we have to a good approximation

$$|K(x_0, y_0; x_1, y_1)| = |K_A(x_1 - x_0, y_1 - y_0)|.$$

The analysis of Dumontex (loc. cit.) shows that as a rule this condition holds to a good approximstion over a considerably larger region of the object plane than the relation (9). Hence the representation of optical imaging as a linear filter has a wider range of validity for incoherent than for coherent illumination. However, to be able to express the frequency response function in terms of the pupil function of the system by a relatively simple formula ((22) below), we restrict ourselves to objects that are so small that the full isoplanatic condition holds.
their "spatial spectra". Then, by the Fourier inversion theorem, we have, in place of equations (11),

$$\mathscr{I}_0(f, g) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I_0(x_0, y_0) e^{2\pi i[fx_0 + \sigma y_0]} dx_0 dy_0 \tag{20a}$$

$$\mathscr{I}_1(f, g) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I_1(x_1, y_1) e^{2\pi i[fx_1 + \sigma y_1]} dx_1 dy_1 \tag{20b}$$

$$\mathscr{L}(f, g) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |K(x, y)|^2 e^{2\pi i[fx + \sigma y]} dx dy \tag{20c}$$

From (19) we obtain by the convolution theorem

$$\mathscr{I}_1(f, g) = \mathscr{I}_0(f, g) \mathscr{L}(f, g). \tag{21}$$

Thus the transformation from the object to the image is again a linear filter, but it is now the spatial spectrum of the intensity and not of the complex amplitude that is transformed in

this way. The frequency response function is the function $\mathscr{L}(f,g)$, and this, by (20c), (10c), and the convolution theorem, may be expressed in the form

$$\mathscr{L}(f,g)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\mathscr{K}\left(f'+f,g'+g\right)\mathscr{K}^{\star}\left(f',g'\right)df'dg' \tag{22}$$

The integral on the right-hand side of (22) is known as the auto-correlation function (of the function $\mathscr{K}$ ) and occurs in the analysis of many physical problems of statistical nature. We shall encounter it again later in connection with the theory of partial coherence.

We have already shown that $\mathscr{K}(f,g)$ is the value of the pupil function $G$ at an appropriate point of the Gaussian reference sphere. If we substitute from (13) into (22) it follows that

$$\mathscr{L}\left(\frac{\xi}{\lambda R},\frac{\eta}{\lambda R}\right)=\frac{1}{(\lambda R)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}G\left(\xi'+\xi,\eta'+\eta\right)G^{\star}\left(\xi',\eta'\right)d\xi'd\eta' \tag{23}$$

and we have thus established the important result that apart from a constant factor, the frequency response function $\mathscr{L}(f,g)$ for incoherent illumination is the auto-correlation function of the pupil function of the system.

Let $\mathscr{A}$ be the area of the exit pupil. Since the pupil function $G\left(\xi',\eta'\right)$ is zero at points outside the boundary of the aperture, the domain of the $\xi'\eta'$-plane over which the integrand of (23) does not vanish is the area common to the aperture $\mathscr{A}$ and to an identical aperture displaced relative to $\mathscr{A}$ by a translation of amounts $\xi$ and $\eta$ in the negative $\xi'$ and $\eta'$ directions respectively (see Fig. 9.12). When $\xi$ and $\eta$ are so large that the two areas do not overlap, the value of the response function is evidently zero ; thus, as in the coherent case, spatial frequencies only up to certain maximum values are transmitted by the system. In particular, for a circular aperture of radius $a$, the areas will have no region in common when $\xi^2+\eta^2\geqslant(2a)^2$, i.e. when

$$f^2+g^2>\left(\frac{2a}{\lambda R}\right)^2 \tag{24}$$

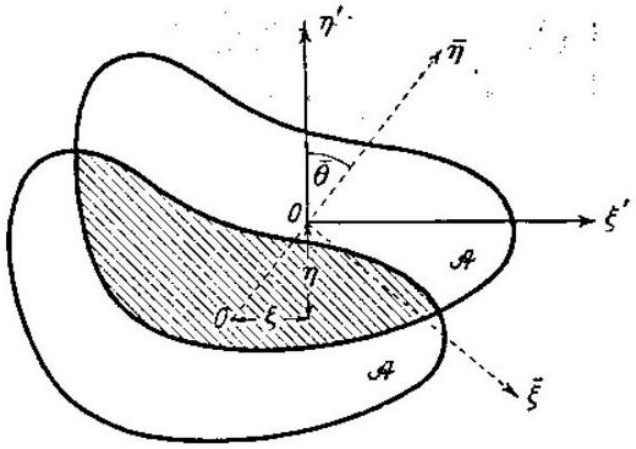



Fig. 9.12. Region of integration (shown shaded) relating to the evaluation of the response function $\mathscr{L}(f, g)$ for incoherent illumination, for the frequency pair $f=\xi/\lambda R, g=\eta/\lambda R$.

By the same argument as that leading from equation (15) to (17), this implies that, with incoherent illumination, an aplanatic system can only transmit information about the spectral components whose period $\Delta Y_0$ is such that

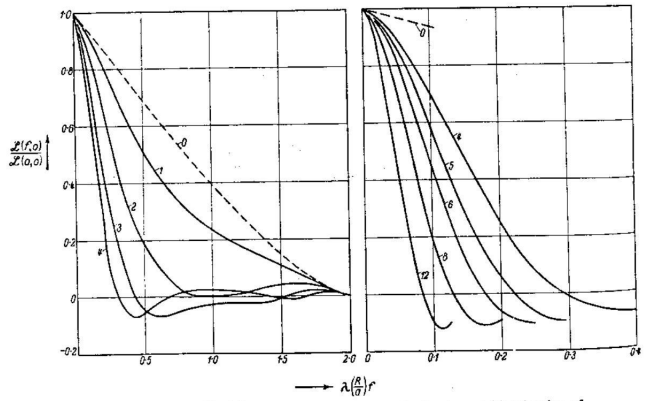


Fig. 9.13. The normalized frequency response curves for incoherent illumination of a system free of geometrical aberrations but suffering from defect of focus. $\Phi=\frac{m \lambda}{\pi} \rho\hat{}\{\}\{2\},|G|=1$. The number on each curve is the value of the parameter $m = \frac{\pi}{2\lambda}\left(\frac{a}{R}\right)^2 z, z$ being the distance between the plane of observation and the Gaussian focal plane.
(After H. H. Hopkins, Proc. Roy. Soc., A, 231 (1955), 98.)

$$\Delta Y_0 > \frac{0.5\lambda_0}{n_0 \sin\theta_0}. \tag{25}$$

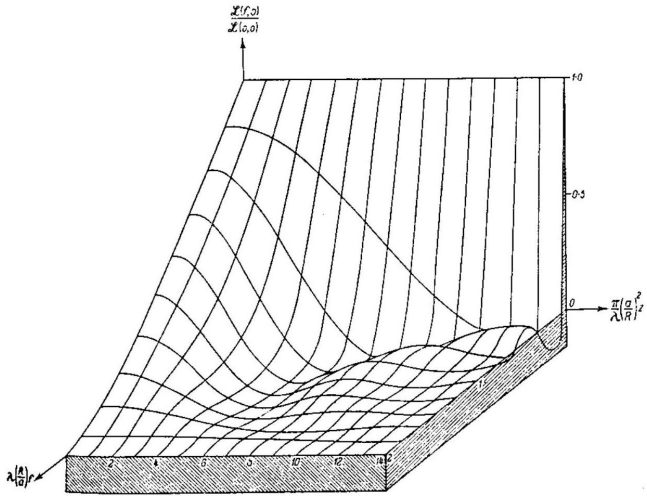


Fig. 9.14. The normalized frequency response curves for incoherent illumination of a system free of geometrical aborrations but suffering from defect of focus, as function of the spatial frequency $f$ and of defocusing $z, |G| = 1$. The curves shown in Fig. 9.13 are the sections of the surface by planes at right angles to the $z$-axis.

(After W. H. Steel, Optica Acta, 3 (1956), 67.)
It is seen that the limiting value is precisely one-half of the value obtained for imaging with coherent light.

Although the response function $\mathscr{L}$ of a given system depends on two variables $f$ and $g$, it is possible, in principle, to deduce all information about it from experiments involving one-dimensional test objects. To see this consider a frequency pair ( $f, g$ ) and introduce polar coordinates such that $f = h\sin 0, g = h\cos\theta$. Suppose now that the axes are rotated in their own plane through an angle $\bar{\theta}$ in the positive 9 -direction. Then $f$ and $g$ transform into $\bar{f} = h\sin(\theta - \vec{\theta})$ and $\bar{g} = h\cos(\theta - \overline{0})$, but the value of $\mathscr{L}$ evidently remains unchanged. Now

we may choose the angle of rotation $\bar{\theta}$ equal to $\tan^{-1} f/g$, which corresponds to taking the new $\eta$ axis ( $O\bar{\eta}$ ) along the line OO (see Fig. 9.12). Then $\bar{f} = 0$, and $\bar{g} = \sqrt{f^2 + g^2}$, and it follows that the

(After G. Black and E. H. Linfoot, Proc. Roy. Soc., A, 239 (1957), 522.)

value of the response function $\mathscr{L}$ of an optical system for the frequency pair ( $f, g$ ), is equal to that for a one-dimensional structure of frequency $\sqrt{f^2 + g^2}$ with its direction of periodicity inclined at the angle $\tan^{-1} f/g$ to the meridional plane $\theta = 0$. This result

(After M. De, Proc. Roy. Soc., A, 233 (1955), 96.)

considerably simplifies the analytical evaluation of the response functions in any given case. A similar result evidently holds in connection with the response function $\mathscr{X}(f, g)$ of coherent objects but the result is of less immediate practical interest.

Let us now consider the frequency response of a centred system that is free of aberrations, but suffers from defect of focus. It follows from the discussion of §9.1.2 that the displacement of the receiving plane by a small amount $z$ in the positive

$z$-direction is formally equivalent to the introduction of a wave aberration of amount

$$\Phi(\xi, \eta) = \frac{1}{2}\left(\frac{a}{R}\right)^2 z\rho^2, \quad \left(\rho^2 = \frac{\xi^2 + \eta^2}{a^2} \leqslant 1\right) \tag{26}$$

so that, if the amplitude of the wave is constant over the Gaussian reference sphere, the pupil function is, apart from a constant factor,

$$G(\xi, \eta) = e^{ik\Phi(\xi,\eta)} = e^{i\frac{1}{2}k\left(\frac{a}{R}\right)^2 z\rho^2}. \tag{27}$$

The response function may be determined from (23) and (27) and the results are shown in Figs. 9.13 and 9.14. It is seen that there is a very rapid deterioration of the response of the system for higher frequencies, with the introduction of a small amount of defect of focus in excess of the value corresponding to $\Phi = \frac{\lambda}{\pi}\rho^2$, i.e. in excess of $z = \frac{2\lambda}{\pi}\left(\frac{R}{a}\right)^2$.

Figs. 9.15 and 9.16 show the response curves for systems suffering from primary spherical aberration and primary astigmatism.

A survey of instruments for measuring of frequency response functions has been given by K. Murata.*

## 9.13 Interference and Diffraction with Partially Coherent Light

### 9.13.1 10.1 Introduction

So far we have been mainly concerned with monochromatic light produced by a point source. Light from a real physical source is never strictly monochromatic, since even the sharpest spectral line has a finite width. Moreover, a physical source is not a point source, but has a finite extension, consisting of very many elementary radiators (atoms). The disturbance produced by such a source may be expressed, according to Fourier's theorem, as the sum of strictly monochromatic and therefore infinitely long wave trains. The elementary monochromatic theory is essentially concerned with a single component of this Fourier representation.

In a monochromatic wave field the amplitude of the vibrations at any point $P$ is constant, while the phase varies linearly with time. This is no longer the case in a wave field produced by a real source: the amplitude and phase undergo irregular fluctuations, the rapidity of which

0

- K. Murata, Progress in Optics. Vol. 5, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1965), p. 199.

depends essentially on the effective width $\Delta\nu$ of the spectrum. The complex amplitude remains substantially constant only during a time interval $\Delta t$ which is small compared to the reciprocal of the effective spectral width $\Delta v$; in such a time interval the change of the relative phase of any two Fourier components is much less than $2\pi$ and the addition of such components represents a disturbance which in this time interval behaves like a monochromatic wave with the mean frequency; however, this is not true for a longer time interval. The characteristic time $\Delta t = 1/\Delta v$ is of the order of the coherence time introduced in §7.5.8.

Consider next the light disturbances at two points $P_1$ and $P_2$ in a wave field produced by an extended quasi-monochromatic source. For simplicity assume that the wave field is in vacuum and that $P_1$ and $P_2$ are many wavelengths away from the source. We may expect that, when $P_1$ and $P_2$ are close enough to each other, the fluctuations of the amplitudes at these points, and also the fluctuations of the phases, will not be independent. It is reasonable to suppose that, if $P_1$ and $P_2$ are so close to each other that the difference $\Delta\mathscr{S} = SP_1 - SP_2$ between the paths from each source point $S$ is small compared to the mean wavelength $\bar{\lambda}$, then the fluctuations at $P_1$ and $P_2$ will effectively be the same; and that some correlation between the fluctuations will exist even for greater separations of $P_1$ and $P_2$, provided that for all source points the path difference $\Delta\mathscr{S}$ does not exceed the coherence length $c\Delta t \sim r/\Delta v = \bar{\lambda}^2/\Delta\lambda$. We are thus led to the concept of a region of coherence around any point $P$ in a wave field.

In order to describe adequately a wave field produced by a finite polychromatic source it is evidently desirable to introduce some measure for the correlation that exists between the vibrations at different points $P_1$ and $P_2$ in the field. We must expect such a measure to be closely related to the sharpness of the interference fringes which would result on combining the vibrations from the two points. We must expect sharp fringes when the correlation is high (e.g. when the light at $P_1$ and $P_2$ comes from a very small source of a narrow spectral range), and no fringes at all in the absence of correlation (e.g. when $P_1$ and $P_2$ each receive light from a different physical source). We described these situations by the terms "coherent" and "incoherent" respectively. In general neither of these situations is realized and we may speak of vibrations which are partially coherent.

The first investigations which have a close bearing on the subject of partial coherence appear to be due to Verdet* who studied the size of the region of coherence for light from an extended primary source. Later researches by Michelson established the connection between the visibility of interference fringes and the intensity distribution on the surface of an extended primary source † (cf. § 7.3.6), and between the visibility and the energy distribution in a spectral line ‡ (cf. §7.5.8). Michelson's results were actually not interpreted in terms of correlations until much later, but his investigations have contributed to the formulation of the modern theories of partial coherence.§ The first quantitative measure of the correlation of light vibrations was introduced by von Laue, || and employed in his researches on the thermodynamics of light beams. Further contributions were made by Berek in who used the concept of correlation in investigations relating to image formation in the microscope.

A new stage in the development of the subject began with the publication of a paper by van Cittert**, who determined the joint probability distribution for the light disturbances at any two points on a screen illuminated by an extended primary source. In a later paper †† he also determined the probability distribution for the light disturbances at any one point, but at two different instants of time. Within the accuracy of his calculations the distributions were found to be Gaussian, and he determined the appropriate correlation coefficients.

A different and simpler approach to problems of partial coherence was introduced by Zernike in an important paper ‡‡ published in 1938. Zernike defined the "degree of coherence" of light vibrations in a manner that is directly related to experiment and established a number of valuable results relating to this quantity. While ZERNIKE's degree of coherence is for most practical purposes equivalent to the correlation factor of van Cittert and is also closely related

to that of von Laue, his methods seem particularly well suited for the treatment of practical problems of instrumental optics. The methods were simplified still further and applied to the study of image formation and resolving power by Hopkins.§§

The investigations mentioned so far bridged the gap between two extreme cases, namely, that of complete coherence and complete incoherence, but the

results obtained were still somewhat restricted in that they mainly applied to quasi-monochromatic light and to situations where the path differences between the interfering beams are sufficiently small. To deal with more complex situations and to formulate the theory on a rigorous basis, a further generalization was necessary. This was carried out by Wolf,* and independently by Blanc-Lapierre and Dumontet † and involved the introduction of more general correlation functions. These correlation functions were found to obey rigorously two wave equations; a result which implies that not only the optical disturbance but also the correlation between disturbances is propagated in the form of waves. In the light of this result many of the theorems established previously obtained a relatively simple interpretation.

The correlation functions mentioned so far characterize the correlation between the light vibrations at two space-time points. Such "second order" correlation functions are entirely adequate for the analysis of the usual optical experiments involving interference and diffraction of light from steady sources. ‡ For the analysis of more sophisticated experiments, higher order correlation functions-i.e. correlation functions involving the field variables in total power higher than the second-may be needed. However, so far such correlation functions have not been used to any appreciable extent.§

An attractive feature of the theory of partial coherence is the fact that it operates with quantities (namely with correlation functions and with time averaged intensities) that may, in principle, be determined from experiment. This is in contrast with the elementary optical wave theory, where the basic quantity is not measurable because of the very great rapidity of optical vibrations. In the present chapter we shall study the properties of partially coherent wave fields and we shall illustrate the results by a

---

0* E. Verdet, Ann. Scientif. l'cole Normale Suprieure, 2 (1865), 291; Leons d'Optique Physique (Paris, L'Imprimerie Impriale), 1 (1869), 106.

† A. A. Michmeson, Phil. Mag. (5), 30 (1890), 1; 31 (1891), 256; Astrophys. J., 51 (1920), 257.

‡ A. A. Michelson, Phil. Mag. (5), 31 (1891), 338; 34 (1892), 280.

§ Cf. F. Zernike, Proc. Phys. Soc., 61 (1948), 158.

|| M. von Laue, Ann. d. Physik (4), 23 (1907), 1, 795.

|| M. Berek, Z. Phys., 36 (1926), 675, 824; 37 (1926), 387; 40 (1926), 420. Experiments related to Berek's investigations were described by C. Lakeman and J. Th. Groosmuller, Physica (Gravenhage), 8 (1928), 193, 199, 305.

** P. H. van Cittert, Physica, 1 (1934), 201.

†† P.H. van Cittert, Physica, 6 (1939), 1129; see also L. Jnossy, Nuovo Cimento, 6 (1957), 111; ibid., 12 (1959), 369.

‡ F. Zernike, Physica, 5 (1938), 785.

§§ H. H. Hopkins, Proc. Roy. Soc., A, 208 (1951), 263; ibid. A, 217 (1953), 408.

Related investigations by D. Gabor and H. Gamo utilize the concept of partial coherence in the study of optical transmission from the standpoint of information theory. [D. Gabor, Proc. Symp. Astr. Optics, ed. Z. Kopal (Amsterdam, North Holland Publishing Company, 1956), 17; Proc. Third Symposium on Information Theory, ed. C. Cherry (London, Butterworths Scientific Publications, 1956), 26; H. Gamo, J. Appl. Phys. Japan, 25 (1956), 431; Progress in Optics, Vol. 3, ed. E. Wolf (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1964), 187.

0

- E. Wolf, Proc. Roy. Soc., A, 230 (1955), 246. See also ibid., A, 225 (1954), 96; Nuovo Cimento, 12 (1954), 884.

  † A. Blanc-Lapierre and P. Dumontet, Rev. d'Optique, 34 (1955), 1.

  ‡ More precisely from sources which give rise to a stationary field in the sense of the definition given on p. 498 below. For the analysis of non-stationary fields similar correlation functions may be employed, but they must be defined in terms of ensemble averages rather than time averages. (For a stationary field the

number of examples of practical interest. We shall only be concerned with the case of light, but our analysis has a close bearing on other fields; in particular, similar considerations apply in connection with correlation techniques for measurements of radio stars* and for the exploration of the ionosphere by radio waves. †

The mathematical techniques employed in connection with partial coherence are also very suitable for the analysis of partial polarization. Here one is concerned with phenomena which can be interpreted in terms of correlation between orthogonal components of the electromagnetic field vectors. Early investigations in this direction are due to G. G. Stoxes. ‡ Modern treatments which employ the concepts of correlation functions and correlation matrices are chiefly due to Wiener,§ Perrin,|| Wolf [4] and Pancharatnam.** This topic will be discussed in the concluding section (§ 10.8) of this chapter.

## 9.13.2 10.2 A Complex Representation of Real Polychromatic Fields

In discussing monochromatic wave fields we have found it useful to regard each real wave function as the real part of an associated complex wave function. In the present chapter we shall be concerned with polychromatic (i.e. non-monochromatic) fields. It will again be useful to employ a complex representation, which may be regarded as a natural generalization of that used with monochromatic fields.

Let $V^{(r)}(t)(-\infty \leqslant t \leqslant \infty)$ represent a real disturbance, for example, a Cartesian component of the electric vector, at a fixed point in space, and assume that $V^{(r)}(t)$ is square integrable. It may be expressed in the form of a Fourier integral

$$V^{(r)}(t) = \int_0^\infty a(\nu)\cos[\phi(\nu) - 2\pi\nu t]d\nu \tag{I}$$

With $V^{(r)}$ we associate the complex function

$$V(t) = \int_0^\infty a(v)e^{i[\phi(v)-2\pi vt]}dv \tag{2}$$

two types of averaging will usually give the same result.) However, since the analysis of non-stationary fields requires a considerably more involved mathematical apparatus and since no experiments relating to coherence effects with non-stationary light have so far been carried out, we restrict our discussion to stationary fields only.

§ When the light originates in $n$ thermal source, such as incandescent matter or a gas discharge, one may assume that the joint probability distribution of the field at $n$ space-time points is, to [4] good approximation, Gaussian. As is well known [see, for example J. J. Fneeman, Frinciples of Noise, New York: J. Wiley & Sons, Inc., 1958, 245-247], such distributions are completely specified by second order correlation functions, which in turn implies that all the higher order correlation functions associated with thermal light may be expressed in terms of the correlation functions of the second order. This, however, is not so for light from non-thermal sources, such as a laser.

Some higher order coherence effects and the appropriate correlation functions were briefly discussed by E. Wolf in Quantum Electronics, 3rd Congress, Vol. 1, ed. N. Bloembergen and P. Grivet (New York, Columbia University Press; Paris, Dunod, 1964), p. 13, and L. Mandel. ibid., p. 101. See also E. Wolf, Proc. Symp. on Optical Masers (New York, Brooklyn Polytechnic Press and J. Wiley & Sons, Inc., 1963), p. 29, where a systematic classification of coherence effects is outlined.

Analogous quantum mechanical correlation functions were introduced by R. J. Glauber, in Qumtum Electronics, 3rd Congress, Vol. 1, ed. N. Blombergen and P. Grivet (New York, Columbia Unversity Press; Paris, Dunod, 1964), p. 111 and Phys. Rev., 130 (1963), 2529. The relation tretween the elassical and quantum treatments is discussed by E.C.G. Sudarshan, Phys. Rev. Lett., 10 (1963), 277 and by J. R. Klauder and E. C. G. Sudarshan, Fundamentals of Quantum Optics (New York, W. A. Benjamin, Inc., 1968). See also L. Mandel and E. Wolf, Rev. Mud. Phys., 37 (1965), 231; this article also contains a review of second and higher order coherence effects.

Then

$$V(t) = V^{(r)}(t) + iV^{(i)}(t), \tag{3}$$

where

$$V^{(i)}(t) = \int_0^\infty a(\nu)\sin[\phi(\nu) - 2\pi\nu t]d\nu \tag{4}$$

The functions $V^{(i)}(t)$ and $V(t)$ are uniquely specified by $V^{(r)}(t), V^{(i)}$ being obtained from $V^{(r)}$ by replacing the phase $\phi(\nu)$ of each Fourier component by $\phi(\nu) - \pi/2$. The

integrals (1) and (4) are said to be allied Fourier integrals, or associated functions (also called conjugate functions) and may be shown* to be Hilbert transforms of each other, i.e.

$$V^{(i)}(t) = \frac{1}{\pi}P\int_{-\infty}^{\infty}\frac{V^{(r)}(t')}{t'-t}dt', \quad V^{(r)}(t) = -\frac{1}{\pi}P\int_{-\infty}^{\infty}\frac{V^{(i)}(t')}{t'-t}dt' \tag{5}$$

where $P$ denotes the Cauchy principal value at $t' = t$.
This complex representation is used frequently in communication theory, where $V$ is called the analytic signal † belonging to $V^{(r)}$. The name derives from the fact that, provided $V^{(r)}$ satisfies certain general regularity conditions, the function $V(z)$, considered as a function of a complex variable $z$ is analytic in the lower half of the $z$-plane. ‡

For future use we note the transition from $V^{(r)}$ to $V$ when $V^{(r)}$ is represented as a Fourier integral of the form

$$V^{(r)}(t) = \int_{-\infty}^{\infty} v(v)e^{-2xiv}dv \tag{6}$$

Since $V^{(r)}$ is real,

$$v(-\nu) = v^\star(\nu). \tag{7}$$

Using (7), we may re-write (6) in the form (1) and we obtain on comparison

$$v(\nu) = \frac{1}{2}a(\nu)e^{i\phi(\nu)}, \quad \nu \geqslant 0, \tag{8}$$

In terms of $v$, (2) becomes

$$V(t) = 2\int_0^\infty v(v)e^{-2\pi ivt}dv \tag{9}$$

---

0

- Cf. R. N. Bracewell, Radio Astronomy Techniques in Encyclopedia of Physice, ed. S. Flgge (Berlin, Springer), 54 (1959), Chapter V.
  † Cf. J. A. Ratcliffe, Rep. Progr. Phys. (London, Physical Society), 19 (1956). 188.
  ‡ G. G. Stokes, Trans. Cambr. Phil. Soc., 9 (1852), 399; reprinted in his Mathematical and Physical Papers, Vol. III (Cambridge University Press, 1901), 233; see also P. Soleilet, Ann. de Physique (10), 12 (1929), 23.
  § N. Wiener, J. Math. and Phys., 7 (1928), 109; J. Franklin Inst., 207 (1929), 525; Acta Math., 55 (1930), § 9, 182.
  || F. Perrin, J. Chem. Phys., 10 (1942), 415.
  |T E. Wolf, Nuovo Cimento, 12 (1954), 884; Proc. Symp. Astr. Optics, ed. Z. Kopal (Ansterdam. North Holland Publishing Company, 1956), 177; Nuovo Cimento, 13 (1959), 1165.
  ** S. Pancharatnam, Proc. Ind. Acad. Sci., A, 44 (1956), 398; ibid., 57 (1963), 218, 231.

Hence $V(t)$ may be derived from $V^{(r)}(t)$ by representing $V^{(r)}$ as a Fourier integral of the form (6), suppressing the amplitudes belonging to the negative frequencies, and multiplying the amplitudes of the positive frequencies by two. For this reason $V$ is also called the complex half-range function associated with $V^{(r)}$. Conversely it is evident that if the Fourier spectrum of a complex function $V$ contains no amplitudes belonging to negative frequencies, then the real and imaginary parts of $V$ are associated

functions. We note the following relations which follow from (6), (7), and (9), by Parseval's theorem and by the use of the relation (3):

$$\int_{-\infty}^{\infty} V^{(r)^1}(t)dt = \int_{-\infty}^{\infty} V^{(i)}(t)dt = \frac{1}{2}\int_{-\infty}^{\infty} V(t)V^{\star}(t)dt = \int_{-\infty}^{\infty} |v(v)|^2 dv = 2\int_{0}^{\infty} |v(v)|^2 dv \quad (10)$$

In most of the applications with which we shall be concerned the spectral amplitudes will only have appreciable values in a frequency interval of width $\Delta v$ which is small compared to the mean frequency $i$. The analytic signal then has a simple interpretation. We express $V$ in the form

$$V(t) = A(t)e^{i[\Phi(t)-2\pi i]}, \quad (11)$$

where $A(\geqslant 0)$ and $\Phi$ are real. According to (9) and (11),

$$\begin{aligned} A(l)e^{i\Phi(l)} &= 2\int_{0}^{\infty} v(v)e^{-2\pi i(v-\bar{\nu})t}dv \\ &= \int_{-\bar{\nu}}^{\infty} g(\mu)e^{-2\pi i\mu t}d\mu \end{aligned} \quad (12)$$

where

$$g(\mu) = 2v(\bar{\nu}+\mu). \quad (13)$$

Now since the spectral amplitudes were assumed to differ appreciably from zero only in the neighbourhood of $v = \bar{\nu}, |g(\mu)|$ will be appreciable only near $\mu = 0$. Hence the integral (12) represents a superposition of harmonic components of low frequencies, and since $\Delta v/\bar{\nu} \leqslant 1, A(t)$ and $\Phi(t)$ will vary slowly* in comparison with $\cos 2\pi\bar{\nu}t$ and $\sin 2\pi\bar{v}t$. Since $V^{(r)}$ and $V^{(t)}$ are the real and imaginary parts of $V$, we have, in terms of $A$ and $\Phi$,

---

0

- See for example E. C. Titcimarse, Introduction to the Theory of Fourier Integrals (Oxford, Clarendon Press, 2nd ed., 1948), Chapter 5.
  † The concept of an analytic signal was introduced by D. Gabor, J. Inst. Electr. Engrs., 93 (1946), Pt. III, 429. See also V. I. Bunimovice, J. Tech. Phys. U.S.S.R., 19 (1949), 1231; J. Ville, Cables et Transmission, 2 (1948), 61 ; ibid., 4 (1950), 9; J. R. Oswald, Trans. Inst. Radio Emyrs., CT-3 (1956), 244.
  Complex functions of a real variable whose real and imaginary parts are connected by the Himent transform relations play an important role in many branches of physics and engineering. In thysics, the Hilbert transform relations are often called dispersion relations as they made their irat appearance in the theory of dispersion of light by atoms [H. A. Kramers, Atti Congr. Internaz. Fuici, Como (Sept. 1927), (V), Bologna, N. Zanichelli, 1928; see aiso J. S. Toll, Phys. Rev., 104 1195f), 1760; J. Hilgevoord, Dispersion Relations and Causal Description (Amsterdam, North Holland Publishing Company, 1960)].
  : C. E. C. Titchmargh, loc. cit., p. 128.

$$V^{(r)}(t) = A(t)\cos[\Phi(t) - 2\pi\bar{\nu}t], \qquad (14)$$
$$V^{(i)}(t) = A(t)\sin[\Phi(t) - 2\pi\bar{\nu}t].$$

These formulae express $V^{(r)}$ and $V^{(i)}$ in the form of modulated signals of carrier frequency $\bar{\nu}$, and we see that the complex analytic signal is intimately connected with the envelope of the real signal †. In terms of the analytic signal $V$, the envelope $4(t)$ and the associated phase factor $\Phi(t)$ are given by

$$\left.\begin{aligned} A(t) &= \sqrt{V^{(r)^2} + V^{(i)}} = \sqrt{VV^\star} = |V|, \\ \Phi(t) &= 2\pi\bar{v}t + \tan^{-1}\tfrac{V^{(i)}}{V^{(r)}} = 2\pi\bar{v}t + \tan^{-1}\left(i\tfrac{V^\star - V}{V^\star + V}\right) \end{aligned}\right\} \qquad (15)$$

We see that $A(t)$ is independent of the exact choice of $\tilde{\nu}$, and that $\Phi(t)$ depends on $\bar{\nu}$ only through the additive term $2\pi\hat{\nu}t$. Evidently we could have chosen in (14) any other frequency $\bar{\nu}'$ in place of $\bar{\nu}$ without affecting the value of $A$; the expression for the new phase factor would differ from that given by (15) only by having $\bar{\nu}'$ written in place of $\bar{\nu}$.

In deriving (14) and (15) we have not made use of the fact that we are dealing with

- Under these circumstances one evidently has, according to (14)

$$V^{(s)}(t) \sim V^{(r)}\left(t + \frac{1}{4\vec{v}^i}\right).$$

† The envelope properties of analytic signals were studied by L. Mander., J. Opt. Soc. Amer., 57 (1967), 613.

a narrow-band signal ( $\Delta v/\bar{v} \ll 1$ ), so that these relations hold quite generally. However, it is only when $\Delta v/\bar{\nu} <\leqslant 1$ that the concept of the envelope is useful.

We have assumed that the "disturbance" $V^{(r)}(t)$ is defined for all values of $t$. In practice the disturbance will exist only during a finite time interval $-T \leqslant t \leqslant T$, but this interval is as a rule so large compared to the physically significant time scales (the mean period $1/\bar{\nu}$ and the coherence time $1/\Delta\nu$ ) that we may idealize the situation by assuming $T \to \infty$. This idealization is mathematically desirable for reasons connected with the assumption of stationarity of the field (cf. §10.3.1). Evidently it is then also necessary to assume that the time average of the intensity (which is proportional to $V^{(r)^2}$ ) tends to a finite value as the averaging interval is indefinitely increased, i.e. that

$$\lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} V^{(r)^2}(l)dl \qquad (I6)$$

is finite. Now if this limit is finite and not zero then obviously $\int_{-\infty}^{\infty} V^{(r)^2}(t)dt$ diverges. Nevertheless.it is possible to utilize the techniques of Fourier analysis.* We define the truncated functions

$$\left.\begin{aligned} V_T^{(r)}(l) &= V^{[r]}(t) \qquad \text{when} \\ &= 0 \qquad\qquad \text{when} \end{aligned}\right. \quad |t| \leqslant T, |\ t \qquad (17)$$

Since each such truncated function may be assumed to be square integrable it may be expressed as a Fourier integral, say

$$V_T^{(r)}(t) = \int_{-\infty}^{\infty} v_T(v)e^{-2\pi ivt}dv \qquad (18a)$$

Let $V_T^{(i)}$ be the associated function and $V_T$ the corresponding analytic signal, i.e.

$$V_{T'}(t) = V_T^{(r)}(t) + iV_T^{(i)}(t) = 2\int_0^\infty v_T(\nu)e^{-2\pi i\nu t}d\nu \tag{I8b}$$

Then the relations (10) hold with $V^{(r)}$ replaced by $V_{1'}^{(r)}$, etc. Hence if we also divide each expression by $2T$, we obtain †

$$\begin{aligned}\frac{1}{2T}\int_{-\infty}^{\infty} V_T^{(r)^2}(t)dt = \frac{1}{2T}\int_{-\infty}^{\infty} V_T^{(i)^2}(t)dt = \frac{1}{2}\frac{1}{2T}\int_{-\infty}^{\infty} V_T(t)V_T^\star(t)dt \\ = \int_{-\infty}^{\infty} G_T(v)dv = 2\int_0^\infty G_T(v)dv\end{aligned} \tag{19}$$

- The problem of analysing functions of time which do not die down as $t$ tends to infinity was encountered at the turn of the twentieth century by physicists who concerned themselves with the study of the nature of white light and noise (notably L. G. Gouy, Lord Rayleioh, and A. Scristffr). Rigorous mathematical techniques were developed chiefly by N. Wiener in his paper on generalized harmonic analysis (Acta Math., 55 (1930), 117). This paper also outlines the history of the problem and includes a very full bibliography. [4] Since the Hilbert transform of a truncated function is not necessarily a truncated function, $f_r$[4] and $V_r$ do not, in general, vanish outside the range $-T \leqslant t \leqslant T$. For this reason, and ass to avoid certain mathematical refinements, the limits of the time integrations in (19) and (23) are taken as $\pm\infty$ rather than $\pm T$.
  where

$$G_T(\nu) = \frac{|v_T(\nu)|^2}{2T} \tag{20}$$

It would seem natural now to proceed to the limit $T \to \infty$. Unfortunately in many cases of practical interest, the function $G_T(\nu)$, known as the periodogram, does not tend to a limit but fluctuates* with increasing $T$. However, one may overcome this difficulty by an appropriate "smoothing procedure". For example, as customary in the theory of random processes, one regards the function $V^{(r)}(t)$ to be a typical member of an ensemble of functions which characterize the statistical properties of the process. Moreover the ensembles that one normally encounters in optics are stationary and ergodic. Stationarity implies that all the ensemble averages are independent of the origin of time, whilst ergodicity implies that each ensemble average is equal to the corresponding time average involving a typical member of the ensemble. We will from now on assume that we are dealing with a stationary ergodic ensemble. † One may then show that the average of $G_T(\nu)$ taken over the ensemble of the functions $V^{(r)}(t)$ tends to a definite limit as $T \to \infty$. Thus if bar denotes the ensemble average, the "smoothed periodogram"

$$\overline{G_T(v)} = \frac{\overline{|v_T(v)|^2}}{2T} \tag{21}$$

will possess a limit ‡

$$G(\nu) = \lim_{T\to\infty} \overline{G_T(\nu)} = \lim_{T\to\infty} \frac{\overline{|v_T(\nu)|^2}}{2T}. \tag{22}$$

Now if sharp brackets denote the time average,

$$\langle F(t)\rangle = \lim_{T\to\infty} \frac{1}{2T}\int_{-\infty}^{\infty} F_T(t)dt \tag{23}$$

then one obtains, in the limit as $T \to \infty$, the following relations analogous to (19):

$$\left\langle V^{(r)^2}(l) \right\rangle = \left\langle V^{(i)^2}(t) \right\rangle = \frac{1}{2} \langle V(t) V^\star(l) \rangle = \int_{-\infty}^{\infty} G(\nu) d\nu = 2 \int_0^{\infty} G(\nu) d\nu \tag{24}$$

In the theory of stationary random processes, the function $G(v)$, defined by (22) is called the power spectrum of the random process, characterized by the ensemble of the functions $V^{(r)}(t)$. In our considerations, where $V^{(r)}(t)$ represents the light disturbance,

$G(v)dv$ is proportional to the contribution to the intensity from the frequency range $(\nu, \nu + d\nu)$; we shall refer to $G(\nu)$ as the spectral density of the light vibrations.

Since $V_T^{(r)}$ is the real part of $V_T$ it follows that, when the operations on $V_T^{(r)}$ are linear, we may operate directly with $V_T$ and take the real part at the end of the calculations. Moreover, just as in the monochromatic case, the relation $\left\langle V^{(r)^*} \right\rangle = \frac{1}{2} \langle V V^* \rangle$ allows us to calculate the time average of the square of the real disturbance directly in terms of the complex disturbance which we have associated with it.

## 9.13.3 10.3 The Correlation Functions of Light Beams

### 10.3.1 Interference of two partially coherent beams. The mutual coherence function and the complex degree of coherence

We have indicated in § 10.1 that for a satisfactory treatment of problems involving light from a finite source and with a finite spectral range it is necessary to specify

the correlation that may exist between the vibrations at two arbitrary points in the wave field. A suitable measure of this correlation is suggested by the analysis of a two-beam interference experiment.

Consider the wave field produced by an extended polychromatic source $\sigma$. For the present we neglect polarization effects, so that we may regard the light disturbance as a real scalar function $V^{(r)}(P,t)$ of position and time. With $V^{(r)}(P,t)$ we associate the analytic signal $V(P,t)$. By observation it is, of course, impossible to determine how these quantities vary with time, since any detector will only record averages over time intervals during which the disturbance will have changed sign very many times. The observable intensity $I(P)$ is proportional to the mean value of $V^{(r)}(P,t)$, so that, apart from an inessential constant,

$$I(P) = 2 \left\langle V^{(r)^2}(P,t) \right\rangle = \langle V(P,t) V^\star(P,t) \rangle \,, \tag{1}$$

0

- See, for example, W. B. Davenport and W. L. Root, An Introduction to the Theory of Random Signals and Noise (New York, McGraw-Hill, 1958), p. 107-108. See also D. Middleton, IRE Trans., CT-3 (1956), 299.

† For a fuller discussion of these concepts, see, for example, W. B. Davenport and W. L. Root, loc. cit.; S. Goldman, Information Theory (New York, Prentice-Hall, Inc. 1953); D. Middleton, An Introduction to Statistical Communication Theory (New York, McGraw-Hill Co., 1960); A. M. Yaglom, An Introduction to the Theory of Stationary Random Functions (Englewood Cliffs, N. J., Prentice-Hall, 1962).

‡ A rigorous proof of the existence of this limit and of some of the relations introduced heuristically in this section would lead us far into ergodic theory and cannot, therefore, be given here. Cf. J. L. Doob, Stochastic Processes (New York, J. Wiley & Sons, Inc., 1953), Chapt. XI; sce also S. Goldman, loc. cit., §8.4, A. M. Yaglom, loc. cit., 43-51 or D. Middleton, loc. cit., § 3.2.

Instead of taking the ensemble average, other smoothing operations may be used [cf. S. Goldman, loc. cit., p. 244 or M. S. Bartlett, An Introduction to Stochastic Processes (Cambridge, University Press, 1955), p. 280-284].

Another way of defining the power spectrum $G(\nu)$, which does not depend on the concept of an ensemble, is indicated in the first footnote on p. 504.

where the relation § 10.2 (24) has been used.
Consider now two points $P_1$ and $P_2$ in the wave field. In addition to measuring the intensities $I(P_1)$ and $I(P_2)$ we may also determine experimentally the interference effects arising on superposition of the vibrations from these points. For this purpose imagine an opaque screen $\mathscr{A}$ to be placed across the field with pinholes at $P_1$ and $P_2$, and consider the intensity distribution on a second screen $\mathscr{B}$ placed some distance from $A$, on the side opposite the source (Fig. 10.1). For simplicity we assume that the medium between the two screens has refractive index unity. Let $s_1$ and $s_2$ be the distances of a typical point $Q$ on the screen $\mathscr{B}$ from $P_1$ and $P_2$. We may regard $P_1$
and $P_2$ as centres of secondary disturbances, so that the complex disturbance at $Q$ is given by

$$V(Q,t) = K_1 V(P_1, t-t_1) + K_2 V(P_2, t-t_2). \tag{2}$$

Here $t_1$ and $t_2$ are the times needed for light to travel from $P_1$ to $Q$ and from $P_2$ to $Q$ respectively, i.e.

$$t_1 = \frac{s_1}{c}, \quad t_2 = \frac{s_2}{c}, \tag{3}$$

where $c$ is the velocity of light in vacuum. The factors $K_1$ and $K_2$ are inversely proportional to $s_1$ and $s_2$, and depend also on the size of the openings and on the geometry of the arrangement (the angles of incidence and diffraction at $P_1$ and $P_2$ ). Since the secondary wavelets from $P_1$ and $P_2$ are out of phase with the primary wave by a quarter of a period (cf. § 8.2, § 8.3), $K_1$ and $K_2$ are pure imaginary numbers.

It follows from (1) and (2) that the intensity at $Q$ is given by*

$$\begin{aligned} I(Q) = K_1 K_1^\star \langle V_1(\ t-t_1)\, V_1^\star(t-t_1)\rangle + K_2 K_2^\star \langle V_2(t-t_2)\, V_2^\star(t-t_2)\rangle \\ + K_1 K_2^\star \langle V_1(t-t_1)\, V_2^\star(t-t_2)\rangle + K_2 K_1^\star \langle V_2(t-t_2)\, V_1^\star(t-t_1)\rangle . \end{aligned} \tag{4}$$

Now the field was assumed to be stationary. We may shift the origin of time in all these expressions, and we have, therefore,

$$\langle V_1(t-t_1)\, V_1^\star(t-t_1)\rangle = \langle V_1(t) V_1^\star(t)\rangle = I_1, \tag{5}$$

and similarly for the other terms. If we also use (3), and remember that $K_1$ and $K_2$ are pure imaginary numbers, (4) may be simplified to give

$$I(Q) = |K_1|^2 I_1 + |K_2|^2 I_2 + 2|K_1 K_2| \Gamma_{12}^{(r)}\left(\frac{s_2 - s_1}{c}\right) \tag{6}$$

where $\Gamma_{12}^{(r)}(\tau)$ is the real part of the function

$$\Gamma_{12}(\tau) = \langle V_1(t+\tau) V_2^\star(t)\rangle . \tag{7}$$

The quantity represented by (7) is basic for the theory of partial coherence. We shall call it the mutual coherence of the light vibrations at $P_1$ and $P_2$, the vibrations at $P_1$ being considered at time $\tau$ later than at $P_2$; and we shall call $\Gamma_{12}(\tau)$ the mutual coherence function † of the wave field. When the two points coincide ( $P_1 = P_2$ ) we obtain

$$\Gamma_{11}(\tau) = \langle V_1(t+\tau) V_1^\star(t)\rangle \tag{8}$$

and we then speak of the self-coherence of the light vibrations at $P_1$; it reduces to ordinary intensity when $\tau = 0$ :

$$\Gamma_{11}(0) = I_1 \quad \Gamma_{22}(0) = I_2.$$

The term $|K_1|^2 I_1$ in (6) is evidently the intensity which would be observed at $Q$ if the pinhole at $P_1$ alone were open ( $K_2 = 0$ ) and the term $|K_2|^2 I_2$ has a similar interpretation. Let us denote these intensities by $I^{(1)}(Q)$ and $I^{(2)}(Q)$ respectively, i.e.,

$$I^{(1)}(Q) = |K_1|^2 I_1 = |K_1|^2 \Gamma_{11}(0), \quad I^{(2)}(Q) = |K_2|^2 I_2 = |K_2|^2 \Gamma_{22}(0). \tag{9}$$

We also normalize $\Gamma_{12}(\tau)$ :

$$\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}} = \frac{\Gamma_{12}(\tau)}{\sqrt{I_1}\sqrt{I_2}}. \tag{10}$$

For reasons which will become apparent shortly, $\gamma_{12}(\tau)$ will be called the complex degree of coherence of the light vibrations. With the aid of (9) and (10), the formula (6) may finally be written in the form

$$I(Q) = I^{(1)}(Q) + I^{(2)}(Q) + 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\gamma_{12}^{(r)}\left(\frac{s_2 - s_1}{c}\right), \tag{11}$$

where $\gamma_{12}^{(r)}$ denotes the real part of $\gamma_{12}$.
The formula (11) is the general interference law for stationary optical fields. It shows that, in order to determine the intensity arising from the superposition of two beams of light, we must know the intensity of each beam and the value of the real part $\gamma_{12}^{(r)}$ of the complex degree of coherence. We shall show later how $\gamma_{12}^{(r)}$ may be calculated from data that specify the source and the transmission properties of the medium.

If the light from $P_1$ and $P_2$ does not reach $Q$ directly, but via an intervening optical system, and if dispersion effects are negligible, (11) retains its validity provided that $s_2 - s_1$ is replaced by the path difference $P_2Q - P_1Q$. With this generalization the formula (11) also holds when the two interfering beams are derived from a primary beam, not by "wave-front division" at $P_1$ and $P_2$, but by "amplitude division" in the immediate neighbourhood of a single point $P_1$, for example in a Michelson interferometer. In this latter case equation (II) will involve $\gamma_{11}^{(\gamma)}(\tau)$ in place of $\gamma_{12}^{(\gamma)}(\tau)$.

Unlike the disturbance $V^{(\tau)}$, the correlation functions $\gamma_{12}^{(\tau)}$ and $\Gamma_{12}^{(\gamma)}$ represent quantities which can be determined from experiment. To find the value of $\gamma_{12}^{(\%)}$ for any prescribed pair of points $P_1$ and $P_2$ and for any prescribed value of $\tau$, one places an opaque screen across the light beam, with pinholes at $P_1$ and $P_2$, as in Fig. 10.1. One then measures the intensity $I(Q)$ at a point $Q$ behind the screen, such that $P_2Q - P_1Q = c\tau$. Next one measures the intensities $I^{(1)}(Q)$ and $I^{(2)}(Q)$ of the light from each pinhole separately. In terms of these three observed values, $\gamma_{12}^{(r)}$ is, according to (11), given by

$$\gamma_{12}^{(r)} = \frac{I(Q) - I^{(1)}(Q) - I^{(2)}(Q)}{2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}}. \tag{12}$$

To determine $\Gamma_{12}^{(q)}$ one must also measure the intensities $I(P_1)$ and $I(P_2)$ at each pinhole. According to (10) and (12) $\Gamma_{12}^{(p)}$ is then given by

0

- From now on, where convenient, we employ a shortened notation, writing $V_1(t)$ in place of $V(P_1, t), \Gamma_{12}(\tau)$ in piace of $\Gamma(P_1, P_2, \tau)$, etc.
  † In the general theory of atationary random processes $\Gamma_{12}(\tau)$ is called the cross-correlation function of $V_1(t)$ and $V_2(t)$ and $\Gamma_{11}(\tau)$ the auto-correlation function of $V_1(t)$.

$$\Gamma_{12}^{(\gamma)} = \sqrt{I\left(P_1\right)}\sqrt{I\left(P_2\right)}\gamma_{12}^{(\gamma)} = \frac{1}{2}\sqrt{\frac{I\left(P_1\right)I\left(P_2\right)}{I^{(1)}(Q)I^{(2)}(Q)}}\left\{I(Q) - I^{(1)}(Q) - I^{(2)}(Q)\right\}. \tag{13}$$

Returning to (10), it is not difficult to see that our normalization ensures that $|\gamma_{12}(\tau)| \leqslant 1$. To show this we introduce, as in § 10.2 (17), the truncated functions

$$\left.\begin{array}{lll} V_T^{(r)}(P,t) & = V^{(r)}(P,t) & \text{when} \\ & = 0 & \text{when} \\ & |t| > T & \end{array}\right\} \tag{14}$$

and denote by $V_T(P,t)$ the associated analytic signal. By the Schwarz inequality*

$$\left|\int_{-\infty}^{\infty} V_T\left(P_1, t+\tau\right) V_T^{\star}\left(P_2, t\right) dt\right|^2 \leqslant \int_{-\infty}^{\infty} V_T\left(P_1, t+\tau\right) V_T^{\star}\left(P_1, t+\tau\right) dt \int_{-\infty}^{\infty} V_T\left(P_2, t\right) V_T^{\star}\left(P_2, t\right) dt \tag{15}$$

In the first integral on the right we may replace $t+\tau$ by $t$. Then dividing both sides by $4T^2$, and proceeding to the limit $T \to \infty$, it follows that
or, by (10),

$$\left|\Gamma_{12}(\tau)\right|^2 \leqslant \Gamma_{11}(0)\Gamma_{22}(0) \tag{16}$$

$$\left|\gamma_{12}(\tau)\right| \leqslant 1 \tag{17}$$

The significance of $\gamma_{12}$ may best be seen by expressing (II) in a somewhat different form. Let $\bar{\nu}$ be a mean frequency of the light and write
where

$$\gamma_{12}(\tau) = \left|\gamma_{12}(\tau)\right| e^{i(\alpha_{12}(\tau) - 2\pi i\tau]} \tag{18}$$

$$\alpha_{12}(\tau) = 2\pi\bar{\tau}\tau + \arg\gamma_{12}(\tau) \tag{19}$$

Then (11) becomes

$$I(Q) = I^{(1)}(Q) + I^{(2)}(Q) + 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\left|\gamma_{12}(\tau)\right|\cos\left[\alpha_{12}(\tau) - \delta\right] \tag{20}$$

where the parameter $\tau$ and the phase difference $\delta$ have the values

$$\tau = \frac{s_2 - s_1}{c}, \quad \delta = 2\pi i\tau = \frac{2\pi}{\bar{\lambda}}\left(s_2 - s_1\right), \tag{21}$$

and $\bar{\lambda}$ is the mean wavelength. If $|\gamma_{12}(\tau)|$ has the extreme value unity, the intensity at $Q$ is the same as would be obtained with strictly monochromatic light of wavelength $\bar{\lambda}$, and with the phase difference between the vibrations at $P_1$ and $P_2$ equal to $\alpha_{12}(\tau)$. In this case the vibrations at $P_1$ and $P_2$ (with the appropriate time delay $\tau$ between them) may be said to be coherent. † If $\gamma_{12}(\tau)$ has the other extreme value, namely zero, the last term in (20) is absent; the beams do not give rise to any interference effects and the vibrations may then be said to be incoherent. If $|\gamma_{12}(\tau)|$ has neither of the two extreme values, i.e. if $0 < |\gamma_{12}(\tau)| < 1$, the vibrations are said to be partiully coherent $|\gamma_{12}(\tau)|$ representing their degree of coherence. ‡

Whatever the value of $|\gamma_{12}|$, the intensity $I(Q)$ may also be expressed in the form

$$I(Q) = |\gamma_{12}(\tau)| \left\{ I^{(1)}(Q) + I^{(2)}(Q) + 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)} \cos\left[\alpha_{12}(\tau) - \delta\right] \right\} + \left\{1 - |\gamma_{12}(\tau)|\right\} \left\{ I^{(1)}(Q) + I^{(2)}(Q) \right\}. \quad (2{:}2)$$

The terms in the first line may be considered to arise from coherent superposition of two beams of intensities $|\gamma_{12}(\tau)|\, I^{(1)}(Q)$ and $|\gamma_{12}(\tau)|\, I^{(2)}(Q)$ and of relative phase difference $\alpha_{12}(\tau) - \delta$; those in the second line from incoherent superposition of two beams

of intensities $\left[1 - |\gamma_{12}(\tau)|\, I^{(1)}(Q)\right.$ and $\left[1 - |\gamma_{12}(\tau)|\right] I^{(2)}(Q)$. Thus the light which reaches $Q$ from both pinholes may be regarded to be a mixture of coherent and incoherent light, with intensities in the ratio
or

$$\frac{I_{\text{coh}}}{I_{\text{incoh}}} = \frac{|\gamma_{12}(\tau)|}{1 - |\gamma_{12}(\tau)|} \quad (23a)$$

$$\frac{I_{\text{coh}}}{I_{\text{tot}}} = |\gamma_{12}|\,, \quad (I_{\text{tot}} = I_{\text{coh}} + I_{\text{incoh}})\,. \quad (23b)$$

We have seen (equation (12)) that $\gamma_{12}^{(\gamma)}$ may be determined from intensity measurements in an appropriate interference experiment. In § 10.4.1 we shall see that in most cases of practical interest the modulus (and in principle also the phase) of $\gamma_{12}$ can likewise be determined from such experiments.

### 10.3.2 Spectral representation of mutual coherence

Let

$$V_T^{(r)}(P,t) = \int_{-\infty}^{\infty} v_T(P,v) e^{-2\pi i r} dv \quad (24)$$

be the Fourier integral representation of the truncated real function $V_T^{(r)}$. Then by the Fourier inversion formula

$$v_T(P,\nu) = \int_{-\infty}^{\infty} V_T^{(r)}(P,t) e^{2\pi i \nu t} dt \quad (25)$$

and it follows that

$$\begin{aligned} \int_{-\infty}^{\infty} V_T^{(r)}(P_1, t+\tau)\, V_T^{(r)}(P_2,t)\, dt &= \int_{-\infty}^{\infty} V_T^{(r)}(P_2,t) \left[ \int_{-\infty}^{\infty} v_T(P_1,v)\, e^{-2\pi i v(t+r)} dv \right] dt \\ &= \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} V_T^{(r)}(P_2,t)\, e^{-2\pi i v t} dt \right] v_T(P_1,v)\, e^{-2\pi i v r} dv \\ &= \int_{-\infty}^{\infty} v_T(P_1,v)\, v_T^{\star}(P_2,v)\, e^{-2\sin r} dv \end{aligned} \quad (26)$$

0

- See, for example, H. Margenau and G. M. Murphy, The Mathematics of Physics and Chemistry (New York, D. van Nostrand Co., 1947), p. 131.
  † General properties of coherent light have been investigated by I. Mandel and E. Wolf, J. Opt. Soc. Amer., 51 (1981), 815.
  ‡ Various methods for measuring the degree of coherence are discussed by M. Franon and S. Mallick in Progress in Optics. Vol. 6. ed. E. Wolf (Amsterdam, North-Holland Publishing Company and Nru York, J. Wiley, 1967), p. 71.

Next we divide both sides of (26) by $2T$ and apply to the quantity $v_T(P_1, v)\, v_T^{\star}(P_2, v)/2T$ a "smoothing operation" such as taking of the ensemble average (denoted by a bar) over the ensemble of the random functions $V^{(r)}$ as explained earlier in connection with eq. (20) of §10.2. Finally proceeding to the limit $T \to \infty$ one may then expect that*

$$\left\langle V^{(r)}(P_1, t+\tau)\, V^{(r)}(P_2, t)\right\rangle = \int_{-\infty}^{\infty} G_{12}(v) e^{-2\pi i v r} dv, \tag{27}$$

where

$$G_{12}(\nu) = \lim_{T\to\infty}\left[\overline{\frac{v_T(P_1, v)\, v_T^{\star}(P_2, v)}{2T}}\right]. \tag{28}$$

The function $G_{12}(\nu)$ may be called the mutual spectral density, or the cross-spectral density, of the light vibrations at $P_1$ and $P_2$. It is a generalization of the spectral density introduced earlier [§ 10.2 (22)] and reduces to it when the two points coincide. The mutual spectral density is the optical analogue of the concept of cross-power spectrum in the theory of stationary random processes. Equation (27) shows that the real correlation function $\left\langle V^{(r)}(P_1, t+\tau)\, V^{(r)}(P, t)\right\rangle$ and the mutual spectral density $G_{12}(v)$ form a Fourier transform pair.*

Let us now pass to the complex representation. Let

$$V(P, t) = 2\int_0^{\infty} v(v) e^{-2\pi i v t} dv \tag{29}$$

be the analytic signal (cf. § 10.2) associated with $V^{(r)}(P, t)$. It follows by an analysis similar to that which leads from (24) to (27), that

$$\Gamma_{12}(\tau) = \left\langle V(P_1, t+\tau)\, V^{\star}(P_2, t)\right\rangle = 4\int_0^{\infty} G_{12}(v) e^{-2\pi i r \tau} dv \tag{30}$$

Since $\Gamma_{12}$ does not contain spectral components belonging to negative frequencies, it is an analytic signal. Hence if $\Gamma_{12}^{(r)}$ and $\Gamma_{12}^{(i)}$ denote its real and imaginary parts, i.e.

$$\Gamma_{12}(\tau) = \Gamma_{12}^{(r)}(\tau) + i\Gamma_{12}^{(i)}(\tau) \tag{31}$$

these functions are connected by the Hilbert transform relations

$$\Gamma_{12}^{(i)}(\tau) = \frac{1}{\pi} P\int_{-\infty}^{\infty} \frac{\Gamma_{12}^{(r)}(\tau')}{\tau' - \tau} d\tau', \quad \Gamma_{12}^{(\tau)}(\tau) = -\frac{1}{\pi} P\int_{-\infty}^{\infty} \frac{\Gamma_{12}^{(i)}(\tau')}{\tau' - \tau} d\tau' \tag{32}$$

It follows [cf. § 10.2 (11)-(15)] that $|\Gamma_{12}|$, considered as a function of $\tau$, is the envelope of $\Gamma_{12}^{(\gamma)}$; and from (30), (31), and (27) it follows that †

$$\Gamma_{12}^{(r)}(\tau) = 2\left\langle V^{(r)}(P_1, t+\tau)\, V^{(r)}(P_2, t)\right\rangle = 2\int_{-\infty}^{\infty} G_{12}(v) e^{-2\pi i r \tau} dv \tag{33}$$

Furthermore $|\gamma_{12}|$ is the envelope of the real correlation factor

$$\gamma_{12}^{(r)}(\tau) = \frac{\Gamma_{12}^{(r)}(\tau)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}} = \frac{\left\langle V^{(r)}(P_1, t+\tau)\, V^{(r)}(P_2, t)\right\rangle}{\sqrt{\left\langle V^{(r)2}(P_1, t)\right\rangle}\sqrt{\left\langle \bar{V}^{(r)2}(P_2, t)\right\rangle}} \tag{34}$$

0

- Similar remarks apply here as those made in footnote ‡ on p. 498.

0

Equation (30) gives the spectral representation of the mutual coherence function $\Gamma_{12}(\tau)$. Equation (33) shows that the real part of $\Gamma_{12}(\tau)$ is equal to twice the crosscorrelation function of the real functions $V^{(r)}(P_1,t)$ and $V^{(r)}(P_2,t)$, and (32) gives the connection between the real and imaginary parts of $\Gamma_{12}(\tau)$.

## 9.13.4 10.4 Interference and Diffraction with Quasi-monochromatic Light

We have seen that, in order to describe adequately interference with partially coherent light, it is in general necessary to know the mutual coherence function $\Gamma_{12}(\tau)$ or, what amounts to the same thing, the ordinary intensities $I_1$ and $I_2$ and the complex degree of coherence $\gamma_{12}(\tau)$. We shall now restrict ourselves to the important case of quasi-monochromatic light, i.e. light consisting of spectral components that cover a frequency range $\Delta\nu$ which is small compared to the mean frequency $\bar{\nu}$. We shall see that the theory takes a simpler form in this case. In particular, we shall find that, under a certain additional assumption which is satisfied in many applications, it is possible to employ in place of $\Gamma_{12}(\tau)$ and $\gamma_{12}(\tau)$ correlation functions which are independent of the parameter $\tau$.

### 10.4.1 Interference with quasi-monochromatic light. The mutual intensity

Let us again consider the interference experiment illustrated in Fig. 10.1. According to equation (20) of §10.3, the intensity at a point $Q$ in the interference pattern is given by

$$I(Q) = I^{(1)}(Q) + I^{(2)}(Q) + 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\,|\gamma_{12}(\tau)|\cos\left[x_{12}(\tau) - \delta\right] \tag{1}$$

where

$$\tau = \frac{s_2 - s_1}{c}, \quad \delta = 2\pi\dot{\nu}\tau = \frac{2\pi}{\bar{\lambda}}\left(s_2 - s_1\right). \tag{2}$$

Suppose now that the light is quasi-monochromatic. Then it follows from equation (18), § 10.3 in the same way as in connection with equation (11) of § 10.2, that $|\ \gamma_{12}(\tau)\ |$ and $\alpha_{12}(\tau)$, considered as functions of $\tau$, will change slowly in comparison to $\cos 2\pi\bar{\nu}\tau$ and $\sin 2\pi\dot{\nu}\tau$. Moreover, if the openings at $P_1$ and $P_2$ are sufficiently small, the intensities $I^{(1)}(Q)$ and $I^{(2)}(Q)$ of the light diffracted from each opening separately will remain sensibly constant throughout a region of the pattern in which $\cos 2\pi\bar{\nu}\tau$ and $\sin 2\pi\bar{\nu}\tau$ change sign many times. It follows that the intensity distribution in the vicinity of any point $Q$ consist of an almost uniform background $I^{(1)}(Q)+I^{(2)}(Q)$ on which a sinusoidal intensity distribution is superimposed, with almost constant amplitude $2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\,|\gamma_{12}(\tau)|$. The behaviour of the total intensity distribution is shown for three typical cases in Fig. 10.2. The intensity maxima and minima near $Q$ are to a good approximation given by

$$\left.\begin{aligned} I_{\max} &= I^{(1)}(Q) + I^{(2)}(Q) + 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\,|\gamma_{12}(\tau)|\,, \\ I_{\min} &= I^{(1)}(Q) + I^{(2)}(Q) - 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\,|\gamma_{12}(\tau)|\,. \end{aligned}\right\} \tag{3}$$

• In the case when $P_1 = P_2$ this result is the optical equivalent of the well-known WIENERKhintchine theorem. [N. Wiener, Acta Math., 55 (1930), 117; A. Khintchine, Math. Ann., 109 (1934), 604.]

Instead of using a smoothing procedure one may use the Fourier inverse of (27) to define the mutual spectral density $G_{12}(\nu)$. This alternative approach is entirely adequate for the purposes of the main part of this chapter.
† It is not difficult to show that ${\Gamma_{12}}^{(r)}(r)$ is also equal to $2\left\langle V^{(1)}(P_1,t+7)\,V^{(1)}(P_2,t)\right\rangle$ and $\Gamma_1^{(t)}(\tau) = 2\left\langle V^{(0)}(P_1,t+\tau)\,V^{(r)}(P_2,t)\right\rangle = -2\left\langle V^{(r)}(P_1,t+\tau)\,V^{(1)}(P_2,t)\right\rangle$. [See, for example, P. Roman and E. Wolf, Nuovo Cimento, 17 (1960), 474-476 or L. Mandel, Progress in Optics, Vol. 2. ed. E. Worr (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1963), 241-242.]

Hence the visibility of the fringes at $Q$ is

$$\mathscr{V}(Q) = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}}{I^{(1)}(Q) + I^{(2)}(Q)} \left|\gamma_{12}(\cdot)\right|. \tag{4}$$

This formula expresses the visibility of the fringes in terms of the intensity of the two beams and of their degree of coherence. If, as is often the case, the two beams are of equal intensity $\left[I^{(1)} = I^{(2)}\right]$, (4) reduces to

$$\mathscr{V}(Q) = \left|\gamma_{12}(\tau)\right|, \tag{5}$$

i.e. the visibility of the fringes is then equal to the degree of coherence.

According to (1) and (2), the positions of the maxima of intensity near $Q$ are given by

$$\frac{2\pi}{\bar{\lambda}}(s_2 - s_1) - \alpha_{12}(\tau) = 2m\pi, \quad (m = 0, \pm 1, \pm 2, \ldots),$$

just as if the opening were illuminated with strictly monochromatic light of wavelength $\bar{\lambda}$ and the phase at $P_1$ were retarded with respect to $P_2$ by $\alpha_{12}(\tau)$. Now according to §7.3 (7) a phase retardation of amount $2\pi$ corresponds to a displacement of the interference pattern in the direction parallel to $P_1P_2$ by an amount $a\bar{\lambda}/d$, where $d$ is the distance between $P_1$ and $P_2$ and $a$ is the distance between the screens $\mathscr{A}$ and $\mathscr{B}$. Hence the quasi-monochromatic fringes are displaced relative to the fringes that would be formed with monochromatic and co-phasal illumination of $P_1$ and $P_2$ by an amount

$$x = \frac{\bar{\lambda}}{2\pi}\frac{a}{d}\alpha_{12}(\tau) \tag{6}$$

in the direction parallel to the line joining the openings.

We see that the amplitude and the phase of the complex degree of coherence of quasi-monochromatic light beams may be determined from measurements of the visibility and the position of interference fringes. These results have a close bearing on Michelson's method, described in § 7.5.8, for determining the intensity distribution in spectral lines from measurements of visibility curves. It follows from equations (10) and (30) of § 10.3 that

$$\gamma_{11}(\tau) = \frac{\int_0^\infty G(v)e^{-2\pi ivr}dv}{\int_0^\infty G(v)dv}.$$

where $G(\nu)$ is the spectral density. Hence, by the Fourier inversion theorem, $G(\nu)$ is proportional to the Fourier transform of $\gamma_{11}(\tau)$. But we have just seen that the modulus of $\gamma_{11}(\tau)$ is essentially the visibility, and the phase of $\gamma_{11}(\tau)$ is simply related to the position of the fringes formed in an appropriate interference experiment. One is thus led to calculating $G$ in exactly the same way as was done by Michelson. The visibility curves exhibited in Figs. 7.54 and 7.55 may be evidently interpreted as representing $|\gamma_{11}|$ as a function of the time delay between the two beams.,

In practice the time delay $\tau$ introduced between the interfering beams is often very small, and it is then possible to simplify the formulae. According to equations

and (10) of § 10.3 we have

$$\left|\Gamma_{12}(\tau)\right| e^{i\alpha_{12}(\tau)} = \sqrt{I_1}\sqrt{I_2}\left|\gamma_{12}(\tau)\right| e^{i\alpha_{12}(\tau)} = 4\int_0^\infty G_{12}(\nu)e^{-2\pi i(\nu-\bar{\nu})t}d\nu. \tag{7}$$

If $|\tau|$ is so small that $|(\nu - \bar{\nu})\tau| <\leqslant 1$ for all the frequencies for which $|G_{12}(\nu)|$ is appreciable, i.e. if

$$|\tau| <\!\!\ll \frac{1}{\Delta v} \tag{8}$$

then evidently only a small error is introduced if the exponential term of the integrand in (7) is replaced by unity. The condition (8) implies, according to §7.5 (105), that $|\ \tau\ |$ must be small compared to the coherence time of the light. When this condition is satisfied, $|\Gamma_{12}(\tau)|, |\gamma_{12}(\tau)|$, and $\alpha_{12}(\tau)$ differ inappreciably from $|\Gamma_{12}(0)|, |\gamma_{12}(0)|$, and $\alpha_{12}(0)$ respectively. It is useful to set*

$$J_{12} = \Gamma_{12}(0) = \langle V_1(t) V_2^{\star}(t) \rangle \tag{9a}$$

$$\mu_{12} = \gamma_{12}(0) = \frac{\Gamma_{12}(0)}{\sqrt{\Gamma_{11}(0)}\sqrt{\Gamma_{22}(0)}} = \frac{J_{12}}{\sqrt{J_{11}}\sqrt{J_{22}}} = \frac{J_{12}}{\sqrt{I_1}\sqrt{I_2}} \tag{9b}$$

$$\beta_{12} = \alpha_{12}(0) = \arg \gamma_{12}(0) = \arg \mu_{12} \tag{9c}$$

Equations (18) and (10) of § 10.3 now give, subject to (8),

$$\gamma_{12}(\tau) \sim |\mu_{12}|\, e^{i(\beta_{12} - 2\pi\bar{\nu}\tau)} = \mu_{12} e^{-2\pi i \bar{\nu}\tau} \tag{10a}$$

$$\Gamma_{12}(\tau) \sim |J_{12}|\, e^{i(\beta_{11} - 2\pi\bar{\nu}\tau)} = J_{12} e^{-2\pi i \bar{\nu}\tau}. \tag{10b}$$

Thus, provided (8) is satisfied, we may replace $\gamma_{12}(\tau)$ and $\Gamma_{13}(\tau)$ in all our formulae by the quantities on the right-hand side of (10a) and (10b) respectively. In particular, the interference law (1) becomes

$$I(Q) \sim I^{(1)}(Q) + I^{(2)}(Q) + 2\sqrt{I^{(1)}(Q)}\sqrt{I^{(2)}(Q)}\, |\mu_{12}| \cos(\beta_{12} - \delta) \tag{11}$$

and is valid as long as the path difference $|s_2 - s_1| = c|\tau|$, introduced between the interfering beams is small, compared to the coherence length $c/\Delta v$, i.e. as long as

$$|\Delta \mathscr{P}| = |s_2 - s_1| = \frac{\bar{\lambda}}{2\pi}\delta <\!\!\ll \frac{\bar{\lambda}^2}{\Delta\lambda} \tag{12}$$

where the relation $c/\Delta\nu = \lambda^2/\Delta\lambda$ has been used.
Equation (11) is the basic formula of an elementary (quasi-monochromatic) theory of partial coherence, which forms the subject matter of the rest of this soction; some applications of this theory will be considered in § 10.5. Within its range of validity (indicated by (8) or (12)) the correlation between the vibrations at any two points $\dot{P}_1$

and $P_2$ in the wave field is characterized by $J_{12}$ rather than by $\Gamma_{12}(\tau)$, i.e. by a quantity which depends on the positions of the two points, but not on the time difference $\tau$. It follows from (10a) that, within the accuracy of this elementary theory,

$$|\gamma_{12}(\tau)| \sim |\mu_{12}|, \tag{13}$$

so that $|\mu_{12}|\ (0 \leqslant |\mu_{12}| \leqslant 1)$ represents the degree of coherence of the vibrations at $P_1$ and $P_2$; and we see from (11) that the phase $\beta_{12}$ of $\mu_{12}$ represents their effective phase difference. $\mu_{12}$, just like $\gamma_{12}(\tau)$ of which it is a special case, is usually called the complex degree of coherence (sometimes the complex coherence factor); and $J_{12}$ is called the mutual intensity.

0

- We again use the shortened notation where convenient, i.e. we write $J_{11}$ in place of $J(P_1, P_2)$, etc

### 10.4.2 Calculation of mutual intensity and degree of coherence for light from an extended incoherent quasi-monochromatic source

(a) The Van Cittert-Zernike theorem

We shall now determine the mutual intensity $J_{12}$ and the complex degree of coherence $\mu_{12}$ for points $P_1$ and $P_2$ on a screen $\mathscr{A}$ illuminated by an extended quasi-monochromatic incoherent source $\sigma$. For simplicity $\sigma$ will be taken to be a portion of a plane parallel to $\mathscr{A}$, and we will assume that the medium between the source and the screen is homogeneous. We also assume that the linear dimensions of $\sigma$ are small

compared to the distance $OO'$ between the source and the screen (Fig. 10.3), and that the angles between $OO'$ and the line joining a typical source point $S$ to $P_1$ and $P_2$ are small.

Imagine the source to be divided into elements $d\sigma_1, d\sigma_2, \ldots$ centred on points $S_1, S_2, \ldots$, of linear dimensions small compared to the mean wavelength $\bar{\lambda}$. If $V_{m1}(t)$ and $V_{m2}(t)$ are the complex disturbances at $P_1$ and $P_2$ due to the element $d\sigma_{1a}$, the total disturbances at these points are

Hence

$$V_1(t) = \sum_m V_{m1}(t), \quad V_2(t) = \sum_m V_{m2}(t) \tag{14}$$

$$J(P_1, P_2) = \langle V_1(t) V_2^\star(t) \rangle = \sum_m \langle V_{m1}(t) V_{m2}^\star(t) \rangle + \sum_{m \neq n} \sum_{m1} \langle V_{m1}(t) V_{n2}^\star(t) \rangle$$

Now the light vibrations arising from different elements of the source may be assumed to be statistically independent (mutually incoherent), and of zero mean value, so that*

$$\langle V_{m1}(t) V_{n2}^\star(t) \rangle = \langle V_{m1}(t) \rangle \langle V_{n2}^\star(t) \rangle = 0 \quad \text{when} \quad m \neq n. \tag{16}$$

If $R_{m1}$ and $R_{m2}$ are the distances of $P_1$ and $P_2$ from the source element $d\sigma_m$, then

$$V_{m1}(t) = A_m \left( t - \frac{R_{m1}}{v} \right) \frac{e^{-2\pi i H(t - R_{m1}/v)}}{R_{m1}}, \quad V_{m2}(t) = A_m \left( t - \frac{R_{m2}}{v} \right) \frac{e^{-2\pi i v(t - R_{m2}/v)}}{R_{m2}}, \tag{17}$$

where $|A_m|$ characterizes the strength and $\arg A_m$ the phase of the radiation from the $m$ th element, † and $v$ is the velocity of light in the medium between the source and the screen. Hence

$$\begin{aligned} \langle V_{m1}(t) V_{m2}^\star(t) \rangle &= \left\langle A_m \left( t - \frac{R_{m1}}{v} \right) A_m^\star \left( t - \frac{R_{m2}}{v} \right) \right\rangle \frac{e^{2\pi i \xi (R_{m1} - R_{m1})/v}}{R_{m1} R_{m2}} \\ &= \left\langle A_m(t) A_m^\star \left( t - \frac{R_{m2} - R_{m1}}{v} \right) \right\rangle \frac{e^{2\pi i \xi (R_{m1} - R_m)/v}}{R_{m1} R_{m2}} \end{aligned} \tag{18}$$

If the path difference $R_{m2} - R_{m1}$ is small compared to the coherence length of the light we may neglect the retardation $(R_{m2} - R_{m1})/v$ in the argument of $A_m^\star$, and we obtain from (15), (16), and (18)

$$J(P_1, P_2) = \sum_m \langle A_m(t) A_m^\star(t) \rangle \frac{e^{2\pi i k R_{m1} - R_m)/p}}{R_{m1} R_{m2}}. \tag{19}$$

The quantity $\langle A_m(l) A_m^\star(l) \rangle$ characterizes the intensity of the radiation from the source element $d\sigma_m$. In any practical case the total number of the source elements may be assumed

to be so large, that we may regard the source to be effectively continuous. Denoting by $I(S)$ the intensity per unit area of the source, i.e. $I(S_m)\,d\sigma_m = \langle A_m(t)A_m^\star(t)\rangle$, (19) becomes ‡

$$J(P_1,P_2) = \int_\sigma I(S)\frac{e^{ik(R_1-R_2)}}{R_1R_2}dS \tag{20}$$

where $R_1$ and $R_2$ denote the distances between a typical source point $S$ and the points $P_1$ and $P_2$, and $\bar{k} = 2\pi\bar{\nu}/v = 2\pi/\bar{\lambda}$ is the wave number in the medium. The complex degree of coherence $\mu(P_1,P_2)$ is, according to (20) and (9b), given by
where

$$\mu(P_1,P_2) = \frac{1}{\sqrt{I(P_1)}\sqrt{I(P_2)}}\int_\sigma I(S)\frac{e^{ik(R_1-R_2)}}{R_1R_2}dS \tag{21}$$

$$I(P_1) = J(P_1,P_1) = \int_\sigma \frac{I(S)}{R_1^2}dS, \quad I(P_2) = J(P_2,P_2) = \int_\sigma \frac{I(S)}{R_2^2}dS \tag{21a}$$

are the intensities at $P_1$ and $P_2$.

We note that the integral (21) is the same as that which occurs in quite a different connection; namely in the calculation, on the basis of the Huygens-Fresnel principle, of the complex disturbance in the diffraction pattern arising from diffraction of a spherical wave on an aperture in an opaque screen. More precisely, (21) implies that the complex degree of coherence, which describes the correlation of vibrations at a fixed point $P_2$ and a variable point $P_1$ in a plane illuminated by an extended quasi-monochromatic incoherent source, is equal to the normalized complex amplitude at the corresponding point $P_1$ in a certain diffraction pattern, centred on $P_2$. This pattern would be obtained on replacing the source by a diffracting aperture of the same size and shape as the source, and on filling it with a spherical wave converging to $P_2$, the amplitude distribution over the wave-front in the aperture being proportional to the intensity distribution across the source. This result was first established by van Cittert* and later in a simpler way by Zernize. † We shall refer to it as the van Cittert-Zernike theorem.

In most applications the intensity $I(S)$ may be assumed to be independent of the position of $S$ on the surface (uniform intensity). The corresponding diffraction problem is then that of diffraction of a spherical wave of uniform amplitude by an aperture of the same size and shape as the source.

Let $(\xi,\eta)$ be the coordinates of a typical source point $S$, referred to axes at $O$, and let ( $X_1, Y_1$ ) and ( $X_2, Y_2$ ) be the coordinates of $P_1$ and $P_2$ referred to parallel axes at $O'$ (Fig. 10.3). Then, if $R$ denotes the distance $OO'$ :

$$R_1^2 = (X_1-\xi)^2 + (Y_1-\eta)^2 + R^2,$$

so that

$$R_1 \sim R + \frac{(X_1-\xi)^2 + (Y_1-\eta)^2}{2R}. \tag{22}$$

0

- Incoherence always implies a finite (though not necessarily wide) spectral range, and equation (16) is, in fact, not valid for the idealized case of strictly monochromatic light. For monochromatic light one has $V_{m1}(t) = U_{m2}e^{-2\pi ivt}, V_{n2}(t) = U_{n2}e^{-2\pi ivt}$, where $U_{m1}$ and $U_{n2}$ are independent of time, so that $\langle V_{m_1}(t)V_{n_2}^*(t)\rangle = U_{m_1}U_{n_2}^*$ and this quantity is in general different from zero.
  † In general $A_m$ also depends on direction, but for simplicity we neglect this dependence.
  # From now on we shall frequently use the notation $dS, dP_1, \ldots$ for surface elements centred on the points $S, P_1, \ldots$

Here only the leading terms in $Y_1/R, Y_1/R, \xi/R$, and $\eta/R$ have been retained. A strictly similar expression is obtained for $R_2$, so that

$$R_1 - R_2 \sim \frac{({X_1}^2 + {Y_1}^2) - ({X_2}^2 + {Y_2}^2)}{2R} - \frac{(X_1 - X_2)\,\xi + (Y_1 - Y_2)\,\eta}{R}. \tag{23}$$

In the denominator of the integrands in (20) and (21), $R_1$ and $R_2$ may to a good approximation be replaced by $R$. We also set

$$\frac{(X_1 - X_2)}{R} = p, \quad \frac{(Y_1 - Y_2)}{R} = q \tag{24}$$

$$\psi = \frac{\bar{k}\,[({X_1}^2 + {Y_1}^2) - ({X_2}^2 + {Y_2}^2)]}{2R}. \tag{25}$$

Then (21) reduces to

$$\mu_{12} = \frac{e^{i\varphi} \iint_\sigma I(\xi, \eta) e^{-ik(pt+q\eta)} d\xi d\eta}{\iint_\sigma I(\xi, \eta) d\xi d\eta} \tag{26}$$

Hence if the linear dimensions of the source and the distance between $P_1$ and $P_2$ are small compared to the distance of these points from the source, the degree of coherence ${\mu_{12}}^3$ is equal to the absolute value of the normalized Fourier transform of the intensity function of the source.

The quantity $\psi$ defined by (25) has a simple interpretation. According to (23) it represents the phase difference $2\pi\,(OP_1 - OP_2)\,/\bar{\lambda}$, and may evidently be neglected when

$$OP_1 - OP_2 \ll \lambda. \tag{27}$$

For a uniform circular source of radius $\rho$ with its centre at $O$, (26) gives on integration (cf. §8.5.2)

$$\mu_{12} = \left(\frac{2J_1(v)}{v}\right) e^{i\varphi}, \tag{28}$$

where

$$\left.\begin{aligned} v &= k\rho\sqrt{p^2 + q^2} = \tfrac{2\pi}{\bar{\lambda}}\,\tfrac{\rho}{R}\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}, \\ \psi &= \tfrac{2\pi}{\bar{\lambda}}\left[\tfrac{({X_1}^2 + {Y_1}^2) - ({X_2}^2 + {Y_2}^2)}{2R}\right] \end{aligned}\right\} \tag{29}$$

$J_1$ being the Bessel function of the first kind and first order.* According to § 8.5.2, $|2J_1(v)/v|$ decreases steadily from the value unity when $v = 0$ to the value zero when $v = 3.83$; thus as the points $P_1$ and $P_2$ are separated more and more, the degree of coherence steadily decreases and there is complete incoherence when $P_1$ and $P_2$ are separated by the distance

$$P_1P_2 = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} = \frac{0.61R\bar{\lambda}}{\rho} \tag{30}$$

A further increase in $v$ re-introduces a small amount of coherence, but the degree of coherence remains smaller than 0.14 , and there is further complete incoherence for $v = 7.02$. Since

0

- P. H. van Cittert, Physica, I (1934), 201.
- F. Zernikf, Physica, 5 (1938), 785.

$J_1(v)$ changes sign as $v$ passes through each zero of $J_1(v)$, the phase $\beta_{12} = \arg \mu_{12}$ changes there by $\pi$; in consequence the position of the bright and dark fringes are interchanged after each disappearance of the fringes.

The function $2J_1(v)/v$ decreases steadily from the value 1 for $v = 0$ to 0.88 when $v = 1$, i.e. when

$$P_1P_2 = \frac{0 \cdot 16R\bar{\lambda}}{\rho} \tag{31}$$

Regarding a departure of 12 per cent from the ideal value unity as the maximum permissible departure, it follows that the diameter of the circular area that is illuminated ulmost coherently by a quasi-monochromatic, uniform, incoherent source of angular radius $\alpha = \rho/R$ is †$0.16\lambda/\alpha$. This result is useful in estimating the size of a source needed in experiments on interference and diffraction.

As an example consider the size of the "area of coherence" around an arbitrary point on a screen illuminated directly by the sun. The angular diameter $2\alpha$ which the sun's disc subtends on the surface of the earth is about $0°32' \sim 0.0093$ radian. Hence, if the variation of brightness across the sun's disc is neglected, the diameter $d$ of the area of coherence is approximately $0 \cdot 16\bar{\lambda}/0 \cdot 0047 \sim 34\bar{\lambda}$. Taking the mean wavelength $\bar{\lambda}$ as $5.5.10^{-5}$ cm this gives $d \sim 0.019$ mm.

In the present context Michelson's method of measuring angular diameters of stars (cf. §7.3.6) appears in a new light. According to (5) and (13), the visibility of the fringes is equal to the degree of coherence of the light vibrations at the two outer mirrors ( $M_1$ and $M_2$ in Fig. 7.16) of the Michelson stellar interferometer. For a uniformly bright circular star disc of angular radius $\alpha$ the smallest separation of the mirrors for which the degree of coherence has zero value (first fringe disappearance) is, according to (30), equal to $0.61\bar{\lambda}/\alpha$ in agreement with §7.3 (43). Moreover, from the measurements of both the visibility and the position of the fringes it is in principle possible to determine not only the stellar diameter, but also the distribution of the intensity over the stellar disc. For, according to § 10.4.1, measurements of the visibility and the position of the fringes are equivalent to determining both the amplitude and the phase of the complex degree of coherence $\mu_{12}$, and, according to ( 26 ), the intensity distribution is proportional to the inverse Fourier transform of $\mu_{12}$.

We mentioned in §7.3.6 the important modification due to Hanbury Brown and Twiss of Michelson's stellar interferometer. In the Hanbury Brown-Twiss system light from the star is focused on two photo-electric detectors $P_1, P_2$ and information about the star is obtained from the study of the correlation in the fluctuations of their current outputs. A full analysis of the performance of this system must take into account the quantum nature of the photo-electric effect,* and requires also some knowledge of electronics, and is thus outside the scope of this book. The principle of the method may, however, be easily understood. Under ideal experimental conditions (absence of noise), the current output of each photo-electric detector is proportional to the instantaneous intensity $I(t)$ of the incident light, and the fluctuation in the current output is proportional to $\Delta I(t) = I(t) - \langle I(t) \rangle$. Hence, in the interferometer of Hanbury Brown and Twiss, the quantity which is effectively being measured is proportional to $\Omega_{12} = \langle \Delta I_1 \Delta I_2 \rangle$. Simple statistical calculations show † that $\Omega_{12}$ is proportional to the square of the degree of coherence, so that the knowledge of $\Omega_{12}$, just like the knowledge of $|\mu_{12}|$, yields

0

- No confusion should arise from the fact that the symbol $J$ is also used for the mutual intensity, A the latter always appears with two suffixes or with several arguments.
† As early as 1865, E. VERDET estimated that the diameter of the "circle of coherence" is somewhat smaller than $0.5R\bar{\lambda}/\rho$. (Ann. Scientif. de l'cole Normale Suprieure, 2 (1865), 991 ; $^{a}|_{so}$ his Leons d'Optique Physique (Paris, L'Imprimerie Impriale), 1 (1869), 106.)

information about the size of the star.

**(b) Hopkins' formula**

In deriving the van Cittert-Zernike formula (21), it was assumed that the medium between the source $\sigma$ and the points $P_1$ and $P_2$ is homogeneous. It is not difficult to generalize the formula to other cases, e.g. when the medium is heterogeneous or consists of a succession of homogeneous regions of different refractive indices.

We again imagine the source to be divided into small elements $d\sigma_1, d\sigma_2 \ldots$, centred on points $S_1, S_2, \ldots$, of linear dimensions small compared to the mean wavelength $\vec{\lambda}$. If, as before, $V_{m1}(t)$ and $V_{m2}(t)$ represent the disturbances at $P_1$ and $P_2$ due to the element $d\sigma_m$, equations (15) and (16) still hold, but in (17) we must replace each factor $e^{i\bar{K}R_{mj}}/R_{mj} (j = 1, 2); \bar{K} = 2\pi\bar{v}/v\big)$ by a more general function. We introduce a transmission function $K(S, P, v)$ of the medium, defined in a similar way as in § 9.5.1; it represents the complex disturbance at $P$, due to a monochromatic point source of frequency $v$, of unit strength and of zero phase, situated at the element $d\sigma$ at $S$. For a homogeneous medium we have, from the Huygens-Fresnel principle, that $K(S, P, v) = -ie^{ikR}/\lambda R$, where $R$ denotes the distance $SP$, it being assumed that

the angle which $SP$ makes with the normal to $d\sigma$ is sufficiently small. It follows that in the more general case the factor $e^{i\bar{k}R_{mj}}/R_{mj}$ must be replaced by $i\bar{\lambda}K\left(S_m, P_{,,\hat{\nu}}\right)$, and we obtain, on passing to a continuous distribution, the following relation in place of (20):

$$J\left(P_1, P_2\right) = \bar{\lambda}^2 \int_\sigma I(S) K\left(S, P_1, \bar{v}\right) K^\star\left(S, P_2, \bar{v}\right) / J_L \Big) \tag{32}$$

According to (32) and (9b),

$$\mu\left(P_1, P_2\right) = \frac{\bar{\lambda}^2}{\sqrt{I\left(\bar{P}_1\right)}\sqrt{I\left(P_2\right)}} \int_\sigma I(S) K\left(S.P_1, \bar{\nu}\right) K^\star\left(S, P_2, \bar{\nu}\right) dS \tag{33}$$

where $l\left(P_1\right) = J\left(P_1, P_1\right)$ and $J\left(P_2\right) = J\left(P_2, P_2\right)$ are the intensities at $P_1$ and $P_2$ respectively.

For the purpose of later applications it will be useful to express (32) and (33) in a slightly different form. We set

$$i\bar{\lambda}K\left(S, P_1, \bar{\nu}\right)\sqrt{I(S)} = U\left(S, P_1\right), \quad i\bar{\lambda}K\left(S, P_2, \bar{\nu}\right)\sqrt{I(S)} = U\left(S, P_2\right). \tag{34}$$

The formulae (32) and (33) become

$$J\left(P_1, P_2\right) = \int_\sigma U\left(S, P_1\right) U^\star\left(S, P_2\right) dS \tag{35a}$$

$$\mu\left(P_1, P_2\right) = \frac{1}{\sqrt{I\left(P_1\right)}\sqrt{I\left(P_2\right)}} \int_u U\left(S, P_1\right) U^\star\left(S, P_2\right) dS \tag{35b}$$

We note that $U(S, P)$, defined by (34), is proportional to the disturbance which would arise at $P$ from a strictly monochromatic point source of frequency $\bar{\nu}$, strength $\sqrt{I(S)}$ and zero phase,

0

- Cf. R. Hanbury Brown and R. Q. Twiss, Proc. Roy. Soc., A, 242 (1957), 300; ibid., A. 243 (1957), 291. See also E. M. Purcell, Nature, 178 (1956), 1449; F. D. Kahn, Optica Acta, 5 (1958), 93 and L. Mandel, Proc. Phys. Soc., 72 (1958), 1037; Progress in Optics, Voi. 2, ed. E. WolF (Amsterdam, North Holland Publishing Company and New York, J. Wiley and Sons, 1963), I81.
  † E. Wolf, Phil. Mag., 2 (1957), 351. See also J. A. Rftcliffe, Rep. Progr. Phys. (London, Physical Society), 19 (1956), 233.

situated at $S$. Thus (35) may be interpreted as expressing the mutual intensity $J(P_1, P_2)$ and the complex degree of coherence $\mu(P_1, P_2)$ due to an extended quasi-monochromatic source, in terms of the disturbances produced at $P_1$ and $P_2$ by each source point of an "associated" monochromatic source.*

The expression (35b) was first suggested by Horkins † from heuristic considerations and is very useful in solving coherence problems of instrumental optics. The main usefulness of the formula arises from the fact that, like the van Cittert-Zernier theorem, it permits the calculation of the complex degree of coherence of light from an incoherent source without the explicit use of an averaging process.

### 10.4.3 An example

We shall illustrate the preceding considerations by the discussion of an experiment. A primary source $\sigma_0$ is imaged by a lens $L_0$ on to a pinhole $\sigma_1$ and the light that emerges from the pinhole is rendered parallel by a lens $L_1$. A second lens $L_2$, exactly similar to $L_1$, brings the interfering beams to a focus $F$ in the focal plane $\mathscr{F}$ of the lens $L_2$.

A plane mirror $M$ is used to reduce the overall length of the instrument (Fig. 10.4). If a diffracting mask (a dark screen $\mathscr{A}$, for example a piece of uniformly blackened film), with apertures of any desired size, shape, and distribution is placed in the parallel beam between $L_1$ and $L_2$, its Fraunhofer diffraction pattern is formed in the focal plane $\mathscr{F}$, and, under normal use, the operator may displace the mask while viewing the diffraction pattern through a microscope.*

If $\sigma_0$ were a quasi-monochromatic point source, it would give rise to coherent illumination in the neighbourhood of its geometrical image in the plane of $\sigma_1$. The

size of this coherently illuminated area is of the order of the effective size of the Alry diffraction pattern $\sigma_4$, formed by the lens $L_0$, of the single source point. The light distribution in the plane of $\sigma_1$ due to a finite incoherent source may be regarded as arising from the incoherent superposition of many such patterns. If, as we assume, the image formed by $L_0$ of this extended source, and the pinhole $\sigma_1$, are both large compared to $\sigma_1$, the illuminated pinhole $\sigma_1$ will itself effectively act as an incoherent source. † According to the van Cittert-Zernike theorem, such a source will give rise to a correlation between vibrations at any two points on the first surface of the lens $L_1$ (and more generally in the plane $\mathscr{A}$ ); and with the usual approximations the complex degree of the coherence is given by the formula (28):

$$\mu_{12} = |\mu_{12}|\, e^{i\beta_{12}} = \frac{2J_1(v)}{v} e^{iv}, \tag{36a}$$

where

$$v = \frac{2\pi}{\bar{\lambda}} \frac{\rho_1 d}{R}, \quad \psi = \frac{2\pi}{\bar{\lambda}} \left( \frac{{r_1}^2 - {r_2}^2}{2R} \right), \tag{36b}$$

where $d = P_1P_2, \rho_1$ is the radius of $\sigma_1, R$ is the distance between $\sigma_1$ and $L_1$, and $r_1$ and $r_2$ are the distances of $P_1$ and $P_2$ from the axis.

If the diffracting mask $\mathscr{A}$ consists of two small circular apertures centred on $P_1$ and $P_2$, the pattern observed in the focal plane $\mathscr{F}$ results from the superposition of

[0]* It must not be assumed that the mutual intensity and the complex degree of coherence of light from such a fictitious source are also given by (35). For, as already explained, the relation (16) used in tho derivation of these formulae is not valid in the limiting case of monochromatic radiation; the degree of coherence of monochromatic light is. in fact. always equal to unity.

† H. H. Hopeins, Proc. Roy. Soc, A, 208 (1951), 263.

[0]

- This apparatus, known as the Diffractometer, is mninly used in connection with optical diffraction meth-

two partially coherent beams, with a degree of coherence $|\mu_{12}|$, emerging from these apertures. We shall investigate the changes in the structure of this pattern as the separation of $P_1$ and $P_2$ is gradually increased, i.e. as the degree of coherence between the two interfering beams is varied.

We assume that $P_1$ and $P_2$ are situated symmetrically about the axis. Then $\psi = 0$ and the intensities $I^{(1)}(Q)$ and $I^{(2)}(Q)$ at a point $Q$ in the focal plane associated with either of the two beams are then equal, and are given by the Fraunhofer formula for

diffraction by a circular aperture (§ 8.5 (14)). If the point $Q$ is the focus for rays diffracted in directions that make an angle $\phi$ with the normal to $\mathscr{A}$, and if $a$ is the radius of each aperture (see Fig. 10.5) then, apart from a normalizing factor,

$$I^{(1)}(Q) = I^{(2)}(Q) = \left(\frac{2J_1(u)}{u}\right)^2, \quad u = \frac{2\pi}{\bar{\lambda}} a \sin\phi \tag{37}$$

The phase difference $\delta$ between the beams diffracted to $Q$ is

$$\delta = \frac{2\pi}{\bar{\lambda}} P_2 N = \frac{2\pi}{\bar{\lambda}} d \sin\phi = Cuv, \quad C = \frac{\bar{\lambda}}{2\pi} \frac{R}{\rho_1 a} \tag{38}$$

Where $N$ is the foot of the perpendicular dropped from $P_1$ on to the ray diffracted at $P_2$. On substituting from (36), (37), and (38) into (11) we finally obtain the following expression for the intensity at the point $Q(\phi)$ in the focal plane, when the apertures at $P_1$ and $P_2$ are separated by a distance $d$ :

$$\left.\begin{aligned} I(\phi, d) = 2\left(\tfrac{2J_1(u)}{u}\right)^2 \left\{1 + \left|\tfrac{2J_1(v)}{v}\right| \cos\left[\beta_{12}(v) - Cuv\right]\right\} \\ \beta_{12}(v) = 0 \quad \text{when} \quad \tfrac{2J_1(v)}{v} > 0 \\ = \pi \quad \text{when} \quad \tfrac{2J_1(v)}{v} < 0 \end{aligned}\right\} \tag{40}$$

In Fig. 10.6 are shown photographs of the patterns observed with such an arrangement, for various separations $d$. The corresponding theoretical curves, computed from the formula (39) are also shown. The chain lines represent the envelopes

$$\left.\begin{aligned} I_{\max}(\phi, d) = 2\left(\frac{2J_1(u)}{u}\right)^2 \left\{1 + \left|\frac{2J_1(v)}{v}\right|\right\}, \\ I_{\min}(\phi, d) = 2\left(\frac{2J_1(u)}{u}\right)^2 \left\{1 - \left|\frac{2J_1(v)}{v}\right|\right\}. \end{aligned}\right\} \tag{41}$$

It is of interest to note that when $\beta = \pi$ (cases (D) and (E)), the intensity at the centre of each pattern has a relative minimum, not a maximum, in agreement with our general considerations. The variation of the degree of coherence with the separation

of the two apertures is shown in Fig. 10.7, where also the six values corresponding tc the photographs of Fig. 10.6 are indicated by the appropriate letters.

ods for solutions of problems of X-ray structure analysis. (Cf. C. A. Taylor, Ir. M. Hinde, and H. Leson, Acta Cryst., 4 (1951), 261; A. W. Hanson, H. Lifson, and C. A. Taylor, Proc. Roy. Soc. A, 218 (1953), 371; W. Hugifs and C. A. Taylor, J. Sci. Instr., 30 (1953), 105.)

‡ This point is discussed quantitatively in § 10.5.1. See also A. T. Forrester, Amer. J. Phys.. 24 (1956), 104.

### 10.4.4 Propagation of mutual intensity

Consider a beam of quasi-monochromatic light from an extended incoherent source $^{\sigma}$, and suppose that the mutual intensity is known for all pairs of points on a fictition $^{3}$ surface $\mathscr{A}$ intercepting the beam. We shall show that it is then possible to determine the mutual intensity for all pairs of points on any other second surface $\mathscr{B}$ illumine ted by the light from $\mathscr{A}$ either directly or via an optical system.

We assume to begin with that the medium between $\mathscr{A}$ and $\mathscr{B}$ is homogeneous and of refractive index unity. Let $U(S,Q_1)$ and $U(S,Q_2)$ be the disturbances at poin $^{SS}Q_1$ and $Q_2$ on $\mathscr{B}$ (Fig. 10.8) due to a typical source point $S$ of the associated monochromatic source. Then, according to (35), the mutual intensity $J(Q_1,Q_2)$ is given by

$$J(Q_1,Q_2)=\int_{\sigma}U(S,Q_1)\,U^{\star}(S,Q_2)\,dS \tag{42}$$

Now $U(S,Q_1)$ and $U(S,Q_2)$ may be expressed in terms of the disturbance at all points of $\mathscr{A}$ by means of the Huygens-Fresnel principle:

$$U(S,Q_1)=\int_{\mathscr{A}}U(S,P_1)\,\frac{e^{iks_1}}{s_1}\Lambda_1 dP_1 \tag{43}$$

Here $s_1$ is the distance from a typical point $P_1$ on $\mathscr{A}$ to $Q_1,\Lambda_1$ is the inclination factor (denoted by $K$ in Chapter VIII) at $P_1$, and $\bar{k}=2\pi\bar{\nu}/c$ is the mean wave number. For small obliquities, $\Lambda_1\sim -i/\lambda$. According to (43) and a similar expression for $U(Q_2)$, we have

$$U(S,Q_1)\,U^{\star}(S,Q_2)=\int_{\mathscr{A}}\int_{\mathscr{S}}U(S,P_1)\,U^{\star}(S,P_2)\,\frac{e^{ik(s_1-s_2)}}{s_1s_2}\Lambda_1\Lambda_2^{\star}dP_1dP_2 \tag{44}$$

where the points $P_1$ and $P_2$ take on independently all positions on the surface $\mathscr{A}$ of integration. Next we substitute from (44) into (42), and change the order of integration. The integration over $\sigma$ gives precisely $J(P_1,P_2)$ and we obtain

$$J(Q_1,Q_2)=\iint_{\mathscr{S}}J(P_1,P_2)\,\frac{e^{ik(s_1-s_2)}}{s_1s_2}\Lambda_1\Lambda_2^{\star}dP_1dP_2 \tag{45}$$

This is the required formula, due to Zernike,* for the propagation of the mutual intensity. We have implicitly assumed in the derivation of (45) that light from every point of the surface $\mathscr{A}$ reaches the points $Q_1$ and $Q_2$. The presence of any diaphragm between the two surfaces may be taken care of by limiting the integration to those portions of the surface $\mathscr{A}$ which send light to $Q_1$ and $Q_2$, unless the diaphragm is so small that the effects of diffraction at its edges cannot be neglected. The diffraction can be taken into account by carrying out the transition from $\mathscr{A}$ to $\mathscr{B}$ in two steps, first from $\mathscr{A}$ to the plane of the diaphragm, and then from the plane of the diaphragm to the surface $\mathscr{B}$.

In the special case when the points $Q_1$ and $Q_2$ coincide, (45) reduces to the following expression for the intensity, when we also substitute for $J(P_1,P_2)$ in terms of the intensities $I(P_1),I(P_2)$ and the complex degree of coherence $\mu(P_1,P_2)$ :

$$I(Q)=\iint_{\mathscr{A}\mathscr{A}}\sqrt{I(P_1)}\sqrt{I(P_2)}\mu(P_1,P_2)\,\frac{e^{i(s_1-s_2)}}{s_1s_2}\Lambda_1\Lambda_2^{\star}dP_1dP_2 \tag{46}$$

This formula expresses the intensity at a point $Q$ as the sum of contributions from each pair of elements $dP_1,dP_2$ of an arbitrary surface $\mathscr{A}$ intercepting the beam

0

- F. Zernike, Physica, 5 (1938), 791.

(see Fig. 10.9). The contribution from each pair of elements depends on the intensities at $P_1$ and $P_2$ and each contribution is weighted by the appropriate value of the complex degree of coherence factor $\mu(P_1, P_2)$. Formula (46) may be regarded as a kind of Huygens-Fresnel principle for the propagation of intensity in a partially coherent field. The resemblance between the formulae just derived and those of the more elementary Huygens-Fresnel theory has a deeper significance which will be brought out in our rigorous formulation of the theory of partial coherence (§ 10.7).
If the light from $\mathscr{A}$ reaches $\mathscr{B}$ via an optical system, factor $\Lambda e^{iks}/s$ must evidently be replaced by an appropriate transmission function $K(P, Q)$. Instead of (45) we then obtain the more general formula

$$J(Q_1, Q_2) = \int_{\mathscr{A}\mathscr{A}} \int_{\mathscr{A}} J(P_1, P_2)\, K(P_1, Q_1)\, K^{\star}(P_2, Q_2)\, dP_1 dP_2 \tag{47}$$

## 9.13.5 10.5 Some Applications

### 10.5.1 The degree of coherence in the image of an extended incoherent quasi-monochromatic source

As a preliminary to the study of image formation with partially coherent light, it will be useful to consider the degree of coherence in the image of an extended incoherent source, formed by a centred optical system. A finite degree of correlation among vibrations in the image plane arises from the fact that, because of diffraction (and in general also because of aberrations), the light from each source point is not concentrated into a point, but spreads over a finite area. Some of these "image patterns" overlap and in consequence points in the image plane that are close enough to each other receive coherent as well as incoherent contributions.

Suppose that $\sigma$ is a uniform quasi-monochromatic incoherent circular source of radius $\rho$ emitting light with mean (vacuum) wavelength $\bar{\lambda}_0$, situated in a homogeneous object space of refractive index $n$. Further let D be the distance between the object plane and the plane of the entrance pupil. Corresponding quantities in the image will be denoted by primed symbols of the same type.

Let $d$ be the distance between two points $P_1$ and $P_2$ in the entrance pupil. We assume that $\rho/\mathrm{D} \ll 1, d/\mathrm{D} < 1$ and $OP_1 - OP_2 \ll \bar{\lambda}_0$, where $O$ is the axial point
of the source*; then the complex degree of coherence $\mu(P_1, P_2)$ is, according to §10.4 (28), given by

$$\mu(P_1, P_2) = \frac{2J_1(v)}{v} \tag{1}$$

$$v = \frac{2\pi}{\bar{\lambda}} d \sin\alpha = \frac{2\pi n}{\bar{\lambda}_0} d \sin\alpha \tag{2}$$

where $\alpha \sim \sin\alpha \sim \rho/D$ is the angular radius of the source as seen from the centre of the entrance pupil (Fig. 10.10).

To determine the complex degree of coherence for any pair of points in the plane of the exit pupil, we could apply the propagation law § 10.4 (47). However, we may

now proceed in a simpler manner as follows:
If $U(S, P_1)$ and $U(S, P_2)$ are the complex disturbances at $P_1$ and $P_2$ due to a source point $S$ of the associated monochromatic source (cf. p. 513), the disturbances due to $S$ at the conjugate points $P_1'$ and $P_2'$ in the exit pupil are, to a good approximation, given by the geometrical optics formulae

$$U(S,P_1') = K_{11}U(S,P_1), \quad U(S,P_2') = K_{22}U(S,P_2). \tag{3}$$

Here $K_{11}$ and $K_{22}$ are transmission factors that characterize, within the accuracy of geometrical optics, the propagation from $P_1$ to $P_1'$ and from $P_2$ to $P_2'$ respectively, along the rays that originate in the source point $S$. By Horkins' formula, § 10.4 (35b)

$$\mu(P_1',P_2') = \frac{1}{\sqrt{I(P_1')}\sqrt{I(P_2')}} \int_\sigma U(S,P_1')\,U^\star(S,P_2')\,dS. \tag{4}$$

Now the intensities $I(P_1')$ and $I(P_1)$ in the two pupil planes are related by

$$I(P_1') = \int_\sigma |U(S,P_1')|^2\,dS = |K_{11}|^2 \int_\sigma |U(S,P_1)|^2\,dS = |K_{11}|^2\,I(P_1), \tag{5}$$

with a similar relation between $I(P_2')$ and $I(P_2)$. From (3), (4), and (5) we have

$$\begin{aligned}\mu(P_1',P_2') &= \frac{K_{11}K_{22}^\star}{|K_{11}|\,|K_{22}|} \frac{1}{\sqrt{I(P_1)}\sqrt{I(P_2)}} \int_\sigma U(S,P_1)\,U^\star(S,P_2)\,dS \\ &= e^{i(\Phi_{11}-\Phi_{22})}\mu(P_1,P_2)\end{aligned} \tag{6}$$

- If this last condition is not satisfied, $|\mu(P_1,P_2)|$ remains unehanged, but according to \$10.4
  (28) the phase of $\mu(P_1,P_2)$ is increased by an amount $\psi = 2\pi\,[OP_1 - OP_2]\,/\bar{\lambda}_0$.
  where $\Phi_{11}$ and $\Phi_{22}$ are the phases of $K_{11}$ and $K_{22}$ respectively. This relation implies that the degree of coherence $|\mu|$ for any two points in the exit pupil is equal to the degree of coherence for the conjugate points in the entrance pupil; and the phases of the corresponding values of the complex degree of coherence for corresponding point pairs differ by the amount $\Phi_{11} - \Phi_{22}$, i.e. by the geometrical phase difference $2\pi\,\{[P_1P_1'] - [P_2P_2']\}\,/\bar{\lambda}_0$.

Let

$$v' = \frac{2\pi n'}{\dot{\lambda}_0} d' \sin\alpha' \tag{7}$$

Since $P_1'$ is the conjugate of $P_1$ and $P_2'$ the conjugate of $P_2$, it follows by the SmithHelmholtz theorem (§ 4.4 (49)) that, within the accuracy of Gaussian optics, $v' = v$.* Hence, by (1) and (6), the complex degree of coherence for pairs of points in the exit pupil may be written as

$$\mu(P_1',P_2') = \left(\frac{2J_1(v')}{v'}\right) e^{i(\Phi_{11}-\Phi_{32})} \tag{8}$$

If, as in § 10.4 (31), we regard the values $|\mu| \geqslant 0.88$ as sufficiently close approximations to full coherence, and remember that $|2J(v)/v| \geqslant 0.88$ when $v \leqslant 1$, it follows that an incoherent quasi-monochromatic uniform circular source will give rise in the exit pupil to coherently illuminated areas of diameter

$$d_{\text{coh}}' \sim \frac{0\cdot 16\lambda_0}{n'\sin\alpha'} \tag{9}$$

where $2\alpha' \sim 2\rho'/D'$ is the angle which the diameter of the image of the source subtends at the centre of the exit pupil and $\bar{\lambda}_0/n' = \bar{\lambda}$ is the mean wavelength of the light in the image space.

We shall express (9) in a somewhat different form. Let $r_A'$ denote the radius of the first dark ring in the AIRY pattern associated with the system,

$$r'_A = \frac{0.61\bar{\lambda}_0}{n' \sin\theta'} \tag{10}$$

where $n' \sin\theta' \sim n'a'/D'$ is the numerical aperture on the image side. Then according to (9) and (10), $d'_{\text{coh}}/r'_A \sim 0.16 \sin\theta'/0.61 \sin\alpha'$, so that

$$d'_{\text{coh}} \sim 0.26a' \left(\frac{r'_4}{\rho'}\right) \tag{11}$$

This formula gives an estimate for the size of the coherently illuminated areas of the exit pupil in terms of the "physical parameters," viz. the radius $r'_4$ of the first dark ring of the associated AIRY pattern, the radius $\rho$ ' of the geometrical image of the source, and the radius $a'$ of the exit pupil.

The exit pupil and hence the image plane will be illuminated almost coherently if $d'_{\text{coh}} \geqslant 2a'$, i.e. if

When $d'_{\text{coh}} \ll 2a'$, i.e. when

$$\rho' \leqslant 0 \cdot 13 r'_\Delta \tag{12}$$

$$\rho' \geqslant\geqslant 0 \cdot 13 r'_A, \tag{13}$$

the coherently illuminated areas of the exit pupil will be small compared to the exit pupil itself so that in this case the illumination of the exit pupil is effectively incoherent. The complex degree of coherence for pairs of points $Q'_1, Q'_2$ in the image plane

will then be essentially the same as that due to an incoherent source; this source has the same size, shape, and position as the exit pupil, and the intensity distribution across this source is the same as the intensity distribution across the exit pupil. Hence according to the van Cittert-Zernike theorem, §10.4 (21),

$$\mu(Q'_1, Q'_2) = \frac{1}{\sqrt{I(Q'_1)}\sqrt{I(Q'_2)}} \int_{\mathscr{A}} I(P') \frac{e^{ik(s_1 - s_2)}}{s_1 s_2} dP' \tag{14}$$

$$I(Q'_1) = \int_{\mathscr{A}} \frac{I(P')}{s_1^2} dP', \quad I(Q'_2) = \int_{\mathscr{S}} \frac{I(P')}{s_2^2} dP' \tag{15}$$

The integration is taken over the exit pupil $\mathscr{A}'$ and $s_1$ and $s_2$ denote the distances from the typical point $P'$ in $\mathscr{A}'$ to the points $Q'_1$ and $Q'_2$ respectively (Fig. 10.11(a)).

The intensity $I(P')$ may be calculated from the intensity $I(P)$ at the conjugate point in the entrance pupil by means of the relation (5). Since the phase of the transmission function does not appear in this relation, $\mu(Q'_1, Q'_2)$ is independent of the aberrations of the system. As a rule $I(P')$ is effectively constant; if, moreover, the points $Q'_1$ and $Q'_2$ are sufficiently close to each other the expression (14) then reduces to

$$\mu(Q'_1, Q'_2) = \frac{2J_1(u')}{u'}, \quad u' = \frac{2\pi n'}{\bar{\lambda}_0} \frac{a'}{\bar{D}'} h', \tag{16}$$

where $h'$ is the distance between $Q'_1$ and $Q'_2$.

In the general case, when neither condition (12) nor condition (13) holds, the exit pupil is illuminated with partially coherent light, characterized by the complex degree of coherence (8).

0

- This means that $v$ and $v'$ represent a particular choice of the Seidel variables (§ 5.2),

The value of the complex degree of coherence for pairs of points in the image plane must then be calculated with the help of the propagation law § 10.4 (45) and leads to the expression

$$\mu(Q_1', Q_2') = \frac{1}{\sqrt{I(Q_1')}\sqrt{I(Q_2')}} \int_{ss's'} \int \sqrt{I(P_1')}\sqrt{I(P_2')} \left(\frac{2J_1(v')}{v'}\right) \frac{e^{i[\Phi_{11}-\Phi_{22}+k(a_1-s_3)]}}{s_1 s_2} \Lambda_1 \Lambda_2^{\star} dP_1' dP_2'. \quad (17)$$

The intensities $I(Q_1')$ and $I(Q_2')$ may also be calculated from this formula, if use is made of the fact that $\mu(Q_1', Q_1') = \mu(Q_2', Q_2') = 1$. We note that, since the integrand contains the phases $\Phi_{11}$ and $\Phi_{22}$ of the transmission function, the complex degree of coherence now depends on the aberrations of the system.

### 10.5.2 The influence of the condenser on resolution in a microscope

In order to examine a small non-luminous object under a microscope, the object has to be illuminated. If, as is usually the case, the object is almost transparent, it is illuminated from behind, or transilluminated, and the light which has passed through the object is then focused on to the image plane of the microscope objective. To obtain a sufficient concentration of light, an auxiliary lens system-a condenseris usually used. Various methods of illumination are employed. We shall briefly describe two commonly used methods, the so-called critical illumination and Köhler's illumination, and discuss the resolving power which can be achieved with them.

#### (a) Critical illumination

In this method of illumination a uniformly bright source is placed close behind the field stop and is imaged by the condenser on to the object plane of the microscope objective (Fig. 10.12). The size of the field stop aperture is adjusted so that its image by the condenser just covers the field.

The illuminated region in the image plane of the condenser (the object plane of the objective) is very much larger than the effective size of the Arry pattern due to a single source point ( $\rho' \geqslant> r_A'$ in the notation of § 10.5.1). It follows from § 10.5.1 that under these circumstances the complex degree of coherence for any pair of points in the object plane of the objective is the same as that due to an incoherent source filling the condenser aperture; moreover, it is independent of the aberrations of the condenser. Now it is evident that the resolving power depends only on the degree of coherence (characterized by this factor) of the light incident upon the object and on the properties of the microscope objective. Hence the aberrations of the condenser have no influence on the resolving power of a microscope. This important result was first established in a different manner by Zernike,* and shows that the widely held view, according to which a highly corrected condenser is of advantage for attaining high resolving power, is incorrect.

To estimate the effect of the size of the condenser on the resolution, consider two pinholes $P_1(X_1, Y_1)$ and $P_2(X_2, Y_2)$ in the plane of the object. With the same assumptions as before, the complex degree of coherence of the light that reaches these pinholes is given by a formula of the form (16):

$$\mu(P_1, P_2) = \frac{2J_1(u_{12})}{u_{12}}, \quad u_{12} = \frac{2\pi}{\bar{\lambda}_0}\sqrt{\left(\bar{X}_1 - \bar{X}_2\right)^2 + \left(Y_1 - \bar{Y}_2\right)^2}\, n_c' \sin\theta_c' \quad (18)$$

0

where $n'_c \sin\theta'_c$ is the numerical aperture of the condenser on the side of the microscope objective.

Let $P(X,Y)$ be any other point in the object plane, and $P'$ its image by the objective. If we assume that the objective is effectively free of aberrations, the intensity distribution, in the image plane of the objective, of the light arriving from $P_1$ alone is an Arry pattern centred on the image $P'_1$ of $P_1$. Hence if $n_0 \sin\theta_0$ is the numerical aperture of the objective, the intensity $I^{(1)}(P')$ due to light that reaches $P'$ from $P_1$ alone is, apart from a constant factor, equal to

$$I^{(1)}(P') = \left(\frac{2J_1(v_1)}{v_1}\right)^2, \quad v_1 = \frac{2\pi}{\bar{\lambda}_0}\sqrt{(X-X_1)^2+(Y-Y_1)^2}\, n_0 \sin\theta_0. \tag{19a}$$

The intensity $I^{(2)}(P')$, due to the light reaching it from the pinhole $P_2$, is given by a similar expression :

$$I^{(2)}(P') = \left(\frac{2J_1(v_2)}{v_2}\right)^2, \quad v_2 = \frac{2\pi}{\bar{\lambda}_0}\sqrt{(X-X_2)^2+(Y-Y_2)^2}\, n_0 \sin\theta_0. \tag{19b}$$

It follows that when the two pinholes are illuminated via the condenser, the intensity $I(P')$ in the image plane of the microscope objective arises from the superposition of two partially coherent beams. The intensity of each beam is given by (19) and the complex degree of coherence of the two beams is given by (18). An expression for $I(P')$ is immediately obtained on substituting from these equations into the formula (11) of § 10.4. This gives, if we also assume that $P'$ is very close to the geometrical images of $P_1$ and $P_2$ (more precisely that $\delta = [P_1P'] - [P_2P'] \ll \bar{\lambda}$ ),

$$I(P') = \left(\frac{2J_1(v_1)}{v_1}\right)^2 + \left(\frac{2J_1(v_2)}{v_2}\right)^2 + 2\left(\frac{2J_1(mv_{12})}{mv_{12}}\right)\left(\frac{2J_1(v_1)}{v_1}\right)\left(\frac{2J_1(v_2)}{v_2}\right), \tag{20}$$

where

$$m = \frac{n'_c \sin\theta'_c}{n_0 \sin\theta_0}, \quad v_{12} = \frac{u_{12}}{m} = \frac{2\pi}{\bar{\lambda}_0}\sqrt{(X_1-X_2)^2+(Y_1-Y_2)^2}\, n_0 \sin\theta_0 \tag{21}$$

Some interesting conclusions may be drawn from (20). When $mv_{12}$ is a root other than $mv_{12} = 0$ of the equation $J_1(mv_{12}) = 0$, the product term is absent and (20) reduces to

$$I(P') = \left(\frac{2J_1(v_1)}{v_1}\right)^2 + \left(\frac{2J_1(v_2)}{v_2}\right)^2. \tag{22}$$

The distribution of the intensity in the image plane is now the same as if $P_1$ and $P_2$ were illuminated incoherently. In particular this will be the case when $m = 1$ and $v_{12}$ is a non-zero root of $J_1(v_{12}) = 0$; that is, when the numerical apertures are equal and the geometrical images of the pinholes are separated by a distance equal to the radius of any dark ring of the Arry pattern of the objective.

When the numerical aperture of the condenser is very small ( $m \to 0$ ), then $\mathbf{2}_1(m\varepsilon_{12})/m\varepsilon_{12} \sim 1$ and (20) reduces to

$$I(P') = \left\{\frac{2J_1(v_1)}{v_1} + \frac{2J_1(v_2)}{v_2}\right\}^2. \tag{23}$$

The distribution of the intensity is now the same as with perfectly coherent illuminatime, whatever the separation of the pinholes.

• F. Zernike, Physica, 5 (1938), 794.

Formula (20) makes it possible to study the dependence of the intensity distribution in the image plane of the microscope objective on the ratio $m$ of the numerical apertures. In particular consider the intensity at the midpoint between $P_1'$ and $P_2'$. We shall regard the pinholes to be just resolved when the intensity at the midpoint is by 26.5 per cent smaller than the intensity at either of the two points. The value 26.5 per cent corresponds to Rayleigh's criterion for a circular aperture in incoherent illumination (cf. §8.6.2). We shall express this limiting separation $(P_1P_2)_{11\ \mathrm{m}}$ in the same form as for incoherent [§ 8.6 (32)] and coherent [§ 8.6 (55)] illuminations:

$$(P_1P_2)_{\mathrm{llm}} = L(m)\frac{\bar{\lambda}_0}{n_0 \sin\theta_0}. \tag{24}$$

The curve $L(m)$ computed from (20) on the basis of this criterion in shown in Fig. 10.13. It is seen that the best resolving power is obtained with $m \sim 1 \cdot 5$, i.e. when the

numerical aperture of the condenser is about 1.5 times that of the objective. The value of $L$ is then slightly smaller than the value 0.61 obtained with incoherent illumination.

### (b) Köhler's illumination

In a method of illumination due to Köhler,* which is illustrated in Fig. 10.14, a converging lens is placed close to the field stop and forms an image of the source $\sigma$ in the focal plane of the condenser, which now contains the condenser diaphragm. The rays from each source point then emerge from the condenser as a parallel beam. This arrangement has the advantage that the irregularities in the brightness distribution on the source do not cause irregularities in the intensity of the field illumination.

To estimate the limit of resolution that is attained with KöHLER's method of illumination we must first determine the complex degree of coherence $\mu$ for pairs of points in the object plane of the microscope objective. Let

$$U(S, P_1) = A_1 e^{i\phi_1}, \quad U(S, P_2) = A_2 e^{i\phi_1}, \tag{25}$$

be the complex disturbances at points $P_1(X_1, Y_1)$ and $P_2(X_2, Y_2)$ of the object plane of the microscope objective, due to a source point $S$ of the monochromatic source associated with $\sigma$ (cf. p. 513). Evidently

$$\phi_1 - \phi_2 = \frac{2\pi}{\lambda_0}\left[p(X_1 - X_2) + q(Y_1 - Y_2)\right], \tag{26}$$

where $p$ and $q$ are the first two ray components of the two parallel rays through $P_1$ and $P_2$ from the source point $S$. If the condenser system suffers from aberrations, the two rays will not be strictly parallel, but as we are only considering points which

are close to each other, this effect may be neglected. On substituting from (25) and (26) into Hopkins' formula § 10.4 (35b) it follows that

$$\left.\begin{aligned} \mu(P_1, P_2) &= \tfrac{1}{\sqrt{I(P_1)}\sqrt{I(P_2)}} \textstyle\int_\sigma A_1 A_2 e^{ik_0\rho(X_1-X_2)+q(Y_1-Y_2)]} dS, \\ I(P_1) &= \textstyle\int_\sigma {A_1}^2 dS, \quad I(P_2) = \int_\sigma {A_2}^2 dS. \end{aligned}\right\} \tag{27}$$

Since to every source point $S(\xi, \eta)$ there corresponds a pair of ray components (p.q), we may transform the integrals over $\sigma$ into integrals over the solid angle

0

• A. Köhler, Zs. f. wiss. Mikrosk., 10 (1893), 433; 16 (1899), 1.

$$p^2 + q^2 \leqslant n_c'^2 \sin^2 \theta_c', \tag{28}$$

formed by the rays that are incident upon the object. Now within the accuracy of Gaussian optics, the relations $\xi = \xi(p,q), \eta = \eta(p,q)$ are linear; in fact, as is easily seen from the formulae of §4.3 (10), $\xi = fp, \eta = fq$ where $f$ is the focal length of the condenser system. Hence the Jacobian $\partial(\xi,\eta)/\partial(p,q)$ of the transformation is constant. Outside the domain of geometrical optics, the Jacobian will in general vary over the region of integration, but this variation may be assumed to be slow compared to the variation of the exponential term and may therefore be neglected. If we also neglect the slow variation of $A_1$ and $A_2$, (27) reduces to

$$\mu(P_1, P_2) = \frac{\iint_\Omega e^{ik_0[p(X_1-X_2)+q(Y_1-Y_q)]} dpdq}{\iint_\Omega dpdq} \tag{29}$$

where $\Omega$ denotes the domain (28). Evaluation of (29) leads to the expression

$$\mu(P_1, P_2) = \frac{2J_1(u_{12})}{u_{12}}, \quad u_{12} = \frac{2\pi}{\bar{\lambda}_0}\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}\, n_c' \sin\theta_c'. \tag{30}$$

This formula is identical with the formula (18) for critical illumination. Hence the complex degree of coherence of the light incident upon the object plane of a microscope is the same whether critical or Köhler's illumination is employed. In view of this result it is somewhat unfortunate that critical illumination is often designated as "incoherent" and Köhler's illumination as "coherent". It follows that formula (20) holds with both types of illumination, and that Fig. 10.13 likewise applies in both cases.

### 10.5.3 Imaging with partially coherent quasi-monochromatic illumination*

#### (a) Transmission of mutual intensity through an optical system

In § 9.5 some general methods were described for the study of imagery of extended objects. The case of completely coherent illumination (§ 9.5.1) and completely incoherent illumination (§ 9.5.2) were studied. In the first case the transmission of the complex amplitude through the system was considered, in the second case the transmission of the intensity. We shall now investigate the more general case of partially coherent quasi-monochromatic illumination. The appropriate quantity to consider in this case is the mutual intensity.

We shall employ the same scale-normalized (Seidel) coordinates as in §9.5 (1), so that the object point and its Gaussian image have the same coordinate numbers. Let $J_0(x_0, y_0; x_0', y_0')$ be the mutual intensity for points $(x_0, y_0), (x_0', y_0')$ in the object plane. If $K(x_0, y_0; x_1, y_1)$ is the transmission function of the system (§ 9.5.1), the mutual intensity in the image plane is, according to the propagation law § 10.4 (47), given by

$$J_1(x_1, y_1; x_1', y_1') = \iiint_{-\infty}^{+\infty} \int J_0(x_0, y_0; x_0', y_0')\, K(x_0, y_0; x_1, y_1)\, K^\star(x_0', y_0'; x_1', y_1')\, dx_0 dy_0 dx_0' dy_0' \tag{31a}$$

The integration extends only formally over an infinite domain, since $J_0$ is zero for all points in the object plane from which no light proceeds to the image plane.

As in §9.5, we assume that the object is so small that it forms an isoplanatic region of the system, i.e. that for all points on it $K(x_0, y_0; x_1, y_1)$ may be replaced to a good approximation by a function depending on the differences. $x_1 - x_0$ and $y_1 - y_0$ only, say $K(x_1 - x_0, y_1 - y_0)$. Equation (3la) then becomes

$$J_1(x_1,y_1;x_1',y_1') = \iiint_{-\infty}^{+\infty} \iint_0 (x_0,y_0;x_0',y_0') K(x_1-x_0,y_1-y_0) \times K^\star(x_1'-x_0',y_1'-y_0')\,dx_0dy_0dx_0'dy_0'. \tag{31b}$$

We represent $J_0, J_1$ and the product $KK^\star$ in the form of four-dimensional Fourier integrals:

$$J_0(x_0,y_0;x_0',y_0') = \iiint_{-\infty}^{+\infty} \int \mathscr{J}_0(f,g,;f',g')\, e^{-2\pi i\left(fx_0+\theta y_0+f'x_0'+g'y_0'\right)} df\,dg df'dg', \tag{32a}$$

$$J_1(x_1,y_1;x_1',y_1') = \iiint_{-\infty}^{+\infty} \int_{-\infty} \mathscr{I}_1(f,g;f',g')\, e^{-2\pi i\left(fx_1+qy_1+f'x_1'+g'y_1'\right)} df\,dg df'dg' \tag{32b}$$

$$K(x,y)K^\star(x',y') = \iiint_{-\infty}^{+\infty} \iint \mathscr{M}'(f,g;f',g')\, e^{-2\pi i(fx+qy+f'x'+o'y')} df\,dg df'dg' \tag{32c}$$

Then by the Fourier inversion formula

$$\mathscr{J}_0(f,g;f',g') = \iiint_{-\infty}^{+\infty} \int J_0(x_0,y_0;x_0',y_0')\, e^{2\pi i\left(fx_0+gy_0+f'x_0'+\sigma'y_0'\right)} dx_0dy_0dx_0'dy_0' \tag{33}$$

and there are strictly analogous relations for $\mathscr{J}_1$ and $\mathscr{M}$.
On applying the convolution theorem to (31b) we obtain the relation

$$\mathscr{J}_1(f,g;f',g') = \mathscr{J}_0(f,g;f',g')\,\mathscr{M}(f,g;f',g'). \tag{34}$$

This formula implies that, if the mutual intensity in the object and image planes are represented as superposition of four-dimensional space-harmonic components of all possible spatial frequencies ( $f,g,f',g'$ ), then each component in the image depends only on the corresponding component in the object, and the ratio of the components is equal to $\mathscr{A}$. Thus within the accuracy of the present approximation the action of the optical system on the mutual intensity is equivalent to the action of a four-dimensional linear filter. $\mathscr{M}$ is called the frequency response function for partially coherent quasimonochromatic illumination.

The frequency response function $\mathscr{A}$ is related in a simple way to the pupil function of the system. If as in §9.5 (10c), we represent $K$ in the form of a two-dimensional Fourier integral

$$K(x,y) = \int_{-\infty}^{+\infty} \int_{\mathscr{K}} \mathscr{K}(f,g) e^{-2\pi i(fx+ov)} df\,dg \tag{35}$$

and substitute for $K$ into the inverse of (32c), we find that

$$\mathscr{M}(f,g;f',g') = \mathscr{K}(f,g)\mathscr{K}^\star(-f',-g'). \tag{36}$$

But, according to §9.5 (13), $\mathscr{K}(f,g)$ is equal to the value of the pupil function $G(\xi,\eta)$ of the system at the point

$$\xi = \lambda Rf, \quad \eta = \lambda Rg \tag{37}$$

0

- The considerations of this section are based in parts on investigations of H. H. Hopkivs. Proc. Roy. Soc., A, 217 (1953), 408 and P. Dumontet, Publ. Sci. Univ. d'Alger, B, 1 (1955), 33.

on the Gaussian reference sphere (radius $R$ ). Hence the frequency response function for partially coherent quasi-monochromatic illumination is connected with the pupil function of the system by the formula

$$\mathscr{M}\left(\frac{\xi}{\lambda R}, \frac{\eta}{\lambda R}; \frac{\xi'}{\lambda R}, \frac{\eta'}{\lambda R}\right) = G(\xi, \eta) G^{\star}\left(-\xi', -\eta'\right). \tag{38}$$

Since the pupil function is zero for points outside the area of the exit pupil, it follows that spectral components belonging to frequencies above certain values are Dot transmitted. If the exit pupil is a circle of radius $a$, then $G(\xi, \eta) G^{\star}\left(-\xi', -\eta'\right)$
vanishes if $\xi^2 + \eta^2 > a^2$ or $\xi'^2 + \eta'^2 > a^2$. Hence spectral components of the mutual intensity belonging to frequencies $(f, g, f', g')$ for which either

$$f^2 + g^2 > \left(\frac{a}{\bar{\lambda} R}\right)^2 \text{ or } f'^2 + g'^2 > \left(\frac{a}{\bar{\lambda} R}\right)^2 \tag{39}$$

are not transmitted.* Here $\bar{\lambda}$ denotes the mean wavelength in the image space.

Table XXIV
The action of an optical system from the standpoint of its response to spatial frequencies. (Isoplanatic object region assumed)

| Illumination | Basic quantity | Transition from object to image | Fre |
|---|---|---|---|
| Coherent | Complex disturbance $U(x, y)$ | $\mathbb{Z}_1(f, g) = \mathbb{W}_0(f, g) \mathscr{X}(f, g)$ | $\mathscr{X}$ |
| Incoherent | Intensity $I(x, y)$ | $\mathscr{L}_1(f, g) = \mathscr{S}_0(f, g) \mathscr{L}(f, g)$ | $\mathscr{L}$ |
| Partially coherent | Mutual intensity $J\left(x, y; x', y'\right)$ | $\mathscr{I}_1\left(f, g; f', g'\right) = \mathscr{L}_0\left(f, g; f', g'\right) \ldots (f, g; f', g')$ | $\mathscr{M}$ |

$\boldsymbol{x}(f, g)$ is the Fourier inverse of the transmission function $K(x, y)$ of the system and is related to the pupil function $G(\xi, \eta)$ by the formula $\mathscr{X}(\xi/\lambda R, \eta/\lambda R) = G(\xi, \eta)$, where $R$ is the radius of the Gaussian reference sphere and $\lambda$ the mean wavelength in the image space.

In Table XXIV, the basic formulae relating to imaging with partially coherent illumination are displayed, together with the corresponding formulae of § 9.5, relating to coherent and incoherent illumination. The formulae relating to incoherent illumination may be derived from the general formulae (34), (36), and (38) by assuming $J_0$ to be of the form $J_0\left(x_0, y_0; x_0', y_0'\right) = I_0\left(x_0, y_0\right) \delta\left(x_0' - x_0\right) \delta\left(y_0' - y_0\right)$, where $\delta$ is the Dirac delta function (see Appendix IV). The calculations are straightforward but somewhat lengthy and will not be given here. Formulae relating to the special case of perfectly monochromatic (and hence completely coherent) illumination may be derived somewhat more easily by noting that in this case the mutual intensity is of the form $J_0\left(x_0, y_0; x_0', y_0'\right) = U_0\left(x_0, y_0\right) U_0^{\star}\left(x_0', y_0'\right)$. The Fourier transform $\mathscr{J}_0$ of $J_0$ now likewise splits into the product of two factors and the application of (34) and (36) shows that each spectral component of the complex disturbance $U_0$ is transmitted through the system in accordance with the formulae displayed in the first row of Table XXIV.

### (b) Images of transilluminated objects

Suppose that a portion of the object plane is occupied by a transparent or semitransparent object which is illuminated with partially coherent quasi-monochromatic

0

- If the angular aperture of the system is small and the sine condition is obeyed, we have, as in δ9.5, p. 483, $a/\lambda R \sim n_0 \sin\theta_0 / M\lambda_0$, where $n_0 \sin\theta_0$ is the numerical aperture of the system, $M$ the Gaussian magnification, and $\lambda_0$ is the mean wavelength in vacuum.

light. It will be assumed that this light originates in an incoherent source and reaches the object plane after the passage through some illuminating system (condenser).

As in §8.6.1, we specify the object by an appropriate transmission function $F_0(x_0, y_0)$. If $U_0(S; x_0, y_0)$ represents the disturbance at the point $(x_0, y_0)$ of the object plane due to a source point $S$ of the associated monochromatic source (cf. p. 513), the disturbance from this source point after the passage through the object is given by

$$U_0(S; x_0, y_0) = U_0(S; x_0, y_0) F(x_0, y_0). \tag{40}$$

Now according to § 10.4 (35a) the mutual intensity of the light incident on the object is given by

$$J_0^-(x_0, y_0; x_0', y_0') = \int_\sigma U_0^-(S; x_0, y_0) U_0^{-\star}(S; x_0', y_0')\, dS \tag{41a}$$

and the mutual intensity of the light emerging from the object is

$$J_0(x_0, y_0; x_0', y_0') = \int_0 U_0(S; x_0, y_0) U_0^\star(S; x_0', y_0')\, dS \tag{41b}$$

so that, because of (40),

$$J_0(x_0, y_0; x_0', y_0') = F(x_0, y_0) F^\star(x_0', y_0') J_0^-(x_0, y_0; x_0', y_0') \tag{42}$$

We shall confine our attention to the important case when the mutual intensity $J_0^-$ of the incident light depends on the four coordinates $x_0, y_0, x_0', y_0'$ through the differences $x_0 - x_0', y_0 - y_0'$ only, i.e. when $J_0$ is of the form

$$J_0(x_0, y_0; x_0', y_0') = J_0(x_0 - x_0', y_0 - y_0'). \tag{43}$$

We learned in § 10.5.2 that this will be the case for both the critical and the Köhler illumination. We retain the earlier assumptions that the object is so small that it forms an isoplanatic area of the system. It then follows from (31b) that the intensity $I_1(x_1, y_1) = J_1(x_1, y_1; x_1, y_1)$ in the image plane is given by

$$\begin{aligned} I_1(x_1, y_1) = \iiint_{-\infty}^{+\infty} J_0(x_0 - x_0', y_0 - y_0') F(x_0, y_0) F^\star(x_0', y_0') K(x_1 - x_0, y_1 - y_0) \\ \times K^\star(x_1 - x_0', y_1 - y_0')\, dx_0 dy_0 dx_0' dy_0' \end{aligned} \tag{44}$$

We represent $F$ and $J_0^-$ in the form of two-dimensional Fourier integrals

$$F(x, y) = \int_{-\infty}^{+\infty} \int^{\infty} \mathscr{F}(f, g) e^{-2\pi i(fx+gy)} df dg \tag{45a}$$

$$J_0^-(x, y) = \int_{-\infty}^{+\infty} \int_0 \mathscr{I}_0(f, g) e^{-2\pi i(fx+gy)} df dg \tag{45b}$$

If we substitute for $F$ and $F^\star$ from (45a) into (44), use the identity $f'x_0 - f''x_0' = (U'' - f'')x_1 - f'(x_1 - x_0) + f''(x_1 - x_0')$ and a similar identity involving $g$ and $y$, and introduce new variables of integration $u' = x_1 - x_0, u'' = x_1 - x_0'$, we obtain the following expression for $I_1$ :

$$I_1(x_1, y_1) = \iiint_{-\infty}^{+\infty} \int \mathscr{T}(f', g'; f'', g'') \mathscr{F}(f', g') \mathscr{F}\star(f^*, g'') e^{-2\pi i[(f'-f'')x_1 + (g'-\theta^*)y_1]} df' dg' df'' dg'' \tag{46}$$

where

$$\begin{aligned}
&\mathscr{T}(f',g';f'',g'')\\
&\quad=\iint_{-\infty}^{+\infty}\iint_{0} J_0(u''-u',v''-v')\,K(u',v')\,K^{\star}(u'',v'')\,e^{2\pi i(f'u'+g'v')-(f''u^{*}+g^{*}v^{*})]}du'dv'du''dv''\\
&\quad=\iiint\iiint \mathscr{J}_0(f,g)K(u',v')\,K^{\star}(u'',v'')\,e^{2\pi[(f+f')u'+(g+g')v'-(f+f'')u''-(g+g'')v'']}\quad df\,dg\,du'dv'du''dv''\\
&\quad=\int_{-\infty}^{+\infty}\int_{-\infty}\mathscr{J}_0(f,g)\mathscr{K}(f+f',g+g')\,\mathscr{K}^{\star}(f+f'',g+g'')\,df\,dg
\end{aligned}\tag{47}$$

In passing from the second to the third lines in (47) we substituted for $J_0$ from (45b) and in passing from the third to the fourth line we used the inverse of (35).

We see that in (46) the influence of the object (characterized by $\mathscr{F}$ ) and the combined effect of the illumination ( $\mathscr{J}_{\overline{0}}$) and of the system ( $\mathscr{K}$ ) are separated. With a uniform illumination ( $I_0$ = constant), the intensity of the light emerging from the object would be proportional to $|F|^2$, and were the imaging perfect, the intensity in the image plane would be given (apart from a constant factor) by

$$\begin{aligned}
\tilde{I}_1(x_1,y_1)&=F(x_1,y_1)\,F^{\star}(x_1,y_1)\\
&=\iiint_{-\infty}^{+\infty}\int\mathscr{F}(f',g')\,\mathscr{F}^{\star}(f'',g'')\,e^{-2\pi i(f'-f'')x_1+(g'-\sigma'')y_1]}df'dg'df''dg''
\end{aligned}\tag{48}$$

Equations (46) and (48) represent the true intensity $I_1$ and the ideal intensity $\tilde{I}_1$ as sums of contributions from all pairs of frequencies $(f',g'),(f'',g'')$ of the spatial spectrum of the object. Each contribution in the former case is $\mathscr{T}$ times that of the latter, and it follows that unless $\mathscr{T}$ is constant for all values $f',g',f'',g''$ for which both the spectral components $\mathscr{F}(f',g'),\mathscr{F}(f'',g'')$ differ from zero, some information about the object will be lost or falsified. The function $\mathscr{T}$ is called the transmission cross-coefficient of the system, working with the given transillumination.

Instead of the intensity itself, let us now consider its spatial spectrum $\mathscr{I}(f,g)$. To derive the appropriate formula we multiply both sides of (46) by $e^{2\pi i(fx_1+gy_1)}$ and integrate with respect to $x_1$ and $y_1$. If next we use the Fourier integral theorem (or more shortly the Fourier integral representation of the Dirac delta function (see Appendix IV), we obtain

$$\mathscr{I}_1(f,g)=\int_{-\infty}^{+\infty}\mathscr{T}(f'+f,g'+g,f',g')\,\mathscr{F}(f'+f,g'+g)\,\mathscr{F}^{\star}(f',g')\,df'dg'\tag{49}$$

For the ideal case represented by (48) we have

$$\tilde{\mathscr{I}}_1(f,g)=\int_{-\infty}^{+\infty}\int_{\mathscr{F}}(f'+f,g'+g)\,\mathscr{F}^{\star}(f',g')\,df'dg'\tag{50}$$

These formulae express $\mathscr{I}_1$ and $\tilde{\mathscr{I}}_1$ as the sum of contributions from each spatial frequency $(f',g')$ of the object structure. It is seen that $\mathscr{T}$ plays a similar role as before. It characterizes the changes which arise in each contribution from the mode
of illumination of the object and from the transmission characteristics of the imageforming system.

Since the response function $\mathscr{K}(f,g)$ is zero when the point $\xi=\bar{\lambda}Rf,\eta=\bar{\lambda}Rg$ lies outside the area of the exit pupil, it follows from (47) that $\mathscr{T}$ vanishes for sufficiently high frequencies. If the exit pupil is a circle of radius $a$, the product $\mathscr{K}(f+f'g+g')\,\mathscr{K}^{\star}(f+f'',g+g'')$ and consequently $\mathscr{T}(f',g';f''g'')$ can only differ from zero if the two circles $C'$ and $C''$, each of

radius $\sqrt{f^2+g^2} = a/\bar{\lambda}R$, centred on the points $O'(-f', -g'), O''(-f'', -g'')$ in the $f, g$-plane have a domain in common (Fig. 10.15).

To illustrate the effect of the illumination, suppose that the illumination is either critical or KöHLER, and that the numerical aperture $n'_c \sin\theta'_c$ of the condenser system is $m$ times the numerical aperture $n_0 \sin\theta_0$ of the system that images the object. Then according to (18) or (30) the mutual intensity of the illuminating beam is

$$J_0(x_0 - x'_0, y_0 - y'_0) = \left(\frac{2J_1(mv)}{mv}\right) I_0 \tag{51}$$

where

$$v = \frac{2\pi}{\bar{\lambda}_0}\sqrt{(x'_0 - x_0)^2 + (y'_0 - y_0)^2}\, n_0 \sin\theta_0, \tag{52}$$

and $I_0$ is the intensity (assumed uniform) of the incident light. Now the function on the right of (51) is a Fourier transform of the function (cf. § 8.5.2)

$$\left.\begin{aligned} \mathcal{F}_0(f,g) &= \text{constant} = \left(\tfrac{\bar{\lambda}_0^2}{\pi m^2 n_0^2 \sin^2\theta_0}\right) I_0^- && \text{when } f^2+g^2 < \tfrac{m^2 n_0^2 \sin^2\theta_0}{\bar{\lambda}_0^2} \\ &= 0 && \text{when } f^2+g^2 > \tfrac{m^2 n_0^2 \sin^2\theta_0}{\bar{\lambda}_0^2}. \end{aligned}\right\} \tag{53}$$

In Fig. 10.15 the circle outside which $\mathscr{J}_0$ vanishes is denoted by $C$ and it follows that for given $(f', g')$ and $(f'', g'')$ only those points of the $f, g$-plane which lie within the area (shown shaded) common to the three circles $C, C'$, and $C''$ contribute to the integral (47) for $\mathscr{T}$.

## 9.13.6 10.6 Some Theorems Relating to Mutual Coherence

In § 10.4 and § 10.5 we considered interference and diffraction of quasi-monochromatic light, and confined ourselves to situations where the time delay $\tau$ was small compared to the coherence time of the light. We have seen that to a good approximation the correlation functions then depend on $\tau$ only through a harmonic term, i.e. they are of the form

$$\Gamma(P_1, P_2, \tau) \sim J(P_1, P_2)\, e^{-2\pi i \sigma \tau}, \quad \gamma(P_1, P_2, \tau) \sim \mu(P_1, P_2)\, e^{-2\pi i \tau \tau}.$$

The elementary theory which makes use of this approximation is adequate to take into account the decrease in the visibility at the "centre" of the pattern ($\tau = 0$), due to a finite extension of the light source. However, it does not take into account the change in the visibility with increasing path difference. To describe adequately situations in which the time delay $\tau$ is not negligible in comparison with the coherence time, it is necessary to use more accurate expressions for the correlation functions. We shall now discuss the appropriate generalization of some of our formulae.

### 10.6.1 Calculation of mutual coherence for light from an incoherent source

Let $V_1(t)$ and $V_2(t)$ be the disturbances at points $P_1$ and $P_2$ in a wave field produced by an extended (not necessarily quasi-monochromatic) incoherent source $\sigma$. To begin with we assume that the medium between $\sigma$ and the points $P_1$ and $P_2$ is homogeneous.

As in § 10.4.2, we imagine the source to be divided into elements $d\sigma_1, d\sigma_2 \ldots$ centred about points $S_1, S_2, \ldots$, with linear dimensions that are small compared to the effective wavelengths. If $V_{m_1}(t)$ and $V_{m_2}(t)$ are the contributions to $V_1$ and $V_2$ from the element $d\sigma_m$, then

$$V_1(t) = \sum_m V_{m1}(t), \quad V_2(t) = \sum_m V_{m2}(t), \tag{1}$$

and the mutual coherence function is given by

$$\Gamma\left(P_1, P_2, \tau\right) = \left\langle V_1(t+\tau) V_2^\star(t)\right\rangle = \sum_m \left\langle V_{m1}(t+\tau) V_{m2}^\star(t)\right\rangle. \tag{2}$$

Terms of the type $\langle V_{m1}(t+\tau)V_{n2}^\star(t)\rangle\,(m \neq n)$ have been omitted on the right hand of (2), since the contributions from the different source elements may be assumed to be mutually incoherent.

We now proceed somewhat differently than in § 10.4. According to § 10.3 (30), each term under the summation sign in (2) may be expressed in the form

$$\left\langle V_{m1}(t+\tau) V_{m2}^\star(t)\right\rangle = 4\int_0^\infty G_m\left(P_1, P_2, v\right) e^{-2\pi i r T} dv \tag{3}$$

where

$$G_m\left(P_1, P_2, v\right) = \mathrm{Lim}_{T\to\infty}\left[\frac{\overline{v_{mT}\left(P_1, v\right) v_{mT}^\star\left(P_2, v\right)}}{2T}\right] \tag{4}$$

is the mutual spectral density of the disturbances $V_{m1} = V_m\left(P_1, t\right), V_{m2} = V_m\left(P_2, t\right)$. Now $v_m$ represents the contribution of the appropriate frequency to the disturbance
arising from the element $d\sigma_m$, and, since the medium is assumed to be homogeneous this contribution is propagated in the form of a spherical wave. Hence

$$v_{mT}\left(P_1, v\right) = a_{mT}(v)\frac{e^{ikR_{m1}}}{R_{m1}}, \quad v_{mT}\left(P_9, v\right) = a_{mT}(v)\frac{e^{ikR_2}}{R_{m2}}, \tag{5}$$

where $R_{m1}$ and $R_{m2}$ (assumed to be large compared to the effective wavelengths) represent the distances of $P_1$ and $P_2$ from the source point $S_m$, and $k = 2\pi\nu/\nu = 2\pi/\lambda$. The amplitude $|a_m(\nu)|$ of $a_m(\nu)$ represents the strength of the component of frequency $\nu$ from the element $d\sigma_m$, and $\arg a_m(\nu)$ represents its phase. Fiom (4) and (5),

$$G_m\left(P_1, P_2, v\right) = \left(\mathrm{Lim}_{T\to\infty}\frac{\overline{\left|a_{mT}(v)\right|^2}}{2T}\right)\frac{e^{ik(R_{mx}-R_m)}}{R_{m1}R_{m2}}. \tag{6}$$

The term in the large bracket on the right represents the spectral density of the light from the source element $d\sigma_m$. As in § 10.4.2, we assume that the number of the source elements is so large that we may treat the source as being effectively continuous. Hence if $I\left(S_m, v\right) d\sigma_m dv = 4\,\mathrm{Lim}_{T\to\infty}\left[\overline{\left|\,a_m T^(v)\right|}^2/2T\right] dv$, i.e. if $I(S, v)$ denotes the intensity per unit area of the source, per unit frequency range, then according to (2), (3), and (6),

$$\Gamma\left(P_1, P_2, \tau\right) = \sqrt{I\left(P_1\right)}\sqrt{I\left(P_2\right)}\gamma\left(P_1, P_2, \tau\right) = \int_0^\infty dv e^{-2\pi i r}\int_\sigma I(S, v)\frac{e^{ik(R_1-R_2)}}{R_1R_2}dS \tag{7}$$

where
$I\left(P_1\right) = \Gamma\left(P_1, P_1, 0\right) = \int_0^\infty dv\int_\sigma \frac{I(S,v)}{{R_1}^2}dS, \quad I\left(P_2\right) = \Gamma\left(P_2, P_2, 0\right) = \int_0^\infty dv\int_\sigma \frac{I(S,v)}{{R_2}^2}dS,$
are the intensities at $P_1$ and $P_2$, and $R_1$ and $R_2$ are the distances of these points from the source point $S$. Equations (7) are generalizations of the van Cittert-Zernike formulae, § 10.4 (20), (21).

If the medium between the source and the points $P_1$ and $P_2$ is not homogeneous, we may proceed as in § 10.4. We only have to replace the factors $e^{ikR_j}/R$, by $icK\left(S,P_j,v\right)/v$, where $K$ is the appropriate transmission function of the medium. In place of (7) we then obtain

$$\begin{aligned}\Gamma\left(P_1,P_2,\tau\right) &= \sqrt{I\left(P_1\right)}\sqrt{I\left(\dot{P}_2\right)}\gamma\left(P_1,P_2,\tau\right)\\ &= c^2\int_0^\infty \frac{dv}{v^2}e^{-2\pi i\tau}\int_\sigma I(S,v)K\left(S,P_1,v\right)K^\star\left(S,P_2,v\right)dS\end{aligned} \tag{9}$$

where

$$I\left(P_1\right)=c^2\int_0^\infty\frac{dv}{v^2}\int_\sigma I(S,v)\left|K\left(S,P_1,v\right)\right|^2dS,\quad I\left(P_2\right)=c^2\int_0^\infty\frac{dv}{v^2}\int_\sigma I(S,v)\left|K\left(S,P_2,v\right)\right|^2dS \tag{10}$$

By analogy with § 10.4, we re-write (9) or (10) in a slightly different form. We set

$$\frac{ic}{v}K\left(S,P_1,v\right)\sqrt{I(S,v)}=U\left(S,P_1,v\right),\quad \frac{ic}{v}K\left(S,P_2,v\right)\sqrt{I(S,v)}=U\left(S,P_2,v\right). \tag{11}$$

Then (9) and (10) become

$$\Gamma\left(P_1,P_2,\tau\right)=\sqrt{I\left(P_1\right)}\sqrt{I\left(P_2\right)}\gamma\left(P_1,P_2,\tau\right)=\int_0^\infty d\nu e^{-2\pi i\tau}\int_\sigma U\left(S,P_1,v\right)U^\star\left(S,P_2,v\right)dS \tag{12}$$

where

$$I\left(P_1\right)=\int_0^\infty d\nu\int_\sigma\left|U\left(S,P_1,v\right)\right|^2dS,\quad I\left(P_2\right)=\int_0^\infty d\nu\int_\sigma\left|U\left(S,P_2,v\right)\right|^2dS \tag{13}$$

The formula (12), which is a generalization of Hopkins' formula § 10.4 (35), expresses the mutual coherence function and the complex degree of coherence in terms of the light distribution arising from an associated fictitious source. For, according to (11), $U(S,P,v)$ may be regarded as the disturbance at $P$ due to a monochromatic point source of frequency $\nu$, of zero phase, and of strength proportional to $\sqrt{I(S,\nu)}$, situated at $S$.

### 10.6.2 Propagation of mutual coherence

As in § 10.4.4 let $\mathscr{A}$ be a fictitious surface which intercepts a beam of light from an extended incoherent source $\sigma$ (Fig. 10.8). We shall show how, from knowledge of the mutual coherence for all pairs of points on $\mathscr{A}$, its value on any other surface $\mathscr{B}$ illuminated by the light from $\mathscr{A}$ can be determined. For simplicity we assume that the medium between $\mathscr{A}$ and $\mathscr{B}$ has refractive index equal to unity.

The mutual coherence for any two points $Q_1$ and $Q_2$ on $\mathscr{B}$ may be calculated by the use of (12). In this formula $U$ represents a monochromatic disturbance and hence may be determined from the knowledge of the disturbance at all points of the surface $\mathscr{A}$ by means of the Huygens-Fresnel principle:

$$U\left(S,Q_1,v\right)=\int_{\mathscr{W}}U\left(S,P_1,v\right)\frac{e^{ik_{v_1}}}{s_1}\Lambda_1dP_1 \tag{14}$$

Here $s_1$ denotes the distance between a typical point $P_1$ on $\mathscr{A}$ and the point $Q_1$, and $\Lambda_1$ is the usual inclination factor. There is a strictly analogous formula for $U\left(S,Q_2,v\right)$. If we

substitute from these formulae into (12) we obtain, after changing the order of integration, the following expression for the mutual coherence:

$$\Gamma(Q_1, Q_2, \tau) = \iint_{\mathscr{A}} \frac{dP_1 dP_2}{s_1 s_2} \int_0^\infty J(P_1, P_2, v)\, \Lambda_1(v) \Lambda_2^\star(v) e^{-2\pi i v\left[\tau - \frac{s_1 - s_2}{c}\right]} dv, \tag{15}$$

where

$$J(P_1, P_2, v) = \int_\sigma U(S, P_1, v)\, U^\star(S, P_2, v)\, dS \tag{16}$$

The repeated integration in (15) means that the points $P_1$ and $P_2$ explore the surface $\mathscr{A}$ independently. Now the inclination factors $\Lambda_1(\nu)$ and $\Lambda_2(\nu)$ depend on the frequency through a multiplicative factor $\nu$ and change slowly with $v$ in comparison with the other terms. If the effective spectral range of the light is sufficiently small, we may replace these factors by $\bar{\Lambda}_1 = \Lambda_1(\bar{\vartheta}), \bar{\Lambda}_2 = \Lambda_2(\bar{\vartheta})$, where $\vec{\nu}$ denotes the mean
frequency of the light. The rest of the $\nu$-integral is, according to (12), precisely $\Gamma(P_1, P_2, \tau - (s_1 - s_2)/c)$. We thus finally obtain the formula

$$\Gamma(Q_1, Q_2, \tau) = \int_{\mathscr{A}\mathscr{A}} \frac{\Gamma\left(P_1, P_2, \tau - \frac{s_1 - s_2}{c}\right)}{s_1 s_2} \bar{\Lambda}_1 \bar{\Lambda}_2^\star dP_1 dP_2 \tag{17}$$

This is the required formula for the mutual coherence function at points $Q_1$ and $Q_2$ on the surface $\mathscr{B}$ in terms of the mutual intensity at all pairs of points on the surface $\mathscr{A}$.

Of particular interest is the case when the two points $Q_1$ and $Q_2$ coincide and $\tau = 0$. Denoting the common point by $Q$, the left-hand side of (17) reduces to the intensity $I(Q)$. Also, if we substitute for $\Gamma$ on the right in terms of the intensities and the correlation factor $\gamma$, the formula reduces to

$$I(Q) = \iint_{\mathscr{A}\mathscr{A}} \frac{\sqrt{I(P_1)}\sqrt{I(P_2)}}{s_1 s_2} \gamma\left(P_1, P_2, \frac{s_2 - s_1}{c}\right) \Pi_1 \bar{\Lambda}_2^\star dP_1 dP_2 \tag{18}$$

This formula expresses the intensity at an arbitrary point $Q$ as the sum of contributions from all pairs of elements of the arbitrary surface $\mathscr{A}$. Each contribution is directly proportional to the geometrical mean of the intensities at the two elements, inversely proportional to the product of their distances from $Q$, and is weighted by the appropriate value of the correlation factor $\gamma$.

The formulae (17) and (18) are generalizations, due to WolF, of the propagation law of Zernike (§ 10.4 (45)) and of the formula § 10.4 (46) for the intensity in a partially coherent wave field.

## 9.13.7 10.7 Rigorous Theory of Partial Coherence*

### 10.7.1 Wave equations for mutual coherence

Some of the theorems relating to the correlation functions, which we derived in the preceding sections are in certain respects similar to theorems relating to the complex disturbance itself. For example, the van Cittert-Zernike formula § 10.4 (21) for the complex degree of coherence in a plane illuminated by an extended quasi-monochromatic incoherent source was seen to be identical with a formula relating to the complex disturbance in a diffraction pattern arising from the diffraction by an aperture of the same size and shape as the source. Other examples are the laws for the propagation of the mutual intensity (§ 10.4 (45)) which was seen to resemble the HoxgensFresnel principle. Now the results relating to the complex disturbance may be regarded as approximate deductions from certain rigorous theorems, namely the formulae of

Helmholtz and KirchhofF (§8.3 (7), (13)), which are a consequence of the fact that the light disturbance is propagated as a wave. This analogy suggests that correlation is also propagated as a wave, and that our theorems are approximate formulations of some associated theorems of the Helmholtz-Kirchhoff type. It is not difficult to show that this in fact is the case.

Consider now a stationary wave field in vacuum and let $V(P_1, t)$ and $V(P_2, t)$

represent the disturbances at points $P_1$ and $P_2$ respectively. It will be convenient to begin by expressing the mutual coherence function in the more symmetrical form

$$\begin{aligned}\Gamma(P_1, P_2, t_1, t_2) &= \langle V(P_1, t_1 + t)\, V^\star(P_2, t_2 + t)\rangle \\ &= \mathrm{Lim}_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} V_T(P_1, t + t_1)\, V_T^\star(P_2, t + t_2)\, dt \end{aligned} \tag{1}$$

Further let

$$\nabla_1^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2} \tag{2}$$

be the Laplacian operator with respect to the Cartesian rectangular coordinates of $P_1$. On applying this operator to (1), and on interchanging the order of the various operations, we obtain

$$\nabla_1^2\Gamma(P_1, P_2, t_1, t_2) = \mathrm{Lim}_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} \left\{\left[\nabla_1^2 V_T(P_1, t + t_1)\right] V_T^\star(P_2, t + t_2)\right\} dt \tag{3}$$

Now the real part $V_T^{(r)}$ of $V_T$ represents the true physical wave field (e.g. a Cartesian component of the electric vector wave) and hence satisfies the wave equation

$$\nabla_1^2 V_T^{(r)}(P_1, t + t_1) = \frac{1}{c^2}\frac{\partial^2}{\partial t_1^2} V_T^{(r)}(P_1, t + t_1) \tag{4a}$$

The imaginary part $V_T^{(i)}$ of $V_T$, and hence $V_T$ itself, also satisfies the wave equation; this result follows immediately on taking the Hilbert transforms of both sides of (4a) and on using the fact that if two functions are Hilbert transforms of each other so are their derivatives. Hence*

$$\nabla_1^2 V_T(P_1, t + t_1) = \frac{1}{c^2}\frac{\partial^2}{\partial t_1^2} V_T(P_1, t + t_1) \tag{4b}$$

It follows that we may replace $\nabla_1{}^2$ by $\partial^2/c^2\partial t_1{}^2$ on the right of (3), and we obtain after again changing the order of the operations,
i.e.

$$\nabla_1^2\Gamma(P_1, P_2, t_1, t_2) = \frac{1}{c^2}\frac{\partial^2}{\partial t_1^2}\,\mathrm{Lim}_{T\to\infty} \frac{1}{2T} \int_{-\infty}^{\infty} V_T(P_1, t + t_1)\, V_T^\star(P_2, t + t_2)\, dt$$

$$\nabla_1^2\Gamma(P_1, P_2, t_1, t_2) = \frac{1}{c^2}\frac{\partial^2\Gamma(P_1, P_2, t_1, t_2)}{\partial t_1^2} \tag{5a}$$

In a strictly similar manner it also follows that

0

- The analysis of this section is in the main part based on investigations of E. Wolf, Proc. Roy. Sise., A, 230 (1955), 246; and Proc. Phys. Soc., 71 (1958), 257.

$$\nabla_2^2\Gamma\left(P_1,P_2,t_1,t_2\right)=\frac{1}{c^2}\frac{\partial^2\Gamma\left(P_1,P_2,t_1,t_2\right)}{\partial t_2^2} \tag{5b}$$

where ${\nabla_2}^2$ is the Laplacian operator with respect to the coordinates of the point $P_2$.

Now for a stationary field $\Gamma$ depends on $t_1$ and $t_2$ only through the difference $t_1-t_2=\tau$, and we then may write as before, $\Gamma\left(P_1,P_2,t_1,t_2\right)=\Gamma\left(P_1,P_2,\tau\right)$. Then $\partial^2/\partial{t_1}^2=\partial^2/\partial{t_2}^2=\partial^2/\partial\tau^2$ and we have from (5)

$$\nabla_1^2\Gamma\left(P_1,P_2,\tau\right)=\frac{1}{c^2}\frac{\partial^2\Gamma\left(P_1,P_2,\tau\right)}{\partial\tau^2} \tag{6a}$$

$$\nabla_2^2\Gamma\left(P_1,P_2,\tau\right)=\frac{1}{c^2}\frac{\partial^2\Gamma\left(P_1,P_2,\tau\right)}{\partial\tau^2} \tag{6b}$$

We see that in vacuum the mutual coherence obeys two wave equations.* Each of them describes the variation of the mutual coherence when one of the points ( $P_2$ or $P_1$ ) is fixed, while the other point and the parameter $\tau$ change. Now $\tau$ represents a time difference between the instants at which the correlation at the two points is considered, and in all experiments enters only in the combination $c\tau=\Delta\mathscr{S}$, i.e. as a path difference. The time itself has thus effectively been eliminated from our final description of the field. This is a particularly attractive feature of the theory of partial coherence, since in optical wave fields true time variations entirely escape detection. The basic entity in this theory, the mutual coherence function $\Gamma\left(P_1,P_2,\tau\right)$ is directly measurable, for example by means of the interference experiments described in § 10.3 and in § 10.4.

### 10.7.2 Rigorous formulation of the propagation law for mutual coherence

We again consider a stationary wave field in vacuum. Let $Q_1$ and $Q_2$ be any two points in the field and let $\mathscr{A}$ be any fictitious surface which surrounds these points. If ${\nabla_1}^2$ denotes the Laplacian operator with respect to the coordinates of $Q_1$, we have, according to (5a),

$$\nabla_1^2\Gamma\left(Q_1,Q_2,t_1,t_2\right)=\frac{1}{c^2}\frac{\partial^2\Gamma\left(Q_1,Q_2,t_1,t_2\right)}{\partial t_1^2} \tag{7}$$

It follows that we may apply to $\Gamma$ the KirchHoff integral formula (13) of §8.3. Thus $\Gamma\left(Q_1,Q_2,t_1,t_2\right)$ may be expressed in terms of the values of $\left[\Gamma\left(P_1,Q_2,t_1,t_2\right)\right]_1$, where $P_1$ takes on all positions on $\mathscr{A}$ and $[..]_1$ denotes retardation with respect to the first time argument, i.e.

$$\left[\Gamma\left(P_1,Q_2,t_1,t_2\right)\right]_1=\Gamma\left(P_1,Q_2,t_1-\frac{s_1}{c},t_2\right) \tag{8}$$

and $s_1$ is the distance between $P_1$ and $Q_1$ (Fig. 10.16). Written out explicitly, Kirchhoff's formula gives

$$\begin{aligned}&\Gamma\left(Q_1,Q_2,t_1,t_2\right)\\&=\frac{1}{4\pi}\int_A\left\{f_1\left[\Gamma\left(P_1,Q_2,t_1,t_2\right)\right]_1+g_1\left[\frac{\partial}{\partial t_1}\Gamma\left(P_1,Q_2,t_1,t_2\right)\right]_1+h_1\left[\frac{\partial}{\partial n_1}\Gamma\left(P_1,Q_2,t_1,t_2\right)\right]_1\right\}dP_1\end{aligned} \tag{9}$$

0

- The following alternative proof may be noted: since $V_T^{(r)}$ satisfies the wave equation, each of its spectral components $v_T(\nu)(-\infty\leqslant\nu\leqslant\infty)$ satisfies the Helmholtz equatic. Now according to eq. (18b) of § 10.2 the spectrum of $V_T=V_T^{(r)}+iV_T^{(i)}$ is $2v_T(v)$ or 0 according as $v\geqslant 0$. Hence each spectral component of $V_T$ also satisfies the Helmholtz equation, and consequently $V_T$ satisfies the wave equation.

$$\nabla_1^2 J(P_1,P_2)+k^2 J(P_1,P_2)=0,\quad \nabla_2^2 J(P_1,P_2)+k^2 J(P_1,P_2)=0$$

Here $\partial/\partial n_1$ denotes differentiation along the inward normal to $\mathscr{A}$ at $P_1$, and

$$f_1=\frac{\partial}{\partial n_1}\left(\frac{1}{s_1}\right),\quad g_1=-\frac{1}{cs_1}\frac{\partial s_1}{\partial n_1},\quad h_1=-\frac{1}{s_1}. \tag{10}$$

Now, according to (5b), we also have

$$\nabla_2^2\Gamma(P_1,Q_2,t_1,t_2)=\frac{1}{c^2}\frac{\partial^2\Gamma(P_1,Q_2,t_1,t_2)}{\partial t_2^2}, \tag{11}$$

where $\nabla_2{}^2$ is the Laplacian operator with respect to the coordinates of $Q_2$. Hence $\Gamma(P_1,Q_2,t_1,t_2)$ which appears on the right of (9) may be expressed in the form of a Kirchioff integral involving the values of $[\Gamma(P_1,P_2,t_1,t_2)]_2$, where $P_2$ takes on all possible positions on $\mathscr{A}$ and $[...]_2$ denotes retardation with respect to the second time argument, e.g.

$$[\Gamma(P_1,P_2,t_1,t_2)]_2=\Gamma\left(P_1,P_2,t_1,t_2-\frac{s_2}{c}\right). \tag{12}$$

and $s_2$ is the distance between $P_2$ and $Q_2$. Written out explicitly, the appropriate formula is

$$\begin{aligned}&\Gamma(P_1,Q_2,t_1,t_2)\\&=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_2\left[\Gamma(P_1,P_2,t_1,t_2)\right]_2+g_2\left[\frac{\partial}{\partial t_2}\Gamma(P_1,P_2,t_1,t_2)\right]_2+h_2\left[\frac{\partial}{\partial n_2}\Gamma(P_1,P_2,t_1,t_2)\right]_2\right\}dP_2\end{aligned} \tag{13}$$

Here $\partial/\partial n_2$ denotes differentiation along the inward normal at $P_2$, and $f_2,g_2,h_2$ are the same quantities as in (10) but with the suffix 1 replaced by 2 . We next differentiate (13) with respect to $t_1$ and $n_1$ and obtain

$$\begin{aligned}\frac{\partial}{\partial t_1}\Gamma(P_1,Q_2,t_1,t_2)&=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_2\left[\frac{\partial}{\partial t_1}\Gamma(P_1,P_2,t_1,t_2)\right]_2+g_2\left[\frac{\partial^2}{\partial t_1\partial t_2}\Gamma(P_1,P_2,t_1,t_2)\right]_2\right.\\&\qquad\left.+h_2\left[\frac{\partial^2}{\partial t_1\partial n_2}\Gamma(P_1,P_2,t_1,t_2)\right]_2\right\}dP_2\\\frac{\partial}{\partial n_1}\Gamma(P_1,Q_2,t_1,t_2)&=\frac{1}{4\pi}\int_{\mathscr{A}}\left\{f_2\left[\frac{\partial}{\partial n_1}\Gamma(P_1,P_2,t_1,t_2)\right]_2+g_2\left[\frac{\partial^2}{\partial n_1\partial t_2}\Gamma(P_1,P_2,t_1,t_2)\right]_2\right.\\&\qquad\left.+h_2\left[\frac{\partial^2}{\partial n_1\partial n_2}\Gamma(P_1,P_2,t_1,t_2)\right]_2\right\}dP_2\end{aligned} \tag{15}$$

We now substitute from (13), (14), and (15) into (9), and obtain the following expression for $\Gamma(Q_1,Q_2,t_1,t_2)$ :

0

- When $\tau$ is small compared to the coherence time, we have, according to § 10.4 (10), $\Gamma(P_1,P_2,\tau)\sim J(P_1,P_2)\,e^{-2\pi i\bar{\nu}}$. It follows from (6) that within the range of validity of the quasimonochromatic theory, the mutual intensity $J_{12}$ in vacuum obeys to a good approximation the Helmholtz equations

$$\Gamma(Q_1,Q_2,t_1,t_2)=\frac{1}{(4\pi)^2}\iint_{\mathscr{A}\mathscr{A}}\left\{f_1f_2[\Gamma]_{1,2}+f_1g_2\left[\frac{\partial}{\partial t_2}\Gamma\right]_{1,2}+f_1h_2\left[\frac{\partial}{\partial n_2}\Gamma\right]_{1,2}\right.$$
$$+g_1f_2\left[\frac{\partial}{\partial t_1}\Gamma\right]_{1,2}+g_1g_2\left[\frac{\partial^2}{\partial t_1\partial t_2}\Gamma\right]_{1,2}+g_1h_2\left[\frac{\partial^2}{\partial t_1\partial n_2}\Gamma\right]_{1,2}$$
$$\left.+h_1f_2\left[\frac{\partial}{\partial n_1}\Gamma\right]_{1,2}+h_1g_2\left[\frac{\partial^2}{\partial n_1\partial t_2}\Gamma\right]_{1,2}+h_1h_2\left[\frac{\partial^2}{\partial n_1\partial n_2}\Gamma\right]_{1,2}\right\}_1 dP_1dP_2 \tag{16}$$

where the first two arguments in $\Gamma$ on the right are $P_1$ and $P_2$ and $[\ldots]_{1,2}$ denotes retardation with respect to both the time arguments, e.g.

$$[\Gamma]_{1.2}=\Gamma\left(P_1,P_2,t_1-\frac{s_1}{c},t_2-\frac{s_2}{c}\right). \tag{17}$$

Finally, we make use of the assumption of stationarity which ensures that $\Gamma$ depends on the two time arguments through their difference only. We write as before $\Gamma(P_1,P_2,t_1,t_2)=\Gamma(P_1,P_2,\tau)$, $\quad\tau=t_1\sim t_2$. Then $\partial/\partial t_1=-\partial/\partial t_2=\partial/\partial\tau$, and (16) becomes*

$$\Gamma(Q_1,Q_2,\tau)=\frac{1}{(\mathbf{4}\pi)^2}\iint_{\mathscr{A},\mathscr{A}}\left\{f_1f_2[\Gamma]-f_1g_2\left[\frac{\partial}{\partial\tau}\Gamma\right]+f_1h_2\left[\frac{\partial}{\partial n_2}\Gamma\right]\right.$$
$$+g_1f_2\left[\frac{\partial}{\partial\tau}\Gamma\right]-g_1g_2\left[\frac{\partial^2}{\partial\tau^2}\Gamma\right]+g_1h_2\left[\frac{\partial^2}{\partial\tau\partial n_2}\Gamma\right]$$
$$\left.+h_1f_2\left[\frac{\partial}{\partial n_1}\Gamma\right]-h_1g_2\left[\frac{\partial^2}{\partial n_1\partial\tau}\Gamma\right]+h_1h_2\left[\frac{\partial^2}{\partial n_1\partial n_2}\Gamma\right]\right\}dP_1dP_2 \tag{18}$$

The first two arguments in $\Gamma$ on the right are $P_1$ and $P_2$, and [. . .] denotes "retardation" by the amount $(s_1-s_2)/c$, e.g.

$$[\Gamma]=\Gamma\left(P_1,P_2,\tau-\frac{s_1-s_2}{c}\right). \tag{19}$$

The formula (18) may be regarded as a rigorous formulation of the law for the propagation of the mutual coherence (§ 10.6 (17)). It expresses the value of the mutual coherence function for any two points $Q_1$ and $Q_2$ in terms of the values of this function and of some of its derivatives at all pairs of points on an arbitrary closed surface which surrounds both these points.

In the special case when $Q_1$ and $Q_2$ coincide and $\tau=0$, we obtain from (18), on substituting $\Gamma_{12}(\tau)=\sqrt{I_1}\sqrt{I_2}\gamma_{12}(\tau)$, the following expressions for the intensity:

$$I(Q)=\frac{1}{(4\pi)^2}\iint_{\mathscr{A}\mathscr{A}}\left(\sqrt{I_1}\sqrt{I_2}\left\{f_1f_2[\gamma]+(f_2g_1-f_1g_2)\left[\frac{\partial}{\partial\tau}\gamma\right]-g_1g_2\left[\frac{\partial^2}{\partial\tau^2}\gamma\right]\right\}\right.$$
$$+\sqrt{I_1}\left\{f_1h_2\frac{\partial}{\partial n_2}\left(\sqrt{I_2}[\gamma]\right)+g_1h_2\frac{\partial}{\partial n_2}\left(\sqrt{I_2}\left[\frac{\partial}{\partial\tau}\gamma\right]\right)\right\}$$
$$+\sqrt{I_2}\left\{f_2h_1\frac{\partial}{\partial n_1}\left(\sqrt{I_1}[\gamma]\right)-g_2h_1\frac{\partial}{\partial n_1}\left(\sqrt{I_1}\left[\frac{\partial}{\partial\tau}\gamma\right]\right)\right\}$$
$$\left.+h_1h_2\frac{\partial^2}{\partial n_1\partial n_2}\left(\sqrt{I_1}\sqrt{I_2}[\gamma]\right)\right)dP_1dP_2 \tag{20}$$

0

where $I_1$ and $I_2$ are the intensities at $P_1$ and $P_2$ respectively, $[\gamma] = \gamma\left(P_1, P_2, (s_2 - s_1)/c\right)$. etc. The formula (20) may be regarded as a rigorous formulation of the theorem expressed by equation (18), § 10.6. It gives the intensity at an arbitrary point $Q$ in terms of the distribution of the intensity and of the complex degree of coherence (and of some of the derivatives of these quantities) on an arbitrary surface surrounding $Q$.

### 10.7.3 The coherence time and the effective spectral width

The concept of the coherence time, which was found useful in many considerations involving polychromatic light, was introduced in §7.5.8 from the study of the disturbance resulting from superposition of identical wave trains of finite duration. We showed from a simple example (a random sequence of periodic wave trains) that the coherence time* $\Delta\tau$ and the effective spectral width $\Delta\nu = c\Delta\lambda/\bar{\lambda}^2$ of the resulting disturbance are connected by the order of magnitude relation

$$\Delta\tau\Delta\nu \sim 1. \tag{21}$$

We also mentioned that a relation of this type holds under more general conditions, provided that $\Delta\tau$ and $\Delta\nu$ are defined as suitable averages. In this section we shall define these quantities and establish the required reciprocity relation rigorously.

Suppose that a beam of light is divided at a point $P$ into two beams which are brought together after the introduction of a path difference $c\tau$ between them. The resulting interference effects are characterized by the self-coherence function

$$\Gamma(\tau) = \langle V(t+\tau)V^{\star}(t)\rangle = 4\int_0^{\infty} G(\nu)e^{-2\pi i\nu\tau}d\nu \tag{22}$$

where $V(t)$ is the complex disturbance at $P$ and $G(\nu)$ is the spectral density.
Since the degree of coherence of the two interfering beams is represented by $|\gamma(\tau)| = |\Gamma(\tau)|/\Gamma(0)$ it is reasonable, and mathematically convenient, to define the coherence time $\Delta\tau$ of the light at $P$ as the normalized root-mean-square width (r.m.s.) of the squared modulus of $\Gamma(\tau)$, i.e. †

$$(\Delta\tau)^2 = \frac{\int_{-\infty}^{\infty}\tau^2|\Gamma(\tau)|^2d\tau}{\int_{-\infty}^{+\infty}|\Gamma(\tau)|^2d\tau}. \tag{23}$$

Next we define the effective spectral width $\Delta\nu$ of the light at $P$ as the normalized r.m.s. width of the spectrum of $\Gamma$, i.e. as the normalized r.m.s. width of the square of the spectral density $G(\nu)$, taken over the range $\nu \geqslant 0$. Thus

$$(\Delta\nu)^2 = \frac{\int_0^{\infty}(\nu-\bar{\nu})^2G^2(\nu)d\nu}{\int_0^{\infty}G^2(\nu)d\nu}, \quad \bar{\nu} = \frac{\int_0^{\infty}\nu G^2(\nu)d\nu}{\int_0^{\infty}G^2(\nu)d\nu}. \tag{24}$$

To establish the required reciprocity relation we set

---

- Equation (18) applies for propagation from a closed surface $\alpha$ of arbitrary form. A much simpler formula exists for propagation from a plane surfuce. [Soe M. J. Beran and G. B. Parrent, Therary of Partial Coherence (Englewood Cliffs, N.J., 1964), $3.3.]

0

- To conform to the notation of the present chapter we now write $\Delta\tau$ in place of $\Delta t$.
† The average value $\bar{\tau} = \int_{-\infty}^{+\infty}|\Gamma(\tau)|^2d\tau/\int_{-\infty}^{+\infty}|\Gamma(\tau)|^2d\tau$ is zero since $|\Gamma(\tau)|$ is an even function of $\tau$.
For another definition of the coherence time, see L. Mandex, Proc. Phys. Soc., 74 (1959), 233. See also L. Mandel and E. Wolf, ibid, 80 (1962), 894.

$$\left.\begin{array}{rll} \xi &= v-\bar{v}, & \\ \Phi(\xi) &= 4G(\bar{v}+\xi) \quad \text{when} \quad \xi > -\bar{v}, & \left.\begin{array}{l}\\ \\ \end{array}\right\} \text{(b) (c)} \\ &= 0 \quad \text{when} \quad \xi < -\bar{v}, & \end{array}\right\} \tag{25}$$

We shall assume that $\Phi(\xi)$ is continuous everywhere $(-\infty < \xi < \infty)$; consequently $\Phi(-\tilde{y}) = G(0) = 0^*$. From (22) it follows that $\Psi$ and $\Phi$ form a Fourier transform pair,

$$\Psi(\tau) = \int_{-\infty}^{\infty} \Phi(\xi) e^{-2\pi i \xi \tau} d\xi, \quad \Phi(\xi) = \int_{-\infty}^{\infty} \Psi(\tau) e^{2\pi i \xi \tau} d\tau \tag{26}$$

The expressions for $\Delta\tau$ and $\Delta\nu$ become

$$(\Delta\tau)^2 = \frac{1}{N} \int_{-\infty}^{\infty} \tau^2 |\Psi(\tau)|^2 d\tau \tag{27}$$

$$(\Delta v)^2 = \frac{1}{N} \int_{-\infty}^{\infty} \xi^2 \Phi^2(\xi) d\xi \tag{28}$$

where

$$N = \int_{-\infty}^{\infty} |\Psi(\tau)|^2 d\tau = \int_{-\infty}^{\infty} \Phi^2(\xi) d\xi \tag{29}$$

Next we express the integral (28) in terms of $\Psi$. We have, on using the second relation in (26),

$$\begin{aligned} (\Delta\nu)^2 &= \frac{1}{N} \int_{-\infty}^{\infty} \xi^2 \Phi(\xi) d\xi \int_{-\infty}^{\infty} \Psi(\tau) e^{2\pi i \xi \tau} d\tau \\ &= \frac{1}{N} \int_{-\infty}^{\infty} \Psi(\tau) d\tau \left(\frac{1}{2\pi i}\right)^2 \frac{\partial^2}{\partial \tau^2} \int_{-\infty}^{\infty} \Phi(\xi) e^{2\pi i \xi \tau} d\xi \\ &= -\frac{1}{4}\pi^2 \frac{1}{N} \int_{-\infty}^{\infty} \Psi(\tau) \frac{\partial^2}{\partial \tau^2} \Psi^\star(\tau) d\tau \\ &= \frac{1}{4\pi^2} \frac{1}{N} \int_{-\infty}^{\infty} \left|\frac{\partial \Psi}{\partial \tau}\right|^2 d\tau \end{aligned} \tag{30}$$

- This condition is, in fact, necessary for the integral in the numerator of (23) to converge. When it is satisfied, as here assumed, the value of (23) remains unchanged when $\Gamma(\tau) = \langle V(t+\tau)V^\star(t)\rangle$ is replaced by the real correlation function $\Gamma^{(s)}(\tau) = \mathscr{R}\Gamma(\tau) = 2\left\langle V^{(s)}(t+\tau)V^{(s)}(t)\right\rangle$. [Cf. E. Wolf, Proc. Phys. Soc., B, 71 (1958), 257 ; R. Silverman, Trans. Inst. Rad. Engrs., CT-5 (1958), 84.] The more general case when $G(0) \neq 0$ was investigated by A. G. Mayer and E. A. Leontovich, Doklady Akad. Nauk SSSR, 4 (1934), 353 and by I. Kay and R. Silverman, Information and Control, 1 (1957), 64, 396. See also A. A. Kharkevich, Spectra and Analysis, Translated from Russian (New York, Consultants Bureau, 1960), § 12.

In passing from the second to the third line, the first relation in (26) was used, together with the relation $\Psi^*(-\tau) = \Psi^\star(\tau)$. The last line follows from the preceding one on integrating by parts and using the fact that $\Psi \to 0$ as $\tau \to \pm\infty$; this is so because the integral $\int_{-\infty}^{\infty} |\Psi(\tau)|^2 d\tau$ in (29) is assumed to be convergent.

It follows from (27), (29), and (30) that

$$(\Delta\tau)^2(\Delta\nu)^2 = \frac{1}{16\pi^2}\left[\frac{4\left(\int_{-\infty}^{\infty}\tau^2|\Psi(\tau)|^2d\tau\right)\left(\int_{-\infty}^{\infty}\left|\frac{\partial\Psi}{\partial\tau}\right|^2 d\tau\right)}{\left(\int_{-\infty}^{\infty}|\Psi(\tau)|^2d\tau\right)^2}\right]. \tag{31}$$

Now by a straightforward algebraical argument given in Appendix VIII, the term in the large brackets on the right side of (31) is greater than or equal to unity for any function $\Psi$ for which the integrals exist. Hence we have established the following reciprocity inequality* for the coherence time and the effective spectral width:

$$\Delta\tau\Delta v \geqslant \frac{1}{4\pi} \tag{32}$$

We recall that when the light is quasi-monochromatic and the intensities of the two interfering beams are equal, the degree of coherence $|\gamma_{11}(\tau)| = |\Gamma_{11}(\tau)|/\Gamma_{11}(0)$ is according to § 10.4 (5) equal to the visibility $\mathscr{V}(\tau)$ of the fringes at a point corresponding to the path difference $c\tau$ between the two beams. Hence (23) may then be written in the form

$$(\Delta\tau)^2 = \frac{\int_{-\infty}^{\infty}\tau^2\mathscr{V}^2(\tau)d\tau}{\int_{-\infty}^{\infty}\mathscr{V}^2(\tau)d\tau} = \frac{\int_{0}^{\infty}\tau^2\mathscr{V}^2(\tau)d\tau}{\int_{0}^{\infty}\mathscr{V}^2(\tau)d\tau} \tag{33}$$

thus when the interfering beams are of equal intensity, the coherence time $\Delta\tau$ is equal to the normalized r.m.s. width of the square of the visibility function.

The present definition of the coherence time is more satisfactory than that given in §7.5.8, for we have now made no special assumptions about the nature of the elementary fields which give rise to the disturbance. In fact we now no longer require the knowledge of the detailed behaviour of the rapidly fluctuating function $V(t)$, our definition being based on the measurable correlation function $\Gamma(\tau)$. If we wish to retain the description of interference phenomena in terins of elementary wave trains,

we may regaid $\Delta\tau$ as the duration of an average wave train; this interpretation must, however, be employed with caution.

Returning to (32) we see that the equality sign only holds when the term in the large brackets on the right-hand side of (31) is equal to unity, and this, according to Appendix VIII, is only possible when $\Psi(\tau)$ is a Gaussian function. Now the Fourier transform of a Gaussian function is again a Gaussian function, and as this function differs from zero for all values of its argument ( $-\infty < \xi < \infty$ ) it does not obey the second condition (25b). Thus the equality sign in (32) never applies. However, when the Gaussian function is centred on a frequency which is large compared to its r.m.s. width, the contribution to $\overline{\boldsymbol{y}}$ and $\Delta\boldsymbol{v}$ from the negative frequency range is negligible and it is evident that, for the high frequency spectra encountered in optics, the value of the product $\Delta\tau\Delta y$ cannot differ appreciably from that which corresponds to the full Gaussian curve. Thus the inequality in (32) may be replaced by the order of magnitude sign, i.e.

$$\Delta\tau\Delta\nu \sim \frac{1}{4\pi} \tag{34}$$

The definition of the coherence time just given is appropriate when the two interfering beams are obtained from a single beam by division at a point $P$. The definition can be extended to

0

- Our derivation is modelled on that given by H. WEYL and W. Pauli in connection with the Heisenberg uncertainty relation. [H. Weyl, The Theory of Groups and Quantum Mechanics (London, Methuen, 1931; also Dover Publications, New York), PP. 77 and 393.]

situations where the two interfering beams are derived by division at two points $P_1$ and $P_2$, for example, in Young's interference experiment. The generalization is straightforward. We have to employ the mutual coherence function $\Gamma_{12}(\tau) = \Gamma(P_1, P_2, \tau)$ in place of the self-coherence function $\Gamma(\tau) = \Gamma(P, P, \tau)$ and the mutual spectral density $G_{12}(v)$ in place of the ordinary spectral density $G(\nu)$. The only difference arises from the fact that $G_{12}(\nu)$ is now complex and that $\Gamma_{12}(\tau)$ is no longer necessarily an even function of $\tau$ and consequently $\tau$ is not necessarily zero. The appropriate definitions are:

$$(\Delta\tau_{12})^2 = \frac{\int_{-\infty}^{\infty}(\tau-\bar{\tau}_{12})^2\,|\Gamma_{12}(\tau)|^2\,d\tau}{\int_{-\infty}^{\infty}|\Gamma_{12}(\tau)|^2\,d\tau}\bar{\tau}_{12} = \frac{\int_{-\infty}^{\infty}\tau\,|\Gamma_{12}(\tau)|^2\,d\tau}{\int_{-\infty}^{\infty}|\Gamma_{12}(\tau)|^2\,d\tau} \tag{35}$$

$$(\Delta\nu_{12})^2 = \frac{\int_{0}^{\infty}(\nu-\bar{\nu}_{12})^2\,|G_{12}(\nu)|^2\,d\nu}{\int_{0}^{\infty}|G_{12}(\nu)|^2\,d\nu},\bar{\nu}_{12} = \frac{\int_{0}^{\infty}\nu\,|G_{12}(\nu)|^2\,d\nu}{\int_{0}^{\infty}|G_{12}(\nu)|^2\,d\nu}$$

$\Delta\tau_{12}$ may be called the mulual coherence time and $\Delta\nu_{12}$ the mutual effective spectral width of the light at $P_1$ and $P_2$. By an obvious modification of the argument given in connection with $\Gamma_{11}(\tau)$ it follows that these quantities satisfy the reciprocity inequality

$$(\Delta\tau_{12})\,(\Delta\nu_{12}) > \frac{1}{4\pi} \tag{37}$$

Finally, for quasi-monochromatic light we now have the following relations as generalization of (33):

$$(\Delta\tau_{12})^2 = \frac{\int_{-\infty}^{\infty}(\tau-\bar{\tau}_{12})^2\,\mathscr{V}_{12}^2(\tau)d\tau}{\int_{-\infty}^{\infty}\mathscr{V}_{12}^2(\tau)d\tau}, \quad \bar{\tau}_{12} = \frac{\int_{-\infty}^{\infty}\tau\mathscr{H}_{12}^2(\tau)d\tau}{\int_{-\infty}^{\infty}\mathscr{V}_{12}^2(\tau)d\tau}. \tag{38}$$

Here $\mathscr{V}_{12}(\tau)$ represents the visibility of the fringes formed by the light from $P_1$ and $P_8$, when the interfering beams have the same intensity.

## 9.13.8 10.8 Polarization Properties of Quasi-monochromatic Light

In the preceding sections of this chapter we have treated the light disturbance as a scalar quantity. We shall now briefly consider some of the vectorial properties of a quasi-monochromatic light wave.

We have learned in § 1.4 that strictly monochromatic light is always polarized, i.e. that with increasing time the end point of the electric (and also of the magnetic) vector at each point in space moves periodically around an ellipse, which may, of course, reduce in special cases to a circle or a straight line. We have also encountered unpolarized light. In this case the end point may be assumed to move quite irregularly, and the light shows no preferential directional properties when resolved in different directions at right angles to the direction of propagation. Like complete coherence and complete incoherence these two cases represent two extremes. In general the variation of the field vectors is neither completely regular, nor completely irregular, and we may say that the light is partially polarized. Such light arises usually from unpolarized light by reflection (cf. §1.5.3) or scattering (cf. § 13.5.2). In this section we shall investigate the main properties of a partially polarized light wave. We shall see that its observable effects depend on the intensities of any two mutually orthogonal components of the electric vector at right angles to the direction of propagation, and on the correlation which exists between them.

### 10.8.1 The coherency matrix of a quasi-monochromatic plane wave*

Consider a quasi-monochromatic light wave of mean frequency $\bar{\nu}$ propagated in the positive $z$-direction. Let

$$E_x(t) = a_1(t)e^{i[\phi_1(t)-2\pi it]}, \quad E_y(t) = a_2(t)e^{i[\phi_x(t)-2\pi\bar{\nu}t]} \tag{1}$$

represent the componerts, at a point $O$, of the electric vector in two mutually orthogonal directions at right angles to the direction of propagation. We again use the complex representation discussed in § 10.2 , in which $E_x$ and $E_y$ are the "analytic signals" associated with the true (real) components $E_x{}^{(r)} = a_1(t)\cos[\phi_1(t) - 2\pi\hat{\nu}l]$, $E_y{}^{(r)} = a_2(t)\cos[\phi_2(t) - 2\pi\bar{\nu}t]$. If the light were strictly monochromatic, $a_1, a_2, \phi_1$ and $\phi_2$ would be constants. For a quasi-monochromatic wave these quantities

depend also on the time $t$, but, as we have seen, they change only by small relative amounts in any time interval that is small compared to the coherence time, i.e. that is small compared to the reciprocal of the effective spectral width $\Delta\nu$ of the light.

Suppose that the $y$-component is subjected to a retardation $\varepsilon$ with respect to the $\dot{x}$-component (this can be done, for example, by means of one of the compensators described in § 14.4.2), and consider the intensity $I(\theta, \varepsilon)$ of the light vibrations in the direction which makes an angle $\theta$ with the positive $x$-direction (Fig. 10.17). This intensity would be observed by sending the light through a polarizer (§ 14.4.1) with the appropriate orientation.

The component of the electric vector in the $\theta$ direction, after the retardation $\varepsilon$ has been introduced, is

$$E(t;\theta,\varepsilon) = E_x\cos\theta + E_y e^{it}\sin 0 \tag{2}$$

so that

$$\begin{aligned} I(\theta,\varepsilon) &= \langle E(t;\theta,\varepsilon)E^{\star}(t;\theta,\varepsilon)\rangle \\ &= J_{xx}\cos^2\theta + J_{vy}\sin^2\theta + J_{xy}e^{-i\varepsilon}\cos\theta\sin\theta + J_{vx}e^{i\varepsilon}\sin\theta\cos\theta \end{aligned} \tag{3}$$

where $J_{xx}, \ldots$ are the elements of the matrix

$$J = \begin{bmatrix} \langle E_x E_z^{\star}\rangle & \langle E_x E_y^{\star}\rangle \\ \langle E_y E_x^{\star}\rangle & \langle E_v E_y^{\star}\rangle \end{bmatrix} = \begin{bmatrix} \langle a_1^2\rangle & \langle a_1 a_2 e^{i(\phi_1-\phi_2)}\rangle \\ \langle a_1 a_2 e^{-i(\phi_1-\phi_2)}\rangle & \langle a_2^2\rangle \end{bmatrix}. \tag{4}$$

The diagonal elements of $J$ are real and are seen to represent the intensities of the components in the $x$ - and $y$-directions. Hence the trace $\mathrm{Tr}\, J$ of the matrix, i.e. the sum of its diagonal elements, is equal to the total intensity of the light,

$$\mathrm{Tr}\, J = J_{xx} + J_{vv} = \langle E_x E_x^{\star}\rangle + \langle E_y E_y^{\star}\rangle. \tag{5}$$

The non-diagonal elements are in general complex, but they are conjugates of each other. (A matrix such as this, which satisfies the relation $J_{ji} = J_i^{\star}$, for all $i$ and $j$ is said to be a Hermitian matrix.)

We shall normalize the mixed term $J_{xy}$ in a similar way as before (cf. § 10.4 (9b)), by setting

$$\mu_{xy} = |\mu_{xy}|\, e^{i\beta_{xy}} = \frac{J_{xy}}{\sqrt{J_{xx}}\sqrt{J_{yv}}} \tag{6}$$

Then it follows by Schwarz' inequality, in the same way as in connection with $10.3 (17), that

0

- The analysis given in § 10.8.1 and 10.8.2 is based on investigations of E. Wolf, Nuoro Cimento, 13 (1959), 1165. Some further developments are described in a paper by G. B. Parrent and P. Roman, ibid, 15 (1960), 370.

$$|\mu_{rv}| \leqslant 1. \tag{7}$$

This complex correlation factor $\mu_{xy}$ has a similar significance as the complex degree of coherence $\mu_{12}$ introduced in § 10.4.1. It is a measure of the correlation between the components of the electric vector in the $x$ - and $y$-directions. The absolute value $|\mu_{xy}|$ is a measure of their "degree of coherence," and its phase $\beta_{xy}$ is a measure of their "effective phase difference." We call J the coherency matrix of the light wave. Since $J_{xx}$ and $J_{yy}$ cannot be negative, (6) and (7) imply that the associated determinant is non-negative, i.e. that

$$|J| = J_{xx}J_{yy} - J_{xy}J_{vx} \geqslant 0. \tag{8}$$

If we use the relation $J_{yx} = J_{xy}^{\star}$, and denote by $\mathscr{R}$ the real part, (3) becomes

$$\begin{aligned} I(\theta, \varepsilon) &= J_{xx}\cos^2\theta + J_{yy}\sin^2\theta + 2\cos\theta\sin\theta\mathscr{R}\left(J_{xy}e^{-i\varepsilon}\right) \\ &= J_{xx}\cos^2\theta + J_{yy}\sin^2\theta + 2\sqrt{J_{xx}}\sqrt{J_{yy}}\cos\theta\sin\theta\,|\mu_{xy}|\cos\left[\beta_{xy} - \varepsilon\right], \end{aligned} \tag{9}$$

where we substituted from (6) on going from the first to the second line. If we set $J_{xx}\cos^2\theta = I^{(1)}, J_{yy}\sin^2\theta = I^{(2)}$, the last formula becomes identical with the basic interference law § 10.4 (11) of quasi-monochromatic wave fields.

Like the coherence functions which we considered earlier, the elements of the coherency matrix of a given wave may be determined by means of relatively simple experiments. This may be done in many different ways. One only needs to measure the intensity for several different values of $\theta$ (orientation of the polarizer) and $\varepsilon$ (delay introduced by the compensator), and solve the corresponding relations obtained from (3). Let $\{\theta, \varepsilon\}$ denote the measurement corresponding to a particular pair, $\theta, \varepsilon$. A convenient set of measurements is the following:

$$\{0^\circ, 0\}, \{45^\circ, 0\}, \{90^\circ, 0\}, \{135^\circ, 0\}, \left\{45^\circ, \frac{\pi}{2}\right\}, \left\{135^\circ, \frac{\pi}{2}\right\}. \tag{10}$$

It follows from (3) that, in terms of the intensities determined from these six measurements, the elements of the coherency matrix are given by

$$\left.\begin{aligned} J_{xx} &= I\,(0^\circ, 0) \\ J_{yy} &= I\,(90^\circ, 0) \\ J_{xy} &= \tfrac{1}{2}\left\{I\,(45^\circ, 0) - I\,(135^\circ, 0)\right\} + \tfrac{1}{2}i\left\{I\left(45^\circ, \tfrac{\pi}{2}\right) - I\left(135^\circ, \tfrac{\pi}{2}\right)\right\}, \\ J_{vx} &= \tfrac{1}{2}\left\{I\,(45^\circ, 0) - I\,(135^\circ, 0)\right\} - \tfrac{1}{2}i\left\{I\left(45^\circ, \tfrac{\pi}{2}\right) - I\left(135^\circ, \tfrac{\pi}{2}\right)\right\}. \end{aligned}\right\} \tag{11}$$

We see that only a polarizer is needed to determine $J_{xx}, J_{yy}$ and the real part of $J_{xy}$ (or $J_{yx}$ ). $J_{xx}$ and $J_{yy}$ may be determined from measurements with a polarizer oriented so as to transmit the components in the azimuths $\theta = 0$ and $\theta = 90^\circ$ respectively. The real part of $J_{xy}$ may be determined from measurements with a polarizer oriented so that it first transmits the component in the azimuth $\theta = 45^\circ$ and then the component in the azimuth $\theta = 135^\circ$. To determine the imaginary part of $J_{xy}$ (or $J_{yz}$ ) we also need, according to the last two relations in (11), a compensator which introduces a phase difference of a quarter period (e.g. a quarter-wave plate, cf. §14.4.2) between the $x$ - and $y$-components; the polarizer is again used, oriented so as to pass first the component in the azimuth $\theta = 45^\circ$ and then the component in the azimuth $0 = 135^\circ$. We shall learn in § 14.4.2, that the last two measurements are those made to detect right-handed and left-handed circular polarization.

It is evident from (9) that two beams of light which have the same coherency matrix are equivalent in the sense that they will yield the same (time averaged) intensity in similar experiments with a polarizer and a compensator.*

Let us now examine how the observed intensity $I(\theta, \varepsilon)$ changes for a given wave when one of the arguments ( $\theta$ or $\varepsilon$ ) is kept fixed, whilst the other varies. Suppose first that we keep $\theta$ fixed and change $\varepsilon$. We see from (9) that the intensity varies sinusoidally between the values and

$$\left.\begin{aligned} I_{\max(\varepsilon)} &= J_{xx}\cos^2\theta + J_{yy}\sin^2\theta + 2\,|J_{xy}|\sin\theta\cos\theta, \\ I_{\min(\varepsilon)} &= J_{xx}\cos^2\theta + J_{yy}\sin^2\theta - 2\,|J_{xy}|\sin\theta\cos\theta. \end{aligned}\right\} \tag{12}$$

Hence

$$\frac{I_{\max(\varepsilon)} - I_{\min(\varepsilon)}}{I_{\max(\delta)} + I_{\min(\varepsilon)}} = \frac{|J_{xy}|\sin 2\theta}{J_{xx}\cos^2\theta + J_{yy}\sin^2\theta^{\circ}} \tag{13}$$

Equation (13) indicates an alternative way of determining the absolute value of $J_{zy}$ (and hence also of $|\mu_{xy}|$ ); it shows that this quantity may be obtained from measurements of $J_{xx}, J_{yy}, I_{\max(e)}$, and $I_{\min(\varepsilon)}$; the phase of $J_{xy}$ may be obtained from measurement of the value of $\varepsilon$ at which the maxima or minima occur. For, according to (9),

$$\left.\begin{aligned} I = I_{\max(\varepsilon)} \quad &\text{when} \quad \varepsilon = \beta_{xy} \pm 2m\pi \qquad (m = 0, 1, 2, \ldots.), \\ I = I_{\min(\varepsilon)} \quad &\text{when} \quad \varepsilon = \beta_{xy} \pm (2m+1)\pi \quad (m = 0, 1, 2, \ldots.). \end{aligned}\right\} \tag{14}$$

To see how the intensity changes when $\varepsilon$ is fixed and $\theta$ is varied, it is convenient to re-write (9) in a somewhat different form. Simple calculation gives
where

$$I(\theta, \varepsilon) = \frac{1}{2}\,(J_{xx} + J_{yy}) + R\cos(2\theta - \alpha), \tag{15}$$

$$\begin{aligned} R &= \frac{1}{2}\sqrt{(J_{xx} - J_{vy})^2 + 4J_{xy}J_{yx}\cos^2(\beta_{xy} - \varepsilon)}, \\ \tan\alpha &= \frac{2\,|J_{xy}|\cos(\beta_{xy} - \varepsilon)}{J_{xx} - J_{vy}}. \end{aligned} \tag{16}$$

We see from (15) that, as $\theta$ changes, the intensity again varies sinusoidally; its extremes are

$$\left.\begin{aligned} I_{\max(\theta)} &= \tfrac{1}{2}\,(J_{xx} + J_{yy}) + R, \\ I_{\min(\theta)} &= \tfrac{1}{2}\,(J_{xx} + J_{yy}) - R. \end{aligned}\right\} \tag{17}$$

On the right of (17) only $R$ depends on $\varepsilon$. It takes its largest value when $\cos^2(\beta_{xy} - \varepsilon) = 1$, i.e. when $\varepsilon$ has one of the values given by (14), and is then equal in

$$\begin{aligned} R_{\max(e)} &= \frac{1}{2}\sqrt{(J_{xx} - J_{yy})^2 + 4J_{xy}J_{yx}} \\ &= \frac{1}{2}\,(J_{xx} + J_{yy})\sqrt{1 - \frac{4|J|}{(J_{xx} + J_{yy})^2}} \end{aligned} \tag{18}$$

So Sof it is only within the range of validity of this theory that the expression (9) for the intensity bolid. The two beams may behave quite differently when the phase delay $\varepsilon$ introduced between Ur. :wo orthogonal components is not negligible in comparison with the coherence length, mairvi in units of the mean wavelength $\lambda$. For a fuller description of the observable properties of a beam it is necessary to introduce more general coherency matrices which characterree the correlations between the components at different times and also at different points. CI. E. Wole, Nuovo Cimento, 12 (1954), 884; also his contribution in Proc.

Symp. Astronom. Optes. ed. Z. Kopal (Amsterdam, North-Holland Publishing Co., 1956), 177: P. Roman and E. Wry. Sumnt.imento, 17 (1960). 462. 477; P. Roman, ibid.. 20 (1961), 750; ibid., 22 (1961), 1065.

where $|J|$ is the determinant (8) of the coherency matrix. It follows that the absolute maxima and minima (with respect to both $\theta$ and $\varepsilon$ ) of the intensity are

$$\left.\begin{aligned} I_{\max(\theta,\varepsilon)} &= \tfrac{1}{2}\left(J_{xx}+J_{yy}\right)\left[1+\sqrt{1-\tfrac{4|J|}{(J_{xx}+J_{yy})^2}}\right], \\ I_{\min(\theta,\varepsilon)} &= \tfrac{1}{2}\left(J_{xx}+J_{yy}\right)\left[1-\sqrt{1-\tfrac{4|J|}{(J_{xx}+J_{yy})^2}}\right]. \end{aligned}\right\} \tag{19}$$

Hence

$$\frac{I_{\max(\theta,\varepsilon)}-I_{\min(0,\varepsilon)}}{I_{\max(0,\varepsilon)}+I_{\min(\theta,\varepsilon)}}=\sqrt{1-\frac{4|J|}{\left(J_{xx}+J_{yy}\right)^2}}. \tag{20}$$

We shall see later that this quantity has a simple physical meaning.

So far we have referred the electric vibrations to arbitrary but fixed rectangular axes $OX, OY$. We shall now consider how the coherency matrix transforms when a new set of axes is chosen. Suppose that we take new rectangular axes $OX', OY'$, again in the plane perpendicular to the direction of propagation, such that $OX'$ makes an angle @ with $OX$ (see Fig. 10.17). In terms of $E_x, E_y$, the components of the electric vector referred to the new axes are

$$\left.\begin{aligned} E_{x'} &= E_x\cos\Theta+E_y\sin\Theta, \\ E_{y'} &= -E_x\sin\Theta+E_y\cos\Theta. \end{aligned}\right\} \tag{21}$$

The elements of the transformed coherency matrix /′ are

$$J_{k'r'}=\left\langle E_{k'}E^{\star}_{v'}\right\rangle, \tag{22}$$

where $k'$ and $l'$ each take on the values $x'$ and $y'$. From (21) and (22) it follows that

$$J'=\begin{bmatrix} J_{xxx}c^2+J_{yy}s^2+\left(J_{xy}+J_{yx}\right)cs & \left(J_{yy}-J_{xx}\right)cs+J_{xy}c^2-J_{yx}s^2 \\ \left(J_{yy}-J_{xx}\right)cs+J_{yx}c^2-J_{xy}s^2 & J_{xxx}s^2+J_{yy}c^2-\left(J_{xy}+J_{yx}\right)cs \end{bmatrix}, \tag{23}$$

where

$$c=\cos\Theta, \quad s=\sin\Theta. \tag{24}$$

It is seen that the trace of the matrix is invariant under rotation of the axes. A straightforward calculation shows that its determinant is also invariant under this transformation. Both these results also follow from well-known theorems of matrix algebra.

We shall now consider the form of the coherency matrix for some cases of particular interest.

(a) Completely unpolarized light (Natural light)

Light which is most frequently encountered in nature has the property that the intensity of its components in any direction perpendicular to the direction of propagation is the same; and, moreover, this intensity is not affected by any previous retardation of one of the rectangular components relative to the other, into which the light may have been resolved. In other words

$$I(\theta,\varepsilon)=\text{ constant} \tag{25}$$

for all values of $\theta$ and $\varepsilon$. Such light may be said to be completely unpolarized and is often also called natural light.

It is evident from (9) that $I(\theta,\varepsilon)$ is independent of $\varepsilon$ and $\theta$ if and only if

$$\mu_{xy}=0 \quad \text{and} \quad J_{xx}=J_{yy}. \tag{268}$$

The first condition implies that $E_x$ and $E_y$ are mutually incoherent. According to (6) and the relation $J_{vx} = J_{xy}^{\star}$, (26a) may also be written as

$$J_{xy} = J_{vx} = 0, \quad J_{xx} = J_{vy} \tag{26b}$$

and it follows that the coherency matrix of natural light of intensity $J_{xx} + J_{yy} = I_0$ is

$$\frac{1}{2} I_0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{27}$$

(b) Completely polarized light

Suppose first that the light is strictly monochromatic. Then the amplitudes $a_1$ and $a_2$, and the phase factors $\phi_1$ and $\phi_2$ in (1) do not depend on the time, and the coherency matrix has the form

$$\begin{bmatrix} {a_1}^2 & a_1 a_2 e^{i\delta} \\ a_1 a_2 e^{-i\delta} & {a_2}^2 \end{bmatrix} \tag{28}$$

where*

$$\delta = \phi_1 - \phi_2. \tag{29}$$

We see that in this case

$$|J| = J_{xx} J_{yy} - J_{xy} J_{yx} = 0, \tag{30}$$

i.e. the determinant of the coherency matrix is zero. The complex degree of coherence of the components $E_x$ and $E_y$ now is

$$\mu_{xy} = \frac{J_{xy}}{\sqrt{J_{xx}}\sqrt{J_{yy}}} = e^{i\delta} \tag{31}$$

i.e. its absolute value is unity (complete coherence) and its phase is equal to the difference between the phases of the two components.

In the special case when the light is linearly polarized we have (cf. § 1.4 (33)), $\delta = m_\pi (m = 0, \pm 1, \pm 2, \ldots)$. Hence the coherency matrix of linearly polarized light is

$$\begin{bmatrix} a_1^2 & (-1)^m a_1 a_2 \\ (-1)^m a_1 a_2 & a_2^2 \end{bmatrix}. \tag{32}$$

The electric vector vibrates in the direction given by $E_\gamma / E_x = (-1)^m a_3 / a_1$. In particular, the matrices

$$I \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad I \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{33}$$

each represent linearly polarized light of intensity $I$, with the electric vector in the $x$-direction ( $a_2 = 0$ ) and the $y$-direction ( $a_1 = 0$ ) respectively; and the matrices

$$\frac{1}{2} I \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \frac{1}{2} I \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{34}$$

each represent linearly-polarized light of intensity $I$ and with the electric vector in directions making angles 45° and 135° with the $x$-direction, respectively ( $a_1 = a_2, m = 0$ and $a_1 = a_2, m = 1$ ).

- For the purpose of later applications of some results of § 1.4 we note that $\phi_1$ and $\phi_2$ correspond
  to $-\delta_1$ and $-\delta_1$ of § 1.4.2, so that (29) is consistent with the earlier definition § 1.4 (16), viz. $\delta = \delta_2 - \delta_1$.

For circularly-polarized light we have (cf. § 1.4 (35), (36)), $a_1 = a_2, \delta = m\pi/2 (m = \pm 1, \pm 3, \dots)$, so that the coherency matrix is

$$\frac{1}{2} I \begin{bmatrix} 1 & \pm i \\ \mp i & 1 \end{bmatrix}, \tag{35}$$

where $I$ is the intensity of the light. By § 1.4 (38) and (40) the upper or lower sign is taken according whether the polarization is right- or left-handed.

The condition (30) may also be satisfied when the light is not monochromatic. For if $a_1, a_2, \phi_1$ and $\phi_2$ depend on time in such a way that the ratio of the amplitudes and the difference in the phases are time independent, so that

$$\frac{a_2(t)}{a_1(t)} = q, \quad \delta = \phi_1(t) - \phi_2(t) = \chi, \tag{36}$$

where $q$ and $\chi$ are constants, then

$$\begin{array}{ll} J_{xx} = \langle a_1^2 \rangle, & J_{xy} = q \langle a_1^2 \rangle e^{i\chi} \\ J_{yx} = q \langle a_1^2 \rangle e^{-i\chi}, & J_{yv} = q^2 \langle a_1^2 \rangle \end{array} \tag{37}$$

and the condition (30) holds. The coherency matrix with the elements (37) is the same as that of strictly monochromatic light with components

$$E_x = \sqrt{\langle a_1^2 \rangle} e^{i[\alpha - 2\pi\bar{\eta} l]}, \quad E_y = q\sqrt{\langle a_1^2 \rangle} e^{i[-x + \alpha - 2\pi\bar{k} l]}, \tag{38}$$

where $\alpha$ is any real constant. It follows that in experiments involving a polarizer and a compensator, the quasi-monochromatic wave which obeys the conditions (36) will behave in exactly the same way as the strictly monochromatic and hence completely polarized wave (38). (It is assumed, of course, that the phase difference introduced by the compensator is small compared to the coherence length of the light, measured in units of the mean wavelength.) The condition (30) may, therefore, be said to characterize a completely polarized light wave.

### 10.8.2 Some equivalent representations. The degree of polarization of a light wave

If several independent light waves which are propagated in the same direction are superposed, the coherency matrix of the resulting wave is equal to the sum of the coherency matrices of the individual waves. To prove this result let $E_x^{(n)}, E_y^{(n)} (n = 1, 2, \dots N)$ be the components of the electric vectors (in the usual complex representation) of the individual waves. The components of the electric vector of the resulting wave are

$$E_x = \sum_{n=1}^{N} E_x^{(n)}, \quad E_y = \sum_{n=1}^{N} E_y^{(n)}, \tag{39}$$

so that the elements of its coherency matrix are given by

$$\begin{aligned} J_{kl} = \langle E_k E_l^\star \rangle &= \sum_{n=1}^{N} \sum_{m=1}^{N} \left\langle E_k^{(n)} E_l^{(m)\star} \right\rangle \\ &= \sum_n \left\langle E_k^{(n)} E_l^{(n)\star} \right\rangle + \sum_{n \neq m} \left\langle E_k^{(n)} E_l^{(m)\star} \right\rangle . \end{aligned} \tag{40}$$

Since the waves are assumed to be independent, each term under the last summation sign is zero, and it follows that

$$J_{kl} = \sum_n J_{kl}^{(n)} \tag{41}$$

where ${J_{kl}}^{(n)} = \left\langle {E_k}^{(n)} {E_l}^{(n)\star} \right\rangle$ are the elements of the coherency matrix of the $n$th wave. Equation (41) shows that the coherency matrix of the combined wave is equal to the sum of the coherency matrices of all the separate waves.

Conversely any wave may be regarded as the sum of independent waves, which' evidently may be chosen in many different ways. One particular choice is of special significance and will now be briefly considered.

We will show that any quasi-monochromatic light wave may be regarded as the sum of a completely unpolarized and a completely polarized wave, which are independent of each other, and that this representation is unique.

To establish this result it is only necessary to show that any coherency matrix / can be uniquely expressed in the form

$$J = J^{(1)} + J^{(2)} \tag{42}$$

where, in accordance with (27) and (30),

$$J^{(1)} = \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}, \quad f^{(2)} = \begin{bmatrix} B & D \\ D^\star & C \end{bmatrix}, \tag{43}$$

with $A \geqslant 0, B \geqslant 0, C \geqslant 0$ and

$$BC - DD^\star = 0 \tag{44}$$

If $J_{xx}, J_{zy}, \ldots$ are the elements of the coherency matrix which characterizes the original wave, we must have, according to (42) and (43),

$$\left.\begin{aligned} A + B &= J_{xx}, & D &= J_{xy} \\ D^\star &= J_{yx}, & A + C &= J_{vy} \end{aligned}\right\} \tag{45}$$

On substituting from (45) to (44) we obtain the following equation for $A$ :

$$(J_{xx} - A)(J_{yy} - A) - J_{xy}J_{yx} = 0 \tag{46}$$

thus $A$ is a characteristic root (eigenvalue) of the coherency matrix $J$. The two roots of (46) are

$$A = \frac{1}{2}(J_{xx} + J_{yy}) \pm \frac{1}{2}\sqrt{(J_{xx} + J_{yy})^2 - 4|J|} \tag{47}$$

where, as before, $|J|$ is the determinant (8). Since $J_{yx} = J_{xy}^\star$, the product $J_{xy}J_{yx}$ is non-negative and it follows from (8) that

$$|J| \leqslant J_{xx}J_{yy} \leqslant \frac{1}{4}(J_{xx} + J_{yv})^2$$

so that both the roots (47) are real and non-negative. Consider first the solution with the negative sign in front of the square root. We then have

$$\left.\begin{aligned} A &= \tfrac{1}{2}\left(J_{xx}+J_{yy}\right) - \tfrac{1}{2}\sqrt{\left(J_{xx}+J_{yy}\right)^2 - 4|J|} \\ B &= \tfrac{1}{2}\left(J_{xx}-J_{yy}\right) + \tfrac{1}{2}\sqrt{\left(J_{xx}+J_{yy}\right)^2 - 4|J|}, \quad D = J_{xy} \\ D^{\star} &= J_{yx}, \quad C = \tfrac{1}{2}\left(J_{yy}-J_{xx}\right) + \tfrac{1}{2}\sqrt{\left(J_{xx}+J_{yy}\right)^2 - 4|J|} \end{aligned}\right\} \text{ Now } \sqrt{\left(J_{xx}+J_{yy}\right)^2 - 4|J|} = \sqrt{\left(J_{xx}-J_{yy}\right)^2 -} \tag{49}$$

Hence $B$ and $C$ are also non-negative, as required. The other root (47) (with the positive sign in front of the square root) leads to negative values of $B$ and $C$ and must therefore be rejected. We have thus obtained a unique decomposition of the required kiud.

The total intensity of the wave is

$$I_{\text{tot}} = \operatorname{Tr} J = J_{xx} + J_{yy}; \tag{50}$$

and the total intensity of the polarized part is

$$I_{\text{pol}} = \operatorname{Tr} J^{(2)} = B + C = \sqrt{\left(J_{xx}+J_{yy}\right)^2 - 4|J|} \tag{51}$$

The ratio of the intensity of the polarized portion to the total intensity is called the degree of polarization $P$ of the wave; according to (50) and (51) it is given by

$$P = \frac{I_{\text{pol}}}{I_{\text{tot}}} = \sqrt{1 - \frac{4|J|}{\left(J_{xx}+J_{vy}\right)^2}}. \tag{52}$$

Since this expression involves only the two rotational invariants of the coherency matrix $J$, the degree of polarization is independent of the particular choice of the axes $OX, OY$, as might have been expected. From (52) and the inequality preceding (48) it follows that

$$0 \leqslant P \leqslant 1 \tag{53}$$

When $P = 1$ there is no unpolarized component, so that the wave is then completely polarized. In this case $|j| = 0$, so that $|\mu_{xy}| = 1$ and consequently $E_x$ and $E_y$ are mutually coherent. When $P = 0$ the polarized component is absent. The wave is then completely unpolarized. In this case $\left(J_{xx}+J_{vy}\right)^2 = 4|J|$, i.e.

$$\left(J_{xx}-J_{yy}\right)^2 + 4J_{xy}J_{yx} = 0 \tag{54a}$$

Since $J_{yx} = J_{xy}^{\star}$ we have the sum of two squares equal to zero, and this is only possible if each vanishes separately, i.e. if

$$J_{xx} = J_{vy} \quad \text{and} \quad J_{xy} = J_{yx} = 0 \tag{54b}$$

in accordance with (26b). $E_x$ and $E_y$ are then mutually incoherent ( $\mu_{xy} = 0$ ). In all other cases ($0 < P < 1$) we say that the light is partially polarized. Comparison of (52) and (20) shows that the quantity $\left(I_{\max(\theta,\varepsilon)} - I_{\min(\theta,\varepsilon)}\right) / \left(I_{\max(\theta,\varepsilon)} + I_{\min(\theta,\varepsilon)}\right)$ is precisely the degree of polarization $P$.

When $E_x$ and $E_y$ are mutually incoherent (but the light not necessarily natural), the expression for the degree of polarization takes a simple form. Since now $J_{xy} = J_{yx} = 0, |J| = J_{r2}J_{vv}$ and (52) reduces to*

$$P = \left|\frac{J_{xx} - J_{vv}}{J_{xx} + J_{vv}}\right|. \tag{55}$$

This expression is in agreement with the formula (42) of § 1.5 employed in connection with polarization of natural light by reflection.

- Since every Hermitian matrix may be made diagonal by a unitary transformation, and since the values of $|j|$ and $\mathrm{Tr}\, j$ remain invariant under this transformation, the degree of polarization may always be expressed in the form

$$P = \left| \frac{A_1 - A_2}{A_1 + A_2} \right|,$$

where $A_1$ and $A_2$ are the two eigenvalues (given by (47)). However, the unitary transformation does not, in general, represent a real rotation of axes about the direction of the propagation of the wave.

It is of interest to note that the eigenvalues $A_1$ and $A_2$ are equal to the values $I_{\max(\theta,\varepsilon)}$ and $I_{\min(\theta,\varepsilon)}$ given by eq. (19).

We note some useful representations of natural light. The coherency matrix (27) of natural light may always be expressed in the form

$$\frac{1}{2}I \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{2}I \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{2}I \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{56}$$

and this implies, according to (33), that a wave of natural light, of intensity $I$, is equivalent to two independent linearly polarized waves, each of intensity $\frac{1}{2}I$, with their electric vectors vibrating in two mutually perpendicular directions at right angles to the direction of propagation.

Another useful representation of natural light is

$$\frac{1}{2}I \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{4}I \begin{bmatrix} 1 & +i \\ -i & 1 \end{bmatrix} + \frac{1}{4}I \begin{bmatrix} 1 & -i \\ +i & 1 \end{bmatrix} \tag{57}$$

and implies, according to (35), that a wave of natural light of intensity $I$ is equivalent to two independent circularly polarized waves, one right-handed, the other lefthanded, each of intensity $\frac{1}{2}I$.

Returning to the general case (partially polarized light), it is to be noted that unlike the degree of polarization, $P$, the degree of coherence $|\mu_{xy}|$ depends on the choice of the $x$ and $y$ directions. One may, however, readily see, that $|\mu_{xy}|$ cannot exceed $P$. For if in (52) we write out the determinant $|J|$ in full and use (6) we find that

$$1 - P^2 = \frac{J_{xx} J_{vy}}{\left[\frac{1}{2}\left(J_{xx} + J_{vy}\right)\right]^2} \left[1 - |\mu_{xy}|^2\right] \tag{58}$$

Since the geometric mean of any two positive numbers cannot exceed their arithmetic mean, it follows that $1 - P^2 \leqslant 1 - |\mu_{xy}|^2$, i.e.

$$P \geqslant |\mu_{xy}|\,. \tag{59}$$

The equality sign in (59) will hold if and only if $J_{xx} = J_{yy}$, i.e. if the (time averaged) intensities associated with the $x$ and $y$ directions are equal. We will show now that a pair of directions always exists, for which this is the case.

If the $xy$-axes are rotated in their own plane, through an angle $\theta$ in the anti-clockwise sense, $J_{xx}$ and $J_{yy}$ are transformed into $J_{x'x'}$ and $J_{y'y'}$ respectively, where, according to (23)

$$\left.\begin{aligned} J_{x'x'} &= J_{xx}\cos^2\theta + J_{yy}\sin^2\theta + (J_{xy} + J_{yx})\cos\theta\sin\theta \\ J_{y'y'} &= J_{xx}\sin^2\theta + J_{yy}\cos^2\theta - (J_{xy} + J_{yx})\cos\theta\sin\theta. \end{aligned}\right\} \tag{60}$$

From (60) it follows that $J_{x'x'} = J_{y'y'}$, if the axes are rotated though the angle $\theta = \Theta$, where

$$\tan 2\Theta = \frac{J_{yy} - J_{xx}}{J_{xy} + J_{yx}} \tag{61}$$

Since $J_{vx} = J^{\star}_{2y}$ and $J_{xx}$ and $J_{yy}$ are both real, the equation (61) has always a real solution for $\Theta$. Thus there always exists a pair of mutually orthogonal directions for which the intensities are equal. For this pair of directions the degree of coherence $|\mu_{xy}|$ of the electric vibrations has its maximum value and this value is equal to the degree of rolarization $P$ of the wave.*

### 10.8.3 The Stokes parameters of a quasi-monochromatic plane wave

We have seen that, in order to characterize a quasi-monochromatic plane wave, four real quantities are in general necessary, for example $J_{xx}, J_{vy}$ and the real and imaginary parts of $J_{xy}$ (or $J_{yx}$ ). In his investigations relating to partially polarized light, Stokes* introduced a somewhat different four-parameter representation which is closely related to the present one. We have already encountered it, in a restricted form, in connection with monochromatic light in § 1.4.2. The general Stokes parameters are the four quantities

$$\left.\begin{aligned} s_0 &= \langle a_1^2 \rangle + \langle a_2^2 \rangle \,, \\ s_1 &= \langle a_1^2 \rangle - \langle a_2^2 \rangle \,, \\ s_2 &= 2 \langle a_1 a_2 \cos\delta \rangle \,, \\ s_3 &= 2 \langle a_1 a_2 \sin\delta \rangle \,, \end{aligned}\right\} \tag{62}$$

where, as before, $a_1$ and $a_2$ are the instantaneous amplitudes of the two orthogonal components $E_x, E_y$ of the electric vector and $\delta = \phi_1 - \phi_2$ is their phase difference. When the light is monochromatic, $a_1, a_2$, and $\delta$ are independent of the time, and (62) reduces to the "monochromatic Stokes parameters" defined in § 1.4 (43).

It follows from (62) and (4) that the Stokes parameters and the elements of the coherency matrix are related by the formulae

$$\left.\begin{aligned} s_0 &= J_{xx} + J_{yv}, \\ s_1 &= J_{xx} - J_{yy}, \\ s_2 &= J_{xy} + J_{yx}, \\ s_y &= i\,(J_{yx} - J_{xy})\,. \end{aligned}\right\} \left\{\begin{aligned} J_{xx} &= \tfrac{1}{2}\,(s_0 + s_1)\,, \\ J_{yy} &= \tfrac{1}{2}\,(s_0 - s_1)\,, \\ J_{xy} &= \tfrac{1}{2}\,(s_2 + i s_3)\,, \\ J_{yx} &= \tfrac{1}{2}\,(s_2 - i s_3)\,. \end{aligned}\right\} \tag{63b}$$

Like the elements of the coherency matrix, the Stokes parameters of any quasi-monochromatic plane wave may be determined from simple experiments. If as before $I(\theta, \varepsilon)$ denotes the intensity of the light vibrations in the direction making an angle $\theta$ with $OX$, when the $y$-component is subjected to a retardation $\varepsilon$ with respect to the $x$-component, then, according to (11) and the relations (63a),

$$\left.\begin{aligned} s_0 &= I\,(0°, 0) + I\,(90°, 0) \\ s_1 &= I\,(0°, 0) - I\,(90°, 0) \\ s_2 &= I\,(45°, 0) - I\,(135°, 0) \\ s_3 &= I\left(45°, \tfrac{\pi}{2}\right) - I\left(135°, \tfrac{\pi}{2}\right). \end{aligned}\right\} \tag{64}$$

The parameter $s_0$ evidently represents the total intensity. The parameter $s_1$ is equal to the excess in intensity of light transmitted by a polarizer which accepts linear polarization in the azimuth $\theta = 0°$, over the light transmitted by a polarizer which

0

• The geometrical significance of this special pair of directions is disussed by **E**. Wolf in Nuovo Cimento, 13 (1959), pp. 1180-1181.

0

accepts linear polarization in the azimuth $\theta = 90°$. The parameter $s_2$ has a similar interpretation with respect to the azimuths $\theta = 45°$ and $\theta = 135°$. Finally, the parameter $s_3$ is equal to the excess in intensity of light transmitted by a device which accepts right-handed circular polarization, over that transmitted by a device which accepts left-handed circular polarization.

If we use the relations (63b) our previous results may be expressed in terms of the Stokes parameters, rather than in terms of the coherency matrix. In particular, the condition (8), viz. $J_{xx}J_{yy} - J_{xy}J_{yx} \geqslant 0$, becomes

$$s_0^2 \geqslant s_1^2 + s_2^2 + s_3^2 \tag{65}$$

For monochromatic light we have, according to (30), $J_{xx}J_{vy} - J_{xy}J_{vx} = 0$ and the equality sign then holds in (65), in agreement with § 1.4 (44).

Let us now consider the decomposition of a given wave into an unpolarized and a polarized portion which are mutually independent, using the Stokes parameter representation. It follows from (41) and (63) that the Stokes parameters of a mixture of independent waves are sums of the respective Stokes parameters of the separate waves. From (27) and (63a) it follows that an unpolarized wave (wave of natural light) is characterized by $s_1 = s_2 = s_3 = 0$. Denoting by a single symbol $s$ the four Stokes parameters $s_0, s_1, s_2, s_3$, the required decomposition of the wave characterized by $s$ evidently is

$$s = s^{(1)} + s^{(2)}, \tag{66}$$

where

$$s^{(1)} = s_0 - \sqrt{s_1^2 + s_2^2 + s_3^2}, 0, 0, 0 \tag{67a}$$

$$s^{(2)} = \sqrt{s_1^2 + s_2^2 + s_3^2}, s_1, s_2, s_3 \tag{67b}$$

$s^{(1)}$ represents the unpolarized and $s^{(2)}$ the polarized part. Hence, in terms of the Stokes parameters, the degree of polarization of the original wave is

$$P = \frac{I_{\text{pol}}}{I_{\text{tot}}} = \frac{\sqrt{s_1^2 + s_2^2 + s_3^2}}{s_0}, \tag{68}$$

as may also be verified by substituting from (63b) into (52). We may also easily write down expressions which give theellipticity and the orientation of the polarization ellipse associated with the polarized part (67b). If, as in § 1.4 (28),

$$\tan\chi = \pm b/a \quad (-\pi/4 < \chi \leqslant \pi/4)$$

• G. G. Stokes, Trans. Cambr. Phil. Soc., 9 (1852), 399. Reprinted in his Mathematical and Physical Papers, Vol. III (Cambridge University Press, 1901), p. 233. See also P. Soleilet, Ann. de Physique (10), 12 (1929), 23; F. Perrin, J. Chem. Phys., 10 (1942), 415; S. Chandrasekhar, Radiative Transfer (Oxford, Clarendon Press, 1950), § 15; M. J. Walker, Amer. J. Phys., 22 (1954), 170; E. Wolf, Nuovo Cimento, 12 (1954), 884; S. Pancharatnam, Proc. Ind. Acad. Sci., A, 44 (1956), 398; ibid., 57 (1963), 218, 231. The Stokes parameters are also employed in quantum mechanical treatments of polarization of elementary particles. Cf. U. Fano, J. Opt. Soc. Amer., 39 (1949), 859; ibid, 41 (1951), 58; Phys. Rev., 93 (1954), 121; D. L. Falkoff and J. E. MacDonald, J. Opt. Soc. Amer., 41 (1951), 861 ; W. H. McMaster, Amer. J. Phys., 22 (1954), 351; J. M. Jauch and F. Rohrlich, The Theory of Photons and Electrons (Cambridge, Mass., Addison-Wesley Publ. Co., 1955), §2.8. See also N. Wiener, Acta Math., 55 (1930), § 9, especially pp. 189-192.

represents the ratio of the minor and the major axes and the sense in which the ellipse is described ($\chi \gtrless 0$ according as the polarization is right- or left-handed), then according to (67b) and § 1.4 (45c),

$$\sin 2\chi = \frac{s_3}{\sqrt{s_1^2 + s_2^2 + s_3^2}}; \tag{69}$$

and the angle $\psi (0 \leqslant \psi < \pi)$, which the major axis makes with $OX$, is according to (67b) and § 1.4 (46) given by

$$\tan 2\psi = \frac{s_2}{s_1} \tag{70}$$

We see that the Stokes parameters, just like the coherency matrix, provide a useful tool for a systematic analysis of the state of polarization of a quasi-monochromatic wave.

## 9.14 Rigorous Diffraction Theory

### 9.14.1 11.1 Introduction

On the basis of Maxwell's equations, together with standard boundary conditions, the scattering of electromagnetic radiation by an obstacle becomes a well-defined mathematical boundary-value problem. In the present chapter some aspects of the theory of diffraction of monochromatic waves are developed from this point of view, and in particular the rigorous solution to the classical problem of diffraction by a perfectly conducting half-plane is given in detail.

In the early theories of Young, Fresnel, and Kirchhoff, the diffracting obstacle was supposed to be perfectly "black"; that is to say, all radiation falling on it was assumed to be absorbed, and none reflected. This is an inherent source of ambiguity in that such a concept of absolute "blackness" cannot legitimately be defined with precision; it is, indeed, incompatible with electromagnetic theory.

Cases in which the diffracting body has a finite dielectric constant and finite conductivity have been examined theoretically, one of the earliest comprehensive treatments of such a case being Mie's discussion in 1908 of scattering by a sphere, which is described in Chapter XIII in connection with the optics of metals. In general, however, the assumption of finite conductivity tends to make the mathematics very complicated, and it is often desirable to accept the concept of a perfectly conducting (and therefore perfectly reflecting) body. This is clearly an idealization, but one which is compatible with electromagnetic theory; furthermore, since the conductivity of some metals (e.g. copper) is very large, it may represent a good approximation if the frequency is not too high, though it should be stressed that the approximation is never entirely adequate at optical frequencies. The simplifying assumption that the diffracting obstacle has infinite conductivity is made in most of the treatments based on a precise mathematical formulation, and the subsequent discussion is confined to this case.

The first rigorous solution of such a diffraction problem was given by Sommerfeld* in 1896, when he treated the two-dimensional case of a plane wave incident on an infinitely thin, perfectly conducting half-plane. The fame of this achievement rests partly on the skill with which the solution was constructed, and partly on the remarkable fact that it could be expressed exactly and simply in terms of the Fresnel integrals which had been such a conspicuous feature of previous approximate theories.

Many mathematicians followed Sommerfeld's lead. Early variants of his problem, dealing with line and point sources, and a noteworthy generalization to the treatment of a wedge rather than a half-plane, are associated with the names of Carslaw, † Macdonald, ‡ and Bromwich.§

Other problems were attacked, and more recently new methods have been introduced, stimulated by the advance in ultra-short-wave

radio techniques. Before proceeding to the main body of the chapter the nature of some of these investigations is very briefly indicated.

If there exists an orthogonal coordinate system, $u_1, u_2, u_3$, say, such that the surface of the diffracting body is identical with one of the surfaces $u_i$ = constant, the classical technique for solving partial differential equations by separation of variables may be appropriate; this, indeed, was Mre's approach in the case of the finitely conducting sphere mentioned above. The solution of the boundary-value problem then appears, in general, as an infinite series, and its utility depends on the ease with which computation of the relevant functions can be carried out and the rapidity with which the series converges. This method has been applied to various cases apart from the sphere, notably to the circular disc or aperture.* It should be mentioned, however, that some of the work only relates to strictly scalar problems, such as those in the theory of small-amplitude sound waves; as is shown later, two-dimensional problems in electromagnetic theory are essentially of this type, but otherwise the vector nature of the electromagnetic field introduces further complications.

Another approach is based on integral equation formulations, a method apparently first considered by Rayleigh. † Certain problems, the simplest being that of the halfplane, yield integral equations that can be solved exactly by the method of Wiener and Hopf, ‡ and the appreciation of this fact by Copson, § Schwinger, and others has led to a number of new solutions in closed form. || Mention should also be made, in this connection, of powerful, if somewhat complicated, variational procedures which can be used to calculate the power diffracted through an aperture. I

For reasons of space the discussion in this chapter is largely confined to one method.** First, some aspects of considerable generality in the theory of the scattering of electromagnetic waves by perfectly conducting structures are developed. Next, a representation of any field as an integral over a spectrum of plane waves is introduced and is shown to lead to the formulation of certain diffraction problems in terms of "dual" integral equations. †† The Sommerfeld half-plane problem is then tractable; the solution of this is obtained and examined in some detail, together with a number of ramifications. Several allied problems are discussed.

## 9.14.2 11.2 Boundary Conditions and Surface Currents

It is well known that an electromagnetic field penetrates but little into a good conductor. The idealization of infinite conductivity, when there is no penetration at all,

0

- A. Sommerfeld, Math. Ann., 47 (1896), 317.
  † H. S. Carslaw, Proc. Lond. Math. Soc., 30 (1899), 121.
  ‡ H. M. Macdonald, Electric Waves (Cambridge University Press, 1902).
  § T. J. I'A. Bromwich, Proc. Lond. Math. Soc., 14 (1916), 450.

0

- C. J. Bouwkamp, Dissertation, Groningen, 1941.
  J. Meixner and W. Andrejewski, Ann. d. Physik, 7 (1950), 157.
  † Lord Rayleigh, Phil. Mag., 43 (1897), 259.
  ‡ For a discussion of the Wiener-Hopf method see E. C. Titchmarsh, Introduction to the Theory of Fourier Integrals (Oxford, Clarendon Press, 1937), p. 339.
  § E. T. Copson, Quart. J. Maths., 17 (1946), 19.
  || J. F. Carlson and A. E. Heins, Quart. Appl. Maths., 4 (1947), 313; 5 (1947), 82. A. E. Heins, Quart. Appl. Maths., 6 (1948), 157, 215. H. Levine and J. Schwinger, Phys. Rev., 73 (1948), 383. For a more comprehensive list of references see J. W. Miles, J. Appl. Phys., 20 (1949), 760, and C. J. Bouwkamp, Rep. Progr. Phys. (London, Physical Society), 17 (1954), 35.

results in the concept of electric currents existing purely on the surface of the conductor, as the following argument shows.
It is a consequence of Maxwell's curl equations (see § 1.1.3) that the tangential component of $E$ is continuous in crossing an infinitely thin electric current sheet whereas that of $\boldsymbol{H}$ is discontinuous; more particularly, the discontinuity in $\boldsymbol{H}$ is in the tangential component normal to the surface current density* $\boldsymbol{J}$, the relative sense of the directions being indicated schematically in Fig. 11.1, and is of amount $4\pi J/c$. Furthermore, in line with the behaviour of the tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$, the normal component of $\boldsymbol{H}$ is continuous across the current sheet whereas that of $E$ is discontinuous, the magnitude of the discontinuity being equal to $4\pi$ times the surface charge density. Hence it is clear that the

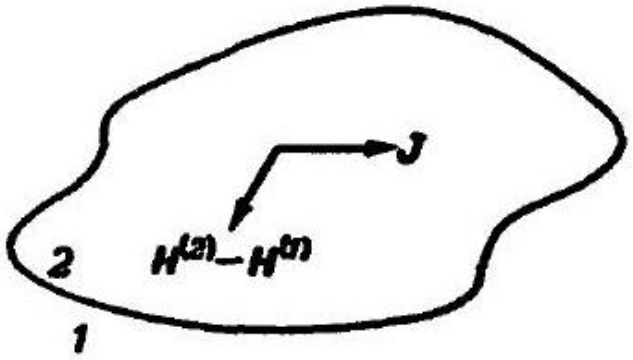



Fig. 11.1. Showing the sense of the discontinuity in $\boldsymbol{H}$ relative to the direction of the surface current density J. $H^{(1)}$ and $H^{(2)}$ are the respective magnetic fields on sides 1 and 2 of the surface.

field in free space exterior to a perfectly conducting body is such that on the surface of the conductor
(a) the tangential component of $\boldsymbol{E}$ is zero;
(b) the tangential component of $\boldsymbol{H}$ is perpendicular to the surface electric current density $\boldsymbol{J}$, in the sense indicated above, and of amount $4\pi J/c$;
(c) the normal component of $\boldsymbol{H}$ is zero;
(d) the outward normal component of $E$ is equal to $4\pi$ times the surface charge density.

The effect of radiation falling on a perfectly conducting body may be interpreted conveniently in terms of the induced surface currents. If $E^{(i)}$ is the electric vector of the incident field and $E^{(s)}$ is that of the "scattered" field due to induced currents, then the total electric vector, everywhere, is $E^{(i)} + E^{(s)}$. The diffraction problem may therefore be stated as follows: given $E^{(i)}$, to find a field $E^{(s)}$ which could arise from a current distribution in the surface of the conductor and which is such that its tangential component on the surface is minus that of $E^{(i)}$. It is worth stressing that the boundary condition (a) above is fundamental and sufficient alone to specify the problem uniquely in the form stated. † As regards the other conditions, (b) is of value in relating the field to the induced currents, but (c) and (d) are of no particular interest.

TT H. Levine and J. Schwinger, Phys. Rev., 74 (1948), 958; 75 (1949), 1423.
** For general accounts of other methods see the article by G. Wolfsonn in Handbuch der Physik, Vol. 20 (Berlin, Springer, 1928), p. 263; B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 1950), Chapters 4 and 5 and an article by H. Hönl, A. W. Maue and K. Westpfahl in Handbuch d. Physik, Vol. 25/1 (Berlin, Springer, 1961). The review by Bouwkamp just mentioned gives in outline a most comprehensive selection of methods and formulae.
†† For a discussion of "dual" integral equations (defined on pp. 564-565 below) see E. C. Titchmarsh, Introduction to the Theory of Fourier Integrals (Oxford, Clarendon Press, 1937), p. 334.

It should be noted that an implication of $E^{(s)} = -E^{(i)}$ at interior points of the conductor is the existence of a unique current density, in any closed surface $S$, which reproduces at all points inside $S$ the field due to sources located outside $S$. Likewise, from a consideration of the case when the boundary of a perfect conductor on which the radiation falls is a complete infinite plane, it follows that there is a unique current density in any plane which reproduces on one side of the plane the field due to sources situated on the other side of the plane.

In problems of diffraction by perfectly conducting screens it is desirable to make the assumption that the screens are infinitely thin; if this is not done, the mathematical difficulties become very great. Of course, the opacity of the screen is maintained, the concept, in fact, being that of a perfect conductor whose thickness tends to zero in the limit. From what has been said, the effect of such a screen can be interpreted in terms of an electric current sheet, with the difference now that the sheet is no longer a closed surface. Of particular interest is the relatively simple case when the sheet is

plane; in this case some important relations satisfied by the field $E^{(s)}, H^{(s)}$ which it radiates may be deduced immediately and are given below.

Suppose that the current sheet occupies part of the plane $y = 0$. Then, by reason of symmetry, it is clear that

$$\left.\begin{array}{ll} E_x^{(s)}(x,y,z) = E_x^{(s)}(x,-y,z), & H_x^{(s)}(x,y,z) = -H_x^{(s)}(x,-y,z) \\ E_y^{(s)}(x,y,z) = -E_y^{(s)}(x,-y,z), & H_y^{(s)}(x,y,z) = H_y^{(s)}(x,-y,z)' \\ E_z^{(s)}(x,y,z) = E_z^{(s)}(x,-y,z), & H_z^{(s)}(x,y,z) = -H_z^{(s)}(x,-y,z). \end{array}\right\} \quad (1)$$

Moreover, if the current.density in the sheet has components $J_x$ and $J_z$, evidently, on $y = 0$,

$$H_x^{(s)} = \mp\frac{2\pi}{c}J_z, \quad H_z^{(s)} = \pm\frac{2\pi}{c}J_x \quad (2)$$

with the upper or lower sign according as to whether $y$ reaches zero through positive or negative values respectively. As discussed in the next section, the application of these simple relations to the interesting problem of diffraction by a plane screen leads to a useful formulation which, in particular, puts in evidence an exact electromagnetic analogue of Babinet's principle.

## 9.14.3 11.3 Diffraction by a Plane Screen: Electromagnetic Form of Babinet's Principle

Suppose that an electromagnetic field $E^{(i)}, H^{(i)}$ is incident on a set of infinitely thin, perfectly conducting laminae lying in the plane $y = 0$. Let $M$ signify those areas of the plane occupied by the metal and $A$ the remaining "apertures", so that $M$ and $A$ together comprise the whole plane. Either $M$ or $A$, or both $M$ and $A$, may be of infinite extent.

As previously explained, a scattered field is sought which satisfies a certain boundary condition on $M$. Now in view of the relations §11.2(1), it is, in fact, only necessary to consider the scattered field in one of the half-spaces $y \geqslant 0, y \leqslant 0$, provided the requirement of continuity across $A$ is explicitly recognized. Hence the problem may be formulated thus: to find, in $y \geqslant 0$ (or in $y \leqslant 0$ ), an electromagnetic field $E^{(s)}, H^{(s)}$, which could be generated by currents in $y = 0$, such that

---

0

- The surface current was denoted by $\boldsymbol{j}$ in § 1.1.3.
† A discussion of proofs of uniqueness, which present some difficulties, is deferred to a later section (§ 11.9).

$$\begin{array}{ll} \text{(I)} & E_x^{(i)} + E_x^{(s)} = E_z^{(i)} + E_z^{(s)} = 0 \quad \text{on } M, \\ \text{(II)} & H_x^{(s)} = H_z^{(s)} = 0 \quad \text{on } A. \end{array}$$

Here, (I) is the fundamental boundary condition for a perfect conductor, whereas (II), which follows from § 11.2 (2), is a convenient way of expressing the fact that there are no induced currents in $A$. If (II) is satisfied by the scattered field in $y \geqslant 0$, and § 11.2 (1) used to deduce the scattered field in $y \leqslant 0$, then continuity across $A$ is achieved.

A form of Babinet's principle for electromagnetic waves and perfectly conducting screens which is exact* may now be easily derived. As in the classical principle (§ 8.3.2), a relation is established between the respective fields existing in the presence of the screen and the "complementary" screen obtained by interchanging the conducting laminae and the apertures; the difference lies in the fact that the field incident

on the complementary screen is no longer the same as that incident on the original screen, but is derived from it by the transformation $\boldsymbol{E} \to \boldsymbol{H}$.

In the first case, then, let the field (suffix 1 ) defined by ${E_1}^{(i)} = F^{(i)}$ be incident, in $y > 0$, on the screen described above. From (I) and (II)

$$\begin{array}{ll} \text{(I}') & {E_{1x}}^{(s)} = -{F_x}^{(i)}, \quad {E_{1z}}^{(s)} = -{F_z}^{(i)} \quad \text{on } M, \\ \text{(II}') & {H_{1x}}^{(s)} = {H_{1z}}^{(s)} = 0 \quad \text{on } A. \end{array}$$

Secondly, let the field (suffix 2) defined by ${\boldsymbol{H}_2}^{(i)} = \boldsymbol{F}^{(i)}$ be incident on the complementary screen. Then, writing the boundary conditions now in terms of the total field,

$$\begin{array}{ll} \text{(I}'') & E_{2x} = E_{2z} = 0 \quad \text{on } A, \\ \text{(II}'') & H_{2x} = F_x^{(i)}, \quad H_{2z} = F_z^{(i)} \quad \text{on } M. \end{array}$$

Since Maxwell's equations in free space are invariant under the transformation $\boldsymbol{E} \to \boldsymbol{H}, \boldsymbol{H} \to -\boldsymbol{E}$, and since there is a unique surface current density in $y = 0$ which would produce the incident field at all points in $y \leqslant 0$, it is clear from a comparison of ( $\mathbf{I}'$ ), ( $\mathbf{II}'$ ) with ( $\mathbf{II}''$ ), ( $\mathbf{I}''$ ), respectively, that

$$H_2 = -E_1^{(s)} \tag{1}$$

in the half-space behind the screen. In terms of the total field $\boldsymbol{E_1}$, (1) gives

$$E_1 + H_2 = F^{(i)} \tag{2}$$

which is the required electromagnetic form of Babinet's principle.

## 9.14.4 11.4 Two-dimensional Diffraction by a Plane Screen

### 11.4.1 The scalar nature of two-dimensional electromagnetic fields

A problem which is completely independent of one Cartesian coordinate, say $z$, is said to be two-dimensional. As already remarked, problems of this type in electromagnetic theory are essentially of a scalar nature in that they can straightway be expressed in terms of a single dependent variable. This will now be shown.

With a time factor $\exp(-i\omega t)$ suppressed and writing $k = \omega/c$, Maxwell's equations in free space are

$$\operatorname{curl} H = -ikE, \quad \operatorname{curl} E = ikH.$$

[0]* H. G. Booker, J. Instn. Elect. Engrs., 93, Pt. III A (1946), 620. L. G. H. Huxley, The Principles and Practice of Waveguides (Cambridge University Press, 1947), p. 284.

Equating to zero all partial derivatives with respect to $z$, these may be split up into the two independent sets

$$\frac{\partial E_z}{\partial y} = ikH_x, \quad \frac{\partial E_z}{\partial x} = -ikH_y, \quad \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = -ikE_z, \tag{1}$$

and

$$\frac{\partial H_z}{\partial y} = -ikE_x, \quad \frac{\partial H_z}{\partial x} = ikE_y, \quad \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} = ikH_z. \tag{2}$$

The first group involves only $H_x, H_y, E_z$, the second only $E_x, E_y, H_z$. Simplicity can therefore be obtained by separating any solution into a linear combination of the two
solutions for which every member of one of the above sets is zero. For the sake of nomenclature we characterize the two types of field as follows:

E-polarization

$$E_x = E_y = H_z = 0$$
$$H_x = \frac{1}{ik}\frac{\partial E_z}{\partial y}, \quad H_y = -\frac{1}{ik}\frac{\partial E_z}{\partial x}$$

and, as is evident on substituting for $H_x$ and $H_y$ into the third equation of (1),

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + k^2 E_z = 0$$

Here the complete field is specified in terms of $E_z$, which, of course, satisfies the twodimensional form of the standard wave equation.

$H$-polarization

$$H_x = H_y = E_z = 0$$

and

$$E_x = -\frac{1}{ik}\frac{\partial H_z}{\partial y}, \quad E_y = \frac{1}{ik}\frac{\partial H_z}{\partial x}.$$

Here the complete field is specified in terms of $H_z$.

### 11.4.2 An angular spectrum of plane waves

For two-dimensional problems, to which the discussion is now confined, it has been shown that the Cartesian components of $\boldsymbol{E}$ and $\boldsymbol{H}$ satisfy the equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + k^2 V = 0 \tag{3}$$

This equation has to be solved subject to the appropriate boundary conditions.
A fundamental elementary solution of (3) is

$$e^{ikr\cos(\theta-\alpha)} = e^{ik(x\cos\alpha + y\sin\alpha)}, \tag{4}$$

$r, \theta (0 \leqslant \theta \leqslant 2\pi)$ being polar coordinates related to $x, y$ by the equations $x = r\cos\theta$, $y = r\sin\theta$. If $\alpha$ is real (4) represents a homogeneous plane wave, that is, one whose equiamplitude and equiphase planes coincide: $\alpha$ is the angle between the direction of propagation and the $x$ axis (Fig. 11.2a). If, on the other hand, $\alpha$ is complex, (4) represents an inhomogeneous plane wave, that is, one whose equiamplitude and equiphase planes do not coincide. In fact, writing $\alpha = \alpha_1 + i\alpha_2$, where $\alpha_1$ and $\alpha_2$ are real, (4) becomes

$$e^{ikr\cosh\alpha_2\cos(\theta-\alpha_1)}e^{-kr\sinh\alpha_2\sin\theta-\alpha_1)} \tag{5}$$

from which it follows that the equiamplitude and equiphase planes are mutually perpendicular (Fig. 11.2b): the direction of phase propagation makes an angle $\alpha_1$ with the $x$ axis, the phase velocity being reduced by the factor sech $\alpha_2$, and there is exponential attenuation, governed by the attenuation factor $k\sinh\alpha_2$, in the direction at right angles.

Now it can be shown* that any solution of (3) can be put in the form of an angular spectrum of plane waves

$$\int f(\alpha)e^{ikr\cos(\theta-\alpha)}d\alpha$$

by a suitable choice of the path of integration and the function $f(\alpha)$. Such a representation is closely linked with the representation of an arbitrary function by means of a Fourier integral, and is likewise of great power in application. Without significant loss of generality a certain fixed path of integration can be prescribed, so that any problem becomes a matter of determining the appropriate $f(\alpha)$. We shall first express the electromagnetic field due to a plane current sheet in this way, and then show that the result leads to a formulation of the problem of diffraction by a plane screen in terms of dual integral equations.

Consider a two-dimensional current sheet in $y = 0$. As already pointed out, it is convenient to deal with $E$-polarization and $H$-polarization separately. We treat,

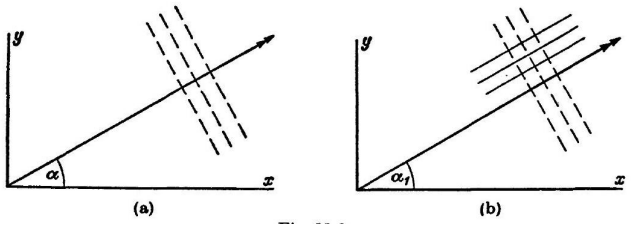



Fig. 11.2.
(a) The homogeneous plane wave (4) when $\alpha$ is real; the dashed lines denote both equiamplitude and equiphase planes.
(b) The inhomogeneous plane wave (4) when $\alpha$ is complex; the full lines denote equiamplitude planes, the dashed lines equiphase planes.

first, the former case, in which the current density has a $z$ component only, $J_z$ say, and begin by asking what particular distribution will radiate the $E$-polarized plane wave

$$\boldsymbol{E} = (0,0,1)e^{ikr\cos(\theta-\alpha)}, \quad \boldsymbol{H} = (\sin\alpha, -\cos\alpha, 0)e^{ikr\cos(\theta-\alpha)} \tag{6}$$

into the half-space $y > 0$. From the first relation in § 11.2 (2) it is, in fact, immediately seen that

$$J_z(\xi) = -\frac{c}{2\pi}e^{ik\xi\cos\alpha}\sin\alpha \tag{7}$$

at the point $(\xi, 0)$. This could, of course, be verified by the standard method of Hertz potentials for finding the field generated by a current distribution, though the evaluation of a quite complicated integral is then necessary.

Now, broadly speaking, any current distribution can be built up by the appropriate superposition of expressions (7) for different values of $\alpha$, and the radiated field will be obtained by

the corresponding superposition of the plane waves (6). More precisely, suppose the current density can be written in the form of a Fourier integral as

$$J_z(\xi) = -\frac{c}{2\pi}\int_{-\infty}^{\infty} P(\mu)e^{ik\xi\mu}d\mu \tag{8}$$

The change of variable $\mu = \cos\alpha$ gives

$$J_z(\xi) = -\frac{c}{2\pi}\int_C \sin\alpha P(\cos\alpha)e^{ik\xi\cos\alpha}d\alpha \tag{9}$$

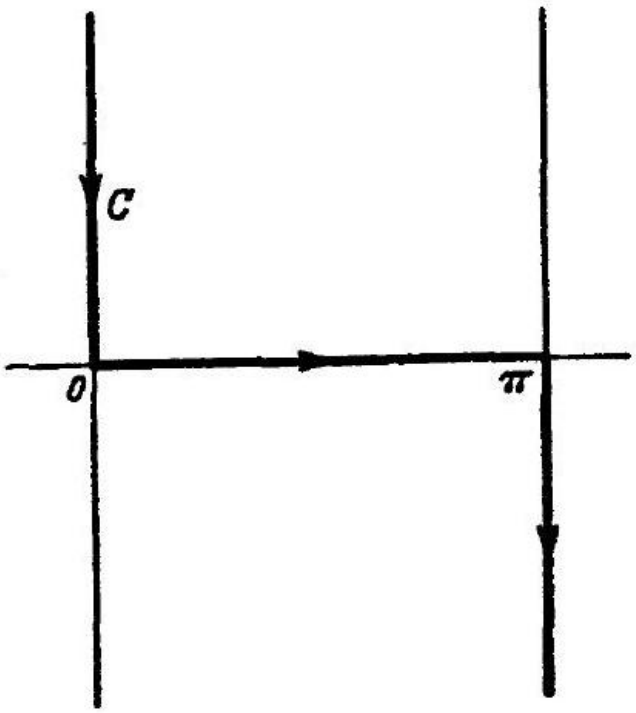



Fig. 11.3. The path $C$ in the complex $\alpha$ plane.

0

- E. T. Whittaker and G. N. Watson, A Course of Modern Analysis (Cambridge University Press, 1927), p. 397.

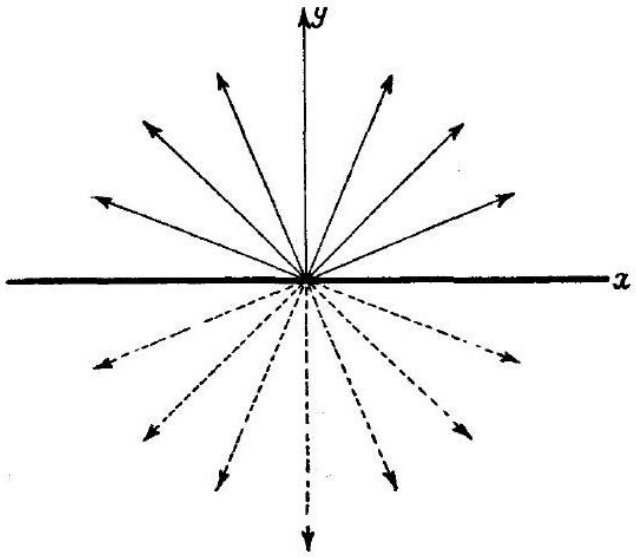



Fig. 11.4. Showing the directions of propagation of the homogeneous waves which radiate into the half-space $y > 0$ (full lines) and the half-space $y < 0$ (interrupted lines).

where $C$ is the path in the complex $\alpha$ plane along which $\cos\alpha$ ranges through real values from $\infty$ to $-\infty$, as shown in Fig. 11.3. The resulting non-zero field components are therefore

$$E_z^{(s)} = \int_C P(\cos\alpha) e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{10}$$

$$H_x^{(s)} = \pm \int_C \sin\alpha P(\cos\alpha) e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{11}$$

$$H_y^{(s)} = -\int_C \cos\alpha P(\cos\alpha) e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{12}$$

with the upper sign for $y \geqslant 0$ and the lower sign for $y \leqslant 0$.
Equations (10), (11), and (12) represent the field in the form of a plane wave spectrum determined by the function $P(\cos\alpha)$. The individual plane waves corresponding to the section of $C$ along the real axis are homogeneous; they radiate into the regions $y > 0$ and $y < 0$, their directions of propagation embracing a range of angles $\pi$ in each region, as illustrated diagrammatically in Fig. 11.4. The plane waves corresponding to the two arms of $C$ on which $\alpha = i\beta$ and $\alpha = \pi - i\beta$ ( $\beta = 0$ to $\infty$ ) are inhomogeneous; all their directions of phase propagation are along the positive or negative $x$ axis, and they are exponentially attenuated in the direction normal to and away from the plane $y = 0$. It can easily be shown, by an examination of the Poynting vector, that on the average no energy is carried away from the plane $y = 0$ by any of these evanescent waves. Their presence is necessary in order to take account of structure in the current distribution which is finer than a wavelength.

For the case of $H$-polarization, the field due to a current density $J_x$ in $y = 0$ would likewise be written

$$H_z^{(s)} = \pm \int_C P(\cos\alpha) e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{13}$$

$$E_x^{(s)} = -\int_C \sin\alpha P(\cos\alpha) e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{14}$$

$$E_y^{(s)} = \pm \int_C \cos\alpha P(\cos\alpha) e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{15}$$

with the upper sign for $y \geqslant 0$, the lower sign for $y \leqslant 0$, where

$$J_x(\xi) = \frac{c}{2\pi} \int_{-\infty}^{\infty} \frac{P(\mu)}{\sqrt{(1-\mu^2)}} e^{ik\xi\mu} d\mu \tag{16}$$

\{}paragraph11.4.3 Formulation in terms of dual integral equations The two-dimensional problem of diffraction by a plane screen can now be formulated in terms of dual integral equations.

Suppose that an electromagnetic field $E^{(i)}, H^{(i)}$ is incident on a set of infinitely thin, perfectly conducting strips lying in $y = 0$; designate by $M$ the ranges of $x$ within

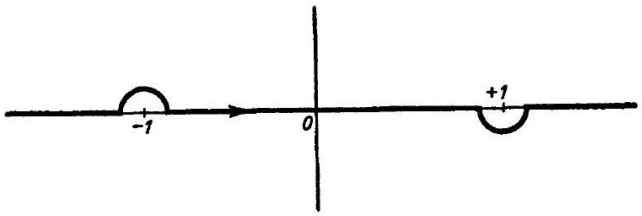



Fig. 11.5. The path of integration from $-\infty$ to $\infty$ in the complex $\mu$ plane.

which there is metal, by $A$ those within which there is not. If the scattered field $E^{(s)}$, $H^{(s)}$ is represented as an angular spectrum of plane waves, in the form § 11.4 (10), (11), (12) or (13), (14), (15) according to the polarization, the conditions (I) and (II) of § 11.3 yield the following integral equations:

E-polarization

$$\int_{-\infty}^{\infty} \frac{P(\mu)}{\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu = -E_z^{(i)} \quad \text{on } M \tag{17}$$

$$\int_{-\infty}^{\infty} P(\mu) e^{ikx\mu} d\mu = 0 \quad \text{on } A \tag{18}$$

H-polarization

$$\int_{-\infty}^{\infty} P(\mu) e^{ikx\mu} d\mu = E_x^{(i)} \quad \text{on } M \tag{19}$$

$$\int_{-\infty}^{\infty} \frac{P(\mu)}{\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu = 0 \quad \text{on } A \tag{20}$$

A consideration of the way in which the complex $\alpha$ plane between $\mathscr{R}\alpha = 0$ and $\mathscr{R}\alpha = \pi$ ( $\mathscr{R}$ denoting the real part) maps into the complete complex $\mu$ plane ( $\mu = \cos\alpha$ ) shows that the path of integration along the real axis avoids the possible branch-points at $\mu = \pm 1$ as illustrated diagrammatically in Fig. 11.5. Integral
equations of this type, in which a single unknown function $P(\mu)$ satisfies different equations for two distinct ranges of the parameter $x$, are called "dual".*

The formulation used by Copson, Schwinger, and others, mentioned in § 11.1, is somewhat different from the one given above in that it only involves a single integral equation. Although it will not be wanted here, its connection with the present method should be pointed out. For the case of $E$-polarization, say, the solution of (16) obtained by taking its Fourier transform could be written in the form

$$P(\mu) = -\frac{k}{c}\int_M J_z(\xi)e^{-ik\mu\xi}d\xi \tag{21}$$

in agreement, of course, with eq. (8) and the fact that $J_z(\xi) = 0$ on $A$. Substituting this value of $P(\mu)$ into (17), and carrying out the integration with respect to $\mu$, we have the integral equation

$$\frac{k}{c}\int_M J_z(\xi)H_0^{(1)}(k|x-\xi|)d\xi = E_z^{(i)} \quad \text{on } M \tag{22}$$

involving the Hankel function $H_0{}^{(1)}$ of the first kind and zero order, to be solved for $J_z(\xi)$. Obviously, the left-hand side of (22) could have been derived from the direct expression of the scattered field in terms of the induced current density.

## 9.14.5 11.5 Two-dimensional Diffraction of a Plane Wave by a Half-plane

### 11.5.1 Solution of the dual integral equations for $E$-polarization

In the next few pages the diffraction of a plane wave by a semi-infinite plane sheet is treated rigorously by obtaining a simple explicit solution of the appropriate dual integral equations.

Consider, first, the $E$-polarized plane wave

$$E_z^{(i)} = e^{-ikr\cos(\theta-\alpha_0)} \tag{1}$$

incident on the perfectly conducting half-plane $y = 0, x > 0$, where it is assumed, for convenience, that $\alpha_0$ is real and $0 < \alpha_0 < \pi$ (Fig. 11.6). Equations (17) and (18) of § 11.4 are now

$$\begin{array}{ll} \int_{-\infty}^{\infty} \frac{P(\mu)}{\sqrt{(1-\mu^2)}} e^{ikx\mu}d\mu = -e^{-ikx\mu_0} & \text{for } x > 0 \\ \int_{-\infty}^{\infty} P(\mu)e^{ikx\mu}d\mu = 0 & \text{for } x < 0 \end{array} \tag{3}$$

where $\mu_0 = \cos\alpha_0$. We proceed to solve these equations by the use of standard techniques in contour integration.

In the integral on the left-hand side of equation (3) $x$ is negative. Hence, by Jordan's lemma, † provided $P(\mu) \to 0$ as $|\mu| \to \infty$ when $0 \geqslant \arg\mu \geqslant -\pi$, we can close the path of integration with an infinite semi-circle below the real axis without

making any additional contribution to the integral. Thus we only require further that $\boldsymbol{P}(\mu)$ should have no singularities in the half-plane below the path of integration for equation (3) to be satisfied, since the integral is then effectively round a closed contour within which the integrand is regular.

Likewise, in the integral on the left-hand side of equation (2) $x$ is positive, and we can close the path of integration with an infinite semi-circle above the real axis without

0

- E. C. Titchmarsh, Introduction to the Theory of Fourier Integrals (Oxford, Clarendon Press, 1937), p. 334.
  † E. T. Whittaker and G. N. Watson, A Course of Modern Analysis (Cambridge University Press, 1920), p. 115.

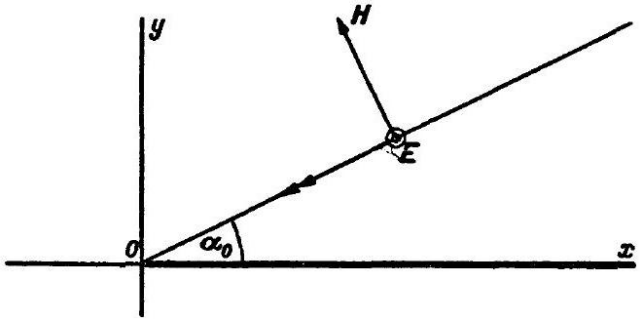



Fig. 11.6. The plane wave incident on the perfectly conducting half-plane.

making any additional contribution to the integral on the assumption that $P(\mu)/\sqrt{(1-\mu^2)} \to 0$ as $|\mu| \to \infty$ when $\pi \geqslant \arg\mu \geqslant 0$. Then if $U(\mu)$ is any function which is free of singularities in the half-plane above the path of integration and has appropriate behaviour as $|\mu| \to \infty$ therein, (2) is clearly satisfied by

$$\frac{P(\mu)}{\sqrt{(1-\mu^2)}} = -\frac{1}{2\pi i}\frac{U(\mu)}{U(-\mu_0)}\frac{1}{(\mu+\mu_0)} \tag{4}$$

if the path of integration be indented below the pole at $\mu = -\mu_0$, as shown diagrammatically in Fig. 11.7. For the only relevant singularity of the function on the right-

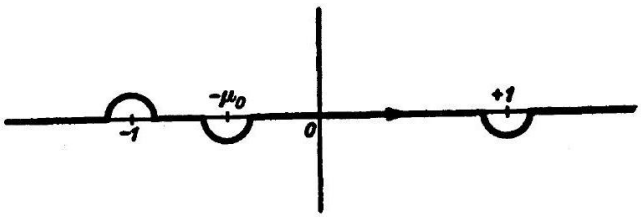



Fig. 11.7. The path of integration in the complex $\mu$ plane.

hand side of (4) is the pole at $\mu = -\mu_0$, with residue $-1/(2\pi i)$, and from Cauchy's residue theorem this will contribute to the integral in (2) precisely the term $-\exp(-ikx\mu_0)$.

If we now rewrite (4) in the form

$$\frac{P(\mu)}{\sqrt{(1-\mu)}}(\mu+\mu_0) = -\frac{1}{2\pi i}\frac{U(\mu)}{U(-\mu_0)}\sqrt{(1+\mu)} \tag{5}$$

it can be argued that each side of (5) is a constant. For the left-hand side is free of singularities in the half-plane below the path of integration and of algebraic growth at infinity therein, whereas the right-hand side has the same characteristics in the halfplane above the path of integration. The function with which both sides are identical is thus free of singularities and of algebraic growth at infinity over the whole complex
$\mu$ plane: it must therefore be a polynomial, and since for some values of $\arg\mu P(\mu) \to 0$ as $|\mu| \to \infty$, this polynomial can contain only a constant term.

The value of the constant is straightway found by putting $\mu = -\mu_0$ in the righthand side of (5), whence

$$P(\mu) = \frac{i}{2\pi}\frac{\sqrt{(1-\mu_0)}\sqrt{(1-\mu)}}{\mu+\mu_0}, \tag{6}$$

or

$$P(\cos\alpha) = \frac{i}{\pi}\frac{\sin\frac{1}{2}\alpha_0 \sin\frac{1}{2}\alpha}{\cos\alpha + \cos\alpha_0}. \tag{7}$$

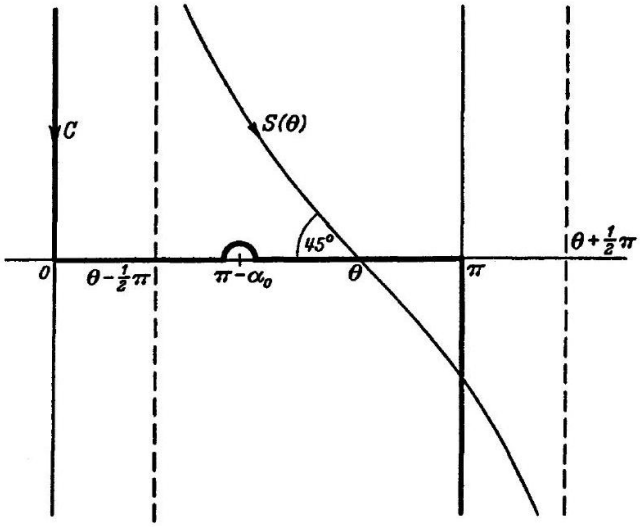



Fig. 11.8. The steepest descents path $S(\theta)$ in the complex $\alpha$ plane.

The significance of the symmetry of (7) in $\alpha$ and $\alpha_0$ is mentioned at the end of § 11.7.1. The components of the scattered field follow from § 11.4 (10), (11), (12) with the value (7) for $P(\cos\alpha)$, and hence the total field is given by

$$E_z = e^{-ikr\cos(\theta-\alpha_0)} - \frac{1}{i\pi}\int_C \frac{\sin\frac{1}{2}\alpha_0 \sin\frac{1}{2}\alpha}{\cos\alpha + \cos\alpha_0} e^{ikr\cos(\theta\mp\alpha)} d\alpha \tag{8}$$

with the upper sign for $y > 0$, and the lower sign for $y < 0$. This completes the actual solution, and it only remains to cast it into a more useful form.

### 11.5.2 Expression of the solution in terms of Fresnel integrals

When $kr$ is large, that is for distances greater than a wavelength or so from the origin, the evaluation of integrals of the general type

$$\int P(\cos\alpha) e^{ikr\cos(\theta-\alpha)} d\alpha \tag{9}$$

may be attempted by the method of steepest descent (see Appendix III). The preliminary step in this procedure is to distort the path of integration (making due allowance for the presence of any singularities in the integrand) into that of steepest descents, $S(\theta)$ say, through the saddle-point at $\alpha = \theta$. The path $S(\theta)$ is shown in Fig. 11.8; along it the new variable

$$\tau = \sqrt{2} e^{\frac{1}{2}i\pi} \sin\frac{1}{2}(\alpha - \theta) \tag{10}$$

goes through real values from $-\infty$ to $\infty$. The integral (9) then appears in the form

$$\sqrt{2} e^{-\frac{1}{2}i\pi} e^{ikr} \int_{-\infty}^{\infty} \frac{P(\cos\alpha)}{\sqrt{\left(1 + \frac{1}{2}i\tau^2\right)}} e^{-krr^2} d\tau, \tag{11}$$

from which asymptotic approximations for $kr \gg 1$ can be obtained.
The application of this procedure to the particular integral in (8) in fact leads, without approximation, to its expression in terms of Fresnel integrals. This will now be shown.

Consider, first, the case $0 < \theta < \pi$. Since (7) can be put in the form

$$P(\cos\alpha) = \frac{1}{4\pi i}\left\{\sec\frac{1}{2}(\alpha-\alpha_0) - \sec\frac{1}{2}(\alpha+\alpha_0)\right\}, \tag{12}$$

it is sufficient to evaluate

$$\int_{S(\theta)} \sec\frac{1}{2}(\alpha-\alpha_0)\, e^{ikr\cos(\theta-\alpha)}d\alpha, \tag{13}$$

as the contribution from $\sec\frac{1}{2}(\alpha+\alpha_0)$ can subsequently be written down by changing the sign of $\alpha_0$. Now by simple transformations (13) is

$$\begin{aligned}
&\int_{S(0)} \sec\frac{1}{2}(\alpha-\alpha_0+\theta)\, e^{ikr\cos\alpha}d\alpha \\
&\quad= \frac{1}{2}\int_{S(0)}\left\{\sec\frac{1}{2}(\alpha-\alpha_0+\theta) + \sec\frac{1}{2}(\alpha+\alpha_0-\theta)\right\} e^{ikr\cos\alpha}d\alpha \\
&\quad= \frac{1}{2}\int_{S(0)} \frac{\cos\frac{1}{2}(\alpha_0-\theta)\cos\frac{1}{2}\alpha}{\cos\alpha+\cos(\alpha_0-\theta)} e^{ikr\cos\alpha}d\alpha
\end{aligned} \tag{14}$$

and using the substitution

$$\tau = \sqrt{2}e^{\frac{1}{2}i\pi}\sin\frac{1}{2}\alpha$$

(14) becomes

$$-2e^{\frac{1}{2}i\pi}e^{ikr}\eta\int_{-\infty}^{\infty}\frac{e^{-kr\tau^2}}{\tau^2-i\eta^2}d\tau, \tag{15}$$

where

$$\eta = \sqrt{2}\cos\frac{1}{2}(\theta-\alpha_0).$$

But

$$\int_{-\infty}^{\infty} e^{-\xi\tau^2}d\tau = \sqrt{\frac{\pi}{\xi}},$$

whence multiplication by $\exp(i\eta^2\xi)$ followed by integration over $\xi$ from $kr$ to infinity gives

$$e^{ikr\eta^2}\int_{-\infty}^{\infty}\frac{e^{-krr^2}}{\tau^2-i\eta^2}d\tau = \sqrt{\pi}\int_{kr}^{\infty}\frac{e^{i\eta^2\xi}}{\sqrt{\xi}}d\xi = \frac{2\sqrt{\pi}}{|\eta|}\int_{|\eta|\sqrt{(kr)}}^{\infty} e^{i\mu^2}d\mu \tag{16}$$

Or, introducing the notation

$$F(a) = \int_a^{\infty} e^{i\mu^2}d\mu \tag{17}$$

for a form of the complex Fresnel integral,*

$$\eta\int_{-\infty}^{\infty}\frac{e^{-krr^2}}{\tau^2-i\eta^2}d\tau = \pm 2\sqrt{\pi}e^{-ikr\eta^2}F\{\pm\eta\sqrt{(kr)}\}, \tag{18}$$

with the upper sign for $\eta > 0$, the lower sign for $\eta < 0$.

Combining these results we finally have, for $y \geqslant 0$,

$$
\begin{aligned}
&-\frac{1}{i\pi}\int_{S(\theta)} \frac{\sin\frac{1}{2}\alpha_0 \sin\frac{1}{2}\alpha}{\cos\alpha + \cos\alpha_0} e^{ikr\cos(\theta-\alpha)} d\alpha \\
&= -\frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}} \left\{ e^{-ikr\cos(\theta-\alpha_0)} F\left[\sqrt{(2kr)}\cos\frac{1}{2}(\theta-\alpha_0)\right]\right. \\
&\left. \mp e^{-ikr\cos(\theta+\alpha_0)} F\left[\pm\sqrt{(2kr)}\cos\frac{1}{2}(\theta+\alpha_0)\right]\right\} \qquad (19)
\end{aligned}
$$

with the upper sign for $\theta + \alpha_0 < \pi$, the lower sign for $\theta + \alpha_0 > \pi$.

To get the complete field from (8) it only remains to take account of the simple pole at $\alpha = \pi - \alpha_0$. For $0 \leqslant \theta \leqslant \pi$, it is easily verified that in distorting the path $C$ to the path $S(\theta)$ the sectors at infinity make no contribution, and it is clear from Fig. 11.8 that the pole is captured if and only if $\pi - \alpha_0 > \theta$. Its contribution, obtained from the residue theorem, is then

$$
e^{-ikr\cos(\theta+\alpha_0)} \qquad (20)
$$

In other words, it is the reflected wave of geometrical optics, the discontinuity in which across $\theta = \pi - \alpha_0$ exactly counterbalances that in the diffraction field (19). In fact, invoking the relation

$$
F(a) + F(-a) = \sqrt{\pi} e^{\frac{1}{2}\pi}, \qquad (21)
$$

the complete field (8) can be written in the form

$$
\begin{aligned}
E_z = \frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}} \left\{ e^{-ikr\cos(\theta-\alpha_0)} F\left[-\sqrt{(2kr)}\cos\frac{1}{2}(\theta-\alpha_0)\right]\right. \\
\left. - e^{-ikr\cos(\theta+\alpha_0)} F\left[-\sqrt{(2kr)}\cos\frac{1}{2}(\theta+\alpha_0)\right]\right\} \qquad (22)
\end{aligned}
$$

which is Sommerfeld's famous result.

When $y < 0$ the integral which has to be evaluated is

$$
\int_C \frac{\sin\frac{1}{2}\alpha_0 \sin\frac{1}{2}\alpha}{\cos\alpha + \cos\alpha_0} e^{ikr\cos(\theta+\alpha)} d\alpha
$$

The appropriate steepest descents path is now $S(2\pi - \theta)$, and the capture of the pole at $\alpha = \pi - \alpha_0$, which occurs only for $\theta > \pi + \alpha_0$, yields minus the incident wave. The complete field is again given by (22).

To obtain the corresponding expressions for the components of $\boldsymbol{H}$ is merely a matter of differentiation, as shown in § 11.4.1. Both the Cartesian components $H_x, H_y$, and the polar components $H_r, H_\theta$ are of interest: in view of the fact that (22) is in terms of $r, \theta$, it is convenient to derive the latter first, from the Maxwell equations

$$
H_r = \frac{1}{ikr}\frac{\partial E_z}{\partial \theta}, \quad H_\theta = -\frac{1}{ik}\frac{\partial E_z}{\partial r} \qquad (23)
$$

0

- This form of the Fresnel integral is more convenient here than those defined in §8.7 (12); the change in limits should be noted.

and then to deduce the former from the relations

$$H_x = \cos\theta H_r - \sin\theta H_\theta, \quad H_y = \sin\theta H_r + \cos\theta H_\theta.$$

The following notation is introduced in order to make the results more compact:

$$u = -\sqrt{(2kr)}\cos\frac{1}{2}(\theta - \alpha_0), \quad v = -\sqrt{(2kr)}\cos\frac{1}{2}(\theta + \alpha_0), \tag{24}$$

$$G(a) = e^{-ia^2}F(a). \tag{25}$$

Note that

$$\frac{dG(a)}{da} = -1 - 2iaG(a)$$

The expression (22) then appears in the form

$$E_z = \frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\{G(u) - G(v)\} \tag{26}$$

and it follows that

$$H_r = \frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\left\{\sin(\theta - \alpha_0)G(u) - \sin(\theta + \alpha_0)G(v) - i\sqrt{\left(\frac{2}{kr}\right)}\sin\frac{1}{2}\alpha_0\cos\frac{1}{2}\theta\right\}, \tag{27}$$

$$H_\theta = \frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\left\{\cos(\theta - \alpha_0)G(u) - \cos(\theta + \alpha_0)G(v) + i\sqrt{\left(\frac{2}{kr}\right)}\sin\frac{1}{2}\alpha_0\sin\frac{1}{2}\theta\right\}, \tag{28}$$

$$H_x = -\frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\left\{\sin\alpha_0[G(u) + G(v)] + i\sqrt{\left(\frac{2}{kr}\right)}\sin\frac{1}{2}\alpha_0\cos\frac{1}{2}\theta\right\},$$

$$H_y = \frac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\left\{\cos\alpha_0[G(u) - G(v)] - i\sqrt{\left(\frac{2}{kr}\right)}\sin\frac{1}{2}\alpha_0\sin\frac{1}{2}\theta\right\}.$$

### 11.5.3 The nature of the solution

We now examine, in some detail, the nature of the results given in §11.5.2. It is evident from their derivation, and can be verified directly, that $G(u)\exp(ikr)$ is itself a solution of the two-dimensional wave equation, for any value of $\alpha_0$; the noteworthy point being that it has periodicity $4\pi$, so that $G(u) - G(v)$ vanishes on $\theta = 0, \theta = 2\pi$, the two faces of the screen, but does not vanish on $\theta = \pi$; Sommerfeld, indeed, arrived at his result (22) by seeking an appropriate solution of the wave equation of period $4\pi$ and combining it with its "image".* Incidentally, it follows from (28) that

$$\cos\frac{1}{2}\theta\frac{e^{ikr}}{\sqrt{(kr)}}, \quad \sin\frac{1}{2}\theta\frac{e^{ikr}}{\sqrt{(kr)}}$$

are also solutions of the two-dimensional wave equation, a result which is well known.
The other aspect of (26) which should be examined is its behaviour as $r \to \infty$. This is a straightforward matter and is the main topic of the subsequent discussion in the present section.

A very attractive feature of the half-plane problem is that the field can be evaluated at any point from tables of the Fresnel integrals. † Furthermore, in two cases of particular interest,

namely $kr \gg 1$ and $kr \ll 1$, simple approximations (mentioned in § 8.4.2) to the Fresnel integrals are available. The former condition is, of course, always satisfied in optical experiments, where the point of observation is likely to be millions of wavelengths from the diffracting edge; the latter condition arises in connection with the behaviour of the field in the vicinity of a sharp edge and can be studied at centimetre radio wavelengths (see § 11.5.6).

$kr \gg 1$. In this case $|u|$ and $|v|$ are large compared to unity except for values of $\theta$ sufficiently close to $\pi + \alpha_0$ and $\pi - \alpha_0$, respectively. To be precise, we introduce five regions as shown in Fig. 11.9. The equations of the curves bounding regions II and IV are taken to be $u^2 = 1$ and $v^2 = 1$, respectively, so that the curves are parabolae with foci at the origin and axes $\theta = \pi + \alpha_0, \theta = \pi - \alpha_0$. Well within region

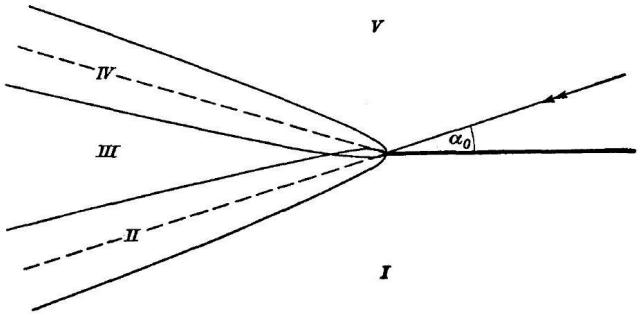



Fig. 11.9. Diffraction of a plane wave by a perfectly conducting half-plane. The five regions in terms of which the behaviour of the field may be described.

II (that is to say, inside a parabola $u^2 = \varepsilon$, where $\varepsilon \ll 1$ ) $|u| \ll 1$; well outside region II (that is to say, outside a parabola $u^2 = \gamma$, where $\gamma \gg 1$ ) $|u| \gg 1$. And similarly with $|v|$ and region IV. Furthermore, for $0 < \theta < \pi - \alpha_0, u$ and $v$ are both negative; for $\pi - \alpha_0 < \theta < \pi + \alpha_0, u$ is negative but $v$ is positive; and for $\pi + \alpha_0 < \theta < 2\pi, u$ and $v$ are both positive.

The regions I, III, and V are obviously closely connected with those which would arise from a geometrical optics treatment according to which the light travels in

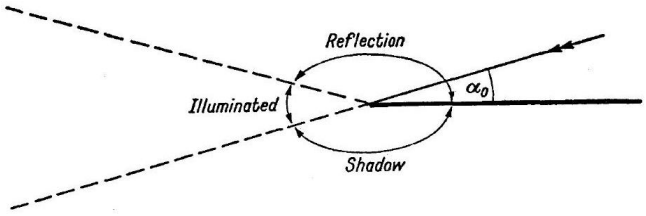



Fig. 11.10. Diffraction of a plane wave by a perfectly conducting half-plane. The three regions of geometrical optics.

straight lines; namely, as illustrated in Fig. 11.10, the shadow sector behind the screen where there is no field at all, the illuminated sector where there is the incident plane wave only, and the reflection sector where there is the incident plane wave together with the reflected plane

0

• A good account of this approach is given in B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 2nd ed., 1950), Chapter 4.
† The most convenient for this purpose seem to be those given by R. A. Rankin, Phil. Trans. Roy. Soc. A, 241 (1949), 457.

wave appropriate to reflection at an infinite screen. In fact, broadly speaking, the regions II and IV are those in which the exact solution effects a smooth transition from the geometrical field in one sector to that in an adjacent sector. To see this in more detail we must pause to derive an asymptotic approximation to the Fresnel integral for large values of the argument.

If $a$ is positive we write

$$G(a) = e^{ia^2} \int_a^\infty \frac{d\left(e^{i\mu^2}\right)}{2i\mu}$$

and integrate by parts twice to get

$$G(a) = \frac{i}{2a} + \frac{1}{4a^3} + \frac{3}{4} e^{-ia^2} \int_a^\infty \frac{e^{i\mu^3}}{\mu^4} d\mu \tag{29}$$

A continuation of this process would, in fact, yield a complete asymptotic expansion of $G(a)$, but for our present purposes we merely observe that the modulus of the integral in the last term of (29) is less than

$$\int_a^\infty \frac{d\mu}{\mu^4} = \frac{1}{3a^3} \tag{30}$$

whence

$$G(a) = \frac{i}{2a} + O\left(\frac{1}{a^3}\right) \tag{31}$$

It is worth noting that this result could also have been obtained from the general method outlined in § 11.5.2, which here consists in expanding the factor $(\tau^2 - i\eta^2)^{-1}$ in the integrand of (18) as a power series in $\tau$, and then integrating term by term.

If $a$ is negative the left-hand side of (30) diverges, but this case can easily be handled by using (21) in conjunction with the result for a positive argument. Thus

$$G(a) = \sqrt{\pi} e^{i\pi} e^{-ia} + \frac{i}{2a} + O\left(\frac{1}{a^3}\right) \tag{32}$$

The fact that the asymptotic approximations (31) for $a$ positive and (32) for $a$ negative are different is a particular example of the Stokes phenomenon.*

Now let us write

$$E_z = E_z^{(g)} + E_z^{(d)}$$

where $E_z{}^{(g)}$ is the geometrical optics field given by

$$E_z^{(g)} = \begin{cases} e^{-ikr\cos(\theta-\alpha_0)} - e^{-ikr\cos(\theta+\alpha_0)} & \text{for } 0 \leqslant \theta < \pi - \alpha_0 \\ e^{-ikr\cos(\theta-\alpha_0)} & \text{for } \pi - \alpha_0 < \theta < \pi + \alpha_0 \\ 0 & \text{for } \pi + \alpha_0 < \theta \leqslant 2\pi \end{cases} \tag{33}$$

and $E_z{}^{(d)}$ is the diffraction field, which is simply the field which must be added to that of geometrical optics to give the complete field. Then, for $kr \gg 1$, the application of (31) and (32) to (26) gives

$$E_z^{(d)} \sim \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}i\pi} \frac{\sin\frac{1}{2}\alpha_0 \sin\frac{1}{2}\theta}{(\cos\theta + \cos\alpha_0)} \frac{e^{-ikr}}{\sqrt{(kr)}} \tag{34}$$

at points not too close to regions II and IV, in the sense indicated above. It is readily seen, either from (23) or (27), that the components of $\boldsymbol{H}^{(d)}$ to the same order of approximation

as (34) are $H_\theta{}^{(d)} = -E_z{}^{(d)}$ and $H_r{}^{(d)} = 0$. Evidently (34) implies that the diffraction field behaves as though it originates from a line-source situated along the diffracting edge whose "polar diagram" varies with angle as specified. This accords with the experimental fact that the diffracting edge, when viewed from the shadow sector for the sake of contrast, appears illuminated.

When $\cos\theta + \cos\alpha_0$ approaches zero, the approximation (34) breaks down and appeal must be made to the exact solution. Since

$$G(0) = \int_0^\infty e^{i\mu^2} d\mu = \frac{1}{2}\sqrt{\pi} e^{\frac{1}{2}i\pi} \tag{35}$$

we see from (26) that, on $\theta = \pi + \alpha_0$,

$$E_z = \frac{1}{2} e^{ikr} + O\left\{\frac{1}{\sqrt{(kr)}}\right\} \tag{36}$$

and on $\theta = \pi - \alpha_0$,

$$E_z = e^{ikr\cos(2\alpha_0)} - \frac{1}{2} e^{ikr} + O\left\{\frac{1}{\sqrt{(kr)}}\right\} \tag{37}$$

Hence, near $\theta = \pi + \alpha_0$ and $\theta = \pi - \alpha_0$ the diffraction field is of the same order as the incident field. In particular, at infinity the transition between the geometrical optics fields in adjacent sectors is via their arithmetic mean.

Interference between the geometrical optics field and the diffraction field in regions where they are comparable gives rise to fringes. These are evident in Fig. 11.11, which is discussed in § 11.5.5.

$kr \ll 1$. In this case $|u|$ and $|v|$ are small compared to unity and series expansions of the Fresnel integral are useful. Writing

$$F(a) = \int_0^\infty e^{i\mu^2} d\mu - \int_0^a e^{i\mu^2} d\mu$$

and expanding the exponential in the integrand of the second integral, we have

$$F(a) = \frac{1}{2}\sqrt{\pi} e^{\frac{1}{2}i\pi} - a + O\left(a^2\right) \tag{38}$$

Hence, from (26) and (28), neglecting powers of $kr$ greater than a half,

$$\left.\begin{aligned} E_z &= 2\sqrt{\tfrac{2}{\pi}} e^{-\frac{1}{2}i\pi} \sqrt{(kr)} \sin\tfrac{1}{2}\alpha_0 \sin\tfrac{1}{2}\theta \\ H_x &= -\sin\alpha_0 - \sqrt{\tfrac{2}{\pi}} e^{-\frac{1}{2}i\pi} \sin\tfrac{1}{2}\alpha_0 \cos\tfrac{1}{2}\theta \left\{\tfrac{i}{\sqrt{(kr)}} + (1 + 2\cos\alpha_0)\sqrt{(kr)}\right\}, \\ H_y &= \sqrt{\tfrac{2}{\pi}} e^{-\frac{1}{2}i\pi} \sin\tfrac{1}{2}\alpha_0 \sin\tfrac{1}{2}\theta \left\{-\tfrac{i}{\sqrt{(kr)}} + (1 + 2\cos\alpha_0)\sqrt{(kr)}\right\} \end{aligned}\right\} \tag{39}$$

It should be noticed that $E_z$ is finite and continuous at $r = 0$, but that $H_x$ and $H_y$ diverge like $r^{-1/2}$, except on $\theta = \pi$ when $H_x = -\sin\alpha_0 \exp(ikr\cos\alpha_0)$, and $\theta = 0$, $2\pi$ when $H_y = 0$. Such behaviour, peculiar in a physical problem, arises of course, from the idealized concept of an infinitely sharp edge. The existence of singularities in the field components in this case must be taken into account in formulating any theorem about the uniqueness of the solution (see § 11.9).

0

- G. G. Stokes, Trans. Camb. Phil. Soc., 10 (1864), 105.

We conclude this investigation into the nature of the solution by examining the current density induced in the diffracting screen. This is - $c/4\pi$ times the difference in $H_x$ at $\theta = 0$ and $\theta = 2\pi$; that is, from (28),

$$\frac{2\pi}{c}J_z = \sin\alpha_0 e^{-ikx\cos\alpha_0} - \frac{e^{-1i\pi}}{\sqrt{\pi}}e^{ikx}\left\{2\sin\alpha_0 G\left[\sqrt{(2kx)}\cos\frac{1}{2}\alpha_0\right] - i\sqrt{\left(\frac{2}{kx}\right)}\sin\frac{1}{2}\alpha_0\right\}. \quad (40)$$

For $\sqrt{(2kx)}\cos\frac{1}{2}\alpha_0 \gg 1$, (40) reads

$$J_z = \frac{c}{2\pi}\sin\alpha_0 e^{-ikx\cos\alpha_0} + O\left\{(kx)^{-3/2}\right\}. \quad (41)$$

This result is of interest in indicating the extent to which it is permissible to assume that the current density is that given by geometrical optics, a standard procedure in problems which cannot be solved exactly. Clearly the assumption is only reasonable for values of $\alpha_0$ not near $\pi$, and benefits from the fact that the "correction" term in (41) tends to zero as $x \to \infty$ like $x^{-3/2}$ rather than $x^{-1/2}$.

On the other hand, for $\sqrt{(2kx)}\cos\frac{1}{2}\alpha_0 \ll 1$,

$$J_z = \frac{c}{\pi\sqrt{(2\pi)}}e^{-\frac{1}{2}i\pi}\sin\frac{1}{2}\alpha_0\left\{\frac{i}{\sqrt{(kx)}} + 4\cos^2\frac{1}{2}\alpha_0\sqrt{(kx)}\right\}e^{ikx}, \quad (42)$$

which diverges at the diffracting edge.

### 11.5.4 The solution for $H$-polarization

The case of $H$-polarization, namely when the incident field is specified by

$$H_z^{(i)} = e^{-ikr\cos(\theta-\alpha_0)}, \quad (43)$$

can be treated in the same way as that of $E$-polarization, the analysis being, in fact, practically identical. Alternatively, the former can be deduced from the latter by invoking the exact electromagnetic form of Babinet's principle, given in § 11.3, because the screen complementary to a half-plane is itself a half-plane. It turns out that the complete field is given by

$$H_z = \frac{e^{-\lambda i\pi}}{\sqrt{\pi}}\left\{e^{-ikr\cos(\theta-\alpha_0)}F[\ -\sqrt{(2kr)}\cos\frac{1}{2}(\theta-\alpha_0)\right] + e^{-ikr\cos(\theta+\alpha_0)}F\left[-\sqrt{(2kr)}\cos\frac{1}{2}(\theta+\alpha_0)\right]\right\}. \quad (44)$$

This differs from the corresponding expression for $E_z$ for an $E$-polarized field, (22), only in the sign of the second term.

Using the notation of § 11.5.2 the non-zero field components appear in the form

$$\left.\begin{aligned} H_z &= \tfrac{e^{-\frac{1}{V}i\pi}}{\sqrt{\pi}}e^{ikr}\{G(u)+G(v)\} \\ E_x &= \tfrac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\left\{\sin\alpha_0[G(u)-G(v)] - i\sqrt{\left(\tfrac{2}{kr}\right)}\cos\tfrac{1}{2}\alpha_0\sin\tfrac{1}{2}\theta\right\}, \\ E_y &= -\tfrac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}}e^{ikr}\left\{\cos\alpha_0[G(u)+G(v)] + i\sqrt{\left(\tfrac{2}{kr}\right)}\cos\tfrac{1}{2}\alpha_0\cos\tfrac{1}{2}\theta\right\} \end{aligned}\right\} \quad (45)$$

Clearly $E_x$ vanishes on $\theta = 0$ and $\theta = 2\pi$, and the behaviour of the field for $kr \gg 1$ can again be interpreted in terms of a diffraction field which appears to originate in a line-source along the diffracting edge for points sufficiently far from $\theta = \pi - \alpha_0$ and $\theta = \pi + \alpha_0$. As

$r \to 0, H_z$ remains finite and continuous, whereas $E_x$ and $E_y$ diverge like $r^{-1/2}$, except in so far as $E_x = 0$ for $\theta = 0, 2\pi$, and $E_y = -\cos\alpha_0 \exp(ikr\cos\alpha_0)$ for $\theta = \pi$.

The current density is given by

$$\frac{2\pi}{c}J_x = e^{-ikx\cos\alpha_0} - \frac{2e^{-1i\pi}}{\sqrt{\pi}}e^{ikx}G\left\{\sqrt{(2kx)}\cos\frac{1}{2}\alpha_0\right\}. \tag{46}$$

For $\sqrt{(2kx)}\cos\frac{1}{2}\alpha_0 \gg 1$,

$$\frac{2\pi}{c}J_x = e^{-ikx\cos\alpha_0} - \frac{e^{\frac{1}{2}i\pi}}{\sqrt{(2\pi)}}\sec\frac{1}{2}\alpha_0\frac{e^{ikx}}{\sqrt{(kx)}}, \tag{47}$$

which approximates less rapidly to the current density of geometrical optics than in the case of $E$-polarization. For $\sqrt{(2kx)}\cos\frac{1}{2}\alpha_0 \ll 1$,

$$J_x = \frac{c}{\pi}\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\pi}\cos\frac{1}{2}\alpha_0\sqrt{(kx)}e^{ikx}. \tag{48}$$

This vanishes at $x = 0$, so that, as might be expected, at the edge itself there is no current normal to the edge.

### 11.5.5 Some numerical calculations

A typical theoretical curve, obtained from (26), is shown in Fig. 11.11. It is for a normally incident $E$-polarized plane wave of amplitude unity, and is a plot of the amplitude of $E_z$ against $x$ at a distance of three wavelengths behind the screen ( $ky = -6\pi$ ). It puts in evidence the diffraction fringes in the illuminated region, and the monotonic decay with deeper penetration into the shadow region.

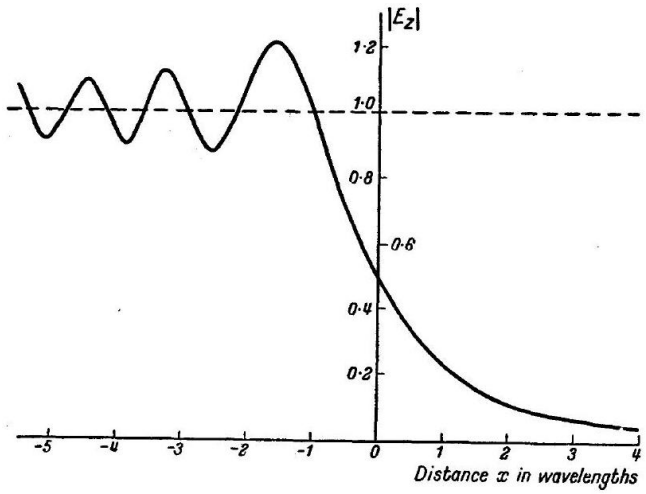



Fig. 11.11. Diffraction of a normally incident $E$-polarized plane wave of amplitude unity by a perfectly conducting half-plane. The variation of $\left|E_{x}\right|$ with $x$ at a distance of three wavelengths behind the screen.

Some interesting calculations have been made by Braunbek and Laukien.* For a normally incident $H$-polarized plane wave of amplitude unity they give contours of equal amplitude (Fig. 11.12) and equal phase (Fig. 11.13) of $H_z$ for the region within about a wavelength of the diffracting edge. They also give the lines of average energy flow (Fig. 11.14) which are orthogonal to the phase contours. That this is the case for any two-dimensional $H$-polarized field is easily proved: write $H_z = he^{i\phi}$, where $h$ and $\phi$ are real; then using the relations

$$E_x = -\frac{1}{ik}\frac{\partial H_z}{\partial y}, \quad E_y = \frac{1}{ik}\frac{\partial H_z}{\partial x},$$

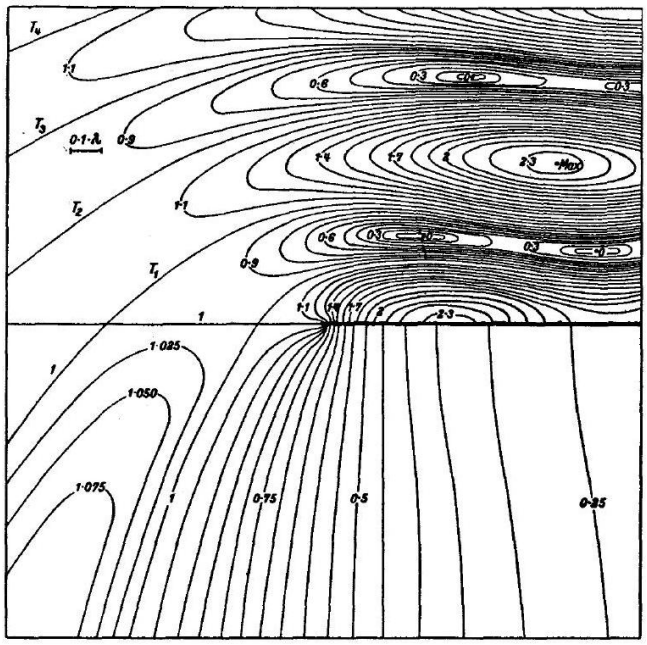


Fig. 11.12. Amplitude contours of $H_{g}$ (amplitude of incident wave is taken as unity).

Diffraction of a normally incident $H$-polarized plane wave by a perfectly conducting half-plane [After W. Braunbek and G. Laukien, Optik, 9 (1952), 174.]

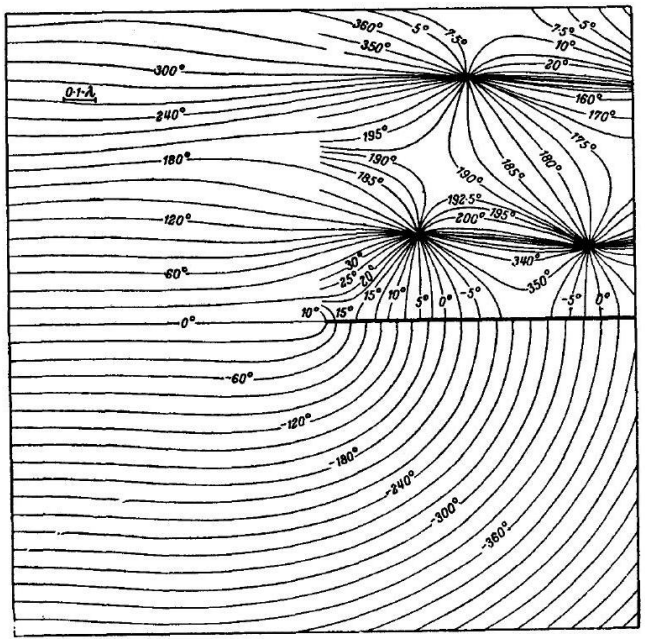


Fig. 11.13. Phase contours of $\boldsymbol{H}_{\boldsymbol{z}}$.

0

- W. Braunbek and G. Laukien, Optik, 9 (1952), 174.

Diffraction of a normally incident $H$-polarized plane wave by a perfectly conducting half-plane [After W. Braunbek and G. Laukien, Optik, 9 (1952), 174.]
the average Poynting vector [§ 1.4 (52)]

$$\frac{c}{8\pi}\mathscr{R}\left(E \wedge H^{\star}\right) = \frac{c}{8\pi}\mathscr{R}\left(E_y H_z^{\star}, -E_x H_z^{\star}, 0\right)$$

is seen to be

$$\frac{c}{8\pi}\frac{h^2}{k}\left(\frac{\partial\phi}{\partial x}, \frac{\partial\phi}{\partial y}, 0\right), \tag{49}$$

which is orthogonal to the surfaces $\phi$ = constant. A corresponding result holds for any $E$-polarized field.

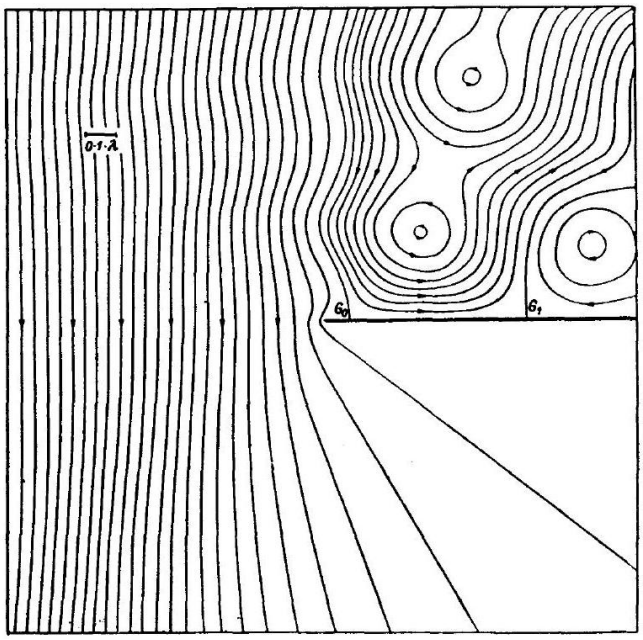



Fig. 11.14. Lines of average energy flow.

Diffraction of a normally incident $H$-polarized plane wave by a perfectly conducting half-plane-
[After W. Braunbek and G. Laukien, Optik, 9 (1952), 174.]

### 11.5.6 Comparison with approximate theory and with experimental results

For points at a great distance from the diffracting edge in the illuminated part of region II (Fig. 11.9), where the fringes appear, the second term in each of the solutions (22) and (44) can be neglected. The intensity for both $E$ - and $H$-polarization, and hence also for unpolarized light, is therefore

$$\frac{1}{2}\left\{\frac{1}{2} + \mathscr{C}\left[2\sqrt{\left(\frac{2r}{\lambda}\right)}\cos\frac{1}{2}\left(\theta - \theta_0\right)\right]\right\}^2 + \frac{1}{2}\left\{\frac{1}{2} + \mathscr{S}\left[2\sqrt{\left(\frac{2r}{\lambda}\right)}\cos\frac{1}{2}\left(\theta - \theta_0\right)\right]\right\}^2, \tag{50}$$

where $\lambda$ is the wavelength and $\mathscr{C}, \mathscr{S}$ are the Fresnel "cosine" and "sine" integrals defined by § 8.7 (12). This should be compared with the analogous result § 8.7 (28) for a black half-plane on the Fresnel-Kirchhoff theory. It has indeed been suggested* that the first term of the exact

solution for the perfectly conducting half-plane could be regarded as giving the solution for a black half-plane.

Well into region I, the shadow region, the $E$-polarization field is given by (34), namely

$$E_z = \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}i\pi} \frac{\sin\frac{1}{2}\alpha_0 \sin\frac{1}{2}\theta}{(\cos\alpha_0 + \cos\theta)} \frac{e^{ikr}}{\sqrt{(kr)}} \tag{51}$$

It can likewise be shown that the $H$-polarization field there is

$$H_z = -\sqrt{\frac{2}{\pi}} e^{i\pi \frac{\cos\frac{1}{2}\alpha_0 \cos\frac{1}{2}\theta}{(\cos\alpha_0 + \cos\theta)}} \frac{e^{ikr}}{\sqrt{(kr)}}. \tag{52}$$

The corresponding ratio of the field strengths is therefore

$$\frac{E\text{-polarization}}{H\text{-polarization}} = -\tan\frac{1}{2}\alpha_0 \tan\frac{1}{2}\theta, \tag{53}$$

and unpolarized incident light will accordingly become partly polarized on diffraction. These results are in broad agreement with optical experiments.*

Developments in microwave radio techniques provide excellent opportunities for the experimental study of the diffraction of electromagnetic waves. In particular, a diffracting screen can be used which is much nearer to the idealization of a perfectly conducting half-plane than can be realized in optical measurements, and the field in the neighbourhood of the diffracting edge can be examined. A number of measurements have been made, mainly on a wavelength of about 3 cm , which show good agreement between theory and experiment. †

## 9.14.6 11.6 Three-dimensional Diffraction of a Plane Wave by a Half-plane

In § 11.5 we solved, in effect, the problem of diffraction by a half-plane of a plane wave which was arbitrary except for the restriction that its direction of propagation was normal to the diffracting edge. It will now be shown that, by a simple device, the previous results can be extended to yield the solution for a completely arbitrary incident plane wave.

Let the incident plane wave be characterized by the phase factor

$$e^{-ikS} = e^{-ik(x\cos\alpha\cos\beta + y\sin\alpha\cos\beta + z\sin\beta)} \tag{1}$$

where, as before, the perfectly conducting screen occupies $y = 0, x > 0$. The angles $\alpha$ and $\beta$, which specify the direction of propagation, are shown in Fig. 11.15.

Now we note that (1) is obtained from the two-dimensional form corresponding to $\beta = 0$ on first replacing $k$ by $k\cos\beta$ and then multiplying by $\exp(-ikz\sin\beta)$. In fact, this procedure applied to any two-dimensional solution of the wave equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + k^2 V = 0 \tag{2}$$

clearly yields a solution of the three-dimensional wave equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} + k^2 V = 0 \tag{3}$$

0

- See the account in B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 2nd ed., 1950), p. 149 et seq.

in which $z$ enters only via the factor $\exp(-ikz\sin\beta)$. Moreover, if $U$, say, is such a solution of (3), then it is easily verified that two electromagnetic fields satisfying Maxwell's equations are

$$\left.\begin{aligned} \boldsymbol{E} &= \left(-\tfrac{i\sin\beta}{k}\tfrac{\partial U}{\partial x}, -\tfrac{i\sin\beta}{k}\tfrac{\partial U}{\partial y}, \cos^2\beta U\right), \\ \boldsymbol{H} &= \left(-\tfrac{i}{k}\tfrac{\partial U}{\partial y}, \tfrac{i}{k}\tfrac{\partial U}{\partial x}, 0\right), \end{aligned}\right\} \tag{4}$$

and

$$\left.\begin{aligned} E &= \left(\tfrac{i}{k}\tfrac{\partial U}{\partial y}, -\tfrac{i}{k}\tfrac{\partial U}{\partial x}, 0\right) \\ H &= \left(-\tfrac{i\sin\beta}{k}\tfrac{\partial U}{\partial x}, -\tfrac{i\sin\beta}{k}\tfrac{\partial U}{\partial y}, \cos^2\beta U\right). \end{aligned}\right\} \tag{5}$$

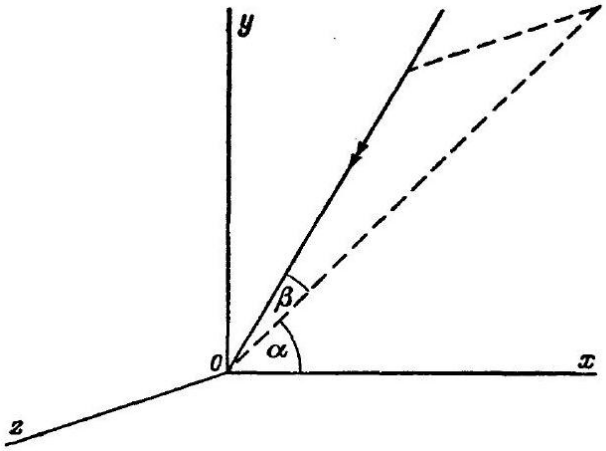



Fig. 11.15. The direction of propagation of the incident plane wave.

When $\beta = 0$, (4) gives a two-dimensional field which is $E$-polarized, (5) one which is $H$-polarized.

If we take the expression (1) for $U$, (4) and (5) yield, respectively, the two plane waves

$$\left.\begin{aligned} \boldsymbol{E} &= (-\cos\alpha\sin\beta, -\sin\alpha\sin\beta, \cos\beta)e^{-ikS}, \\ \boldsymbol{H} &= (-\sin\alpha, \cos\alpha, 0)e^{-ikS}, \end{aligned}\right\} \tag{6}$$

and

$$\left.\begin{aligned} \boldsymbol{E} &= (\sin\alpha, -\cos\alpha, 0)e^{-ikS} \\ \boldsymbol{H} &= (-\cos\alpha\sin\beta, -\sin\alpha\sin\beta, \cos\beta)e^{-ikS}\} \end{aligned}\right\} \tag{7}$$

where a factor $\cos\beta$ has been removed throughout. Now any plane wave with space variation (1) is specified by two components of $\boldsymbol{E}$ (or $\boldsymbol{H}$ ), because the third component would follow from $\operatorname{div} E = 0$ (or $\operatorname{div} H = 0$ ). Hence any plane wave can be formed by suitable superposition of (6) and (7); with the consequence that, in the diffraction problem, attention can be confined, without loss of generality, to the two cases in which they are the respective incident fields.

---

0

- See Wolfsoun's article in Handbuch der Physik, Vol. 20 (Berlin, Springer, 1928), 275, and J. Savornin, Ann. de Physique, 11 (1939), 129.
  † C. W. Horton and R. B. Watson, J. Appl. Phys., 21 (1950), 16; B. N. Harden, Proc. Inst. Elec. Engrs., 99, Pt. III (1952), 229; R, D, Kodis, J. Appl. Phys., 23 (1952), 249; R, V, Row, J. Appl. Phys., 24 (1953), 1448.

It should now be clear that the solution to the diffraction problem with (6) as the incident wave is given by (4) with $U$ got from the known expression for $E_z$ sec $\beta$ in the two-dimensional case on first replacing $k$ by $k\cos\beta$ and secondly multiplying by $\exp(-ikz\sin\beta)$ : for the fact that $U = 0$ on $y = 0, x > 0$ implies also that $\partial U/\partial x = 0$
there, whence from (4) $E_x = E_z = 0$ on the screen, as required. Explicitly, we have, from § 11.5 (24) and § 11.5 (26)

$$U = \frac{e^{-\frac{1}{2}\pi}}{\sqrt{\pi}} \sec\beta e^{ik(r\cos\beta - z\sin\beta)}\{G(p) - G(q)\}, \tag{8}$$

where

$$p = -\sqrt{(2kr\cos\beta)}\cos\frac{1}{2}(\theta - \alpha), \quad q = -\sqrt{(2kr\cos\beta)}\cos\frac{1}{2}(\theta + \alpha). \tag{9}$$

Thus, from (4),

$$\left.\begin{aligned}
E_z &= \tfrac{e^{-\frac{1}{2}\pi}}{\sqrt{\pi}}\cos\beta e^{ik(r\cos\beta - z\sin\beta)}\{G(p) - G(q)\},\\
H_x &= -\tfrac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}} e^{ik(r\cos\beta - z\sin\beta)}\left\{\sin\alpha[G(p) + G(q)] + i\sqrt{\left(\tfrac{2}{kr\cos\beta}\right)}\sin\tfrac{1}{2}\alpha\cos\tfrac{1}{2}\theta\right\},\\
H_y &= \tfrac{e^{-\frac{1}{2}i\pi}}{\sqrt{\pi}} e^{ik(r\cos\beta - z\sin\beta)}\left\{\cos\alpha[G(p) - G(q)] - i\sqrt{\left(\tfrac{2}{kr\cos\beta}\right)}\sin\tfrac{1}{2}\alpha\sin\tfrac{1}{2}\theta\right\},\\
E_x &= -H_y\sin\beta, \quad E_y = H_x\sin\beta, \quad H_z = 0.
\end{aligned}\right\} \tag{10}$$

When $\beta = 0$ the expressions in (10) reduce at once to the corresponding expressions in § 11.5 (26) and (28).

Similar results can likewise be obtained for the case when the incident wave is given by (7), the appropriate two-dimensional solution being that for $H$-polarization, namely, the expression for $H_z$ in § 11.5 (44). As already pointed out, the solution for a quite arbitrary incident plane wave can therefore be derived. Furthermore, a simple generalization of the argument of § 11.5.2 shows that the field radiated by any source distribution can be represented as a spectrum of plane waves. Thus, in principle, the solution of the diffraction problem for any source distribution can be built up from the solutions for the individual plane waves. Two cases of interest are a line-source parallel to the diffracting edge (a two-dimensional problem) and a point-source: these are treated in the next section.

## 9.14.7 11.7 Diffraction of a Localized Source by a Half-plane

### 11.7.1 A line-current parallel to the diffracting edge

We shall consider a line-source situated at $T$, the point $(\ r_0, \theta_0\ )$ where $0 \leqslant \theta_0 \leqslant \pi$, which would, in free-space, radiate the $E$-polarized cylindrical wave

$$E_z^{(i)} = \sqrt{\frac{\pi}{2}} e^{\frac{1}{2}i\pi} H_0^{(1)}(kR) \sim \frac{e^{ikR}}{\sqrt{(kR)}}, \tag{1}$$

where $H_0{}^{(1)}$ is the Hankel function of the first kind and zero order and $R$ is the distance measured from $T$ (see Fig. 11.16).

The source is, in fact, an electric current through $T$ flowing parallel to the $z$ axis and oscillating everywhere with the same phase. It is well known that (1) is the fundamental solution of the two-dimensional wave equation representing a diverging wave which depends only on radial distance.

In order to write (1) as an angular spectrum of plane waves we adopt, in essence, Sommerfeld's integral representation of the Hankel function.* Since we require

the individual plane waves of the spectrum to be incident on the screen, we consider the representation in the half-space $y < r_0 \sin\theta_0$, namely

$$H_0^{(1)}(kR) = \frac{1}{\pi}\int_{S\left(\frac{1}{2}\pi\right)} e^{ikr_0\cos(\theta_0-\alpha)}e^{-ikr\cos(\theta-\alpha)}d\alpha \tag{2}$$

The particular path $S\left(\frac{1}{2}\pi\right)$ is chosen here because (apart from the fact that it is valid for all points in $y < r_0\sin\theta_0$ ) it is parallel to the path of steepest descents, which proves a convenience in the subsequent analysis. The phase factor $\exp\{ikr_0\cos(\theta_0-\alpha)\}$

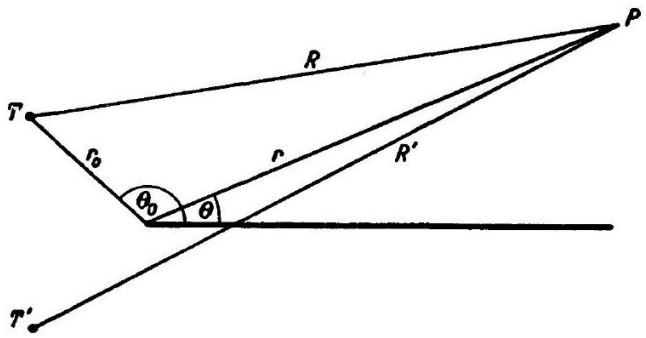



Fig. 11.16. The configuration for a line-source $T$ in the presence of a diffracting half-plane.

in the integrand of (2) appears because the plane waves of the spectrum must all have zero phase at $T$.

It is apparent, therefore, that the solution of the diffraction problem for the incident field (1) is obtained from that for an incident field $\exp\{-ikr\cos(\theta-\alpha)\}$ on multiplying by the factor

$$\frac{e^{ti\pi}}{\sqrt{(2\pi)}}e^{ikr_0\cos(\theta_0-\alpha)} \tag{3}$$

and integrating with respect to $\alpha$ over the path $S\left(\frac{1}{2}\pi\right)$.
Now the solution for the incident field $\exp\{-ikr\cos(\theta-\alpha)\}$, as was shown in § 11.5, could be written

$$E_z^{(p)} = E_z^{(pg)} + E_z^{(pd)} \tag{4}$$

(the affix $p$ being introduced here to indicate that the incident wave is plane), where $E_z{}^{(pg)}$ is given by § 11.5 (33) and (see § 11.5 (19))

$$E_z^{(pd)} = \frac{i}{\pi}\int_{S(0)} \frac{\sin\frac{1}{2}\alpha\sin\frac{1}{2}(\theta+\beta)}{\cos\alpha+\cos(\theta+\beta)}e^{ikr\cos\beta}d\beta \tag{5}$$

The use of this form for $E_z{}^{(pd)}$ (rather than that in terms of Fresnel integrals) is adopted because, as will shortly be seen, it brings out the symmetry between $\alpha$ and $\beta$. The solution for the incident field (1) is thus

0

• See, for example, J. A. Stratton, Electromagnetic Theory (New York, McGraw-Hill, 1941), p. 367.

$$E_z = \frac{e^{\frac{1}{2}i\pi}}{\sqrt{(2\pi)}} \int_{S\left(\frac{1}{2}\pi\right)} E_z^{(pg)} e^{ikr_0\cos(\theta_0-\alpha)} d\alpha + \frac{e^{\frac{1}{2}i\pi}}{\sqrt{(2\pi)}} \int_{S\left(\frac{1}{2}\pi\right)} E_z^{(pd)} e^{ikr_0\cos(\theta_0-\alpha)} d\alpha \tag{6}$$

In order to separate (6) into geometrical optics and diffraction terms, the path of integration for $\alpha$ in the second integral must be displaced to $S(\theta_0)$. In this transposition of the path the poles of $E_z{}^{(pd)}$, regarded as a function of $\alpha$, must be taken into account: from (5) they occur where $\cos\alpha = -\cos(\theta+\beta)$ for all $\beta$ on $S(0)$, and it is easy to verify that the contribution of their residues combines with the first expression in (6) to yield the geometrical optics terms

$$E_z^{(g)} = \begin{cases} \sqrt{\left(\frac{1}{2}\pi\right)}\, e^{\frac{1}{2}i\pi} \left\{ H_0^{(1)}(kR) - H_0^{(1)}(kR') \right\} & \text{for } 0 \leqslant \theta \leqslant \pi - \theta_0 \\ \sqrt{\left(\frac{1}{2}\pi\right)}\, e^{\frac{1}{2}i\pi} H_0^{(1)}(kR) & \text{for } \pi - \theta_0 < \theta < \pi + \theta_0 \\ 0 & \text{for } \pi + \theta_0 < \theta \leqslant 2\pi \end{cases} \tag{7}$$

where $R'$ is distance measured from $T'$, the image of $T$ in the plane $y = 0$ (Fig. 11.16). The diffraction term may be written

$$E_z{}^{(d)} = \frac{e^{-\frac{1}{d}i\pi}}{\pi\sqrt{(2\pi)}} \int_{S(0)} \int_{S(0)} \frac{\sin\frac{1}{2}(\alpha+\theta_0)\sin\frac{1}{2}(\beta+\theta)}{\cos(\alpha+\theta_0)+\cos(\beta+\theta)} e^{ik(r_0\cos\alpha + r\cos\beta)} d\alpha d\beta \tag{8}$$

It is convenient to express (8) as

$$E_z^{(d)} = \mathscr{I}(\theta_0) - \mathscr{I}(-\theta_0) \tag{9}$$

where

$$\mathscr{I}(\theta_0) = -\frac{e^{-\frac{1}{2}i\pi}}{4\pi\sqrt{(2\pi)}} \int_{S(0)} \int_{S(0)} \frac{e^{ik(r_0\cos\alpha + r\cos\beta)}}{\cos\frac{1}{2}(\alpha+\beta+\theta_0+\theta)} d\alpha d\beta \tag{10}$$

The final step is to reduce $\mathscr{I}(\theta_0)$ to a single integral. Multiplying top and bottom of the integrand in (10) by $4\cos\frac{1}{2}(\alpha-\beta+\theta_0+\theta)$ and discarding that part which is an odd function of $\beta$, we have

$$\mathscr{I}(\theta_0) = -\frac{e^{-\frac{1}{2}i\pi}}{8\pi\sqrt{(2\pi)}} \int_{S(0)} \int_{S(0)} \left\{ \frac{1}{\cos\frac{1}{2}(\alpha+\theta_0+\theta) - \sin\frac{1}{2}\beta} + \frac{1}{\cos\frac{1}{2}(\alpha+\theta_0+\theta) + \sin\frac{1}{2}\beta} \right\} \cos\frac{1}{2}\beta e^{ik(r_0\cos\alpha + r\cos\beta)} d\alpha d\beta \tag{11}$$

In the second term of the integrand in (11) change $\alpha$ to $-\alpha$ and then recombine the two terms. This gives

$$\mathscr{I}(\theta_0) = -\frac{e^{-\frac{1}{2}i\pi}}{2\pi\sqrt{(2\pi)}} \times \int_{S(0)} \int_{S(0)} \frac{1}{N} \left\{ \cos\frac{1}{2}(\theta_0+\theta)\cos\frac{1}{2}\alpha\cos\frac{1}{2}\beta e^{ik(r_0\cos\alpha + r\cos\beta)} \right\} d\alpha d\beta \tag{12}$$

where

$$N = (\cos\alpha - 1) + (\cos\beta - 1) - 4\sin\frac{1}{2}(\theta_0+\theta)\sin\frac{1}{2}\alpha\sin\frac{1}{2}\beta + 2\cos^2\frac{1}{2}(\theta_0+\theta).$$

In (12) make the "steepest descent" substitutions

$$\xi = \sqrt{2e^{\frac{1}{i}\pi}} \sin\frac{1}{2}\alpha, \quad \eta = \sqrt{2e^{\frac{1}{i}\pi}} \sin\frac{1}{2}\beta,$$

and write $R_1 = r_0 + r$, to get

$$\begin{aligned}\mathscr{I}(\theta_0) =& \frac{e^{-\frac{1}{2}i\pi}}{\pi\sqrt{(2\pi)}} e^{ikR_1} \cos\frac{1}{2}(\theta_0+\theta) \\ &\times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{e^{-k\left(r_0\xi^2+r\eta^2\right)}}{\xi^2+\eta^2+2\sin\frac{1}{2}(\theta_0+\theta)\,\xi\eta - 2i\cos^2\frac{1}{2}(\theta_0+\theta)} d\xi d\eta \end{aligned} \tag{13}$$

Next, make the polar substitutions

$$\xi = \sqrt{(R_1/r_0)}\,\rho\cos\phi, \quad \eta = \sqrt{(R_1/r)}\,\rho\sin\phi$$

Then

$$\mathscr{I}(\theta_0) = \frac{e^{-\frac{1}{2}i\pi}}{\pi\sqrt{(2\pi)}} e^{ikR_1} \cos\frac{1}{2}(\theta_0+\theta) \int_0^{\infty} \rho K(\rho) e^{-kR_1\rho^2} d\rho \tag{14}$$

where

$$\begin{aligned}K(\rho) = \int_0^{2\pi} \Bigg\{ \rho^2 \left[\sqrt{\frac{r}{r_0}}\cos^2\phi + \sqrt{\frac{r_0}{r}}\sin^2\phi + 2\sin\frac{1}{2}(\theta_0+\theta)\sin\phi\cos\phi\right] \\ -2i\frac{\sqrt{(rr_0)}}{R_1}\cos^2\frac{1}{2}(\theta_0+\theta)\Bigg\}^{-1} d\phi \end{aligned} \tag{15}$$

Now $K(\rho)$ can be evaluated by the standard technique of putting $z = \exp(i\phi)$, when the path of integration becomes the unit circle and it is only necessary to calculate the residues of the enclosed poles. It is found that

$$K(\rho) = 2\pi \left|\sec\frac{1}{2}(\theta_0+\theta)\right| \left\{\rho^4 - 2i\rho^2 - \frac{4rr_0}{R_1^2}\cos^2\frac{1}{2}(\theta_0+\theta)\right\}^{-1/2} \tag{16}$$

where the branch of the square root is that which lies (for real values of $\rho$ ) in the fourth quadrant of the complex plane. Hence (10) becomes

$$\mathscr{I}(\theta_0) = \pm\sqrt{\frac{2}{\pi}} e^{-1i\pi} e^{ikR_1} \int_0^{\infty} \frac{\rho e^{-kR_1\rho^2}}{\sqrt{\left\{\left[\rho^2 - i\left(R_1-R'\right)/R_1\right]\left[\rho^2 - i\left(R_1+R'\right)/R_1\right]\right\}}} d\rho \tag{17}$$

with the $\begin{matrix}\text{upper}\\ \text{lower}\end{matrix}$ sign for $\cos\frac{1}{2}(\theta_0+\theta) \gtrless 0$. Finally the substitution

$$\mu^2 = ikR_1\rho^2 + k\left(R_1 - R'\right)$$

gives the required result, namely

$$\mathscr{I}(\theta_0) = \pm\sqrt{\frac{2}{\pi}} e^{-\frac{1}{4}i\pi} e^{ikR'} \int_{\sqrt{\{k(R_1-R')\}}}^{\infty} \frac{e^{i\mu^2}}{\sqrt{\left(\mu^2+2kR'\right)}} d\mu \tag{18}$$

with the $\begin{matrix}\text{upper}\\ \text{lower}\end{matrix}$ sign for $\cos\frac{1}{2}(\theta_0+\theta)\gtrless 0$.

Since the incident field (1) can be written in the form

$$\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}\pi}e^{ikR}\int_{-\infty}^{\infty}\frac{e^{i\mu^2}}{\sqrt{(\mu^2+2kR)}}d\mu \tag{19}$$

the geometrical optics term (7) combines with the diffraction term (9) to give the total field

$$E_z=\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}i\pi}\left\{e^{ikR}\int_m^{\infty}\frac{e^{i\mu^2}}{\sqrt{(\mu^2+2kR)}}d\mu-e^{ikR'}\int_{m'}^{\infty}\frac{e^{i\mu^2}}{\sqrt{(\mu^2+2kR')}}d\mu\right\} \tag{20}$$

where

$$\left.\begin{aligned}m&=-2\sqrt{\left(\tfrac{krr_0}{R_1+R}\right)}\cos\tfrac{1}{2}(\theta_0-\theta)=\mp\sqrt{\{k(R_1-R)\}}\,,\quad \mp\text{ for }\cos\tfrac{1}{2}(\theta_0-\theta)\gtrless 0,\\ m'&=-2\sqrt{\left(\tfrac{krr_0}{R_1+R'}\right)}\cos\tfrac{1}{2}(\theta_0+\theta)=\mp\sqrt{\{k(R_1-R')\}}\,,\quad \text{for }\cos\tfrac{1}{2}(\theta_0+\theta)\gtrless 0.\end{aligned}\right\} \tag{21}$$

The solution was first given in essentially the form (20) by Macdonald,* who obtained it via a transformation of an earlier solution due to Carslaw. † It is very

similar in type to that of Sommerfeld for an incident plane wave which, indeed, is immediately recovered on multiplying by $\sqrt{(kr_0)}\exp(-ikr_0)$ and letting $r_0\to\infty$. The solution for $H$-polarization differs only in that the two terms in (20) are added instead of being subtracted.

If $kR_1\gg 1,\mu$ may be replaced by its lower-limit value in the non-exponential factor in the integrand of (18) to give the approximate result

$$\mathscr{I}(\theta_0)=\pm\sqrt{\frac{2}{\pi}}e^{-1i\pi}\frac{e^{ikR'}}{\sqrt{\{k(R_1+R')\}}}F\left[\sqrt{\{k(R_1-R')\}}\right] \tag{22}$$

thus, using also the corresponding approximation for $\mathscr{I}(-\theta_0)$, the diffraction field is expressed in terms of Fresnel integrals to a degree of accuracy which is only inadequate if both the source and the point of observation are well within a wavelength of the diffracting edge.

Furthermore, if $k(R_1-R')\gg 1$, the asymptotic approximation § 11.5 (31) may be applied to (22) to give

$$\mathscr{I}(\theta_0)=\frac{e^{\frac{1}{2}i\pi}}{2\sqrt{(2\pi)}}\sec\frac{1}{2}(\theta_0+\theta)\frac{e^{ikr_0}}{\sqrt{(kr_0)}}\frac{e^{ikr}}{\sqrt{(kr)}} \tag{23}$$

and similarly, if $k(R_1-R)\gg 1$,

$$\mathscr{I}(-\theta_0)=\frac{e^{\frac{1}{2}i\pi}}{2\sqrt{(2\pi)}}\sec\frac{1}{2}(\theta_0-\theta)\frac{e^{ikr_0}}{\sqrt{(kr_0)}}\frac{e^{ikr}}{\sqrt{(kr)}} \tag{24}$$

The diffraction field therefore resembles the field of a certain line-source located at the diffracting edge for all points well outside the two hyperbolae $k(R_1-R')=1$ and $k(R_1-R)=1$, the axes of which are $\theta+\theta_0=\pi$ and $\theta-\theta_0=\pi$ respectively. These hyperbolae are the counterparts, for an incident cylindrical wave, of the parabolae discussed in § 11.5.3 in connection with the solution for an incident plane wave.

0

- H. M. Macdonald, Proc. Lond. Math. Soc., 14 (1915), 410.
† H. S. Carslaw, Proc. Lond. Math. Soc., 30 (1899), 121.

Finally, attention should be drawn to the fact that the solution (20) is reciprocal in the sense that it is unaltered by the respective interchange of $r_0, \theta_0$, and $r, \theta$. This, of course, is a particular example of the general theorem on reciprocity* implicit in Maxwell's equations: the present analysis shows that it is here associated with the fact that the spectrum function § 11.5 (7) is symmetrical in $\alpha$ and $\alpha_0$.

### 11.7.2 A dipole

The most elementary source of electromagnetic waves which is localized at a point is a dipole, electric or magnetic. The problem of a dipole in the presence of a halfplane can be solved by representing the undisturbed dipole field as a (three-dimensional) spectrum of plane waves and applying the results of § 11.6 to each plane wave. The case of an electric dipole with its axis normal to the diffracting sheet has been treated in this way by Senior, † and the analysis is sketched in here.

The configuration is shown in Fig. 11.17 with, as before, Cartesian coordinates $x, y, z$ and cylindrical polar coordinates $r, \theta, z$, where the diffracting sheet occupies $y = 0, x > 0$. The dipole at $T, (x_0, y_0, z_0)$ or $(r_0, \theta_0, z_0)$, is taken to be parallel to the $y$ axis; $T'$ is the image of $T$ in the plane $y = 0$, and $R, R'$ are the distances of the point of observation $P$ from $T$ and $T'$ respectively.

With a convenient choice of the magnitude of the dipole moment the undisturbed field of the dipole (cf. § 2.2) may be taken as

$$\boldsymbol{E} = \left( \frac{\partial^2 \Pi}{\partial x \partial y}, \frac{\partial^2 \Pi}{\partial y^2} + k^2 \Pi, \frac{\partial^2 \Pi}{\partial y \partial z} \right), \quad \boldsymbol{H} = ik \left( \frac{\partial \Pi}{\partial z}, 0, -\frac{\partial \Pi}{\partial x} \right), \tag{25}$$

where

$$\Pi = \frac{e^{ikR}}{kR} \tag{26}$$

The required resolution of (25) into plane waves is then obtained from the formula*

$$\frac{e^{ikR}}{kR} = \frac{i}{2\pi} \int_{S\left(\frac{1}{2}\pi\right)} \int_{S\left(\frac{1}{2}\pi\right)} \cos\beta e^{-ik((x-x_0)\cos\alpha\cos\beta + (y-y_0)\sin\alpha\cos\beta - (z-z_0)\sin\beta)} d\alpha d\beta \tag{27}$$

The procedure, therefore, is to write down the total field in the presence of the diffracting sheet arising from the incidence on it of each plane wave implicit in the integrand of (27), and then to carry out the integration.

0

- L. G. H. Huxley, The Principles and Practice of Waveguides (Cambridge University Press' 1947), § 7.17. † T. B. A. Sentor, Quart. J. Mech. Appl. Maths., 6 (1953), 101. For other methods see A. E. Heins, Trans. Inst. Radio Engrs., AP-4 (1956), 294; B. D. Woods, Quart. J. Mech. Appl. Maths., 10 (1957), 90; W. E. Whliams, Quart, J. Mech, Appl. Maths., 10 (1957), 210,

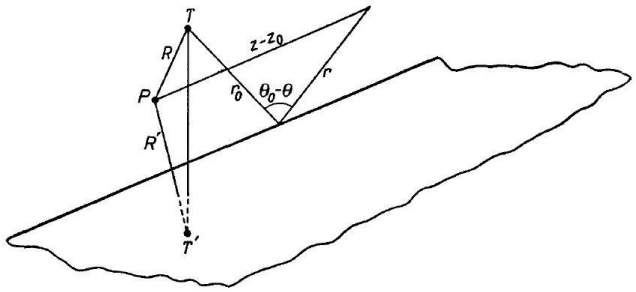



Fig. 11.17. The configuration for a dipole at $T$ in the presence of a diffracting half-plane.

In view of the subsequent integration it is convenient to use (cf. §11.7 (5)) the basic solutions for incident plane waves in the form § 11.5 (8) (for $E$-polarization) with the corresponding form for $H$-polarization. These expressions are modified to yield the three-dimensional solutions corresponding to the incident plane waves implicit in (27) in the manner explained in § 11.6. The component of $E$ parallel to the dipole then appears in the form

$$E_y = E_y^{(g)} + E_y^{(d)}, \tag{28}$$

where $E_y{}^{(g)}$ is the field of geometrical optics and

$$E_y^{(d)} = \frac{ik}{2\pi}\left\{A\left(\theta_0\right) + A\left(-\theta_0\right)\right\} - \frac{i}{2\pi k}\left(\frac{\partial^2}{\partial y \partial y_0} - k^2\right) B\left(\theta_0\right) + \frac{i}{2\pi k}\left(\frac{\partial^2}{\partial y \partial y_0} + k^2\right) B\left(-\theta_0\right) \tag{29}$$

where

$$A\left(\theta_0\right) = \frac{ik}{2\pi}\int_{S(0)}\int_{S\left(\frac{1}{2}\pi\right)}\int_{S(0)} \cos\beta \cos\frac{1}{2}\left(\alpha + \gamma + \theta - \theta_0\right)$$
$$B\left(\theta_0\right) = \frac{ik}{2\pi}\int_{S(0)}\int_{S\left(\frac{1}{2}\pi\right)}\int_{S(0)} \cos\beta \sec\frac{1}{2}\left(\alpha + \gamma + \theta - \theta_0\right) \tag{30}$$
$$\times\, e^{ik(r\cos\gamma\cos\beta + r_0\cos\alpha\cos\beta + (z-z_0)\sin\beta)} d\alpha d\beta d\gamma$$

By means of an analysis similar to that of Carslaw* and Macdonald † it can be shown that

$$A\left(\theta_0\right) = \frac{\pi}{\sqrt{\left(rr_0\right)}} \cos\frac{1}{2}\left(\theta - \theta_0\right) H_0^{(1)}\left(kR_1\right), \tag{32}$$

where

$$R_1^2 = \left(r + r_0\right)^2 + \left(z - z_0\right)^2; \tag{33}$$

$$B\left(\theta_0\right) = \pm\pi k \int_{\pm m}^{\infty} H_1^{(1)}(kR\cosh\mu)d\mu, \quad \pm \text{ for } \theta - \theta_0 \lessgtr \pi \tag{34}$$

$$B\left(-\theta_0\right) = \pm\pi k \int_{\pm m'}^{\infty} H_1^{(1)}\left(kR'\cosh\mu\right)d\mu, \quad \pm \text{ for } \theta + \theta_0 \lessgtr \pi \tag{35}$$

0

- This is a straightforward modification of a formula given by H. Weyl, Ann. d. Physik, 60 (1919), 481.

where

$$m = \sinh^{-1}\left\{2\frac{\sqrt{(rr_0)}}{R}\cos\frac{1}{2}(\theta-\theta_0)\right\}, \quad m' = \sinh^{-1}\left\{2\frac{\sqrt{(rr_0)}}{R'}\cos\frac{1}{2}(\theta+\theta_0)\right\}. \tag{36}$$

These results, together with the formula

$$\frac{1}{2}i\int_{-\infty}^{\infty} H_1^{(1)}(kR\cosh\mu)d\mu = \frac{e^{ikR}}{kR} \tag{37}$$

enable (28) to be written

$$E_y = \frac{ik}{\sqrt{(rr_0)}}\cos\frac{1}{2}\theta\cos\frac{1}{2}\theta_0 H_0^{(1)}(kR_1) - \frac{1}{2}i\left(\frac{\partial^2}{\partial y\partial y_0}-k^2\right)\mathscr{I} + \frac{1}{2}i\left(\frac{\partial^2}{\partial y\partial y_0}+k^2\right)\mathscr{I}' \tag{38}$$

where

$$\mathscr{I} = \int_{-m}^{\infty} H_1^{(1)}(kR\cosh\mu)d\mu, \quad \mathscr{I}' = \int_{-m'}^{\infty} H_1^{(1)}(kR'\cosh\mu)\,d\mu \tag{39}$$

The remaining field components can likewise be expressed, exactly, in terms of $\mathscr{I}$ and $\mathscr{I}'$ as follows

$$E_x = -\frac{ik}{\sqrt{(rr_0)}}\sin\frac{1}{2}\theta\cos\frac{1}{2}\theta_0 H_0^{(1)}(kR_1) - \frac{1}{2}i\frac{\partial^2\mathscr{I}}{\partial x\partial y_0} + \frac{1}{2}i\frac{\partial^2\mathscr{I}'}{\partial x\partial y_0}, \tag{40}$$

$$E_z = -\frac{1}{2}i\frac{\partial^2\mathscr{I}}{\partial z\partial y_0} + \frac{1}{2}i\frac{\partial^2\mathscr{I}'}{\partial z\partial y_0}, \tag{41}$$

$$H_x = \frac{k(z-z_0)}{R_1\sqrt{(rr_0)}}\cos\frac{1}{2}\theta\cos\frac{1}{2}\theta_0 H_1^{(1)}(kR_1) + \frac{1}{2}k\frac{\partial\mathscr{I}}{\partial z_0} + \frac{1}{2}k\frac{\partial\mathscr{I}'}{\partial z_0}, \tag{42}$$

$$H_y = \frac{k(z-z_0)}{R_1\sqrt{(rr_0)}}\sin\frac{1}{2}\theta\cos\frac{1}{2}\theta_0 H_1^{(1)}(kR_1), \tag{43}$$

$$H_z = -\frac{1}{2}k\frac{\partial\mathscr{I}}{\partial x_0} - \frac{1}{2}k\frac{\partial\mathscr{I}'}{\partial x_0}. \tag{44}$$

It is interesting to note that in the complete field there is a non-zero component of $\boldsymbol{H}$ parallel to the dipole, so that the analysis cannot be formulated in terms of a single component Hertz vector.

It is again possible, under a relatively trivial restriction, to express the solution in

terms of Fresnel integrals. If $kR_1 \gg 1$, it is not difficult to establish the asymptotic approximation

$$\mathscr{I} = -\frac{2e^{\frac{1}{2i\pi}}}{k}\sqrt{\left\{\frac{2}{\pi R_1(R+R_1)}\right\}}e^{ikR}F\left\{-2\sqrt{\left(\frac{krr_0}{R+R_1}\right)}\cos\frac{1}{2}(\theta-\theta_0)\right\}, \tag{45}$$

with a similar result for $\mathscr{I}'$.

0

- H. S. Carslaw, Proc. Lond. Math. Soc., 30 (1899), 121.
† H. M. Macdonald, Proc. Lond. Math. Soc., 14 (1915), 410.

## 9.14.8 11.8 Other Problems

In this section several other diffraction problems are briefly reviewed.

### 11.8.1 Two parallel half-planes

The problem of diffraction by two parallel half-planes which are perpendicular to the common plane through their edges is tractable by the method used in this chapter for the single half-plane.*

Consider $E$-polarization, with the problem precisely as in § 11.5.1 except that the diffracting obstacle now consists of two sheets; one (sheet 1) occupying $y = 0, x > 0$; the other (sheet 2), $y = -a, x > 0$. It is convenient to introduce the additional coordinates $r', \theta'$, measured from the edge $(0, -a)$ of sheet 2 .

The scattered fields due to the induced currents in sheets 1 and 2 may be written

$$\begin{aligned} E_z^{(s1)} &= \int_C P_1(\cos\alpha) e^{ikr\cos(\theta \mp \alpha)} d\alpha, \quad \text{for } y \gtrless 0 \\ E_z^{(s2)} &= \int_C P_2(\cos\alpha) e^{ikr'\cos(\theta' \mp \alpha)} d\alpha, \quad \text{for } y \gtrless -a \end{aligned} \tag{2}$$

respectively. The continuity of ${H_x}^{(s1)}$ and ${H_x}^{(s2)}$ across the regions $y = 0, x < 0$, and $y = -a, x < 0$, respectively, is ensured by taking $P_1(\mu)$ and $P_2(\mu)$ to be free of singularities below the path of integration, which is that shown in Fig. 11.7. Furthermore, the boundary condition that $E_z$ should vanish on the two sheets leads to the integral equations

$\int_{-\infty}^{\infty} \frac{P_1(\mu)}{\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu + \int_{-\infty}^{\infty} \frac{P_2(\mu)}{\sqrt{(1-\mu^2)}} e^{ika\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu = -e^{-ikx\mu_0}$,

$\int_{-\infty}^{\infty} \frac{P_1(\mu)}{\sqrt{(1-\mu^2)}} e^{ika\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu + \int_{-\infty}^{\infty} \frac{P_2(\mu)}{\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu = -e^{ika\sqrt{(1-{\mu_0}^2)}} e^{-ikx\mu_0}$,

which must hold for $x > 0$.

If we write

$$P_1(\mu) + P_2(\mu) = Q_1(\mu), \quad P_1(\mu) - P_2(\mu) = Q_2(\mu) \tag{5}$$

addition and subtraction, respectively, of (3) and (4) give

$$\int_{-\infty}^{\infty} \frac{Q_1(\mu)}{\sqrt{(1-\mu^2)}} \left\{1 + e^{ika\sqrt{(1-\mu^2)}}\right\} e^{ikx\mu} d\mu = -\left\{1 + e^{ika\sqrt{(1-{\mu_0}^2)}}\right\} e^{-ikx\mu_0} \tag{6}$$

$$\int_{-\infty}^{\infty} \frac{Q_2(\mu)}{\sqrt{(1-\mu^2)}} \left\{1 - e^{ika\sqrt{(1-\mu^2)}}\right\} e^{ikx\mu} d\mu = -\left\{1 - e^{ika\sqrt{(1-{\mu_0}^2)}}\right\} e^{-ikx\mu_0} \tag{7}$$

for $x > 0$.

The form of each of the equations (6) and (7) is similar to § 11.5 (2), and to obtain solutions analogous to § 11.5 (4) the paths of integration must be closed by infinite semi-circles above the real axis. To this end, that branch of $\sqrt{(1-\mu^2)}$ is chosen which has a positive imaginary part. The required solution of (6) is then

$$\frac{Q_1(\mu)}{\sqrt{(1-\mu^2)}} \left\{1 + e^{ika\sqrt{(1-\mu^2)}}\right\} = \frac{i}{2\pi} \left\{1 + e^{ika\sqrt{(1-\mu_0^2)}}\right\} \frac{U(\mu)}{U(-\mu_0)} \frac{1}{\mu + \mu_0}, \tag{8}$$

where $U(\mu)$ is any function of $\mu$ which is free of singularities in the half-plane above the path of integration and tends to zero as $|\mu| \to \infty$ therein.

0

- Solutions for an incident plane wave were first given, using the method introduced by Schwinger (mentioned in § 11.1), by A. E. Heins, Quart. Appl. Maths., 6 (1948), 157, and L. A Vainstein, Izvestiya Akad. Nauk SSSR, Ser. Fiz. (Bull. Acad. Sci. USSR), 12 (1948), 144 and 166.

It remains to be seen how $Q_1(\mu)$ and $U(\mu)$ can be chosen to satisfy (8), remembering that $Q_1(\mu)$ has no singularities below the path of integration. The procedure is to factorize the coefficient of $Q_1(\mu)$ in (8) in the form

$$\frac{1+e^{ika\sqrt{(1-\mu^2)}}}{\sqrt{(1-\mu^2)}} = U_1(\mu)L_1(\mu) \tag{9}$$

where $U_1(\mu)$ is free of singularities and zeros in the half-plane above the path of integration, and is of algebraic growth at infinity therein, while $L_1(\mu)$ has similar characteristics in the half-plane below the path of integration. That such factorization is possible is known from the general theory of Wiener and Hopf,* and explicit expressions for $U_1(\mu)$ and $L_1(\mu)$ are given by Heins. † Then clearly

$$Q_1(\mu) = \frac{i}{2\pi}\sqrt{\left(1-\mu_0^2\right)}\,U_1\left(\mu_0\right)\frac{1}{L_1(\mu)\left(\mu+\mu_0\right)} \tag{10}$$

where the relation $U_1(\mu) = L_1(-\mu)$, which is implicit in (9) (apart from arbitrary constant factors), has been invoked.

Similarly, if

$$\frac{1-e^{ika\sqrt{(1-\mu^2)}}}{\sqrt{(1-\mu^2)}} = U_2(\mu)L_2(\mu) \tag{11}$$

we have

$$Q_2(\mu) = \frac{i}{2\pi}\sqrt{\left(1-\mu_0^2\right)}\,U_2\left(\mu_0\right)\frac{1}{L_2(\mu)\left(\mu+\mu_0\right)} \tag{12}$$

The total scattered field is obtained by adding (1) and (2). In $y > 0$, for instance, it is therefore

$$\begin{aligned} E_z^{(s)} &= \int_O \left\{P_1(\cos\alpha) + P_2(\cos\alpha)e^{ika\sin\alpha}\right\} e^{ikr\cos(\theta-\alpha)}d\alpha \\ &= \frac{1}{2}\int_C \left\{Q_1(\cos\alpha)\left(1+e^{ika\sin\alpha}\right) + Q_2(\cos\alpha)\left(1-e^{ika\sin\alpha}\right)\right\} e^{ikr\cos(\theta-\alpha)}d\alpha \\ &= \int_C P(\cos\alpha)e^{ikr\cos(\theta-\alpha)}d\alpha, \qquad \text{say} \end{aligned} \tag{13}$$

where

$$\begin{aligned} P(\mu) &= \frac{1}{2}\sqrt{\left(1-\mu^2\right)}\left\{Q_1(\mu)U_1(\mu)L_1(\mu) + Q_2(\mu)U_2(\mu)L_2(\mu)\right\} \\ &= \frac{i}{4\pi}\frac{\sqrt{\left(1-\mu_0^2\right)}\sqrt{\left(1-\mu^2\right)}}{\mu+\mu_0}\left\{U_1\left(\mu_0\right)U_1(\mu) + U_2\left(\mu_0\right)U_2(\mu)\right\} \end{aligned} \tag{14}$$

Again the symmetry of $P(\mu)$ in $\mu$ and $\mu_0$ should be noted. Also that, when $a = 0$, $U_1(\mu) = \sqrt{\{2/(1+\mu)\}}, U_2(\mu) = 0$, so that (14) reduces, as it should, to § 11.5 (6).

0

- E. C. Tifchmarsh, Introduction to the Theory of Fourier Integrals (Oxford, Clarendon Press, 1937), p. 339.
  † A. E. Heins, Quart. Appl. Maths., 6 (1948), 157.

### 11.8.2 An infinite stack of parallel, staggered half-planes

In this problem* we have an infinite set of diffracting sheets, with the $n$th sheet occupying $y = na, x > nb$, where $n = 0, \pm 1, \pm 2, \ldots$.

Then, for an $E$-polarized incident plane wave as before, the scattered field due to currents induced in the $m$ th sheet may be written

$$E_z^{(sm)} = \int_C P_m(\cos\alpha) e^{-ikm(a\cos\alpha \pm b\sin\alpha)} e^{ikr\cos(\theta \mp \alpha)} d\alpha \tag{15}$$

with the upper sign for $y > ma$, and the lower sign for $y < ma$. All the $P_m(\mu)$ must be free of singularities below the path of integration in the $\mu = \cos\alpha$ plane.

The boundary conditions on the $n$th sheet yield the integral equation

$$\sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{P_m(\mu)}{\sqrt{(1-\mu^2)}} e^{ika|n-m|\sqrt{(1-\mu^2)}} e^{ik(x-mb)\mu} d\mu = -e^{-ikx\mu_0} e^{-ikna\sqrt{(1-\mu_0{}^2)}} \tag{16}$$

which must hold for $x > nb$.

From the periodicity of the problem it is clear that

$$P_m(\mu) = P_0(\mu) e^{-ikm\left\{b\mu_0 + a\sqrt{(1-\mu_0^2)}\right\}} \tag{17}$$

and on putting $n - m = q$, (16) therefore becomes

$$\sum_{q=-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{P_0(\mu)}{\sqrt{(1-\mu^2)}} e^{ikq\left(b\mu_0 + a\sqrt{(1-\mu_0{}^2)}\right\}} e^{ika|q|\sqrt{(1-\mu^2)}} e^{ik(qb+x)\mu} d\mu = -e^{-ikx\mu_0} \tag{18}$$

for $x > 0$.

The infinite sum over $q$ can be put in closed form, again leaving an integral equation which can be solved by the use of Cauchy's residue theorem. As in the previous problem, it is necessary to "split" a certain function into a pair of factors, one of which is free of singularities and zeros in the upper half-plane and of algebraic growth at infinity therein, the other having similar characteristics in the lower half-plane: the factors are given explicitly, together with further details, in the papers of Carlson and Heins to which reference has already been made.

### 11.8.3 A strip

Another problem of intrinsic interest is the deceptively simple one in which the diffracting obstacle is a perfectly conducting plane strip, infinitely long with parallel edges; or the corresponding one with the complementary "screen", a slit in an infinite plane. Various methods of solution have been given, † but none yields a result in closed form. The dual integral equation approach has been used ‡ to obtain the

0

- J. F. Carlson and A. E. Heins, Quart. Appl. Maths., 4 (1947), 313 and 5 (1947), 82.
  † Lord Rayleigh, Phil. Mag., 43 (1897), 259 [reprinted in Scientific Papers, 4, 283]; K. Schwarzschild, Math. Ann., 55 (1902), 177; P. M. Morse and P. J. Rubenstein, Phys. Rev., 54 (1938), 895; A. Sommerfeld, Optics (New York, Academic Press, 1954), p. 273; B. B. Baker and E. T. Copson, The Mathematical Theory of Huygens' Principle (Oxford, Clarendon Press, 2nd ed., 1950), p. 177 et seq.; S. Skavlem, Arch. Math. Naturvid., 51 (1951), 61; E. B. Moullin and F. M. Phillips, Proc. Inst. Elec. Engrs., 99, Pt. IV (1952), 137; R. Mller and K. Westpfahl, Z. *f*. Phys., 134 (1953), 245; P. C. Clemmow, Trans. Inst. Radio Engrs., AP-4 (1956), 282; S. N. Karp and A. Russek, J. Appl. Phys., 27 (1956), 886; R. F. Mllar, Proc. Camb. Phil. Soc., 54 (1958), 479, 497.
  ‡ E. Groschwitz and H. Hönl, Z. f. Phys., 131 (1952), 305; H. Hönl and E. Zimmer, Z. f. Phys., 135 (1953), 196; C. J. Tranter, Quart. J. Mech. Appl. Maths., 7 (1954), 317.

first two terms of a series solution in powers of $ka$, where $2a$ is the width of the strip, for a normally incident plane wave, in the following manner.

For a strip occupying $y = 0, |x| < a$, and a normally incident $H$-polarized plane wave, the integral equations § 11.4 (19) and § 11.4 (20) are

$$\begin{array}{ll} \int_{-\infty}^{\infty} P(\mu) e^{ikx\mu} d\mu = 1 & \text{for } |x| < a \\ \int_{-\infty}^{\infty} \frac{P(\mu)}{\sqrt{(1-\mu^2)}} e^{ikx\mu} d\mu = 0 & \text{for } |x| > a \end{array} \tag{20}$$

or, since the symmetry of the problem implies $P(\mu) = P(-\mu)$,

$$\begin{array}{ll} \int_{0}^{\infty} P(\mu) \cos(kx\mu) d\mu = \frac{1}{2} & \text{for } |x| < a \\ \int_{0}^{\infty} \frac{P(\mu)}{\sqrt{(1-\mu^2)}} \cos(kx\mu) d\mu = 0 & \text{for } |x| > a \end{array} \tag{22}$$

A solution is sought of the form

$$\frac{P(\mu)}{\sqrt{(1-\mu^2)}} = \sum_{m=0}^{\infty} c_m \frac{J_{2m+1}(ka\mu)}{\mu} \tag{23}$$

since (22) is satisfied by each term of the series. Substitution into (21) shows that the $c_m$ must be found such that

$$\sum_{m=0}^{\infty} c_m \Phi_m = \frac{1}{2} \quad \text{for } |x| < a \tag{24}$$

where

$$\Phi_m = \int_0^{\infty} \frac{\sqrt{(1-\mu^2)}}{\mu} J_{2m+1}(ka\mu) \cos(kx\mu) d\mu \tag{25}$$

Now it can be shown that, for $|ka| \ll 1$,

$$\Phi_m = i \int_0^{\infty} J_{2m+1}(ka\mu) \cos(kx\mu) d\mu + O(ka) = \frac{i \cos\left[(2m+1) \sin^{-1} \frac{x}{a}\right]}{ka\sqrt{(1 - x^2/a^2)}} + O(ka) \tag{26}$$

whence, to the first order, (24) is

$$\frac{i}{\sqrt{(1-x^2/a^2)}} \sum_{m=0}^{\infty} c_m \cos\left[(2m+1) \sin^{-1} \frac{x}{a}\right] = \frac{1}{2} ka \tag{27}$$

giving

$$c_0 = \frac{ka}{2i}, \quad c_m = 0 \quad \text{for } m = 1, 2, 3, \ldots \tag{28}$$

Thus, to this approximation,

$$\frac{P(\mu)}{\sqrt{(1-\mu^2)}} = \frac{ka}{2i} \frac{J_1(ka\mu)}{\mu} \tag{29}$$

and the current density, derived from § 11.4 (16), is

$$J_x(x) = \frac{cka}{2\pi i} \int_0^{\infty} \frac{J_1(ka\mu)}{\mu} \cos(kx\mu) d\mu = \frac{cka}{2\pi i} \sqrt{\left(1 - x^2/a^2\right)} \tag{30}$$

The next approximation, involving terms in $(ka)^3$, is considerably more complicated. Identical expressions for it have been obtained independently by several authors.*

### 11.8.4 Further problems

There are a number of other interesting problems, allied to those which we have discussed, which are capable of solution, but we cannot do more than mention them here.

Diffraction by a two-dimensional wedge, reducing to a half-plane when the exterior wedge angle is $2\pi$, was solved many years ago. † It involves angular spectra of period $2\pi/n$, where $\pi/n$ is the exterior wedge angle.

The problem of a half-plane in the plane interface between two different homogeneous media was first considered by Hanson. ‡ It is tractable by the method of this chapter and has been applied§ in the theory of the propagation of radio waves over the surface of the earth.

Two investigations have been made of the effect of a half-plane under less idealized conditions. In the first, ‖ finite though large conductivity is introduced, which necessitates the use of approximate boundary conditions; in the second, 9 the plate is assumed to be perfectly conducting but to have a finite, though small, thickness.

## 9.14.9 11.9 Uniqueness of Solution

In § 11.2 we saw that the general diffraction problem with which we were concerned could be stated in the following form: given a field $E^{(i)}$ incident on a perfectly conducting surface $S$, to find a field $E^{(s)}$ which could arise from an electric current distribution in $S$ and which is such that its tangential component on $S$ is minus that of $E^{(i)}$.

It is, of course, vital that the formulation should yield a unique solution,** but the demonstration that there cannot, in fact, be more than one field $E^{(s)}$ satisfying the stated conditions is by no means straightforward, particularly when the possibilities of $S$ being infinite and the field containing plane waves are taken into account. Only comparatively recently, it appears, has the result been satisfactorily established, †† although it has long been tacitly accepted.

A further difficulty arises in the special, though commonest, type of diffraction problem in which the obstacle may be assumed to have an infinitely sharp edge and to which the discussion in this chapter has, indeed, been confined. The reason for the

additional complication is that the solution then contains, as we have seen, a singularity at the edge, thus violating an assumption necessary to the uniqueness proof just mentioned.

It can easily be seen that, if an arbitrary edge singularity is permitted, an infinite sequence of solutions can be got by a process of differentiation.* For example, differentiation of the $E$-polarization solution of the half-plane problem § 11.5(22) with respect to $x$ gives an essentially new expression which also satisfies the wave equation and vanishes on the screen; or, again, differentiation of §11.5(22) with respect to $y$ yields an expression which would apparently suffice for the $H$-polarization solution, but which differs from § 11.5 (44).

0

- R. Mller and K. Westpfahl, Z.f. Phys., 134 (1953), 245; C. J. Tranter, Quart. J. Mech* Appl. Maths., 7 (1954), 317. See also C. J. Bouwkamp, Rep. Progr. Phys. (London, Physical Society), 17 (1954), 73.
  † H. M. Macdonald, Electric Waves (Cambridge University Press, 1902); H. S. Carslaw, Proc. Lond. Math. Soc., 18 (1919), 291.
  ‡ E. T. Hanson, Phil. Trans. Roy. Soc. A, 237 (1938), 35.
  § P. C. Clemmow, Phil. Trans. Roy. Soc. A, 246 (1953), 1.
  || T. B. A. Senior, Proc. Roy. Soc. A, 213 (1952), 436.
  IT D. S. Jones, Proc. Roy. Soc. A, 217 (1953), 153.
  ** To establish the existence of a solution is perhaps less important, since this question is settled when one is found in any particular case. But see C. Mller, Math. Ann., 123 (1951), 345, and W. K. Saunders, Proc. Nat. Acad. Sci. U.S.A., 38 (1952), 342.
  †† F. Rellich, Jahr. Deut. Math. Ver., 53 (1943), 57. See the account in A. Sommerfeld, Partial Differential Equations in Physics (New York, Academic Press, 1949), § 28. W. K. Saunders, Proc. Nat. Acad. Sci. U.S.A., 38 (1952), 342.

Each differentiation introduces a singularity of higher order at the diffracting edge. Evidently, in order to ensure uniqueness, some restriction on the nature of the singularity must be specified. The appropriate restriction, and the various ways in which it may be formulated, have been the subject of a number of papers. † The reader must be referred to these for details, but broadly speaking it may be said that the solution involving the singularity of lowest possible order is to be taken as representing the answer to the physical problem, and that this, in fact, rules out any singularities of order greater than $r^{-1/2}$ as $r \to 0$ at the diffracting edge. In particular, it can be established that a solution is unique which implies for the induced current its integrability over the diffracting surface and the vanishing at the edge of its component normal to the edge. The behaviour of the components of $\boldsymbol{E}$ and $\boldsymbol{H}$ near the edge can be deduced from these conditions.

Finally, the possibility of an infinite number of "solutions" having been seen, it may be asked why the method used in this chapter apparently yields a unique solution, which is, moreover, the correct one judged by the above-mentioned criteria. The answer is because of the assumption that the field components in the plane of the diffracting screen can be expressed as convergent Fourier integrals: this precludes them from having singularities of too high an order.

## 9.15 Diffraction of Light by Ultrasonic Waves

In Chapters I and II it was shown that the propagation of electromagnetic waves may be studied either by using Maxwell's equations, supplemented by the material equations, or by means of certain integral equations which utilize the polarization properties of the medium. In particular, either of these methods may also be applied to the study of the propagation of light through a medium whose density depends on space coordinates and on time. Though the former method has been used extensively in the past, the latter has only more recently been applied to such studies. In this chapter we shall apply the integral equation method to the problem of diffraction of light by a transparent homogeneous medium, disturbed by the passage of ultrasonic waves. It will be useful, however, to give first a qualitative description of this diffraction phenomenon and a brief summary of the theoretical work on this problem based on Maxwell's differential equations.

### 9.15.1 12.1 Qualitative Description of the Phenomenon and Summary of Theories Based on Maxwell's Differential Equations

#### 12.1.1 Qualitative description of the phenomenon

Ultrasonic waves are sound waves whose frequencies are higher than those of waves normally audible to the human ear. The angular frequencies of the ultrasonic waves produced in laboratories lie from about $10^5 \text{sec}^{-1}$ to about $3 \times 10^9 \text{sec}^{-1}$, the former value representing the limit of audibility of the human ear. The corresponding range of wavelengths $\Lambda$ of course depends on the velocity $v$ of these waves in the medium in which they travel. For example, in water $v = 1.2 \times 10^5$ cm/sec and the above frequency range corresponds to the wavelength range $\Lambda = 7.5$ cm to $2.5 \times 10^{-4}$ cm.*

In 1921 Brillound † predicted that a liquid traversed by compression waves of short wavelengths, when irradiated by visible light, would give rise to a diffraction phenomenon similar

0

- C. J. Bouwkamp, Physica, 12 (1946), 467.
  † J. Meixner, Ann. d. Physik, 6 (1949), 1; A. W. Maue, Z. f. Phys., 126 (1949), 601; E. T. Copson, Proc. Roy. Soc. A, 202 (1950), 277; D. S. Jones, Quart. J. Mech. Appl. Maths., 3 (1950), 420 ; A. E. Heins and S. Silver, Proc. Camb. Phil. Soc., 51 (1955), 149; ibid., 54 (1958), 131.

to that due to a grating. In order to see this, consider a fluid lying between two infinite planes $y = 0$ and $y = d$, and let a plane compression wave of wavelength $\Lambda$ progress through it along the positive $x$ direction. This creates periodic stratifications of matter along the $x$ axis, the distance between two successive planes of maximum density being $\Lambda$.

Let a monochromatic plane light wave of angular frequency $\omega$ and wavelength $\bar{\lambda}$ inside the medium be incident, with its wave-normal lying in the $xy$-plane and making an angle $\bar{\theta}$ with the $y$ axis (see Fig. 12.1). Further let $\bar{\phi}$ denote the angle which a diffracted ray makes with the $y$ axis. Since the velocity $v$ of the compression waves is always very much smaller than the velocity of light, we may, to a first approximation,

consider the stratification of matter to be stationary. Then the directions $\bar{\phi}$ in which there is an appreciable intensity are determined by the condition that the optical path difference between the rays from two successive planes distance $\Lambda$ apart shall be an integral multiple of $\bar{\lambda}$. This condition gives a relation between $\bar{\lambda}, \bar{\theta}$ and the directions of propagation $\bar{\phi}_l$ of the waves of various orders in the diffracted spectrum:

$$BC - AD = \Lambda \left(\sin \bar{\phi}_l - \sin \bar{\theta}\right) = l\bar{\lambda} \quad (l = 0, \pm 1, \pm 2, \ldots), \tag{1}$$

$AB$ and $CD$ being portions of wave fronts associated with the refracted and diffracted rays. It will be convenient to rewrite (1) in terms of the angles $\theta$ and $\phi$, and the wavelength $\lambda$ outside the medium. If we use in (1) the law of refraction

$$\frac{\sin \bar{\theta}}{\sin \theta} = \frac{\sin \bar{\phi}}{\sin \phi} = \frac{\bar{\lambda}}{\lambda},$$

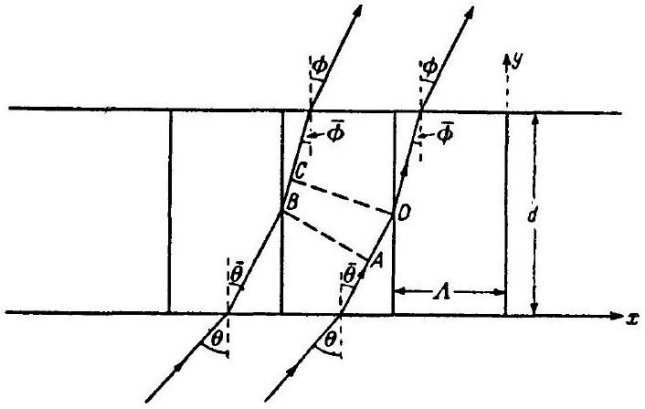



Fig. 12.1. Ultrasonic waves regarded as a diffraction grating.

we obtain

$$\Lambda \left(\sin \phi_l - \sin \theta\right) = l\lambda \quad (l = 0, \pm 1, \pm 2, \ldots). \tag{2}$$

From (2) one has for the angular separation between successive orders

$$\sin \phi_l - \sin \phi_{l-1} \simeq \phi_l - \phi_{l-1} = \frac{\lambda}{\Lambda}.$$

0

- For methods of generation of ultrasonic waves and their many uses, see for example, L. Bergmann, Der Ultraschall (Zrich, Hirzel, 1954). † L. Brillouin, Ann. de Physique, 17 (1921), 103.

Thus, for a given $\lambda$, the angular separation decreases with increasing $\Lambda$. If $\Lambda$ is sufficiently large, the principal lines will be so close together that they will not be resolved in the observing instrument, and for this reason diffraction effects are not observed when ordinary sound waves are irradiated by visible light.

It was nearly a decade after Brillouin's prediction that Debye and Sears,* and Lucas and Biquard, † independently, observed the diffraction of light by ultrasonic waves. Since then, many investigators have studied this phenomenon under a variety of experimental conditions obtained by varying one or more of the following quantities: (a) the angle of incidence $\theta$, (b) the wavelength $\Lambda$ of the ultrasonic wave, (c) the wavelength $\lambda$ of the incident light, (d) the amplitude of the ultrasonic waves, (e) the width $d$ of the ultrasonic beam.

Naturally the positions of the various orders on the screen, and their number and relative intensities, depend on one or more of these factors. ‡ Fig. 12.3, p. 608, shows

in a typical case the number of orders appearing on either side of the transmitted beam at different angles of incidence $\theta$. The usual experimental arrangement for studying diffraction spectrum is schematically drawn in Fig. 12.2.

It will be convenient at this stage to define some of the symbols and sign conventions used in this chapter.

The number density of molecules (atoms) of a medium will be denoted by $N(r,t)$. For an isotropic homogeneous medium traversed by a plane compression wave propagated in the positive $x$ direction, $N(r,t)$ may be written in the form*

$$N(r,t) = N_0[1 + \Delta \cos(Kx - \Omega t)], \tag{3}$$

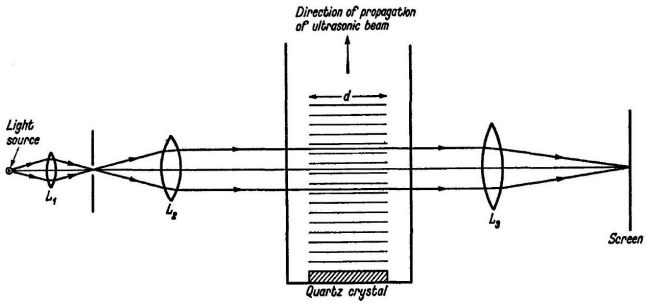



Fig. 12.2. Experimental arrangement for observing diffraction of light by ultrasonic waves.

where $N_0$ is the average number density of the medium, $N_0\Delta$ (usually of the order of $10^{-4}N_0$ ) the amplitude of the compression wave, $K = 2\pi/\Lambda$ its wave number (the magnitude of the wave-vector) and $\Omega = Kv$ the angular frequency of the ultrasonic disturbance. In such a medium the dielectric constant $\varepsilon$ will also be a function of the space and time coordinates; this dependence may be assumed to be of the form

$$\varepsilon = \varepsilon_0 + \varepsilon_1 \cos(Kx - \Omega t). \tag{4}$$

There is, of course, a relation between $\varepsilon_1$ and $\Delta$ which we shall write in the form

0

• P. Debye and F. W. Sears, Proc. Nat. Acad. Sci., Wash., 18 (1932), 409.
† R. Lucas and P. Biquard, J. Phys. Radium, 3 (1932), 464.
‡ For a semi-quantitative discussion of the dependence of the diffracted spectrum on these factors, see G. W. Willard, J. Acoust. Soc. Amer., 21 (1949), 101.

$$\varepsilon_1 = \gamma\Delta. \tag{5}$$

If we assume the Lorentz-Lorenz law (cf. § 2.3 (17))

$$\frac{\varepsilon - 1}{\varepsilon + 2} \cdot \frac{1}{N} = \text{ constant }, \tag{6}$$

differentiate the logarithm of (6) and remember that both $\Delta$ and $\varepsilon_1/\varepsilon_0$ are very much less than unity, we obtain

$$\varepsilon_1 \left[\frac{1}{\varepsilon_0 - 1} - \frac{1}{\varepsilon_0 + 2}\right] - \frac{N_0\Delta}{N_0} = 0$$

or

$$\gamma = \frac{1}{3}(\varepsilon_0 - 1)(\varepsilon_0 + 2); \tag{7}$$

$\gamma$ is of the order of magnitude of unity for most liquids.
We also set

$$\left.\begin{array}{lll} n = \sqrt{\varepsilon_0}, & k = \frac{2\pi}{\lambda}, & \bar{k} = nk \\ \delta = \frac{\Delta\Lambda^2}{\lambda^2}, & \xi = \frac{\Lambda}{\lambda}\sin\theta, & \beta = \frac{\pi\lambda}{n\Lambda^2} \end{array}\right\} \tag{8}$$

Finally the angles $\theta, \phi$, etc., will be measured clockwise from the positive $y$ direction to the direction along which light advances (cf. Fig. 12.1). It may be assumed that $0 \leqslant \theta < \pi/2$.

### 12.1.2 Summary of theories based on Maxwell's equations

In regions free of currents and charges Maxwell's equations for a non-magnetic, non-conducting medium, whose dielectric constant $\varepsilon$ may be a function of space and time coordinates, are

$$\begin{array}{lll} \operatorname{curl} E = -\frac{1}{c}\frac{\partial H}{\partial t}, & & \operatorname{curl} H = \frac{1}{c}\frac{\partial D}{\partial t} \\ \operatorname{div} H = 0, & \operatorname{div} D = 0. & \end{array} \tag{9b}$$

Eliminating $\boldsymbol{H}$ from (9a), and making use of the relations $\boldsymbol{D} = \varepsilon\boldsymbol{E}$, $\operatorname{div}\boldsymbol{D} = 0$ and curl curl $\equiv -\nabla^2 +$ grad div, we find that

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}(\varepsilon E) = \nabla^2 E + \operatorname{grad}(E.\operatorname{grad}\log\varepsilon) \tag{10}$$

If we now use (4) and consider $\boldsymbol{E}$ as a superposition of plane waves of wavelengths $\bar{\lambda} \sim \lambda/n$, we find that the second term on the right-hand side of (10) is of the order $\varepsilon_1(\lambda/\Lambda)$ times the first term. Since under the usual experimental conditions both $\varepsilon_1$ and $\lambda/\Lambda$ are very much less than unity, we may neglect this term in (10), and obtain

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}(\varepsilon E) = \nabla^2 E \tag{11}$$

Now consider the physical situation illustrated in Fig. 12.1, and let the incident monochromatic plane electromagnetic wave be linearly polarized with its electric vector perpendicular to

0

- For simplicity, we shall be concerned only with plane progressive ultrasonic waves. The diffraction of light by standing waves has also been studied experimentally (cf. Bergmann, loc. cit.); the corresponding generalization of the theory is straightforward.

It should be mentioned that completely plane ultrasonic waves are hard to produce experimentally; in general, however, it is possible to regard the wave-fronts to be plane over regions of linear dimensions much greater than $\Lambda$.

the plane of incidence ( $E$-polarization), i.e. along the $z$ axis. It then follows from the preceding discussion that inside the medium the components $E_x$ and $E_y$ of $E$ will be small quantities of order $\varepsilon_1(\lambda/\Lambda)$ times $E_z$ and may, therefore, be neglected. Consequently $E_x$ and $E_y$ outside the medium are also negligible.

From the symmetry of the problem it is clear that $E_z$ will be independent of the $z$-coordinate; hence using (4) and (11), we obtain for $E_z$ the equation

$$\frac{\partial^2 E_z}{\partial x^2}+\frac{\partial^2 E_z}{\partial y^2}-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\left\{\left[\varepsilon_0+\frac{1}{2}\varepsilon_1\left(e^{i(Kx-\Omega t)}+e^{-i(Kx-\Omega t)}\right)\right]E_z\right\}=0 \tag{12}$$

In order to solve (12) assume $E_z$ to be of the form

$$E_z=\sum_l V_l(y)e^{i[(k\sin\theta+lK)x-(\omega+l\Omega)t]} \tag{13}$$

where the summation is over all integral (positive, negative, and zero) values of $l$.
By substituting (13) into (12) and equating the coefficient of each exponential to zero, we obtain the following recurrence relations for $V_l(y)$

$$\begin{aligned}V_l''(y)+&\left\{\varepsilon_0 c^{-2}(\omega+l\Omega)^2-(k\sin\theta+lK)^2\right\}V_l(y)\\&=-\frac{1}{2}\varepsilon_1 c^{-2}(\omega+l\Omega)^2\left\{V_{l-1}(y)+V_{l+1}(y)\right\}\quad(l=0,\pm1,\pm2,\ldots)\end{aligned} \tag{14}$$

where a prime on $V_l(y)$ denotes differentiation with respect to $y$. These equations have to be solved subject to the boundary conditions*
and

$$\left.\begin{array}{ll}V_0(0)=B, & \text{the amplitude of the incident light wave}\\ V_l(0)=0 & \text{for all } l\neq 0.\end{array}\right\} \tag{15}$$

Before solving (14) we observe that (13) represents a superposition of waves of frequencies $\omega_l=\omega+l\Omega(l=0,\pm1,\pm2,\ldots)$. Moreover, the $x$-component of the wave-vector for the wave of frequency $\omega_l$ is $k\sin\theta+lK$. Therefore the sine of the angle $\phi_l$ which the wave of frequency $\omega_l$ makes with the $y$ axis beyond the scattering medium is given by

$$\begin{aligned}\sin\phi_l&=\frac{c(k\sin\theta+lK)}{\omega+l\Omega}\\&\sim\sin\theta+l\frac{\lambda}{\Lambda},\qquad\text{since }\frac{\Omega}{\omega}\ll 1,\end{aligned} \tag{16}$$

in agreement with (2). Further the intensity of a particular order $l$ may be taken to be $|V_l(d)|^2$.

We first solve (14) by assuming that $V_0\gg V_{\pm1}\gg V_{\pm2}\ldots$ Remembering that one has to consider only those solutions of (14) which correspond to light waves travelling in directions along which $y$ increases, we may put all $V_l$, except $V_0$, equal to zero in (14), make use of (15), and have in the first approximation

$$V_0^{(0)}(y)=Be^{i\sqrt{\left(\varepsilon_0-\sin^2\theta\right)}ky}+O\left(\delta^2\right) \tag{17}$$

Similarly, by putting all $V_l$, except $V_{\pm1}$ and $V_0$, equal to zero in (14), we obtain after a straightforward calculation

$$V_{\pm1}(y)=\frac{1}{4}\gamma\delta\frac{1-e^{-2i\beta\left(\xi\pm\frac{1}{2}\right)y}}{\xi\pm\frac{1}{2}}V_0^{(0)}(y). \tag{18}$$

Here use has also been made of (5) and (8). If we now substitute (18) into the equations (14) for $l = 0$ and for $l = \pm 2$, we obtain a correction term to $V_0^{(0)}(y)$ and expressions for $V_{\pm 2}(y)$; both are proportional to $\delta^2$ in this approximation. In this manner we obtain expressions for the intensities of any order $l$ in the form of a series in ascending powers of $\delta$. Solutions in terms of such power series were first derived by Brillouin, † who used somewhat more intricate analysis than outlined

here, while David,* following the above procedure, gave explicit expressions for the intensities of the first and second order lines on the assumption that the intensities of the higher orders are negligible. The latter formulae are quoted in § 12.2.5 [eq. (38)].

Brillouin's approximation (and also David's) is convenient if either $\delta \ll 1$ or $\delta/\xi \ll 1$, for then the method of successive approximations and the power series converge rapidly. When these conditions are satisfied, it is clear that only the first few orders will have appreciable intensity.

The explanation of the simultaneous appearance of many orders and approximate expressions for their intensities were first given by Raman and Nath. † They solved equations (14) in the following manner. Setting in (14)

$$V_l(y) = e^{\pm il\pi} \times e^{i\bar{k}(\cos\delta)y} \times U_l(y), \tag{19}$$

and remembering that $\Omega/\omega \sim 10^{-5}$ or less, $k = nk$ and $k \sin\theta = \bar{k} \sin\bar{\theta}$, we obtain the following recurrence relations for $U_\iota(y)$

$$\begin{aligned} U_l''(y) + 2i\bar{k}\cos\bar{\theta}U_l'(y) - \left(2l\bar{k}K\sin\bar{\theta} + l^2K^2\right)U_l(y) \\ + \frac{1}{2}i(\bar{k})^2\varepsilon_1\varepsilon_0^{-1}\left(U_{l+1}(y) - U_{l-1}(y)\right) = 0 \quad (l = 0, \pm 1, \pm 2, \ldots) \end{aligned} \tag{20}$$

Introducing a new variable

$$\chi = \frac{1}{2}yk\varepsilon_1\varepsilon_0^{-1}\sec\bar{\theta}$$

equations (20) become

$$\begin{aligned} 2U_l'(\chi) + U_{l+1}(\chi) - U_{l-1}(\chi) = \left(\frac{1}{2}i\varepsilon_0^{-1}\varepsilon_1\right)\sec^2\bar{\theta}U_l''(\chi) \\ - 2i\varepsilon_0\varepsilon_1^{-1}\left\{2lK(\bar{k})^{-1}\sin\bar{\theta} + l^2K^2(\bar{k})^{-2}\right\}U_l(\chi) \end{aligned} \tag{21}$$

where now the prime on $U$ denotes differentiation with respect to $\chi$. Since $\lambda/d \ll 1$, the first term on the right is usually of the order $\varepsilon_1 U$, i.e. $10^{-4}U$, and may be neglected. Moreover if, following Raman and Nath, we put the second term also equal to zero, the resulting set of equations are the recurrence relations ‡ satisfied by Bessel functions of integral order. On making use of the boundary conditions (15), we have for the intensity of the $l$ th order wave the expression $B^2J_l^2\left(\frac{1}{2}\varepsilon_1{\varepsilon_0}^{-1}kd\sec\bar{\theta}\right)$.

It will be noticed that the approximation made by Raman and Nath essentially consists in neglecting $l^2/\delta$ and $l\xi/\delta$ for all $l$. Therefore, if $\delta$ is sufficiently large compared to unity, this

0

- These boundary conditions are, of course, correct only if the intensities of the reflected waves are negligible; this is so in the present problem since the angle of incidence $\theta$, for which the amplitudes of the diffracted waves are appreciable, is at most about 3° (see also § 12.2.4).
  † L. Brilloun, La Diffraction de la Lumire par des Ultrasons (Paris, Hermann, 1933).

approximation will be a good one for the intensities of the lower orders. The BeSSEL function expressions, however, overestimate the intensities of higher orders. This has been shown by Extermann and Wannier§ by numerical calculation of the intensities for three values of the parameter $\delta$. In the work of these authors, the solution of (12) is ultimately determined by equations essentially similar to (18) and (19) of the next section.

Finally we mention that Nath* and Aggarwal † have obtained solutions of (21) in the form of power series in ascending powers of $1/\delta$; these series appear to have rather limited applicability as they converge very slowly. Another treatment based on Maxwell's equations in which the diffraction is treated as a boundary value problem was given by Wagner. ‡

## 9.16 12.2 Diffraction of Light by Ultrasonic Waves as Treated by the Integral Equation Method

It was pointed out in § 2.4 that the integral equations § 2.4 (4) for the effective electric field $\boldsymbol{E}'(\boldsymbol{r},t)$ and the accompanying formula § 2.4 (5) for $\boldsymbol{H}'$ are equivalent to Maxwell's equations for isotropic non-magnetic substances. It was assumed there that the density of the medium is independent of time but an extension to cover the more general case of time dependence may easily be made. As before, we shall assume the medium to be non-magnetic and non-conducting.

We recall that the main content of the integral equation method is that the influence of matter on the propagation of an electromagnetic wave is equivalent to the effect of electric dipoles embedded in vacuum, the dipole moment induced in any physically infinitesimal volume element $dr'$ of linear dimensions much smaller than $\lambda$§ being proportional to the field $\boldsymbol{E}'(\boldsymbol{r}',t)$ acting on it and to the number of molecules (atoms) in that volume. Associated with such a dipole at $r'$ is the Hertzian vector

$$\Pi_b = \alpha N\left(r', t-\frac{R}{c}\right)\frac{E'(t-R/c,r')}{R}dr'$$

from which the field at a point $r$ and at time $t$ may be derived by operating on it with (cf. § 2.2 (43))

$$-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}+\ \text{grad div}.$$

Here the various symbols have the same significance as in § 2.4; thus $R=|\boldsymbol{r}-\boldsymbol{r}'|$ and the operator grad div acts on the variables $r(x,y,z)$. Then by an argument similar to that which led to the equation § 2.4 (4), one now obtains the following integral equation for $\boldsymbol{E}'$ within the medium ‖

$$\boldsymbol{E}'(\boldsymbol{r},t)=\boldsymbol{E}^{(i)}(\boldsymbol{r},t)+\alpha\iiint\left\{-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}+\text{grad div}\right\}\left\{N(\boldsymbol{r}',t-R/c)\frac{\boldsymbol{E}'(\boldsymbol{r}',t-R/c)}{R}\right\}d\boldsymbol{r}'. \tag{1}$$

0

- E. David, Phys. Z., 38 (1937), 587.
  † C. V. Raman and N. S. N. Nath, Proc. Ind. Acad. Sci. A, 2 (1935), 406, 413; ibid., 3 (1936), 75, 119.

It may be mentioned that Raman and Nath in their first two papers (loc. cit.), following the work of Rayleigh (Proc. Roy. Soc., A, 79 (1907), 399) on phase gratings, found that the amplitudes of the diffracted waves can be represented by Bessel functions. This method considers only the changes in phase of the plane light wave as it traverses the ultrasonic beam.

‡ See, for example, E. T. Whittaker and G. N. Watson, A Course of Modern Analysis (Cambridge University Press, 4th ed., 1946), p. 360.

§ R. Extermann and G. Wannier, Helv. Phys. Actas 9 (1936), 520.

As in § 2.4 (4), the integration extends throughout the whole medium, except for a small domain occupied by the atom at the point of observation $r(x, y, z)$.

This is the basic integral equation of the present theory. When it is solved for $\boldsymbol{E}'$ at all points inside the medium the field outside the medium is calculated by adding together the incident field $\boldsymbol{E}^{(i)}(\boldsymbol{r}, t)$ and the dipole field $\boldsymbol{E}^{(d)}(\boldsymbol{r}, t)$ given by the integral

in (1) but extending now throughout the whole medium. It will be noted that this treatment of the propagation of light through media, in contrast to the usual method in which one sets up Maxwell's equations for the medium and the vacuum, avoids explicit introduction of boundary conditions at the refracting surfaces but instead brings in the dimensions of the medium through the process of integration throughout the medium. Moreover, whereas in the Maxwell equations the variation in density of the medium is taken into account through its influence on the dielectric constant $\varepsilon$, in the integral equations (1) the density function $N(r, t)$ occurs explicitly.

Equation (1) is valid only under certain restrictive conditions. First, the polarizability $\alpha$ per molecule in general depends on the frequency of $\boldsymbol{E}'$, so that $\boldsymbol{E}'$ should be strictly monochromatic. However, unless one is too near the dispersion frequencies, the variation of $\alpha$ with the frequency of the external field is small. Hence, provided that all the component frequencies of $\boldsymbol{E}'$ lie close to each other, we may still use (1). Even when the incident field $\boldsymbol{E}^{(i)}$ is strictly monochromatic, the field $\boldsymbol{E}'$ acting on a molecule and producing a dipole will not necessarily be monochromatic when $N(r, t)$ depends on time, on account of thermal agitation or some other cause of disorder; the spread of frequencies depends on the time variations of $N(\boldsymbol{r}, t)$. Hence only when the variation in time of $N$ is slow in comparison with that of $E^{(i)}$, may (1) be used with confidence*; fortunately this restriction is not severe, since in problems of scattering and diffraction of light the required condition is almost always satisfied.

Further, we have taken $\alpha$ to be a scalar, an assumption fully justified for atoms and for molecules having special symmetries, but holding also more generally whenever the molecules are oriented at random as already mentioned in § 2.3. Finally, the absorption of light by the medium has been assumed to be negligible here; it could be taken into account by allowing $\alpha$ to be complex.

We shall now apply the integral equation method to the problem of diffraction of light by a fluid traversed by ultrasonic waves, following the analysis of Bhatia and Noble. †

## 9.16.1 12.2.1 Integral equation for $\boldsymbol{E}$-polarization

Let us again consider the physical situation described in § 12.1.1, and assume the incident light to be linearly polarized with its electric vector perpendicular to the plane of incidence; then the components of the electric vector $E^{(i)}(r, t)$ of the incident light wave are (the real part, as usual, representing the physical quantity),

$$\left.\begin{aligned} E_x^{(i)} &= E_y^{(i)} = 0 \\ E_z^{(i)} &= Be^{i(kx\sin\theta + ky\cos\theta - \omega t)}, \end{aligned}\right\} \tag{2}$$

0

- N. S. N. Nath, Proc. Ind. Acad. Sci., A, 4 (1936), 222; ibid., A, 8 (1938), 499.
† R. R. Aggarwal, Ph.D. Thesis, Delhi University, India (1954).
‡ E. H. Wagner, Z. f. Phys., 141 (1955), 604, 622.
§ Such volume elements, namely elements much larger than the volume of an individual atom (molecule) but of linear dimensions small compared to $\lambda$, can be nearly always chosen for optical wavelengths.
|| Equation (1) differs from § 2.4 (4) in that the operator curl curl in the latter has been replaced by ( $-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}+$ grad div ) and that the density $N$ which now depends on space and time coordinates has also to be taken-just like $\boldsymbol{E}'$-at the retarded times $t - R/c$.

From the arguments given on p. 594, it may be concluded that the effective field $\boldsymbol{E}'(\boldsymbol{r},t)$ inside the medium will then be nearly parallel to the $\boldsymbol{z}$ axis, so that we may assume $E'_x = E'_y = 0$, and the vector integral equation (1) for $E'$ reduces to a single integral equation for $E'_z$. Remembering that $N(r,t)$ is now given by §12.1(3), we may write the integral equation for $E'_z$ in the form

$$E'_z(r,t) = Be^{i(kx\sin\theta + ky\cos\theta - \omega t)} + \frac{\tau_0}{4\pi}\iiint\left(-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial z^2}\right) \times \left[\frac{1}{R}\left\{1 + \frac{1}{2}\Delta\left(e^{i[Kx' - \Omega(t-R/c)]} + e^{-i[Kx' - \Omega(t-R/c)]}\right)\right\} \times E'_z\left(r', t - R/c\right)\right] dr' \tag{3}$$

where for convenience we have written*

$$4\pi N_0 \alpha = \tau_0$$

From § 2.3 (17)

$$\tau_0 = \frac{3\left(n^2-1\right)}{n^2+2}. \tag{4}$$

## 9.16.2 12.2.2 The trial solution of the integral equation

Since all planes perpendicular to the $z$ axis are physically equivalent, we take as a trial solution of our integral equation (3) an expression of the form †

$$E'_z = \sum_{l,m} N_{lm} e^{-i(\omega_{lm}t - p_l x - q_m y)} \tag{5}$$

where $l$ and $m$ are integers (positive, negative, and zero). (5) is seen to represent a doubly infinite sheaf of plane waves; this form of a possible solution is suggested by the multiple reflections and refractions to be expected in an infinite slab of stratified medium with parallel plane faces. It will be seen presently that the various unknowns $N_{lm}, \omega_{lm}, p_l$, and $q_m$ may be determined from the condition that (5) satisfies the integral equation (3).

To solve (3), integrals $\mathscr{J}(\omega,p,q)$ must be evaluated, defined by

$$\begin{aligned}\mathscr{J}(\omega,p,q) \times e^{\{-i(\omega t - px - qy)\}} &= \frac{1}{4\pi}\iiint_{V'}\left[-\frac{1}{c^2}\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial z^2}\right]\frac{e^{-i[\omega(t-R/c) - px' - qy']}}{R}dx'dy'dz' \\ &= \frac{1}{4\pi}\iiint_{V'} e^{-i(\omega t - px' - qy')}\left\{\frac{\omega^2}{c^2} + \frac{\partial^2}{\partial z^2}\right\}\frac{e^{i\omega R/c}}{R}dx'dy'dz'\end{aligned} \tag{6}$$

where $\omega^2 > c^2p^2$. If the point of observation $x, y, z$ is outside the scattering medium, the volume $V'$ extends throughout the medium ($-\infty < x' < \infty, 0 \leqslant y' \leqslant d,\ -\infty < z' < \infty$ ). If the point of observation is within the scattering medium, the integral extends over the same

0

- The use of the relation $D = \varepsilon E$ in Maxwell's equations for heterogeneous media is in fact also justified only under similar restrictions on $N$.
† W. J. Noble, Ph.D. Thesis, University of Edinburgh (1952); A. B. Bhatia and W. J Noble, Proc. Roy. Soc., A, 220 (1953), 356, 369.

volume, except for a small sphere of radius $a$ (which is eventually taken to the limit $a \to 0$ ) around the point of observation. Setting

$$x_1 = x' - x, \quad y_1 = y' - y, \quad z_1 = z' - z$$

(6) may be written in the form, after cancelling the factor $e^{-i\omega t}$ on both sides,

$$\mathscr{J}(\omega, p, q) = \mathscr{J}_1 + \mathscr{J}_2 \tag{7}$$

where

$$\mathscr{J}_1 = \frac{1}{4\pi} \iiint_{V_1} \left\{ e^{i(px_1+qy_1)} \frac{\partial^2}{\partial z_1^2} \left( \frac{e^{i\omega R/c}}{R} \right) \right\} dx_1 dy_1 dz_1 \tag{8}$$

$$\mathscr{J}_2 = \frac{1}{4\pi} \frac{\omega^2}{c^2} \iiint_{V_1} \left\{ e^{i(px_1+qy_1)} \left( \frac{e^{i\omega R/c}}{R} \right) \right\} dx_1 dy_1 dz_1 \tag{9}$$

the scattering medium now extending over the volume $V_1 : -\infty < x_1 < \infty - y \leqslant y_1 \leqslant d - y, -\infty < z_1 < \infty$. These integrals are evaluated in Appendix IX and give

(a) when $(x, y, z)$ lies inside the scattering medium*

$$\mathscr{J}(\omega, p, q) = \frac{1}{\sigma(\omega, p, q)} - \frac{\omega^2}{2c^2} \frac{\exp\{}{g(\omega, p, q) \left[\omega^2 c^{-2} - p^2\right]^{1/2}} + \frac{\omega^2}{2c^2} \frac{\exp\{-ih(\omega, p, q)(y - d)\}}{h(\omega, p, q) \left[\omega^2 c^{-2} - p^2\right]^{1/2}} \tag{10a}$$

(b) when ( $x, y, z$ ) lies beyond the scattering medium,

$$\mathscr{J}(\omega, p, q) = \frac{\omega^2 \exp[-ig(\omega, p, q)y]}{2c^2 g(\omega, p, q) \left[\omega^2 c^{-2} - p^2\right]^{1/2}} (\exp[ig(\omega, p, q)d] - 1) \tag{10b}$$

(c) when ( $x, y, z$ ) lies in front of the scattering medium (i.e. on the same side as the incident light),

$$\mathscr{J}(\omega, p, q) = \frac{\omega^2 \exp[-ih(\omega, p, q)y]}{2c^2 h(\omega, p, q) \left[\omega^2 c^{-2} - p^2\right]^{1/2}} (\exp[ih(\omega, p, q)d] - 1) \tag{10c}$$

In these expressions,

$$\left.\begin{aligned} \sigma(\omega, p, q) &= 3\left(p^2 + q^2 - \omega^2 c^{-2}\right)\left(p^2 + q^2 + 2\omega^2 c^{-2}\right)^{-1} \\ g(\omega, p, q) &= q - \left[\omega^2 c^{-2} - p^2\right]^{1/2} \\ h(\omega, p, q) &= q + \left[\omega^2 c^{-2} - p^2\right]^{1/2} \end{aligned}\right\} \tag{11}$$

If we now substitute (5) into (3) and make use of the relations (6) and (10a), we obtain in a straightforward manner

0

- The macroscopic quantity $\tau(r, t) = 4\pi\alpha N(r, t)$ is sometimes called the scattering index of the medium.
† No confusion should arise between the amplitudes $N_{lm}$ and the number density $N(r, t)$ of the molecules since from here onwards the latter quantity does not appear explicitly in our equations.

$$
\begin{aligned}
&- \sum_{l,m} N_{lm} \exp\left[(-i)\left(\omega_{lm} t - p_l x - q_m y\right)\right] + B \exp[(-i)(\omega t - kx \sin\theta - ky\cos\theta)] \\
&+ \tau_0 \sum_{l,m} N_{lm} \left[ \frac{\exp\left[(-i)\left(\omega_{lm} t - p_l x - q_m y\right)\right]}{\sigma\left(\omega_{lm}, p_l, q_m\right)} \right. \\
&+ \frac{1}{2}\Delta \sum_{+,-} \frac{\exp\left[(-i)\left\{\left(\omega_{lm} \pm \Omega\right) t - \left(p_l \pm K\right) x - q_m y\right\}\right]}{\sigma\left(\omega_{lm} \pm \Omega, p_l \pm K, q_m\right)} \\
&- \frac{{\omega_{lm}}^2 \exp\left[(-i)\left\{\omega_{lm} t - p_l x - \left[{\omega_{lm}}^2 c^{-2} - {p_l}^2\right]^{1/2} y\right\}\right]}{2c^2 g\left(\omega_{lm}, p_l, q_m\right)\left[{\omega_{lm}}^2 c^{-2} - {p_l}^2\right]^{1/2}} \\
&\quad \left(\omega_{lm} \pm \Omega\right)^2 \exp\left[(-i)\left\{\left(\omega_{lm} \pm \Omega\right) t - \left(p_l \pm K\right) x\right.\right. \\
&\quad \left.\left. - \frac{1}{2}\Delta \sum_{+,-} \frac{e^{-2}\left(\omega_{lm} \pm \Omega\right)^2 - \left(p_l \pm K\right)^2\right]^{1/2} y\right\}\right] \\
&\quad \frac{}{2c^2 g\left(\omega_{lm} \pm \Omega, p_l \pm K, q_m\right)\left[c^{-2}\left(\omega_{lm} \pm \Omega\right)^2 - \left(p_l \pm K\right)^2\right]^{1/2}} \\
&+ \frac{{\omega_{lm}}^2 \exp\left[(-i)\left\{\omega_{lm} t - p_l x + \left[{\omega_{lm}}^2 c^{-2} - {p_l}^2\right]^{1/2} y\right\}\right] \times \exp\left[ih\left(\omega_{lm}, p_l, q_m\right) d\right]}{2c^2 h\left(\omega_l, p_l, q_m\right)\left[{\omega_{lm}}^2 c^{-2} - {p_l}^2\right]^{1/2}} \\
&\quad + \frac{1}{2}\Delta \sum_{+,-} \frac{\left(\omega_{lm} \pm \Omega\right)^2 \exp\left[(-i)\left\{\left(\omega_{lm} \pm \Omega\right) t - \left(p_l \pm K\right) x\right.\right.}{2c^2 h\left(\omega_{lm} \pm \Omega, p_l \pm K, q_m\right)} \\
&\quad \left.\left. \times \frac{\exp\left[ih\left(\omega_{lm} \pm \Omega, p_l \pm K, q_m\right) d\right]}{\left[c^{-2}\left(\omega_{lm} \pm \Omega\right)^2 - \left(p_l \pm K\right)^2\right]^{1/2}} \right] \equiv 0, \qquad (12)
\end{aligned}
$$

- Throughout § 12.2 the positive square root of an expression is to be taken, unless otherwise stated.
  where the summation symbol $\sum_{+,-}$ in front of any expression is to be interpreted as follows:

$$\sum_{+,-} F(a \pm b, c \pm d) \equiv F(a+b, c+d) + F(a-b, c-d)$$

In order that (12) shall be satisfied at all times and at all points within the scattering medium, the coefficient of each exponential differing from all the others in any of the variables ( $x, y, t$ ) must separately vanish. We see from (12) that $\omega_{lm}$ changes in steps of $\Omega$ and is always accompanied by a change of $p_l$ in steps of $K$. The coefficients of $y$ in the various exponentials, however, either remain unchanged ($q_m$) or are always the same function of the corresponding $\omega$ 's and $p$ 's. Hence we may take $\omega_{lm}$ to depend only on the index $l$. Moreover, since we may assume without loss of generality that $\omega_0$ is the frequency $\omega$ of the incident light, we have

$$\left.\begin{array}{ll} \omega_0 = \omega, & \omega_l = \omega + l\Omega \\ p_0 = k\sin\theta, & p_l = k\sin\theta + lK \end{array}\right\} \quad (l = 0, \pm 1, \pm 2, \ldots) \qquad (13a)$$

Using these relations in (12) and regrouping the various terms, (12) becomes

$$\sum_{l,m}\left[N_{lm}\left(\tau_0/\sigma_{lm}-1\right)+\frac{1}{2}\Delta\left(\tau_0/\sigma_{lm}\right)\left(N_{l-1,m}+N_{l+1,m}\right)\right]\times\exp\left[(-i)\left(\omega_l t-p_l x-q_m y\right)\right]$$
$$+\sum_l\left(B\delta_{l,0}-G_l\right)\exp\left[(-i)\left(\omega_l t-p_l x-\left[\omega_l^2c^{-2}-p_l^2\right]^{1/2}y\right)\right]$$
$$+\sum_l H_l\exp\left[(-i)\left(\omega_l t-p_l x+\left[\omega_l^2c^{-2}-p_l^2\right)^{1/2}y\right)\right]\equiv 0 \tag{14}$$

where $\delta_{l,l'}$ is the Kronecker delta symbol* (i.e. $\delta_{l,l'}=0$ when $l\neq l'$, and $\delta_{ll}=1$ ) and $G_l$ and $H_l$ are given by

$$G_l=\tau_0\omega_l^2\sum_m\left[N_{lm}+\frac{1}{2}\Delta\left(N_{l-1,m}+N_{l+1,m}\right)\right]\times\left\{2c^2g_{lm}\left[\omega_l^2c^{-2}-p_l^2\right]^{1/2}\right\}^{-1} \tag{15}$$

$$H_l=\tau_0\omega_l^2\sum_m\left[N_{lm}+\frac{1}{2}\Delta\left(N_{l-1,m}+N_{l+1,m}\right)\right]\left[\exp\left(ih_{lm}d\right)\right]\left\{2c^2h_{lm}\left[\omega_l^2c^{-2}-p_l^2\right]^{1/2}\right\}^{-1} \tag{16}$$

Here we have also used the abbreviation (cf. (11))

$$\sigma_{lm}=\sigma\left(\omega_l,p_l,q_m\right),\quad g_{lm}=g\left(\omega_l,p_l,q_m\right)\quad\text{and}\quad h_{lm}=h\left(\omega_l,p_l,q_m\right) \tag{17}$$

Equating to zero the coefficient of each exponential in (14), we obtain the following sets of equations for the permissible values of $q_m$ and the amplitudes $N_{lm}$ :

$$\left.\begin{aligned}N_{lm}\left(1-\sigma_{lm}/\tau_0\right)+\tfrac{1}{2}\Delta\left(N_{l-1,m}+N_{l+1,m}\right)&=0,\quad\text{for all } l \text{ and } m\\ B\delta_{l,0}-G_l&=0,\\ H_l&=0\end{aligned}\right\}\ \text{for all } l. \tag{19}$$

## 9.16.3 12.2.3 Expressions for the amplitudes of the light waves in the diffracted and reflected spectra

Before discussing the solution of the equations (18)-(20), we write down the expressions for the total light disturbance at a point ( $x,y,z$ ) beyond the scattering medium. For this purpose one has to substitute (5) into the integrand on the right-hand side of (3), integrate over the scattering medium, and add to this result the incident field.

One again encounters the integrals $\mathscr{J}(\omega,p,q)$ discussed in the preceding section; remembering that for a point beyond the scattering medium $\mathscr{J}(\omega,p,q)$ is given by (10b) and making use of (18), we obtain the following expression for the only nonvanishing component of the total transmitted electric field:
where

$$E_z=\sum_l B_l\exp\left[(-i)\left\{\omega_l t-p_l x-\left[\omega_l{}^2c^{-2}-p_l{}^2\right]^{1/2}y\right\}\right], \tag{21}$$

$$B_l=\omega_l^2\sum_m\sigma_{lm}N_{lm}\left\{\exp\left[ig_{lm}d\right]\right\}\times\left\{2c^2g_{lm}\left[\omega_l^2c^{-2}-p_l^2\right]^{1/2}\right\}^{-1} \tag{22}$$

0

- No confusion should arise between the Kronecker symbol $\delta_{l,l'}$, and the parameter $\delta$ introduced in § 12.1 (8) since the former always occurs with subscripts.

According to (21) and (22) the transmitted wave may be regarded as consisting of many plane waves, each with a different frequency and a different direction of propagation. By substituting (13) in the exponential of (21), the expressions for the frequencies $\omega_l$ and the angles $\phi_l$ may be easily obtained; these are the same as those given in § 12.1.2.

With the help of (10c) one may similarly write down the expressions for the amplitudes ${B_l}^{(r)}$ in the reflected spectrum also. ${B_l}^{(r)}$ are given by

$$B_l^{(r)} = -\omega_l^2 \sum_m \sigma_{lm} N_{lm} \left\{ 2c^2 h_{lm} \left[ \omega_l^2 c^{-2} - p_l^2 \right]^{1/2} \right\}^{-1} \tag{23}$$

We shall, however, be concerned here only with points beyond the scattering medium, since in the present problem the intensities of the waves of the various orders in the reflected spectrum are generally very small.

## 9.16.4 12.2.4 Solution of the method of successive approximations

In § 12.2.4 and § 12.2.5, on the assumption that the amplitude $\Delta$ of the ultrasonic wave is small, equations (18)-(20) will be solved and approximate expressions for the intensities of the first and second order lines in the transmitted spectrum will be obtained. The case for which this approximation fails will be qualitatively discussed in § 12.2.6, and finally in § 12.2.7 equations (18)-(20) will be solved by an approximation essentially equivalent to that of Raman and Nath.

First consider the set of equations (18). The suffix $m$ in these equations distinguishes quantities referring to different permissible values of $q$. Therefore, we may suppress this suffix and write (18) in the form

$$f_l\left(q^2\right) N_l\left(q^2\right) - \frac{1}{2}\Delta\left(N_{l-1}\left(q^2\right) + N_{l+1}\left(q^2\right)\right) = 0 \quad (l = 0, \pm 1, \pm 2, \ldots) \tag{24}$$

with

$$f_l\left(q^2\right) = \frac{\sigma_l(q)}{\tau_0} - 1 = \frac{3\left({p_l}^2 + q^2 - n^2 {\omega_l}^2 c^{-2}\right)}{\left(n^2 - 1\right)\left({p_l}^2 + q^2 + 2{\omega_l}^2 c^{-2}\right)} \tag{25}$$

Equations (24) form an infinite set of linear homogenous equations for the amplitudes $N_l\left(q^2\right)$. The condition for the existence of a non-trivial solution, i.e. $N_l \not\equiv 0$ for all $l$, is that the determinant formed by the coefficients of $N_l$ vanishes. The roots of this determinantal equation give the permissible values of $q^2$; let these be denoted by ${q_m}^2 (m = 0, \pm 1, \pm 2, \ldots)$. Corresponding to each such $q^2$ there are, of course, two values of $q$, viz. $+|q|$ and $-|q|$ and two sets of amplitudes, i.e. $N_l(+|q|) \equiv N_l^+\left(q^2\right)$ and $N_l(-|q|) \equiv N_l^-\left(q^2\right), (l = 0, \pm 1, \ldots)$. Then for a given permissible value of $q^2$, say ${q_m}^2$, the recurrence relations (24) determine all the ${N_l}^+\left({q_m}^2\right), (l = 0, \pm 1, \ldots)$ in terms of one of them, say $N_m^+\left({q_m}^2\right)$. In this way, one may obtain all the amplitudes $N_l^{\pm}\left(q^2\right)$ in terms of $N_m^{\pm}\left({q_m}^2\right) (m = 0, \pm 1, \pm 2, \ldots)$. These latter sets of amplitudes are to be determined from the equations (19) and (20), which are just enough in number for this purpose. (Note that in expressions (5), (12), (15), etc., the symbol $\sum_m$ implies sum over both $N^+_{\ldots,m}$and $N^-_{\ldots,m}$terms.)

We shall obtain here approximate solutions of (24) by using a perturbation method. Regarding $\Delta$ as a small parameter and following the usual perturbation procedure, we expand $\eta\,(= q^2)$ and $N_l$ in powers of $\frac{1}{2}\Delta$ :

$$N_l(\eta) = N_l^{(0)} + \frac{1}{2}\Delta N_l^{(1)} + \left(\frac{1}{2}\Delta\right)^2 N_l^{(2)} + \ldots, \tag{26a}$$

$$\eta = \eta^{(0)} + \frac{1}{2}\Delta\eta^{(1)} + \left(\frac{1}{2}\Delta\right)^2 \eta^{(2)} + \ldots \tag{26b}$$

Making use of (26b), $f_l(\eta)$ may be written as

$$f_l(\eta) = f_l\left(\eta^{(0)}\right) + \frac{1}{2}\Delta\eta^{(1)} f_l'\left(\eta^{(0)}\right) + \left(\frac{1}{2}\Delta\right)^2 \left[\eta^{(2)} f_l'\left(\eta^{(0)}\right) + \frac{1}{2}\left(\eta^{(1)}\right)^2 f_l''\left(\eta^{(0)}\right)\right] + \ldots \tag{27}$$

where a prime on $f$ denotes differentiation with respect to $\eta$. (Note that $f_l'(\eta)$, $f_l''(\eta)\ldots$ are non-zero for every positive real value of $\eta$.) Substituting (26) and (27) in (24), we have ( $l = 0, \pm1, \pm2, \ldots$ )

$$\begin{aligned}
&\left\{f_l\left(\eta^{(0)}\right) + \frac{1}{2}\Delta\eta^{(1)} f_l'\left(\eta^{(0)}\right) + \left(\frac{1}{2}\Delta\right)^2 \left[\eta^{(2)} f_l'\left(\eta^{(0)}\right) + \frac{1}{2}\left(\eta^{(1)}\right)^2 f_l''\left(\eta^{(0)}\right)\right] + \ldots\right\} \\
&\times \left\{N_l^{(0)} + \frac{1}{2}\Delta N_l^{(1)} + \left(\frac{1}{2}\Delta\right)^2 N_l^{(2)} + \ldots\right\} - \frac{1}{2}\Delta\left[N_{l-1}^{(0)} + \frac{1}{2}\Delta N_{l-1}^{(1)}\right. \\
&\left. + \ldots + N_{l+1}^{(0)} + \frac{1}{2}\Delta N_{l+1}^{(1)} + \ldots\right] = 0.
\end{aligned} \tag{28}$$

Equating first to zero those terms in (28) which are independent of $\Delta$, we obtain in the zero-order approximation

$$f_l\left(\eta_l{}^{(0)}\right) N_l{}^{(0)}(\eta) = 0 \quad (l = 0, \pm1, \pm2, \ldots). \tag{29}$$

Equations (29) have as solutions,

$$\text{either } f_l\left(\eta^{(0)}\right) = 0, \quad N_l^{(0)} \neq 0 \quad \text{or} \quad N_l^{(0)} = 0, \quad f_l\left(\eta^{(0)}\right) \neq 0 \tag{30}$$

Denoting by $\eta_l{}^{(0)}$ the value of $\eta^{(0)}$ given by $f_l\left(\eta^{(0)}\right) = 0$, we find that (30) gives

$$\eta_l^{(0)} = n^2\omega_l^2 c^{-2} - p_l^2, \quad N_l^{(0)}\left(\eta_l^{(0)}\right) \equiv N_{ll} \neq 0, \tag{31}$$

and

$$N_{l'}^{(0)}\left(\eta_l^{(0)}\right) \equiv N_{l'l^{(0)}}^{(0)} = N_{ll}\delta_{l'l'} \tag{32}$$

Next equating to zero the coefficient of $\Delta$ in (28), we obtain

$$f_l\left(\eta_m{}^{(0)}\right) N_l{}^{(1)}\left(\eta_m{}^{(0)}\right) + \eta_m{}^{(1)} f_l'\left(\eta_m{}^{(0)}\right) N_l{}^{(0)}\left(\eta_m{}^{(0)}\right) = N_{l-1}{}^{(0)}\left(\eta_m{}^{(0)}\right) + N_{l+1}{}^{(0)}\left(\eta_m{}^{(0)}\right). \tag{33}$$

Putting $l = m, m+1, m-1, m+2, m-2, \ldots$ successively in (33) and making use of (32), we find

$$\eta_m^{(1)} = 0 \tag{34a}$$

$$N_{m\pm1,m}^{(1)} = \frac{N_{mm}}{f_{m\pm1}\left(\eta_m^{(0)}\right)} \tag{34b}$$

and

$$N_{m\pm j,m}{}^{(1)} = 0 \quad \text{for } j \geqslant 2. \tag{34c}$$

Similarly, by equating to zero the coefficient of $\Delta^2$ in (28), we obtain the following expressions for the correction to $\eta_m{}^{(0)}$ and for the amplitudes $N_{lm}$ up to second order:*

$$\eta_m^{(2)} = \frac{1}{f'_m\left(\eta_m^{(0)}\right)}\left[\frac{1}{f_{m+1}\left(\eta_m^{(0)}\right)} + \frac{1}{f_{m-1}\left(\eta_m^{(0)}\right)}\right] \tag{35a}$$

$$N_{m\pm2,m}{}^{(2)} = \frac{N_{mm}}{f_{m\pm1}\left(\eta_m^{(0)}\right) f_{m\pm2}\left(\eta_m^{(0)}\right)} \tag{35b}$$

and

$$N_{m\pm1,m}{}^{(2)} = 0, \quad N_{m\pm j,m}{}^{(2)} = 0 \quad \text{for } j \geqslant 3 \tag{35c}$$

It will be seen from the foregoing calculations that in the zero-order perturbation calculation only the quantities $N_{mm}(m = 0, \pm1, \ldots)$ are different from zero; in the first-order calculation $N_{mm}$ and $N_{m\pm1,m}$ differ from 0 , while, in the second order, $N_{m\pm2,m}$ are also different from zero. Likewise, by pursuing the perturbation calculation to still higher orders one obtains more and more non-diagonal amplitudes (i.e. amplitudes whose two suffixes are not equal) which are different from zero. Thesis calculations are lengthy and cannot be given here; but it may be assumed that whenever the perturbation procedure is valid one may neglect the higher order terms.

The non-diagonal amplitudes (34b) and (35b) are completely determined once the diagonal amplitudes which form the zero-order solution of (24) are known. We determine the latter from (19) and (20). First, however, it will be instructive to examine their solutions in the simple case $\Delta = 0$, for which an exact solution of equations (18)-(20) is easily obtained. In this case the only possible non-zero ampli. tudes are the diagonal amplitudes $N^{\pm}_{m,m}(m = 0, \pm1, \ldots)$. If we put all the nondiagonal amplitudes equal to zero in (19) and (20), we find that all the amplitudes $N^{\pm}_{m,m}$are identically equal to zero except $N^{\pm}_{0,0}$, which are given by*

$$N^{+}_{0,0} = (2B/\sigma_{0,0})\cos\theta\left[\left(n^2 - \sin^2\theta\right)^{1/2} - \cos\theta\right]\left(1 + \rho^2 - 2\rho\cos 2q_0 d\right)^{-1/2} e^{i\varphi}$$

$$N^{-}_{0,0} = N^{+}_{0,0}\,(g_{0,0}/h_{0,0})\,e^{2iq_0 d}$$

where

$$\rho = \left|\frac{N^{-}_{0,0}}{N^{+}_{0,0}}\right|^2, \quad \psi = \tan^{-1}\left[\frac{\rho\sin 2q_0 d}{1 - \rho\cos 2q_0 d}\right] \tag{36}$$

For normal incidence ($\theta = 0$), (11), (31), and (36) give

$$\rho_0 = \left|\frac{N^{-}_{0,0}}{N^{+}_{0,0}}\right|^2_{\theta=0} = \frac{(n-1)^2}{(n+1)^2} \tag{37}$$

0

- The formulae given here are valid only if $f_{m\pm1}\left(\eta_m{}^{(0)}\right), f_{m\pm2}\left(\eta_m{}^{(0)}\right), \ldots$ are all different from zero. If this is not so, the above perturbation procedure has to be modified to take this degeneracy into account. In perturbation calculations up to second order, the degeneracy comes into play when light is incident at angles $\theta = 0$ and $\theta = \sin^{-1}(\lambda/2\Lambda)$. However, we shall not pursue this matter further here, but will give the results for these two cases in § 12.2.5.

With the help of (36), (27), and (23), one easily obtains for the reflectivity at normal incidence on a plane parallel plate the expression

$$\left|\frac{B_0^{(r)}}{B}\right|^2 = \frac{4\rho_0 \sin^2 q_0 d}{1+\rho_0^2 - 2\rho_0 \cos 2q_0 d}$$

in agreement with §7.6(9), since $q_0 = nk$ for $\theta = 0$.
When $\Delta$ differs from zero but is still sufficiently small for the perturbation method to apply, the above solution for $\Delta = 0$ suggests that

$$N_{0,0} \gg N_{\pm1,\pm1} \gg N_{\pm2,\pm2} \cdots$$

We may, therefore, solve equations (19) and (20) for the diagonal amplitudes by successive approximations. Moreover, since for normal or near normal incidence

( $\theta$ is at most about 3° in the ultrasonic diffraction experiments) the ratio of a given $N^-$ to the corresponding $N^+$will be small (cf. (37)), we may neglect the $N^-$altogether and determine the $N^+$from equations (19) alone.* It will be recalled that a similar approximation is also implicit in the use of the boundary conditions § 12.1 (15).

The expressions for the diagonal amplitudes $N^+_{0,0}, N^\pm_{\pm1,\pm1}$, and $N^\pm_{\pm2,\pm2}$may now be obtained by using (34b), (34c), (35b), and (35c) in (19). Further, the expressions for the intensities of the first and second order lines in the transmitted spectrum can be easily written down with the help of (22). These expressions are given in the next section.

## 9.16.5 12.2.5 Expressions for the intensities of the first and second order lines for some special cases

(a) $\delta/\xi \ll 1$ and $\xi$ large compared to unity

This is the case considered in detail in the previous section. The intensities $I_{\pm1}$ and $I_{\pm2}$ of the first and second order lines are respectively

$$I_{\pm1} = |B_{\pm1}|^2 = \frac{1}{4}B^2\gamma^2\delta^2 \frac{\sin^2\left[\beta d\left(\xi \pm \frac{1}{2}\right)\right]}{\left(\xi \pm \frac{1}{2}\right)^2} \tag{38a}$$

and

$$I_{\pm2} = |B_{\pm2}|^2 = \tfrac{1}{64}B^2\gamma^4\delta^4 \left\{\frac{1}{2(\xi\pm1)\left(\xi\pm\frac{1}{2}\right)\left(\xi\pm\frac{3}{2}\right)}\right\} \left[\frac{\sin^2\left[\beta d\left(\xi\pm\frac{1}{2}\right)\right]}{\left(\xi\pm\frac{1}{2}\right)} + \frac{\sin^2\left[\beta d\left(\xi\pm\frac{3}{2}\right)\right]}{\left(\xi\pm\frac{3}{2}\right)} - \frac{\sin^2[2\beta d(\xi\pm1)]}{2(\xi\pm1)}\right]. \tag{38b}$$

These equations are to be understood with either all upper signs or all lower signs.
We quote without proof the expressions for the intensities for two other cases considered by Bhatia and Noble (loc. cit.)
(b) $\xi \sim \frac{1}{2}, \delta \ll 1$

0

- There are some misprints in the expressions for $N^+_{0,0}$and $N^-_{0,0}$given by Bhatia and Noble (loc. cit.); these have been corrected here.

$$I_0 = \frac{1}{4}\frac{B^2\delta^2\gamma^2}{\left(\xi-\frac{1}{2}\right)^2+\frac{1}{4}\delta^2\gamma^2}\left[\frac{\left(\xi-\frac{1}{2}\right)^2}{\frac{1}{4}\delta^2\gamma^2}+\cos^2\left[\beta d\left\{\left(\xi-\frac{1}{2}\right)^2+\frac{1}{4}\delta^2\gamma^2\right\}^{1/2}\right]\right] \tag{39a}$$

$$I_{-1} = \frac{1}{4}\frac{B^2\delta^2\gamma^2}{\left(\xi-\frac{1}{2}\right)^2+\frac{1}{4}\delta^2\gamma^2}\sin^2\left[\beta d\left\{\left(\xi-\frac{1}{2}\right)^2+\frac{1}{4}\delta^2\gamma^2\right\}^{1/2}\right] \tag{39b}$$

The expressions for $I_1$ and $I_{-2}$ are more complicated. For $\xi = \frac{1}{2}$, however, they take the simple form

$$I_1 = \frac{1}{16}B^2\delta^2\gamma^2\left[-\sin^2\left(\frac{\pi d\Delta\gamma}{2\lambda n}\right)+2\sum_{+,-}\sin^2\left\{\beta d\left(1\pm\frac{1}{4}\delta\gamma\right)\right\}\right] \tag{39c}$$

and

$$I_{-2} = \frac{1}{16}B^2\delta^2\gamma^2\sin^2\left(\frac{\pi d\Delta\gamma}{2\lambda n}\right). \tag{39d}$$

(c) Normal incidence ( $\xi = 0$ ), $\delta \ll 1$

$$I_1 = I_{-1} = B^2\delta^2\gamma^2\sin^2\left[\frac{1}{2}\beta d\left(1+\frac{1}{8}\delta^2\gamma^2\right)\right], \tag{40a}$$

$$I_2 = I_{-2} = \frac{1}{48}B^2\delta^4\gamma^4\left[-\frac{1}{4}\sin^2 2\beta d+\sin^2\left\{\frac{1}{2}\beta d\left(1+\frac{1}{8}\delta^2\gamma^2\right)\right\}\right.$$
$$\left.+\frac{1}{3}\sin^2\left\{\frac{3}{2}\beta d\left(1-\frac{1}{24}\delta^2\gamma^2\right)\right\}\right]. \tag{40b}$$

- For a detailed discussion on these points see Bhatia and Noble (loc. cit.). In particular, it is shown there that the effect of the $N^-$on the amplitudes ${B_1}^{(r)}$ of the reflected spectrum is generally not negligible.

If we neglect the quantity $\delta^2\gamma^2$ occurring in the argument of the sines in (40), the resulting expressions may also be obtained from (38) by putting $\xi = 0$ in the latter.

As already mentioned, Brillouin* and David, † and also Rytov, ‡ derived the expressions (38) for the intensities of the first and second order lines in the manner outlined in § 12.1. Aggarwal§ has derived the expressions (38) from the Raman and Nath differential equations § 12.1 (21). Phariseau || has shown that expressions (39) for the intensities of the first and second order lines when $\xi \sim \frac{1}{2}$ can also be derived from equations (21). [f] This is, of course, as it should be, since the method based on Maxwell's differential equations and the integral equation method of the present section are equivalent.

## 9.16.6 12.2.6 Some qualitative results

It is clear from the expressions for the intensities given in the previous section that, for values of $\delta$ and $\xi$ such that either (a) $\delta \ll 1$ or (b) $\delta/\xi \ll 1$, only the first few lower orders will appear on each side of the transmitted beam and their intensities will diminish rapidly with increasing orders. When, however, neither (a) nor (b) is satisfied, i.e. when $\delta$ and $\delta/\xi$ are both large compared to unity, many more orders will in general appear. For this case the solution of the equations (18)-(20), and hence the calculation of the intensities of the various orders, is

more difficult. By examining the conditions for the validity of the perturbation method of § 12.2.4, it may be shown** that in solving (18)-(20), one may consider only those amplitudes $N_{lm}$ as non-zero for which both the suffixes $l$ and $m$ lie between the numbers - $M_1$ and $M_2$ defined approximately by ( $0 \leqslant \xi < \delta^{2/3}$ )

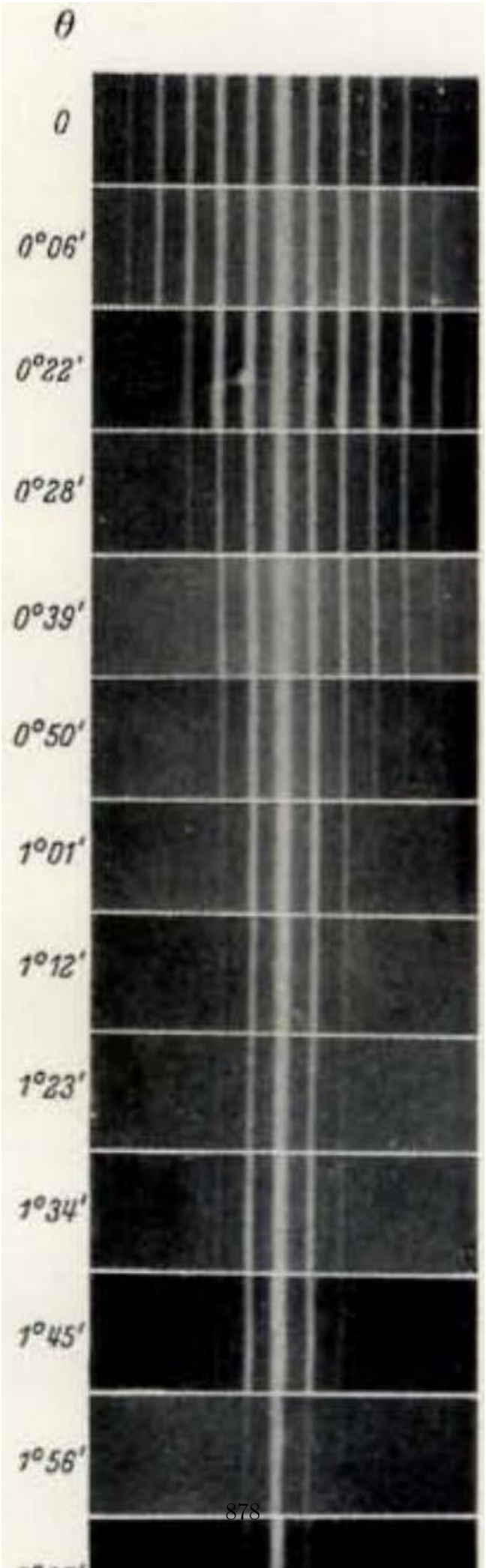
$$\left.\begin{aligned} M_1 &\sim \xi + \delta^{2/3} + 1 \\ M_2 &\sim -\xi + \delta^{2/3} + 1 \end{aligned}\right\} \tag{41}$$

Since, in general, $\Delta$ cannot be increased much beyond $10^{-4}$ and the maximum value of $\lambda/\Lambda$ for which diffraction phenomena can be observed is restricted by practical limitations of resolving power, etc., the maximum possible value of $\delta\,(=\Delta\Lambda^2/\lambda^2)$ is about 100. Hence, even under extreme experimental conditions, one has only to solve at most about 20 linear simultaneous equations from each of the infinite sets (18), (19), (20). But even with this simplification the calculations are necessarily tedious and have not been performed.

Qualitatively, the numbers $M_1$ and $M_2$ also represent the number of orders likely to appear on the two sides of the direct transmitted beam. According to (41), the number of orders appearing on the two sides of the direct transmitted beam should become different as $\xi = (\Lambda \sin\theta)/\lambda$ increases from zero, more lines appearing on the side which can be reached by light reflected from the wave-fronts of the ultrasonic wave. Parthasarathy †† has studied experimentally the diffracted spectrum as a function of the angle of incidence $\theta$, and we reproduce in Fig. 12.3 a plate from his paper. In this experiment $(\lambda/\Lambda) = 3 \times 10^{-3}$; assuming $\Delta \sim 10^{-4}$, we have from (41) for normal incidence $M_1 = M_2 = 5$. Table XXV gives the number of lines actually observed

0

- L. Brillouin, La Diffraction de la Lumire par des Ultrasons (Paris, Hermann, 1933).
  † E. David, Phys. Z., 38 (1937), 587.
  ‡ S. Rytov, Diffraction de la Lumire par les Ultra-sons (Paris, Hermann, 1938).
  § R. R. Aggarwal, Proc. Ind. Acad. Sci., A, 31 (1950), 417.
  || P. Phariseau, Proc. Ind. Acad. Sci., A, 44 (1956), 165.
  II A discussion of the experimental results in relations to the various expressions for the intensities given in this section may be found in the papers by Bhatia and Noble (loc, cit.).
  ** Bhatia and Noble (loc. cit.).
  †† S. Parthasarathy, Proc. Ind. Acad. Sci., A, 3 (1936), 442.


θ
0
0°06'
0°22'
0°28'
0°39'
0°50'
1°01'
1°12'
1°23'
1°34'
1°45'
1°56'

to appear on each side of the direct transmitted beam at different angles of incidence. Within brackets are the corresponding theoretical numbers as given by (41). (Of course, once $\delta/\xi$ becomes much less than unity one or two orders only will appear on each side of the direct transmitted beam).

Similar experimental results have also been obtained by Nomoto.* His curves of $M_1$ and $M_2$ against $\theta$, though in qualitative agreement with (41), show also a slight periodic variation with $\xi$ in the range $0 \leqslant \xi < \delta^{2/3}$. Equations (41) are based on too crude a consideration to explain this latter feature.

Table XXV Diffraction of light by ultrasonic waves: number of orders observed, and (in brackets) the number predicted by theory, for different values of the angle of incidence $\theta$

$$\left(\lambda/\Lambda = 3 \times 10^{-3}, \Delta \sim 10^{-4}\right)$$

| $\theta$ | 0 | **0°06′** | 0°22′ | **0°39′** | 1°01′ | **1°23′** | 1°45′ | 2°07′ |
|---|---|---|---|---|---|---|---|---|
| $\xi$ | 0 | 0.6 | 2 | 4 | 6 | 8 | 10 | 13 |
| $M_1$ | 5 (5) | 5 (6) | 6 (7) | 6 (9) | 3 | 2 | 1 | 1 |
| $M_2$ | 5 (5) | 5 (4) | 3 (3) | 2 (2) | 2 | 1 | 1 | 1 |

## 9.16.7 12.2.7 The Raman-Nath approximation

Finally it will be shown that the Bessel function expressions for the intensities obtained by Raman and Nath may also be deduced from the solution of (18) to (20). Neglecting the small variation of the frequencies $\omega_l$ with $l$ and remembering that ${q_m}^2 \simeq (n^2{\omega_m}^2c^{-2} - {p_m}^2)$, we can take, to a good approximation, ${p_l}^2 + q^2 + 2{\omega_l}^2c^{-2} \simeq k^2\,(n^2+2)$ in the denominator of (25). Hence (24) may be written in the form

$$\frac{q^2 - n^2k^2 + k^2\sin^2\theta}{\frac{1}{2}\Delta\gamma k^2}N_l(q) - N_{l+1}(q) - N_{l-1}(q) = -\frac{l^2 + 2l\xi}{2\gamma\delta}N_l(q)$$
$$(l = 0, \pm1, \pm2, \ldots). \tag{42}$$

Now the Bessel function expressions for the intensities were obtained by Raman and Nath by neglecting the terms $[(l^2 + 2l\xi)\,/\delta] \times U_l$ in their equations §12.1(21). To the same approximation we may also neglect the right-hand side of (42). (42) may then be written as

$$\frac{q^2 - b^2}{\mu^2}N_l(q) = N_{l+1}(q) + N_{l-1}(q) \quad (l = 0, \pm1, \ldots) \tag{43}$$

where

$$b^2 = k^2\left(n^2 - \sin^2\theta\right) \qquad \text{and} \qquad \mu^2 = \frac{1}{2}\Delta\gamma k^2. \tag{44}$$

To solve (43)† let us assume $N_l(q)$ to be of the form

$$N_l(q) = Ne^{2\pi ilm/M}, \tag{45}$$

0

- O. Nomoto, Proc. Phys.-Math. Soc. Japan, 24 (1942), 380, 613.
† Equations (43) are similar to the relations which determine the normal modes of vibration of a linear chin of atoms (cf, M, Born and Th. v. Krmn, Phys., Z., 13 (1912), 297).

where $M$ is some very large integer and $m$ is an integer such that $0 \leqslant m < M$. (As will be seen presently the final results are independent of $M$.) Substituting from (45) in (43), we have for the permissible values of $q^2$,

$$q_m^2 = b^2 + 2\mu^2 \cos(2\pi m/M), \quad 0 \leqslant m < M \tag{46}$$

or remembering that $\mu^2 \ll b^2$,

$$\pm |q_m| = \pm \left[b + \left(\mu^2/b\right) \cos(2\pi m/M)\right]. \tag{47}$$

Thus for each $m$ there are two values of $q$, namely $\pm |q_m|$; the corresponding amplitudes are given by

$$N_l (\pm q_m) \equiv N_{lm}^{\pm} = N^{\pm} e^{2\pi i l m/M} \tag{48}$$

The two constants $N^+$and $N^-$in (48) are now to be determined from (19) and (20). As in § 12.2.4, we shall neglect the amplitudes $N_{lm}^-$and determine the constant $N^+$ from (19) alone. Noting that for not too large a value of $l, [\sigma_l(q)/g_l(q)] \sim \left[\sigma_0 \left({q_0}^{(0)}\right)/g_0 \left({q_0}^{(0)}\right)\right]$, we may verify that the equations (19) are identically satisfied by taking

$$N^+ = \frac{1}{M} 2Bc^2\omega^{-2} \left[\omega^2 c^{-2} - {p_0}^2\right]^{1/2} \left(g_{0,0}/\sigma_{0,0}\right) \tag{49}$$

in (48). Substituting (48) and (49) in (22) and putting $({\omega_l}^2 c^{-2} - {p_l}^2) = (\omega^2 c^{-2} - {p_0}^2)$, we obtain for the amplitude $B_l$ of the $l$ th order diffracted wave

$$B_l \simeq \frac{B}{M} \sum_{m=0}^{M-1} e^{i\left[2\pi l m/M + \left(\mu^2 d/b\right) \cos(2\pi m/M)\right]} \tag{50}$$

Here a phase factor independent of $m$ has been omitted. Since by hypothesis $M$ is a very large integer, we may replace the series in (50) by an integral. Putting $2\pi m/M = \psi'$ and $d\psi' = 2\pi/M$, we may write (50) as

$$B_l \simeq \frac{B}{2\pi} \int_0^{2\pi} e^{i\left[l\psi' + \left(\mu^2 d/b\right) \cos\psi'\right]} d\psi' \tag{51}$$

which is independent of $M$. By splitting the integral in (51) into two parts (i) from 0 to $\frac{3}{2}\pi$ and (ii) from $\frac{3}{2}\pi$ to $2\pi$, and putting in them $\psi' = \frac{1}{2}\pi - \psi$ and $\psi' = \frac{5}{2}\pi - \psi$ respectively, we obtain*

$$\begin{aligned} B_l &\simeq \frac{B}{2\pi} \int_{-\pi}^{\pi} e^{i\left[\frac{1}{l} l\pi - l\psi + \left(\mu^2 d/b\right) \sin\psi\right]} d\psi \\ &= B e^{\frac{1}{i} ll\pi} J_l \left(\mu^2 d/b\right) \end{aligned} \tag{52}$$

Hence the intensities $I_l = |B_l|^2$ are just $B^2 {J_l}^2 (\mu^2 d/b)$. Moreover, it may be seen with the help of (44), § 12.1 (5) and § 12.1 (8) that the argument ( $\mu^2 d/b$ ) of the Bessel functions $J_l$ is the same as that occurring in the expressions of Raman and Nath, given on p. 598.

0

- H. and B. S. Jeffreys, Methods of Mathematical Physics (Cambridge University Press, 1946), p. 547.

## 9.17 Optics of Metals

So far we have been concerned with the propagation of light in non-conducting, isotropic media. We now turn our attention to optics of conducting media, more particularly to metals. An ordinary piece of metal is a crystalline aggregate, consisting of small crystals of random orientation. Single crystals of appreciable size are rare, but can be produced artificially; their optical properties will be studied in Chapter XIV. A mixture of randomly oriented crystallites behaves evidently as an isotropic substance, and as the theory of light propagation in a conducting isotropic medium is much simpler than in a crystal, we shall consider it here in some detail.

According to § 1.1, conductivity is connected with the appearance of Joule heat. This is an irreversible phenomenon, in which the electromagnetic energy is destroyed, or more precisely transformed into heat, and in consequence an electromagnetic wave in a conductor is attenuated. In metals, on account of their very high conductivity, this effect is so large that they are practically opaque. In spite of this, metals play an important part in optics. Strong absorption is accompanied by high reflectivity, so that metallic surfaces act as excellent mirrors. Because of the partial penetration of light into a metal, it is possible to obtain information about the absorption constants and the mechanism of absorption from observations of the reflected light, even though the depth of penetration is small.

We shall first consider the purely formal results arising from the existence of conductivity, and then briefly discuss a simple, somewhat idealized, physical model for this process, based on the classical theory of the electron. This model accounts only roughly for some of the observed effects; a more precise model can only be obtained with the help of quantum mechanics and is thus outside the scope of this book. The formal theory will be illustrated by applications to two problems of practical interest: the optics of stratified media containing an absorbing element, and the diffraction of light by a metallic sphere.

A particularly attractive mathematical feature of the theory is that the existence of conductivity may be taken into account simply by introducing a complex dielectric constant (or complex index of refraction), instead of a real one. In metals the imaginary part is preponderant.

### 9.17.1 13.1. Wave Propagation in a Conductor

Consider a homogeneous isotropic medium of dielectric constant $\varepsilon$, permeability $\mu$, and conductivity $\sigma$. Using the material equations § 1.1 (9)-(11), viz.: $j = \sigma E$, $D = \varepsilon E, B = \mu H$, Maxwell's equations take the form

$$\operatorname{curl} H - \frac{\varepsilon}{c}\dot{E} = \frac{4\pi}{c}\sigma E, \tag{1}$$

$$\operatorname{curl} E + \frac{\mu}{c}\dot{H} = 0, \tag{2}$$

$$\operatorname{div} E = \frac{4\pi}{\varepsilon}\rho, \tag{3}$$

$$\operatorname{div} H = 0. \tag{4}$$

It is easy to see that for an electromagnetic disturbance, incident from outside on the conductor, one may replace (3) by $\operatorname{div} E = 0$. For if we take the divergence of (1) and use (3) we obtain

$$-\frac{\varepsilon}{c}\operatorname{div}\dot{E} = \frac{4\pi\sigma}{c}\frac{4\pi}{\varepsilon}\rho$$

Also, differentiation of (3) with respect to time gives

$$\operatorname{div} \dot{E} = \frac{4\pi}{\varepsilon}\dot{\rho}.$$

Eliminating div $\dot{E}$ between the last two equations one obtains

$$\dot{\rho} + \frac{4\pi\sigma}{\varepsilon}\rho = 0, \tag{5}$$

giving on integration

$$\rho = \rho_0 e^{-\frac{t}{\tau}} \quad \text{where } \tau = \frac{\varepsilon}{4\pi\sigma}. \tag{6}$$

Any electric charge density $\rho$ is thus seen to fall off exponentially with time. The relaxation time $\tau$ is exceedingly small for any medium having appreciable conductivity. For metals, this time is very much less than the periodic time of vibration of the wave; for example for light in the orange region of the visible spectrum the periodic time is $2 \times 10^{-15}$sec whereas for copper $\tau/\varepsilon$ is of the order of $2 \times 10^{-19}$sec. For any value of $\varepsilon$ which may reasonably be expected, $\tau$ is so small compared with the periodic time of the light wave that we may assume that $\rho$ in a metal is always sensibly zero. Equation (3) may therefore be written as

$$\operatorname{div} E = 0. \tag{7}$$

From (1) and (2) it follows by elimination of $H$ and the use of (7) that $E$ satisfies the wave equation

$$\nabla^2 \boldsymbol{E} = \frac{\mu\varepsilon}{c^2}\ddot{\boldsymbol{E}} + \frac{4\pi\mu\sigma}{c^2}\dot{\boldsymbol{E}}. \tag{8}$$

The term in $\dot{E}$ implies that the wave is damped, i.e. it suffers a progressive attenuation as it is propagated through the medium.

If the field is strictly monochromatic, and of angular frequency $\omega$, i.e. if $E$ and $H$ are of the form $E = E_0 e^{-i\omega t}$, we have $\partial/\partial t \equiv -i\omega$ so that equations (1) and (2) may be re-written as

$$\operatorname{curl} \boldsymbol{H} + \frac{i\omega}{c}\left(\varepsilon + i\frac{4\pi\sigma}{\omega}\right)\boldsymbol{E} = 0 \tag{9}$$

$$\operatorname{curl} \boldsymbol{E} - \frac{i\omega\mu}{c}\boldsymbol{H} = 0 \tag{10}$$

and (8) becomes

$$\nabla^2 E + \hat{k}^2 E = 0, \tag{11}$$

where

$$\hat{k}^2 = \frac{\omega^2\mu}{c^2}\left(\varepsilon + i\frac{4\pi\sigma}{\omega}\right) \tag{12}$$

These equations are formally identical with the corresponding equations for nonconducting media if in the latter the real dielectric constant $\varepsilon$ is replaced by

$$\hat{\varepsilon} = \varepsilon + i\frac{4\pi\sigma}{\omega} \tag{13}$$

The analogy with non-conducting media becomes closer still if, in addition to the complex wave number $\hat{k}$ and the complex dielectric constant $\hat{\varepsilon}$, we also introduce a complex phase

velocity $\hat{v}$ and a complex refractive index $\hat{n}$ which, in analogy with § 1.2 (8), § 1.2 (12), and § 1.3 (21), are defined by

$$\hat{v} = \frac{c}{\sqrt{\mu\hat{\varepsilon}}}, \quad \hat{n} = \frac{c}{\hat{v}} = \sqrt{\mu\hat{\varepsilon}} = \frac{c}{\omega}\hat{k}. \tag{14}$$

We set

$$\hat{n} = n(1 + i\kappa), \tag{15}$$

where $n$ and $\kappa$ are real, and we call $\kappa$ the attenuation index.* The quantities $n$ and $\kappa$ may easily be expressed in terms of the material constants $\varepsilon, \mu$, and $\sigma$. Squaring (15) we have

$$\hat{n}^2 = n^2\left(1 + 2i\kappa - \kappa^2\right).$$

Also, from (14) and (13),

$$\hat{n}^2 = \mu\hat{\varepsilon} = \mu\left(\varepsilon + i\frac{4\pi\sigma}{\omega}\right).$$

Equating the real and imaginary parts in these two expressions for $\hat{n}^2$ we obtain

$$n^2\left(1 - \kappa^2\right) = \mu\varepsilon \tag{16a}$$

$$n^2\kappa = \frac{2\pi\mu\sigma}{\omega} = \frac{\mu\sigma}{v} \tag{16b}$$

From these equations it follows that

$$n^2 = \frac{1}{2}\left\{\sqrt{\mu^2\varepsilon^2 + \frac{4\mu^2\sigma^2}{\nu^2}} + \mu\varepsilon\right\}, \tag{17a}$$

$$n^2\kappa^2 = \frac{1}{2}\left\{\sqrt{\mu^2\varepsilon^2 + \frac{4\mu^2\sigma^2}{\nu^2}} - \mu\varepsilon\right\}. \tag{17b}$$

The positive sign of the square roots is taken here, since $n$ and $n\kappa$ are real, and consequently $n^2$ and $n^2\kappa^2$ must be positive.

Equation (11) is formally identical with the wave equation for a non-conducting medium, but the wave number is now complex. The simplest solution is that of a plane, time-harmonic wave

$$\boldsymbol{E} = \boldsymbol{E}_0 e^{i[\hat{k}(r.s) - \omega t]} \tag{18}$$

If, in accordance with (14) and (15), we substitute for $\hat{k}$ from the relation $\hat{k} = \omega\hat{n}/c = \omega n(1 + i\kappa)/c$, (18) becomes

$$E = E_0 e^{-\frac{\omega}{c}nk(r.s)} e^{i\omega\left[\frac{n}{c}(r.s) - t\right]}.$$

The real part of this expression, viz.:

$$E = E_0 e^{-\frac{\omega}{c}n\kappa(r.s)} \cos\left\{\omega\left[\frac{n}{c}(r.s) - t\right]\right\}, \tag{19}$$

0

- The term "extinction coefficient" is also used.

which represents the electric vector, is a plane wave with wavelength $\lambda = 2\pi c/\omega n$ and with attenuation given by the exponential term. Since the energy density $w$ of the wave is proportional to the time average of $E^2$, it follows that $w$ decreases in accordance with the relation

$$w = w_0 e^{-\chi(r.s)}, \tag{20}$$

where

$$\chi = \frac{2\omega}{c} n\kappa = \frac{4\pi v}{c} n\kappa = \frac{4\pi}{\lambda_0} n\kappa = \frac{4\pi}{\lambda}\kappa \tag{21}$$

$\lambda_0$ being the wavelength in vacuum and $\lambda$ the wavelength in the medium. The constant $\chi$ is called the absorption coefficient.

The energy density falls to $1/e$ of its value after the wave has advanced a distance $d$, where

$$d = \frac{1}{\chi} = \frac{\lambda_0}{4\pi n\kappa} = \frac{\lambda}{4\pi\kappa} \tag{22}$$

This quantity is usually a very small fraction of the wavelength (see Table XXVI).* Returning to equations (17) we see that, when $\sigma = 0$, the first equation correctly reduces to Maxwell's relation § 1.2 (14) $n^2 = \mu\varepsilon$, and the second gives $\kappa = 0$. For metals $\sigma \neq 0$ and is in fact so large that in (17) $\varepsilon$ may be neglected in comparison with $2\sigma/v$. To get an idea of the orders of magnitude involved let it be remarked that for most metals $\sigma$ is of the order of $10^{17}\text{sec}^{-1}$, so that with $\lambda = 5500\text{Å}$ (corresponding to $\nu = 5 \cdot 5 \cdot 10^{14}\text{sec}^{-1}$ ), $\sigma/\nu \sim 200$. The dielectric constant $\varepsilon$ of a metal cannot be measured directly, but as we shall see it can be deduced from optical experiments. However, as the mechanism of electric polarization in metals is not fundamentally different from that of a dielectric, it may be assumed that $\varepsilon$ is of the same order of magnitude. Hence, provided the wavelength is not too short, one may suppose that

$$\frac{\mu\sigma}{\nu} = n^2\kappa \gg \mu\varepsilon. \tag{23}$$

Equations (17) and (22) now reduce to

$$n \sim n\kappa = \sqrt{\frac{\mu\sigma}{v}}. \tag{24}$$

$$d \sim \frac{\lambda_0}{4\pi}\sqrt{\frac{v}{\mu\sigma}} = \frac{1}{4\pi}\sqrt{\frac{c\lambda_0}{\mu\sigma}}. \tag{25}$$

A perfect conductor is characterized by infinitely large conductivity ( $\sigma \to \infty$ ). Since according to (16), $\varepsilon/\sigma = (1-\kappa^2)/\nu\kappa$, we have in this limiting case $\kappa^2 \to 1$, or by (16a), $n \to \infty$. Such a conductor would not permit the penetration of an electromagnetic wave to any depth at all and would reflect all the incident light (cf. §13.2 below).

Whilst the refractive index of transparent substances may easily be measured from the angle of refraction, such measurements are extremely difficult to carry out for metals, because a specimen of the metal which transmits any appreciable fraction of incident light has to be exceedingly thin. Nevertheless Kundt* succeeded in constructing metal prisms that enabled

0

- This phenomenon of penetration to a depth that is a small fraction of the wavelength, is well known in the conduction of alternating currents and is known to engineers as the "skin effect".

Table XXVI
The "penetration depth" $d$ for copper for radiation of wavelengths in the familiar regions of the spectrum ( $\sigma \sim 5 \cdot 14.10\hat{}\{\}\{17\}\ \mathrm{sec}\hat{}\{\}\{-1\},\ \mu=1$ )

| Radiation | Ultra-violet | Infra-red | Microwaves | Long |
|---|---|---|---|---|
| $\lambda_0 d$ | $1000\text{Å} = 10^{-5}$ cm$6 \cdot 2.10^{-8}$ cm | $10\mu = 10^{-3}$ cm$6 \cdot 2 \cdot 10^{-7}$ cm | 10 cm $6 \cdot 2 \cdot 10^{-5}$ cm | 1000 |

direct measurements of the real and imaginary parts of the complex refractive index to be made. Usually, however, the optical constants of metals are determined by means of katoptric rather than dioptric experiments, i.e. by studying the changes which light undergoes on reflection from a metal, rather than by means of measurements on the light transmitted through it.

## 9.17.2 13.2. Refraction and Reflection at a Metal Surface

We have seen that the basic equations relating to the propagation of a plane timeharmonic wave in a conducting medium differ from those relating to propagation in a transparent dielectric only in that the real constants $\varepsilon$ and $k$ are replaced by complex constants $\hat{\varepsilon}$ and $\hat{k}$. It follows that the formulae derived in Chapter I, as far as they involve only linear relations between the components of the field vectors of plane monochromatic waves, apply also in the present case. In particular, the boundary conditions for the propagation of a wave across a surface of discontinuity and hence also the formulae of § 1.5 relating to refraction and reflection remain valid.

Consider first the propagation of a plane wave from a dielectric into a conductor, both media being assumed to be of infinite extent, the surface of contact between them being the plane $z = 0$. By analogy with § 1.5 (8) the law of refraction is

$$\sin\theta_t = \frac{1}{\hat{n}} \cdot \sin\theta_i. \tag{1}$$

Since $\hat{n}$ is complex, so is $\theta_t$, and this quantity therefore no longer has the simple significance of an angle of refraction.

Let the plane of incidence be the $xz$-plane. The space-dependent part of the phase of the wave in the conductor is given by $\hat{k}\left(r.s^{(t)}\right)$ where (cf. § 1.5 (4))

$$s_x^{(t)} = \sin\theta_t, \quad s_y^{(t)} = 0, \quad s_z^{(t)} = \cos\theta_t. \tag{2}$$

From (1) and (2) and § 13.1 (15)

$$s_x{}^{(t)} = \sin\theta_t = \frac{\sin\theta_i}{n(1+i\kappa)} = \frac{1-i\kappa}{n\left(1+\kappa^2\right)}\sin\theta_i \tag{3a}$$

$$s_z{}^{(t)} = \cos\theta_t = \sqrt{1-\sin^2\theta_t}$$

$$= \sqrt{1 - \frac{\left(1-\kappa^2\right)}{n^2\left(1+\kappa^2\right)^2}\sin^2\theta_i + i\frac{2\kappa}{n^2\left(1+\kappa^2\right)^2}\sin^2\theta_i} \tag{3b}$$

It is convenient to express $s_z{}^{(t)}$ in the form

0

- A. Kundt, Ann. d. Physik, 34 (1888), 469.

$$s_z^{(t)} = \cos\theta_t = qe^{i\gamma} \tag{4}$$

( $q, \gamma$ real). Expressions for $q$ and $\gamma$ in terms of $n, \kappa$ and $\sin\theta_i$ are immediately obtained on squaring (3b) and (4) and equating real and imaginary parts. This gives

$$\left.\begin{aligned} q^2\cos 2\gamma &= 1 - \tfrac{1-\kappa^2}{n^2(1+\kappa^2)^2}\sin^2\theta_i \\ q^2\sin 2\gamma &= \tfrac{2\kappa}{n^2(1+\kappa^2)^2}\sin^2\theta_i \end{aligned}\right\} \tag{5}$$

It follows that

$$\begin{aligned} \hat{k}\left(r.s^{(t)}\right) &= \frac{\omega}{c}n(1+i\kappa)\left(xs_x^{(t)} + zs_z^{(t)}\right) \\ &= \frac{\omega}{c}n(1+i\kappa)\left[\frac{x(1-i\kappa)}{n\left(1+\kappa^2\right)}\sin\theta_i + z(q\cos\gamma + iq\sin\gamma)\right] \\ &= \frac{\omega}{c}\left[x\sin\theta_i + znq(\cos\gamma - \kappa\sin\gamma) + inzq(\kappa\cos\gamma + \sin\gamma)\right]. \end{aligned} \tag{6}$$

We see that the surfaces of constant amplitude are given by

$$z = \text{ constant} \tag{7}$$

and are, therefore, planes parallel to the boundary. The surfaces of constant real phase are given by

$$x\sin\theta_i + znq(\cos\gamma - \kappa\sin\gamma) = \text{ constant }, \tag{8}$$

and are planes whose normals make an angle $\theta_t'$ with the normal to the boundary, where

$$\left.\begin{aligned} \cos\theta_t' &= \frac{nq(\cos\gamma - \kappa\sin\gamma)}{\sqrt{\sin^2\theta_i + n^2q^2(\cos\gamma - \kappa\sin\gamma)^2}} \\ \sin\theta_t' &= \frac{\sin\theta_i}{\sqrt{\sin^2\theta_i + n^2q^2(\cos\gamma - \kappa\sin\gamma)^2}} \end{aligned}\right\} \tag{9}$$

Since the surfaces of constant amplitude and the surfaces of constant phase do not in general coincide with each other, the wave in the metal is an inhomogeneous wave.

If we denote the square root in (9) by $n'$, the equation for $\sin\theta_t'$ may be written in the form $\sin\theta' = \sin\theta_i/n'$, i.e. it has the form of Snell's law. However, $n'$ depends
now not only on the quantities that specify the medium, but also on the angle of incidence $\theta_i$.

We may also derive expressions for the amplitude and the phase of the refracted and reflected waves by substituting for $\theta_t$ the complex value given by (1) in the Fresnel formulae (§ 1.5.2). The explicit expressions will be given in § 13.4.1 in connection with the theory of stratified conducting media. Here we shall consider how the optical constants of the metal may be deduced from observation of the reflected wave.

Since we assumed that the first medium is a dielectric, the reflected wave is an ordinary (homogeneous) wave with a real phase factor. As in § 1.5 (21a) the amplitude components $A_\parallel, A_\perp$ of the incident wave and the corresponding components $R_\parallel, R_\perp$ of the reflected wave are related by

$$\left.\begin{aligned} R_\parallel &= \tfrac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)}A_\parallel \\ R_\perp &= -\tfrac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}A_\perp \end{aligned}\right\} \tag{10}$$

Since $\theta_t$ is now complex, so are the ratios $R_\parallel/A_\parallel$and $R_\perp/A_\perp$, i.e. characteristic phase changes occur on reflection; thus incident linearly polarized light will in general become elliptically

polarized on reflection at the metal surface. Let $\phi_\parallel$and $\phi_\perp$ be the phase changes, and $\rho_\parallel$and $\rho_\perp$ the absolute values of the reflection coefficients, i.e.

$$r_\parallel = \frac{R_\parallel}{A_\parallel} = \rho_\parallel e^{i\phi_\parallel}, \quad r_\perp = \frac{R_\perp}{A_\perp} = \rho_\perp e^{i\phi_\perp}. \tag{11}$$

Suppose that the incident light is linearly polarized in the azimuth $\alpha_i$ i.e.

$$\tan\alpha_i = \frac{A_\perp}{A_\parallel} \tag{12}$$

and let $\alpha_r$ be the azimuthal angle (generally complex) of the light that is reflected. Then*

$$\tan\alpha_r = \frac{R_\perp}{R_i} = -\frac{\cos(\theta_i - \theta_t)}{\cos(\theta_i + \theta_t)} \tan\alpha_i = Pe^{-i\Delta}\tan\alpha_i \tag{13}$$

where

$$P = \frac{\rho_\perp}{\rho_u}, \quad \Delta = \phi_\parallel - \phi_\perp \tag{14}$$

We note that $\alpha_r$ is real in the following two cases:
(1) For normal incidence ($\theta_i = 0$); then $P = 1$ and $\Delta = -\pi$, so that $\tan\alpha_r = -\tan\alpha_i$.
(2) For grazing incidence ($\theta_i = \pi/2$); then $P = 1$ and $\Delta = 0$, so that $\tan\alpha_r = \tan\alpha_i$.

It should be remembered that in the case of normal incidence the directions of the incident and reflected rays are opposed; thus the negative sign implies that the

- We write $-i\Delta$ rather than $+i\Delta$ in the exponent on the right-hand side of (13) to facilitate comparison with certain results of § 1.5.
  azimuth of the linearly polarized light is unchanged in its absolute direction in space. It is also unchanged in its absolute direction when the incidence is grazing.

Between the two extreme cases just considered, there exists an angle $\bar{\theta}_i$ called principal angle of incidence which is such that $\Delta = -\pi/2$. At this angle of incidence, linearly polarized light is, in general, reflected as elliptically polarized light, but as may be seen from § 1.4 (31b) (with $\delta = \pi/2$ ), the axes of the vibration ellipse are parallel and perpendicular to the plane of incidence. If, moreover, $P\tan\alpha_i = 1$, then according to (13) $\tan\alpha_r = -i$, and the reflected light is circularly polarized.

Suppose that with linearly polarized incident light an additional phase difference $\Delta$ is introduced between $R_\parallel$and $R_\perp$ by means of a suitable compensator (cf. § 14.4.2).

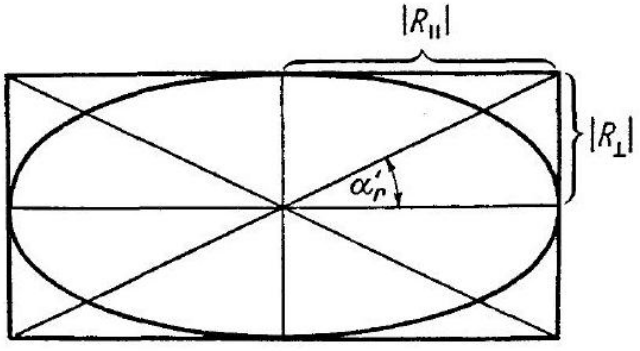



Fig. 13.1. Vibration ellipse of light reflected from a metal at the principal angle of incidence.

The total phase difference is then zero, and, according to (13) the reflected light is linearly polarized in an azimuth $\alpha'_r$ such that

$$\tan \alpha'_r = P \tan \alpha_i. \tag{15}$$

The angle $\alpha'_\tau$ is, for obvious reasons, called the angle of restored polarization, though it is usually defined only with incident light that is linearly polarized in the azimuth $\alpha_i = 45°$. The values of $\alpha'_r$ and $P$ relating to the principal angle of incidence $\theta_i = \bar{\theta}_i$ will be denoted by $\bar{\alpha}'_r$ and $\bar{P}$ respectively. If we imagine a rectangle to be circumscribed round the vibration ellipse of the (uncompensated) reflected light obtained from light that is incident at the principal angle, with its sides parallel and perpendicular to the plane of incidence, then the sides are in the ratio $\bar{P} \tan \alpha_i$ and the angle between a diagonal and the plane of incidence is $\bar{\alpha}'_r$ (see Fig. 13.1).

For the purpose of later calculations it is useful to introduce an angle $\psi$ such that

$$\tan \psi = P \tag{16}$$

the value of $\psi$ corresponding to the principal angle of incidence will be denoted by $\bar{\psi}$. Using (10) and (1) we can compute the quantities $P(= \tan \psi)$ and $\Delta$ in terms of $\theta_i$, if the constants $n$ and $\kappa$ of the metal are known. Fig. 13.2 (a) shows their dependence on $\theta_i$ in a typical case. In Fig. 13.2 (b) analogous curves relating to reflection from a transparent dielectric are displayed for comparison. The sudden discontinuity from $-\pi$ to 0 in the value of $\Delta$ which occurs when light is reflected from a transparent dielectric at the polarizing angle is absent when light is reflected from a metal surface. The sharp cusp when $\tan \psi$ becomes infinite is likewise absent, and the curve is replaced by a smooth curve with a comparatively broad maximum. The angle of incidence at which this maximum occurs is sometimes called the quasi-polarizing angle; it is nearly equal to the principal angle of incidence $\bar{\theta}_i$. It is commonly assumed that this maximum is actually at $\bar{\theta}_i$, which is almost exactly true if $n^2 (1 + \kappa^2) > 1$, as is usually the case [cf. Table XVII]. In general the two angles are, however, different; for example, in the case of silver at the ultraviolet wavelength 3280Å; the quantity $n^2 (1 + \kappa^2)$ is small; then $\bar{\theta}_i = 47.8°$ and $\psi = 31.8°$, whereas $\psi_{\max} = 29.5°$ and occurs at $\theta_i = 40°$, approximately.

Generally speaking the problem is not to find $\psi$ and $\Delta$ from known values of $n$ and $\kappa$, but to determine $n$ and $\kappa$ from experimental observations of the amplitude and phase of light reflected from the metal.

As the quantities $R_\parallel, R_\perp, \phi_\parallel, \phi_\perp, \psi$ and $\Delta$ are all functions of $\theta_i$, and of $n$ and $\kappa$, measurement of any two of these quantities for a specific value of the angle of incidence $\theta_i$ will in general permit the evaluation of $n$ and $\kappa$. Since in many experiments one determines the last two of these quantities, we shall derive the fundamental expressions for $n$ and $\kappa$ in terms of $\psi$ and $\Delta$. From (13) and (1)

$$\frac{1 - Pe^{-i\Delta}}{1 + Pe^{-i\Delta}} = -\frac{\cos \theta_i \cos \theta_t}{\sin \theta_i \sin \theta_t} = -\frac{\sqrt{\hat{n}^2 - \sin^2 \theta_i}}{\sin \theta_i \tan \theta_i}. \tag{17}$$

0

- Then $\alpha'_f$ is equal to the angle $\psi$ introduced in (16).

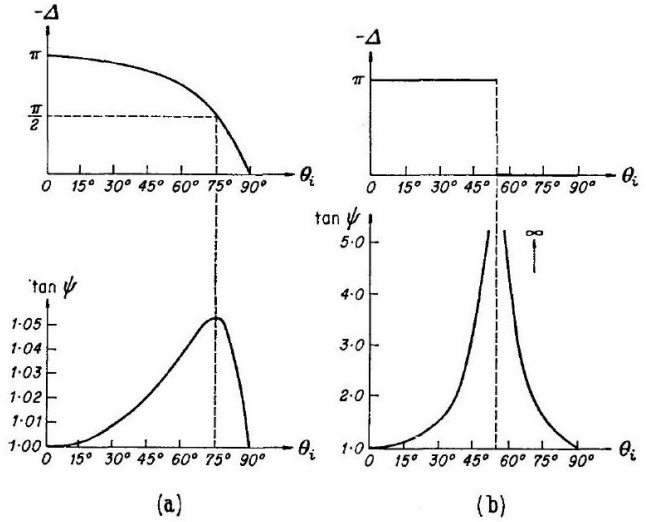



Fig. 13.2. The quantities $-\Delta=\phi_{\perp}-\phi_{\mathrm{l}}$ and $P = \tan\psi = \rho_\perp/\rho_\mathrm{l}$, which characterize the change in the state of polarization of light on reflection from a typical metal (a) and from a transparent dielectric (b).

Since $P = \tan\psi$, the left-hand side of equation (17) may be expressed in the form

$$\frac{1 - Pe^{-i\Delta}}{1 + Pe^{-i\Delta}} = \frac{1 - e^{-i\Delta}\tan\psi}{1 + e^{-i\Delta}\tan\psi} = \frac{\cos 2\psi + i\sin 2\psi\sin\Delta}{1 + \sin 2\psi\cos\Delta}. \tag{18}$$

From (17) and (18),

$$\frac{\sqrt{\hat{n}^2 - \sin^2\theta_i}}{\sin\theta_i\tan\theta_i} = -\frac{\cos 2\psi + i\sin 2\psi\sin\Delta}{1 + \sin 2\psi\cos\Delta}. \tag{19}$$

Now if, as is usually the case in the visible region,

$$n^2\left(1 + \kappa^2\right) \geqslant> 1, \tag{20}$$

$\sin^2\theta_i$ may be neglected in comparison with $\hat{n}^2$ and we obtain

$$\frac{\hat{n}}{\sin\theta_i\tan\theta_i} = \frac{n(1 + i\kappa)}{\sin\theta_i\tan\theta_i} \sim -\frac{\cos 2\psi + i\sin 2\psi\sin\Delta}{1 + \sin 2\psi\cos\Delta}. \tag{21}$$

Equating the real parts, we obtain

$$n \sim -\frac{\sin\theta_i\tan\theta_i\cos 2\psi}{1 + \sin 2\psi\cos\Delta}. \tag{22a}$$

Equating the imaginary parts and using (22a) we find that

$$\kappa \sim \tan 2\psi\sin\Delta. \tag{22b}$$

These expressions permit the calculation of the optical constants $n$ and $\kappa$ from measurements of $\psi$ and $\Delta$ at any angle of incidence. In the particular case of observation at the principal angle of incidence $\bar{\theta}_i$ one has $\Delta = -\pi/2, \psi = \bar{\psi}$ and equations (22a) and (22b) reduce to

$$n \sim -\sin\bar{\theta}_i\tan\bar{\theta}_i\cos 2\bar{\psi} \tag{23a}$$

$$\kappa \sim -\tan 2\bar{\psi}. \tag{23b}$$

Other formulae relating to $n$ and $\kappa$ are sometimes useful. Without assuming (20) we have, on squaring (19),

$$\frac{\hat{n}^2 - \sin^2\theta_i}{\sin^2\theta_i \tan^2\theta_i} = \frac{\cos^2 2\psi - \sin^2 2\psi \sin^2\Delta + i \sin 4\psi \sin\Delta}{(1 + \sin 2\psi \cos\Delta)^2} \tag{24}$$

If we substitute $\hat{n}^2 = n^2(1-\kappa^2) + 2in^2\kappa$ and equate real and imaginary parts we obtain

$$n^2\left(1-\kappa^2\right) = \sin^2\theta_i \left\{1 + \frac{\tan^2\theta_i \left(\cos^2 2\psi - \sin^2 2\psi \sin^2\Delta\right)}{(1+\sin 2\psi\cos\Delta)^2}\right\} \tag{25a}$$

$$2n^2\kappa = \frac{\sin^2\theta_i \tan^2\theta_i \sin 4\psi \sin\Delta}{(1+\sin 2\psi\cos\Delta)^2} \tag{25b}$$

In particular, at the principal angle of incidence ( $\theta = \bar{\theta}_i, \Delta = -\pi/2$ ), these equations reduce to*

$$n^2\left(1-\kappa^2\right) = \sin^2\bar{\theta}_i \left(1 + \tan^2\bar{\theta}_i \cos 4\bar{\psi}\right) \tag{26a}$$

$$2n^2\kappa = -\sin^2\bar{\theta}_i \tan^2\theta_i \sin 4\bar{\psi} \tag{26b}$$

The formulae (25) do not yield $n$ and $\kappa$ directly but in the combinations $n^2(1-\kappa^2)$ and $n^2\kappa$. On reference to § 13.1 (16) we see that these quantities have a direct physical significance. With $\mu = 1$ (as is always the case at optical wavelengths), $n^2(1-\kappa^2)$ is the dielectric constant, and $n^2\kappa^2$ the ratio of the conductivity and the frequency. From the magnitude of these quantities, and particularly from their variation with frequency, information may be obtained about the structure of the metal (cf. §13.3 below).

So far our analysis has centred round the amplitudes of the components of the reflected light, but, as we shall see shortly, useful information may also be obtained from comparison of the intensity of the reflected light with that of the incident light, especially at long wavelengths. If we consider normal incidence ( $\theta_i = 0$ ), the distinction between $R_\parallel$and $R_\perp$ disappears, the plane of incidence then being undetermined, and we may write

$$\mathscr{R} = \left|\frac{R_{\text{in}}}{A_{\text{il}}}\right|^2 = \left|\frac{R_\perp}{A_\perp}\right|^2 \tag{27}$$

If we substitute from (1) and (10) (or if we replace $n$ by $\hat{n}$ in §1.5(23) ), we obtain

$$\mathscr{R} = \left|\frac{\hat{n}-1}{\hat{n}+1}\right|^2 = \frac{n^2(1+\kappa^2)+1-2n}{n^2(1+\kappa^2)+1+2n} \tag{28}$$

The optical constants of many metals have been determined from measurements of reflected light. In Table XXVII values of the constants as found by various observers are given for a wavelength in the yellow region of the visible spectrum. The metals are arranged in order of

0

- Equations such as (23) and (26), which involve measurements only at the principal angle of incidence, are simpler than the more general expressions (22) or (25), and for this reason alone many experimenters have restricted themselves to measurements at this angle. At other angles of incidence experimental accuracy may be greater. Convenient choice of the angle of incidence is discussed in P. Drude, Ann. d. Physik, 39 (1890), 504; J. R. Collins and R. O. Bock, Rev. Sci. Instr., 14 (1943), 135; I. Simon, J. Opt. Soc. Amer., 41 (1951), 336; D. G. Avery, Proc. Phys. Soc., 65 (1952), 425; R. W. Ditchburn, J. Opt. Soc. Amer., 45 (1955), 743.

their reflectivity $\mathscr{R}$. We note that in all cases $n < n\kappa$ so that according to § 13.1 (16a) $\mu\varepsilon$ and consequently (since $\mu \sim 1$ at optical wavelengths) $\varepsilon$ is negative. At first sight it might appear that no physical significance can be attached to a negative dielectric constant. We shall see later that this is not the case and that the negative value of $\varepsilon$ can be explained from certain simple assumptions concerning the electron mechanism of conductivity. From the table it would appear that values* $n < 1$ are associated with high reflectivity but in general this is not the case.

Condensed from H. H. Landolt and R. Börnstein, Phys. Chem. Tabellen, (5 Aufl., Berlin, 1923; 1-3 Ergnzungsb, Berlin, 1927-1936).

Table XXVII
The optical constants of metals, for light of wavelength $\lambda = 5893$Å} (Sodium D lines)

| Metal | $n$ | $n\kappa$ | $\mathscr{R}$ | Observer | |
|---|---|---|---|---|---|
| Sodium, solid | 0.044 | $2 \cdot 42$ | 0.97 | Duncan | 1913 |
| Silver, massive | 0.20 | 3.44 | 0.94 | Oppitz | 1917 |
| Magnesium, massive. | 0.37 | 4.42 | 0.93 | Drude | 1890 |
| Potassium, molten | 0.084 | 1.81 | 0.92 | Nathanson | 1928 |
| Cadmium, massive | $1 \cdot 13$ | 5.01 | 0.84 | Drude | 1890 |
| Aluminium, massive. | $1 \cdot 44$ | $5 \cdot 23$ | 0.83 | Drude | 1890 |
| Tin, massive | 1.48 | 5.25 | 0.83 | Drude | 1890 |
| Gold, electrolytic | 0.47 | 2.83 | 0.82 | Meier | 1910 |
| Mercury, liquid | 1.60 | 4.80 | 0.77 | Lowery and Moore | 1932 |
| Zinc, massive | 1.93 | 4.66 | 0.75 | Meier | 1910 |
| Copper, massive | 0.62 | 2.57 | 0.73 | Oppitz | 1917 |
| Gallium, single crystal | $3 \cdot 69$ | 5.43 | 0.71 | Lange | 1935 |
| Antimony, massive | 3.04 | 4.94 | 0.70 | Drude | 1890 |
| Cobalt, massive | 2.12 | $4 \cdot 04$ | 0.68 | Minor | 1904 |
| Nickel, electrolytic | 1.58 | 3.42 | 0.66 | Meier | 1910 |
| Manganese, massive | 2.41 | 3.88 | 0.64 | Littleton | 1911 |
| Lead, massive | 2.01 | 3.48 | 0.62 | Drude | 1890 |
| Platinum, electrolytic | 2.63 | 3.54 | 0.59 | Meier | 1910 |
| Rhenium, massive | 3.00 | 3.44 | 0.57 | Lange | 1935 |
| Tungsten, massive | 3.46 | 3.25 | 0.54 | Littleton | 1912 |
| Bismuth, massive | 1.78 | 2.80 | 0.54 | Meier | 1910 |
| Iron, evaporated | 1.51 | 1.63 | 0.33 | Meier | 1910 |

The values displayed in Table XXVII are not in agreement with the formula § 13.1 (16b) or (24). For example, for copper $\sigma = 5 \cdot 14.10^{17}\text{sec}^{-1}$ so that for light of wavelength 5893Å ($\nu \sim 5.10^{14}\text{sec}^{-1}$) $\sigma/\nu \sim 10^3$, whereas, according to the table, $n^2\kappa = 1.57$. Moreover, study of the dependence of the optical constants on frequency shows a much more complicated behaviour than our formula predicts (see Fig. 13.3). Thus we must conclude that our theory is inadequate when applied to radiation in the

visible region of the electromagnetic spectrum. This discrepancy between theory and experiment is perhaps not so surprising if we recall that even for transparent media the relation

0

- When $n < 1$, the real phase velocity $c/n$ exceeds the velocity of light in vacuum, but as explained at the end of § 1.3.3 this is not in contradiction with the theory of relativity.

connecting the material constants with the refractive index (Maxwell's relation $\mu\varepsilon = n^2$ ) is of restricted validity. The explanation is similar to that given before: we are not justified in assuming that $\varepsilon, \mu$, and $\sigma$ are true constants, but should consider them to be functions of frequency so that the refractive index and the attenuation index also depend on the frequency. The only difference in the mechanism of the dispersion is that whilst in a transparent medium the dispersion is associated with forced vibrations of bound electrons, in a metal it is connected with forced vibrations of electrons that are free. This will be discussed in § 13.3; here we only remark that if $\varepsilon$ is interpreted as the static dielectric constant and $\sigma$ the static conductivity our theory can only be expected to be valid when the vibrations are sufficiently slow (long wavelengths). It was shown by Hagen and Rubens* that for infra-red radiation of wavelength $\lambda > 1 \cdot 2.10^{-3}$ cm, the reflectivity, as calculated from the static conductivity, is in good agreement with experiment.

If we substitute for $n$ and $n\kappa$ from § 13.1 (24), eq. (28) becomes (taking $\mu = 1$ ),

$$\mathscr{R} = \frac{2\frac{\sigma}{\nu} + 1 - 2\sqrt{\frac{\sigma}{\nu}}}{2\frac{\sigma}{\nu} + 1 + 2\sqrt{\frac{\sigma}{\nu}}} \tag{29}$$

When $\nu/\sigma$ is sufficiently small, we may neglect 1 in comparison with the other terms and may develop (29) in powers of $\sqrt{v/\sigma}$. We then obtain

$$\mathscr{R} \sim 1 - 2\sqrt{\frac{v}{\sigma}} + \ldots \tag{30}$$

Hagen and Rubens found that, at wavelength $\lambda = 1 \cdot 2 \cdot 10^{-3}$ cm, one has for copper $1 - \mathscr{R} = 1 \cdot 6 \cdot 10^{-2}$, whilst on substituting for the electrical conductivity, (30) gives $1 - \mathscr{R} = 1 \cdot 4 \cdot 10^{-2}$.

As the wavelength is increased further, $\mathscr{R}$ becomes so nearly equal to unity that it is difficult to measure $1 - \mathscr{R}$ with any accuracy. Hagen and Rubens obtained, however, useful estimates by an indirect method. According to Kirchhoff's law of heat radiation † the ratio of the emissive power $E_\nu$ to the absorptive power $A_\nu$ of a body ‡ depends only on the frequency $\nu$ and on the temperature $T$ of the body and not on the nature of the body, i.e.

$$\frac{E_v}{A_v} = K(v, T), \tag{31}$$

where $K(\nu, T)$ is a universal function of $\nu$ and $T$. Evidently $K$ is equal to the emissive power of a body whose absorptive power is unity, a so-called black body. Now suppose

that radiation falls on a metal specimen of such a thickness that all the incident energy that is not reflected is absorbed in its interior. Then

$$A_y = 1 - \mathscr{R}, \tag{32}$$

and from (30), (31) and (32)

$$A_v = \frac{E_v}{K(v, T)} = 2\sqrt{\frac{v}{\sigma}}, \tag{33}$$

0

- E. Hagen and H. Rubens, Ann. d. Physik, (4), 11 (1903), 873.
  † See, for example, M. Planck, Theory of Heat (London, Macmillan, 1932), p. 189; or A. Sommerfeld, Thermodynamics and Statistical Mechanics, edited by F. Bopp and J. Meixner (New York, Academic Press, 1956), p. 136.
  ‡ By emissive power is meant the radiant energy emitted by the body per unity time, by absorptive power the fraction which the body absorbs of the radiant energy which falls upon it.

or,

$$\sqrt{\sigma}E_v = 2\sqrt{v}K(v,T). \tag{34}$$

The right-hand side of this equation is independent of the nature of the metal. It is a well-known function of $v$ and $T$, the function $K(v,T)$ being precisely known both from experiment and theory, and represented by the celebrated formula of Planck.*

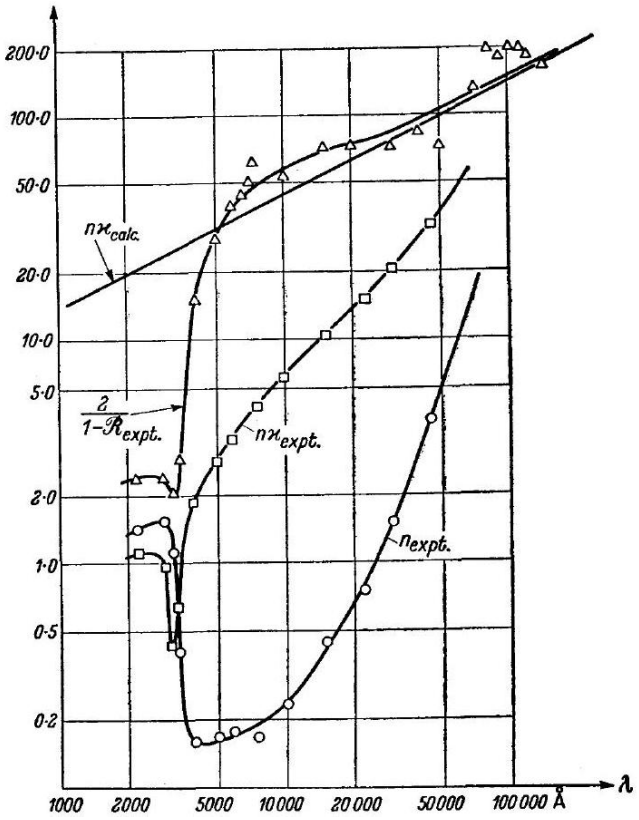



Fig. 13.3. The optical constants of silver as functions of the wavelength. Subscript "expt" refers to data obtained from experiment. The scales are logarithmic.

It follows that the validity of the formula (30) may be tested even when $\mathscr{R}$ is very close to unity by determining the conductivity $\sigma$ and the emissive power $E_\nu$ as functions of the frequency and temperature and examining whether the product $\sqrt{\sigma}E$, satisfies (34). Hagen and Rubens confirmed that this is so at long infra-red wavelengths, using for this purpose the so-called residual rays. These are rays left over from a wider spectral range after repeated reflections from certain crystals, e.g. fluorite, rock-salt or sylvine. These substances have pronounced absorption maxima in the

spectral region $\lambda = 22 \cdot 9\mu$ to $63\mu$, and hence [cf. (28)] highly selective reflectivity for such wavelengths.

In Fig. 13.3 curves are given, illustrating, for the case of silver, the dependence of $n$ and $n\kappa$ on the wavelength, as determined from experiment. For comparison the theoretical curve computed from the formula §13.1(24) is also shown. The scales are logarithmic so that the theoretical curve is the straight line

0

• See, for example, M. Born, Atomic Physics (London and Glasgow, Blackie and Son, 5th ed., 1951), p. 238.

$$\log n \sim \log n\kappa \sim \frac{1}{2}\log\lambda + C$$

where $C = \log\sqrt{\mu\sigma/c}$. From § 13.1 (24) and from (30) we may also express $n$ and $n\kappa$ in terms of the reflectivity (for long waves):

$$n \sim n_\kappa \sim \frac{2}{1-\mathscr{R}}. \tag{35}$$

The function $2/(1-\mathscr{R})$ is also displayed in the figure for comparison. We see that the experimental curve for $n\kappa$ has a sharp minimum near $\lambda = 3000$Å and that the curve for $n$ has a much flatter minimum near $\lambda = 5000$Å. At about $\lambda = 3300$Å the reflectivity of the silver is seen to be very poor.

With increasing wavelength the experimental curves approach the theoretical curve calculated from the conductivity measured electrically.

## 9.17.3 13.3. Elementary Electron Theory of the Optical Constants of Metals

The formal theory developed in the preceding sections, which treats a metal like a continuous conducting medium characterized by the dielectric constants $\varepsilon$, the magnetic permeability $\mu$, and the electric conductivity $\sigma$ is, as we have seen, not very satisfactory in the optical range of the electromagnetic spectrum. An improved theory is obtained by considering a metal not as a structureless medium, but rather as an assembly of fixed ionized metallic atoms and of electrons which move freely (or almost freely) between them. In conformity with § 13.1 (7) there cannot be a total free charge, but one may suppose that it is the mobility of the free electrons that causes the optical constants of the metal to vary with the frequency of the incident radiation. If we regard the electrons as particles obeying the laws of classical mechanics, the theory is, however, still not adequate to account for all the observed features. Only a treatment based on quantum mechanics can give a complete agreement with experiment. Such a treatment is outside the scope of this book. We shall only briefly indicate how the classical atomistic model accounts, at least to some extent, for some of the discrepancies which we have encountered.

Consider a number of electrons moving through the lattice formed by the ions. Their motion is practically free of forces apart from short periods of collisions with the ions, during which small amounts of energy are exchanged between the ions and the electrons. We cannot enter into a detailed discussion of this process which has to be treated by means of statistical methods of the kinetic theory of gases. The very plausible result is that the averaged total effect is the same as that of a damping force proportional and opposite in direction to the velocity of a model electron that represents the average behaviour of the whole set of electrons. The equation of motion of this model electron in an electric field $\boldsymbol{E}$ is, therefore,

$$m\ddot{r} + m\beta\dot{r} = eE, \tag{1}$$

where $m$ is the mass, $e$ the charge of the electron and $\beta$ the damping constant referred to unit mass. In order to understand the meaning of $\beta$ consider first the case where no electric field is present. If $\boldsymbol{E} = 0$, we have

$$\ddot{r} + \beta\dot{r} = 0, \tag{2}$$

with the solution

$$r = r_0 - \frac{1}{\beta}v_0 e^{-\beta t}, \quad \dot{r} = v = v_0 e^{-\beta t}; \tag{3}$$

we see that in this case the model electron starting with the velocity $\boldsymbol{v}_0$ is slowed down in an exponential way, with $\beta$ as decay constant. The time $\tau = 1/\beta$ is called the decay time.

Let us now assume a time-harmonic field $E = E_0 e^{-i\omega t}$. The solution of (1) is then the sum of two terms, one representing the decaying motion (solution of the homogeneous equation (2)) and the other representing a periodic motion

$$r = -\frac{e}{m\left(\omega^2 + i\beta\omega\right)} E \tag{4}$$

The periodic motion contributes the amount $p = er$ to the polarization of the medium. If there are $N$ free electrons per unit volume, then the total polarization $P$ due to all the electrons is

$$\boldsymbol{P} = N\boldsymbol{p} = -\frac{Ne^2}{m\left(\omega^2 + i\beta\omega\right)} \boldsymbol{E} \tag{5}$$

Since $\hat{\varepsilon}E = E + 4\pi P$ it follows that

$$\hat{\varepsilon} = 1 - \frac{4\pi Ne^2}{m\left(\omega^2 + i\beta\omega\right)} \tag{6}$$

Now (with $\mu = 1$ ), $\hat{\varepsilon} = n^2\left(1 - \kappa^2\right) + 2in^2\kappa$, so that (6) gives, on equating real and imaginary parts and on using equations (16) of § 13.1,

$$\varepsilon = n^2\left(1 - \kappa^2\right) = 1 - \frac{4\pi Ne^2}{m\left(\omega^2 + \beta^2\right)} \tag{7a}$$

$$\frac{\sigma}{v} = n^2\kappa = \frac{2\pi Ne^2\beta}{m\omega\left(\omega^2 + \beta^2\right)} \tag{7b}$$

Now if $\beta$ is small enough, $\varepsilon$ becomes negative for sufficiently small values of $\omega$. The critical value $\omega_c$ for which $\varepsilon$ changes sign is given by

$$\omega_c^2 = \frac{4\pi Ne^2}{m} - \beta^2 \tag{8}$$

We may re-write equations (7) in terms of this critical value and obtain

$$\varepsilon = n^2\left(1 - \kappa^2\right) = 1 - \frac{\omega_c^2 + \beta^2}{\omega^2 + \beta^2} \tag{9a}$$

$$\frac{\sigma}{\nu} = n^2\kappa = \frac{\beta\left(\omega_c^2 + \beta^2\right)}{2\omega\left(\omega^2 + \beta^2\right)} \tag{9b}$$

We shall now assume that ${\omega_c}^2$ is much larger than $\beta^2$, so that in place of (8) we may write

$$\omega_c^2 \sim \frac{4\pi Ne^2}{m} \tag{10}$$

If we also restrict ourselves to sufficiently high frequencies ( $\omega^2 \gg \beta^2$ ), we obtain, in place of (9), the simpler formulae

$$\varepsilon = n^2\left(1 - \kappa^2\right) \sim 1 - \left(\frac{\omega_c}{\omega}\right)^2, \tag{11a}$$

$$\frac{\sigma}{\nu} = n^2\kappa \sim \frac{\beta}{2\omega}\left(\frac{\omega_c}{\omega}\right)^2. \tag{11b}$$

It follows from (11a) that when $\omega < \omega_c$, then $\varepsilon$ is negative and $\kappa > 1$. We now see that a negative value for the real part of the dielectric constant is not without meaning, but is actually to be expected, for it indicates that the effect of the vibration of the electrons is in antiphase with the exciting field. The reflectivity, which is given by § 13.2 (28) becomes very large. When, however, $\omega > \omega_c$ then $\varepsilon$ is positive again, $\kappa$ becomes less than unity, and the reflectivity falls to quite a low value; the medium becomes fairly transparent, like a dielectric.
The alkali metals exhibit precisely these phenomena, for in the long wavelength region they are opaque and highly reflecting, whereas at some critical wavelength in the visible or ultra-violet they become transparent, and have comparatively low absorption. Table XXVIII shows in the second row the experimentally determined wavelengths at which this transition occurs. The third row contains these critical wavelengths $\lambda_c = 2\pi c/\omega_c$ determined from the approximate formula (10), where the number of free electrons is taken to be the same as the number $N$ of atoms in the unit volume. It is seen that the values in the two rows are different, except for sodium. The last row gives the ratio of number $N_{\text{eff}}$ of electrons that are "effective" and the number of atoms, determined from the formula

$$\frac{N_{\text{eff}}}{N} = \frac{(\lambda_c)^2_{\text{calc}}}{(\lambda_c)^2_{\text{obs}}} \tag{12}$$

It is seen that this number is of the order of unity though (except for sodium) considerably smaller. Thus the elementary theory gives the correct order of magnitude of the parameters, but does not describe the phenomena in detail.

Table XXVIII The critical wavelengths $\lambda_c$ below which the alkali metals become transparent, and above which they are opaque and highly reflecting

| Metal . | Lithium | Sodium | Potassium | Rubidium | Caesium |
|---|---|---|---|---|---|
| $(\lambda_c)_{\text{obs}}$ . | 2050 A | 2100Å | 3150Å | 3600 A | 4400 A |
| $(\lambda_0)_{\text{calc}}$ | 1500 A | 2100Å | 2900 A | 3200 A | 3600Å |
| $\frac{N_{\text{eff}}}{N} N$ | 0.54 | 1.00 | 0.85 | 0.79 | 0.67 |

The theory can be improved by using, in place of the crude approximations (11) the more exact formulae (9) which contain the decay constant $\beta$. The calculation of $\beta$ is, however, beyond the powers of classical theory and requires quantum mechanics. A further indication of the inadequacy of the classical model is the appearance of
another region of opacity at short wavelengths, which is not predicted by our formulae. It can be explained on the basis of quantum mechanics as the effect of internal photoelectric absorption; the electrons are raised from their normal orbits to excited states, the energy being supplied by the incident light. However, these considerations are outside the frame of this book.

## 9.17.4 13.4. Wave Propagation in a Stratified Conducting Medium. Theory of Metallic Films

In § 1.6 we have studied the propagation of electromagnetic waves in stratified dielectric media, that is, in dielectric media with optical properties depending on one Cartesian coordinate only. We shall now briefly discuss the extension of the theory to stratified media that contain absorbing elements. Thus we assume that in addition to $\varepsilon$ and $\mu$ being functions of only one coordinate, there may be a finite conductivity $\sigma$ which likewise is a function of this coordinate alone.

As explained at the beginning of 13.2, the formulae of Chapter I as far as they involve only linear relations between the components of the field vectors of a time harmonic wave retain their validity for conducting media, provided that the real dielectric constant $\varepsilon$ and the real

wave number $k$ are replaced by the complex dielectric constant $\hat{\varepsilon} = \varepsilon + i4\pi\sigma/\omega$ and by the complex wave number $\hat{\varepsilon} = \omega\sqrt{\mu(\varepsilon + i4\pi\sigma/\omega)}/c$ respectively. Hence we may take over the basic formulae of the theory of stratified dielectric media as developed in § 1.6, provided we make this formal change in the appropriate formulae. It follows, in particular, that a stratified absorbing medium may be characterized by a two-by-two matrix. In contrast to the case of a dielectric stratified medium, the elements of this matrix are no longer real or pure imaginary numbers but are complex numbers that contain both real and imaginary parts.

We shall illustrate the theory by studying in detail two cases of practical interest.

### 13.4.1 An absorbing film on a transparent substrate

Consider a plane parallel absorbing film situated between two dielectric media (Fig. 13.4). The formulae relating to the reflection and transmission of a plane mono-

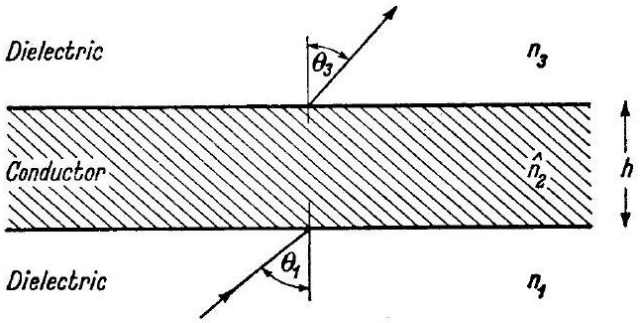



Fig. 13.4. An absorbing film situated between two dielectric media.

chromatic wave by the film are obtained from equations (55)-(58) of § 1.6 on replacing $n_2$ by $\hat{n}_2 = n_2(1 + i\kappa_2)$. It is convenient to set

$$\hat{n}_2 \cos\theta_2 = u_2 + iv_2, \tag{1}$$

where $u_2$ and $v_2$ are real. We can easily express $u_2$ and $v_2$ in terms of the angle of incidence and the constants which characterize the optical properties of the first and
the second medium. It follows, on squaring (1) and using the law of refraction $\hat{n}_2 \sin\theta_2 = n_1 \sin\theta_1$, that

$$(u_2 + iv_2)^2 = \hat{n}_2^2 - n_1^2 \sin^2\theta_1. \tag{2}$$

On equating real and imaginary parts this gives

$$\left.\begin{aligned} u_2^2 - v_2^2 &= n_2^2(1 - \kappa_2^2) - n_1^2 \sin^2\theta_1, \\ u_2 v_2 &= n_2^2 \kappa_2. \end{aligned}\right\} \tag{3}$$

From (3) we find that

$$\left.\begin{aligned} 2u_2^2 &= n_2^2(1 - \kappa_2^2) - n_1^2 \sin^2\theta_1 + \sqrt{\left[n_2^2(1 - \kappa_2^2) - n_1^2 \sin^2\theta_1\right]^2 + 4n_2^4\kappa_2^2}, \\ 2v_2^2 &= -\left[n_2^2(1 - \kappa_2^2) - n_1^2 \sin^2\theta_1\right] + \sqrt{\left[n_2^2(1 - \kappa_2^2) - n_1^2 \sin^2\theta_1\right]^2 + 4n_2^4\kappa_2^2}. \end{aligned}\right\} \tag{4}$$

Next we must evaluate the reflection and transmission coefficients for the interfaces 1-2 and 2-3 respectively, for these coefficients enter the formulae for the reflection and transmission coefficients of the film. We consider separately the cases when the electric vector of the incident wave is perpendicular, or parallel, to the plane of incidence.

**Electric vector perpendicular to the plane of incidence (TE wave)**

In this case we have, on replacing $n_2 \cos\theta_2$ by $\hat{n}_2 \cos\theta_2 = u_2 + iv_2$ in § 1.6 (55),

$$r_{12} = \rho_{12} e^{i\phi_{12}} = \frac{n_1 \cos\theta_1 - (u_2 + iv_2)}{n_1 \cos\theta_1 + (u_2 + iv_2)} \tag{5}$$

We shall later need explicit expressions for the amplitude $\rho_{12}$ and the phase change $\phi_{12}$. From (5) we have:

$$\rho_{12}^2 = \frac{(n_1 \cos\theta_1 - u_2)^2 + v_2^2}{(n_1 \cos\theta_1 + u_2)^2 + v_2^2}, \quad \tan\phi_{12} = \frac{2v_2 n_1 \cos\theta_1}{u_2^2 + v_2^2 - n_1^2 \cos^2\theta_1} \tag{6}$$

For transmission at the first interface, we have from § 1.6 (56)

$$t_{12} = \tau_{12} e^{i\chi_{12}} = \frac{2n_1 \cos\theta_1}{n_1 \cos\theta_1 + u_2 + iv_2} \tag{7}$$

which gives

$$\tau_{12}^2 = \frac{(2n_1 \cos\theta_1)^2}{(n_1 \cos\theta_1 + u_2)^2 + v_2^2}, \quad \tan\chi_{12} = -\frac{v_2}{n_1 \cos\theta_1 + u_2} \tag{8}$$

In a strictly analogous way we obtain the following expressions relating to reflection and transmission at the second interface:

$$\begin{aligned} \rho_{23}^2 &= \tfrac{(n_3\cos\theta_3 - u_2)^2 + v_2^2}{(n_3\cos\theta_3 + u_2)^2 + v_2^2}, \quad \tan\phi_{23} = \tfrac{2v_2 n_3 \cos\theta_3}{u_2^2 + v_2^2 - n_3^2\cos^2\theta_3} \\ \tau_{23}^2 &= \tfrac{4\left(u_2^2 + v_2^2\right)}{(n_3\cos\theta_3 + u_2)^2 + v_2^2}, \quad \tan\chi_{23} = \tfrac{v_2 n_3 \cos\theta_3}{u_2^2 + v_2^2 + u_2 n_3 \cos\theta_3} \end{aligned} \tag{10}$$

Since, according to the law of refraction $n_1 \sin\theta_1 = \hat{n}_2 \sin\theta_2, \hat{n}_2 \sin\theta_2 = n_3 \sin\theta_3$, the angle $\theta_3$ is determined from $\theta_1$ by means of the formula

$$n_3 \sin\theta_3 = n_1 \sin\theta_1 \tag{11}$$

**Electric vector parallel to the plane of incidence (TM wave)**

As explained in § 1.6.3 the formulae for the reflection and transmission coefficients for a TM wave can be obtained from those for a TE wave simply by replacing the quantities $p_j = n_j \cos\theta_j$ by $q_j = \cos\theta_j / n_j$, it being assumed that the media are nonmagnetic. The quantities $r$ and $t$ now refer to the ratios of the magnetic, not the electric vectors. In particular we have from § 1.6 (55),

$$\begin{aligned} r_{12} = \rho_{12} e^{i\phi_{12}} &= \frac{\frac{1}{n_1}\cos\theta_1 - \frac{1}{\hat{n}_2}\cos\theta_2}{\frac{1}{n_1}\cos\theta_1 + \frac{1}{\hat{n}_2}\cos\theta_2} = \frac{\hat{n}_2^2 \cos\theta_1 - n_1 \hat{n}_2 \cos\theta_2}{\hat{n}_2^2 \cos\theta_1 + n_1 \hat{n}_2 \cos\theta_2} \\ &= \frac{\left[n_2^2\left(1 - \kappa_2^2\right) + 2in_2^2\kappa_2\right]\cos\theta_1 - n_1\left(u_2 + iv_2\right)}{\left[n_2^2\left(1 - \kappa_2^2\right) + 2in_2^2\kappa_2\right]\cos\theta_1 + n_1\left(u_2 + iv_2\right)} \end{aligned} \tag{12}$$

From (12) we find after a straightforward calculation

$$\left.\begin{aligned} \rho_{12}^2 &= \frac{\left[n_2^2\left(1-\kappa_2^2\right)\cos\theta_1 - n_1 u_2\right]^2 + \left[2n_2^2\kappa_2\cos\theta_1 - n_1 v_2\right]^2}{\left[n_2^2\left(1-\kappa_2^2\right)\cos\theta_1 + n_1 u_2\right]^2 + \left[2n_2^2\kappa_2\cos\theta_1 + n_1 v_2\right]^2} \\ \tan\phi_{12} &= 2n_1 n_2^2 \cos\theta_1 \frac{2\kappa_2 u_2 - \left(1-\kappa_2^2\right)v_2}{n_2^4\left(1+\kappa_2^2\right)^2\cos^2\theta_1 - n_1^2\left(u_2^2 + v_2^2\right)}. \end{aligned}\right\} \tag{13}$$

For the ratio $t_{12}$ we obtain from §1.6(56) on replacing $n_j \cos\theta_j$ by $\cos\theta_j / n$,

$$t_{12} = \tau_{12} e^{i\chi_{12}} = \frac{\frac{2}{n_1}\cos\theta_1}{\frac{1}{n_1}\cos\theta_1 + \frac{1}{\hat{n}_2}\cos\theta_2} = \frac{2\left[n_2^2\left(1-\kappa_2^2\right)+2in_2^2\kappa_2\right]\cos\theta_1}{\left[n_2^2\left(1-\kappa_2^2\right)+2in_2^2\kappa_2\right]\cos\theta_1 + n_1\left(u_2+iv_2\right)} \tag{14}$$

From (14) we find that

$$\left.\begin{aligned} \tau_{12}^2 &= \frac{4n_2^4\left(1+\kappa_2^2\right)^2\cos^2\theta_1}{\left[n_2^2\left(1-\kappa_2^2\right)\cos\theta_1+n_1u_2\right]^2+\left[2n_2^2\kappa_2\cos\theta_1+n_1v_2\right]^2} \\ \tan\chi_{12} &= \frac{n_1\left[2\kappa_2u_2-\left(1-\kappa_2^2\right)v_2\right]}{n_2^2\left(1+\kappa_2^2\right)^2\cos\theta_1+n_1\left[\left(1-\kappa_2^2\right)u_2+2\kappa_2v_2\right]} \end{aligned}\right\} \tag{15}$$

In a similar way we obtain the following formulae for the reflection and transmission coefficients relating to the second interface:

$$\left.\begin{aligned} \rho_{23}^2 &= \frac{\left[n_2^2\left(1-\kappa_2^2\right)\cos\theta_3-n_3u_2\right]^2+\left[2n_2^2\kappa_2\cos\theta_3-n_3v_2\right]^2}{\left[n_2^2\left(1-\kappa_2^2\right)\cos\theta_3+n_3u_2\right]^2+\left[2n_2^2\kappa_2\cos\theta_3+n_3v_2\right]^2}, \\ \tan\phi_{23} &= 2n_3n_2^2\cos\theta_3\frac{2\kappa_2u_2-\left(1-\kappa_2^2\right)v_2}{n_2^4\left(1+\kappa_2^2\right)^2\cos^2\theta_3-n_3^2\left(u_2^2+v_2^2\right)}, \end{aligned}\right\} \tag{16}$$

and

$$\left.\begin{aligned} \tau_{23}^2 &= \frac{4n_3^2\left(u_2^2+v_2^2\right)}{\left[n_3u_2+n_2^2\left(1-\kappa_2^2\right)\cos\theta_3\right]^2+\left[n_3v_2+2n_2^2\kappa_2\cos\theta_3\right]^2}, \\ \tan\chi_{23} &= \frac{n_2^2\left[(1-\kappa_2)^2v_2-2\kappa_2u_2\right]\cos\theta_3}{n_3\left[u_2^2+v_2^2\right]+n_2^2\left[(1-\kappa_2)^2u_2+2\kappa_2v_2\right]\cos\theta_3}. \end{aligned}\right\} \tag{17}$$

From the knowledge of the quantities $\rho_{12}, \phi_{12}$, etc., the complex reflection and transmission coefficients of the film may immediately be evaluated. It will be useful to set

$$\eta = \frac{2\pi}{\lambda_0}h, \tag{18}$$

so that

$$\beta = \frac{2\pi}{\lambda_0}\hat{n}_2h\cos\theta_2 = \left(u_2+iv_2\right)\eta. \tag{19}$$

The equations § 1.6 (57)-(58) now become

$$r = \rho e^{i\delta_r} = \frac{\rho_{12}e^{i\phi_{13}}+\rho_{23}e^{-2v_2\eta}e^{i(\phi_{33}+2u_2\eta)}}{1+\rho_{12}\rho_{23}e^{-2v_2\eta}e^{i(\phi_{12}+\phi_{23}+2u_2\eta)}} \tag{20}$$

$$t = \tau e^{i\delta_t} = \frac{\tau_{12}\tau_{23}e^{-v_2\eta}e^{i[\chi_{12}+\chi_{23}+u_2\eta]}}{1+\rho_{12}\rho_{23}e^{-2v_2\eta}e^{i[\phi_{12}+\phi_{23}+2u_2\eta]}} \tag{21}$$

From (20) we obtain, after straightforward calculation, the following expressions for the reflectivity $\mathscr{R}$ and for the phase change $\delta_r$ on reflection:

$$\mathscr{R} = |r|^2 = \frac{{\rho_{12}}^2e^{2v_2\eta}+{\rho_{23}}^2e^{-2v_2\eta}+2\rho_{12}\rho_{23}\cos\left[\phi_{23}-\phi_{12}+2u_2\eta\right]}{e^{2v_2\eta}+{\rho_{12}}^2{\rho_{23}}^2e^{-2v_2\eta}+2\rho_{12}\rho_{23}\cos\left[\phi_{12}+\phi_{23}+2u_2\eta\right]}, \tag{22}$$

$$\tan\delta_r = \frac{\rho_{23}\left(1-{\rho_{12}}^2\right)\sin\left(2u_2\eta+\phi_{23}\right)+\rho_{12}\left[e^{2v_2\eta}-{\rho_{23}}^2e^{-2v_2\eta}\right]\sin\phi_{12}}{\rho_{23}\left(1+{\rho_{12}}^2\right)\cos\left(2u_2\eta+\phi_{23}\right)+\rho_{12}e^{2v_2\eta}+{\rho_{23}}^2e^{-2v_2\eta}\right]\cos\phi_{12}}. \tag{23}$$

These formulae are valid for a TE wave as well as for a TM wave. In the former case one must substitute for $\rho$ and $\phi$ the values given by (6) and (9), in the latter case those given by (13) and (16).

In a similar way we obtain from (21) the following expressions for the transmissivity $\mathscr{T}$ and for the phase change $\delta_t$ on transmission:

$$\mathscr{T} = \frac{n_3 \cos\theta_3}{n_1 \cos\theta_1}|t|^2 = \frac{n_3 \cos\theta_3}{n_1 \cos\theta_1} \frac{{\tau_{12}}^2 {\tau_{23}}^2 e^{-2v_2\eta}}{1 + {\rho_{12}}^2 {\rho_{23}}^2 e^{-4v_2\eta} + 2\rho_{12}\rho_{23} e^{-2v_2\eta} \cos\left[\phi_{12} + \phi_{23} + 2u_2\eta\right]}, \quad (24)$$

$$\tan\left[\delta_t - \chi_{12} - \chi_{23} + u_2\eta\right] = \frac{e^{2v_2\eta} \sin 2u_2\eta - \rho_{12}\rho_{23} \sin}{e^{2v_2\eta} \cos 2u_2\eta + \rho_{12}\rho_{23} \cos\left(\phi_{12} + \phi_{23}\right)}. \quad (25)$$

For a TM wave the factor $n_3 \cos\theta_3 / n_1 \cos\theta_1$ must be replaced by $(\cos\theta_3/n_3) \,/\, (\cos\theta_1/n_1)$. For a TE wave the values given by (6), (8), (9), and (10) are substituted in these formulae, and for a TM wave those given by (13), (15), (16), and (17).

It may be worthwhile to recall that the phase change on reflection ( $\delta_r$ ) is referred to the first boundary (1-2), whereas the phase change on transmission ( $\delta_t$ ) is referred to the second boundary.

Equations (22)-(25) allow the computations of the four basic quantities that characterize reflection and transmission by an absorbing film of known optical properties and of prescribed thickness. Figs. 13.5 and 13.6 illustrate, for some typical cases, the dependence of the reflectivity and transmissivity on the thickness of the film.

For a non-absorbing film $\mathscr{R}$ and $\mathscr{T}$ are periodic functions of the film thickness $h$, with a period of one wavelength. Absorption is seen to reduce the amplitude of the successive maxima and to give rise to a displacement of the maxima in the direction of smaller thickness. At optical wavelengths absorption of metals is so large that the thickness at which there is appreciable transmission is well below a
quarter wavelength* (cf. Table XXVI, p. 615). With transmitted light maxima and minima are therefore not observed.

In optics metal films are chiefly used to attain high reflectivities, for example in connection with the Fabry-Perot interferometer (§7.6.2). Such films used to be produced by chemical deposition but this method has in more recent times been superseded by the techniques of high vacuum evaporation. †

Finally let us briefly consider reflection and transmission with a "thick" film. If the thickness $h$ and consequently the parameter $\eta$ are sufficiently large all the terms

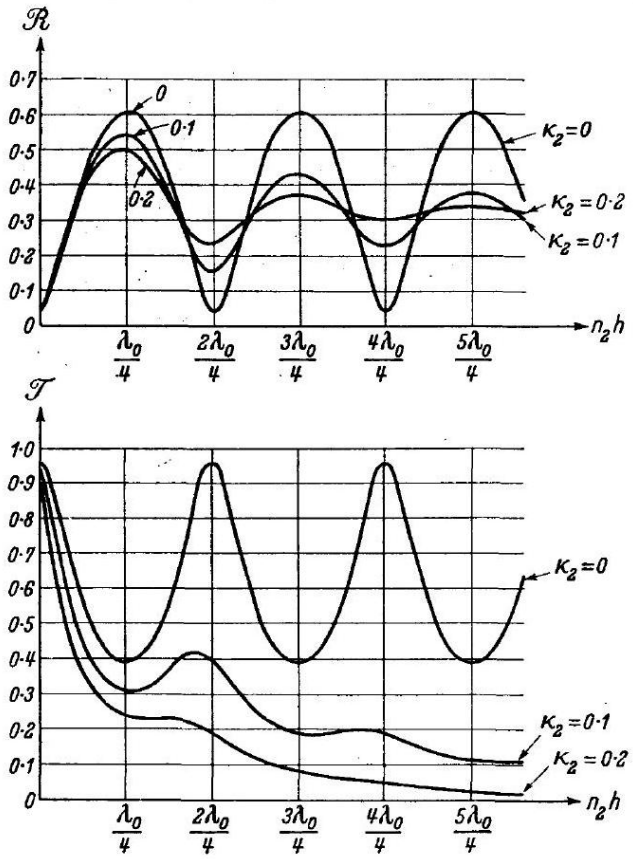



Fig. 13.5. The reflectivity $\mathscr{R}$ and transmissivity $\mathscr{T}$ of a metallic film as functions of its optical thickness. $[n_1 = 1, n_2 = 3.5, n_3 = 1.5, \kappa_1 = \kappa_3 = 0; \theta_1 = 0.]$
(After K. Hammer, Z. Tech. Phys., 24 (1943), 169.)

in (22)-(25) which do not contain the multiplicative factor $\exp(2v_2\eta)$ may be neglected. For example, if $\exp(2v_2\eta) \gtrsim 100$ this neglection does not, as a rule, involve an error of more than a few per cent. For such a film, one has at normal incidence $4\pi h n_2 \kappa_2/\lambda \geqslant \log_e 100 = 4 \cdot 61$, or (dropping the suffix 2 ),

$$\frac{h}{\lambda} \gtrsim \frac{0 \cdot 37}{n\kappa}. \tag{26}$$

For a silver film, for example, $n\kappa \sim 3.67$ at $\lambda = 5780$Å, and (26) gives $h \gtrsim \lambda/10 \sim 5 \cdot 8 \cdot 10^{-6}$ cm.

For a thick film, we have from (22) and (24)

$$\mathscr{R} \sim {\rho_{12}}^2, \quad \mathscr{T} = \frac{n_3 \cos\theta_3}{n_1 \cos\theta_1} {\tau_{12}}^2 {\tau_{23}}^2 e^{-4v_2\eta}. \tag{27}$$

0

- Simplified formulae relating to such thin films may be obtained by expanding the numerator and denominator (22)-(25) into series in powers of the film thickness, and retaining terms in the first few powers only (cf. F. Abels, Rev. d'Optique, 32 (1953), 257).

The optical properties of thin metallic films in the visible and infra-red spectral region are thoroughly discussed by L. N. Hadley and D. M. Dennison, J. Opt. Soc. Amer., 37 (1947), 451; 38 (1948), 483.
† See, for example, S. Tolansky, Multiple-beam Interferometry of Surfaces and Films (Oxford University Press, 1948), p. 26; or O. S. Heavens, Optical Properties of Thin Solid Films (London, Butterworths, 1955).

We see that the reflectivity of a "thick" film is almost that of an infinitely thick one, and that its transmissivity decreases exponentially with the thickness. The phase changes are immediately obtained from (23) and (25):

$$\delta_r \sim \phi_{12}, \quad \delta_t \sim \chi_{12} + \chi_{23} + u_2\eta. \tag{28}$$

Formulae (27) and (28) interpret our definition of a "thick" film in somewhat more physical terms, implying that in such a film the effect of multiple beam interference is negligible.

### 13.4.2 A transparent film on an absorbing substrate

As a second example, consider reflection from a transparent film on an absorbing substrate (Fig. 13.6).

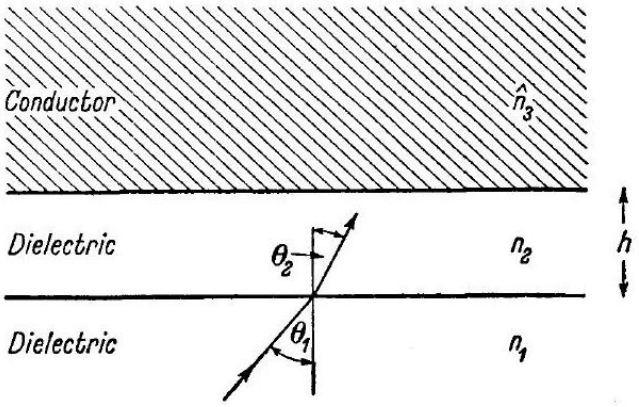



Fig. 13.6. A transparent film on an absorbing substrate.

In this case $r_{12}$ is real whilst $r_{23}$ is complex. The amplitude ratio $\rho_{23}$ and the phase change $\phi_{23}$ are given by (6) or (13), with suffixes 1 and 2 replaced by 2 and 3 respectively. According to § 1.6 (57) we now have

$$r = \frac{r_{12} + \rho_{23}e^{i(\phi_{23}+2\beta)}}{1 + r_{12}\rho_{23}e^{i(\phi_{23}+2\beta)}} \tag{29}$$

This expression is identical with the formula (57) of § 1.6 if $2\beta$ is replaced by $2\beta + \phi_{23}$ and $r_{23}$ by $\rho_{23}$. Thus without any calculation we may at once write down the expression for the reflectivity and the phase change on reflection, simply by making this substitution in equations (59) and (61) of § 1.6; we then obtain

$$\mathscr{R} = \frac{{r_{12}}^2 + {\rho_{23}}^2 + 2r_{12}\rho_{23}\cos(\phi_{23} + 2\beta)}{1 + {r_{12}}^2{\rho_{23}}^2 + 2r_{12}\rho_{23}\cos(\phi_{23} + 2\beta)}, \tag{30}$$

and

$$\tan\delta_r = \frac{\rho_{23}(1 - {r_{12}}^2)\sin(\phi_{23} + 2\beta)}{r_{12}(1 + {\rho_{23}}^2) + \rho_{23}(1 + {r_{12}}^2)\cos(\phi_{23} + 2\beta)}. \tag{31}$$

Thin transparent films on absorbing substrates have many practical uses. They are employed, for example, to protect metallic mirrors and to increase their reflectivity.

They may also be used to reduce the reflectivity of a metal surface. We have mentioned on p. 65 that one may design a polarizer consisting of a dielectric film on a dielectric substrate for which $\mathscr{R}_\mathrm{I} = 0$ and $\mathscr{R}_\perp$ is quite large. With a metallic substrate one may have either $\mathscr{R}_1 = 0$ or $\mathscr{R}_\perp = 0$.*

## 9.17.5 13.5. Diffraction by a Conducting Sphere; Theory of Mie

Metals exhibit marked optical characteristics not only in bulk, but also when they are finely divided, as in colloidal suspensions. We may recall the brilliant ruby colours of colloidal gold, whether in liquids or glasses. These phenomena are of great interest, as refraction, absorption, and diffraction take place here side by side.

Were the metallic particles perfect conductors, one would be dealing with a problem of pure diffraction. We have, however, not discussed the subject from this standpoint in the chapters on diffraction, since it is the effects that are due to partial penetration of light into the particles that are of particular physical interest. Absorption then plays an important part and it is, therefore, more appropriate to treat the subject in the present chapter; the corresponding results relating to dielectric spheres are contained in our considerations as a limiting case ( $\kappa \to 0$ ).

Of the early workers who studied the optical properties of metallic particles, mention must be made of Maxwell Garnett. † He considered the passage of light through a dielectric medium containing many small metallic spheres in a volume of linear dimensions of a wavelength. With the help of the Lorentz-Lorenz formula (§ 2.3 (17)) Maxwell Garnett showed that such an assembly is equivalent to a medium of a certain complex refractive index $\hat{n}' = n'(1 + i\kappa')$ and he found formulae for $n'$ and $\kappa'$ in terms of the indices $n$ and $\kappa$ that characterize the metallic spheres. By means of these considerations he was able to account for some of the observed features.

In a paper published in 1908, G. Mif ‡ obtained on the basis of the electromagnetic theory a rigorous solution for the diffraction of a plane monochromatic wave by a homogeneous sphere of any diameter and of any composition situated in a homogeneous medium. An equivalent solution of the same problem was published shortly afterwards by P. Debye§ in a paper concerned with light pressure (i.e. the mechanical force exerted by light) on a conducting sphere, and since then the subject has been treated in its different aspects by many writers. ||

The solution due to Mie, though derived for diffraction by a single sphere, also applies to diffraction by any number of spheres, provided that they are all of the same diameter and composition and provided also that they are randomly distributed and separated from each other by distances that are large compared to the wavelength. Under these circumstances there are no coherent phase relationships between the light that is scattered by the different spheres and the total scattered energy is then equal to the energy that is scattered by one sphere multiplied by their total number. It is particularly in this connection that Mie's solution is of great practical value and may be applied to a variety of problems: in addition to the question of colours exhibited

by metallic suspensions, we may mention applications such as the study of atmospheric dust, interstellar particles or colloidal suspensions, the theory of the rainbow, the solar corona, the effects of clouds and fogs on the transmission of light, etc.

Before deriving Mie's formulae it will be helpful to explain briefly the method employed. We are concerned with finding the solution of Maxwell's equations which describe the field arising from a plane monochromatic wave incident upon a spherical surface, across which the properties

0

- Cf. H. Schopper, Optik, 10 (1953), 426.
  † J. C. Maxwell Garnett, Phil. Trans. Roy. Soc., A. 203, (1904), 385; ibid. 205 (1906), 237.
  ‡ G. Mie, Ann. d. Physik (4), 25 (1908), 377.
  § P. Debye, Ann. d. Physik (4), 30 (1909), 57.
  || Of later investigations that deal with the basic theory, particular mention may be made of contributions by T. J. I'A. Bromwich, Phil. Trans. Roy. Soc., A, 220 (1920), 175 and by H. C. van de Hulst, Rech. Astron. Observ. Utrecht, XI, Pt. 1 (1946).

of the medium change abruptly. An appropriate system of curvilinear coordinates (spherical polar coordinates) is introduced and the field is represented as the sum of two "subfields"; one of the subfields is such that its electric vector has no radial component while the other has a magnetic vector with this property. In the spherical polar coordinates Maxwell's equations together with the boundary conditions separate into a set of ordinary differential equations, which are then solved for the two subfields in the form of infinite series. Section 13.5.1 is concerned with the derivation of this solution and in § 13.5.2 the main consequences of the solution are discussed. The final section (§ 13.5.3) is concerned with some general results relating to the total amount of energy that is scattered and absorbed by an obstacle of arbitrary shape, and the case of a spherical obstacle is discussed in detail.

### 13.5.1 Mathematical solution of the problem

(a) Representation of the field in terms of Debye's potentials

We consider the diffraction of a plane, linearly polarized monochromatic wave by a sphere of radius $a$, immersed in a homogeneous, isotropic medium. We assume that the medium in which the sphere is embedded is a non-conductor and that both this medium as well as the sphere are non-magnetic.

Assuming, as usual, the time dependence $\exp(-i\omega t)$, the time independent parts of the electric and magnetic vectors both outside and inside the sphere satisfy Maxwell's equations in their time-free form:

$$\begin{aligned} \operatorname{curl} H &= -k_1 E, \\ \operatorname{curl} E &= k_2 H, \end{aligned} \tag{1}$$

(a) (b))

where

$$\left.\begin{aligned} k_1 &= \tfrac{i\omega}{c}\left(\varepsilon + i\tfrac{4\pi\sigma}{\omega}\right), \quad \text{(a)} \\ k_2 &= \tfrac{i\omega}{c} \qquad\qquad\quad \text{(b)} \end{aligned}\right\} \tag{2}$$

The square of the usual wave number $k$ (real outside and complex inside the sphere) is

$$k^2 = -k_1 k_2. \tag{3}$$

Quantities which refer to the medium surrounding the sphere will be denoted by superscript I, those referring to the sphere by superscript II. As the medium surrounding the sphere is assumed to be non-conducting, $\sigma^{(\mathrm{I})} = 0$.

We take a rectangular system of coordinates with origin at the centre of the sphere with the $z$-direction in the direction of propagation of the wave and with the $x$ direction in the direction of its electric vector (Fig. 13.7).

If the amplitude of the electric vector of the incident wave is normalized to unity, i.e.

$$\left|E^{(i)}\right| = \left|e^{ik^{(1)}}\right| = 1,$$

the six components of the field vectors are

$$\left.\begin{aligned} E_x^{(i)} &= e^{ik^{(1)}z}, \\ H_y^{(i)} &= \tfrac{ik^{(1)}}{k_2^{(1)}} e^{ik^{(1)}z}, \\ E_y^{(i)} &= E_z^{(i)} = H_x^{(i)} = H_z^{(i)} = 0. \end{aligned}\right\} \tag{4}$$

As regards boundary conditions, we only demand, in accordance with § 1.1.3, that the tangential components of $\boldsymbol{E}$ and $\boldsymbol{H}$ shall be continuous across the surface of the sphere:

$$\left.\begin{aligned} E_{\text{tang}}^{(\text{I})} &= E_{\text{tang}}^{(\text{II})}, \\ H_{\text{tang}}^{(\text{I})} &= H_{\text{tang}}^{(\text{II})} \end{aligned}\right\} \text{ when } r = a. \tag{5}$$

The condition that the radial components of $\varepsilon E$ and $H$ shall also be continuous across the surface then follows from (5) and from Maxwell's equations.

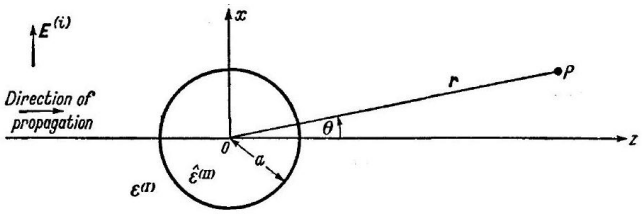



Fig. 13.7. Diffraction by a conducting sphere: notation.

In order to satisfy the boundary conditions, we must assume that apart from the incident field $E^{(i)}, H^{(i)}$ and the field $E^{(w)}, H^{(w)}$ within the sphere there is a secondary (scattered or diffracted) field $\boldsymbol{E}^{(s)}, \boldsymbol{H}^{(s)}$ in the medium surrounding the sphere. Thus the total electric field in the two regions is written as

$$\left.\begin{aligned} E &= E^{(i)} + E^{(s)} \qquad \text{outside the sphere} \\ &= E^{(w)} \qquad\qquad\quad \text{within the sphere,} \end{aligned}\right\} \tag{6}$$

with similar expressions for the magnetic vector. The fields $E^{(s)}, H^{(s)}$ and $E^{(w)}, H^{(w)}$ may be considered to be analogous to the reflected field and the transmitted field respectively, for propagation involving a plane boundary (§ 1.5.1); this analogy is, however, appropriate only when the diameter of the sphere is large compared with the wavelength. Since the boundary conditions must hold for all time, all the six vectors must have the same time dependence $[\exp(-i\omega t)]$.

The curvilinear coordinates appropriate to the present problem are the spherical polar coordinates, $r, \theta$, and $\phi$ defined by

$$\left.\begin{aligned} x &= r\sin\theta\cos\phi, \\ y &= r\sin\theta\sin\phi, \\ z &= r\cos\theta. \end{aligned}\right\} \tag{7}$$

The components of any vector $A$ are transformed from the Cartesian system to this new system according to the rule*

$$\left.\begin{aligned} A_r &= A_x\sin\theta\cos\phi + A_y\sin\theta\sin\phi + A_z\cos\theta \\ A_\theta &= A_x\cos\theta\cos\phi + A_y\cos\theta\sin\phi - A_z\sin\theta \\ A_\phi &= -A_x\sin\phi + A_y\cos\phi. \end{aligned}\right\} \tag{8}$$

Applying these formulae to the vector curl $A$ we obtain

$$\left.\begin{aligned} (\operatorname{curl} A)_r &= \frac{1}{r^2\sin\theta}\left\{\frac{\partial\left(rA_\phi\sin\theta\right)}{\partial\theta} - \frac{\partial(rA_\theta)}{\partial\phi}\right\} \\ (\operatorname{curl} A)_\theta &= \frac{1}{r\sin\theta}\left\{\frac{\partial A_r}{\partial\phi} - \frac{\partial\left(rA_\phi\sin\theta\right)}{\partial r}\right\} \\ (\operatorname{curl} A)_\phi &= \frac{1}{r}\left\{\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial\theta}\right\} \end{aligned}\right\} \tag{9}$$

In the spherical polar coordinates, the field equations (1) become

$$\left.\begin{aligned} -k_1 E_r &= \tfrac{1}{r^2\sin\theta}\left\{\frac{\partial\left(rH_\phi\sin\theta\right)}{\partial\theta} - \frac{\partial(rH_\theta)}{\partial\phi}\right\}, \\ -k_1 E_\theta &= \tfrac{1}{r\sin\theta}\left\{\frac{\partial H_r}{\partial\phi} - \frac{\partial\left(rH_\phi\sin\theta\right)}{\partial r}\right\}, \\ -k_1 E_\phi &= \tfrac{1}{r}\left\{\frac{\partial(rH_\theta)}{\partial r} - \frac{\partial H_r}{\partial\theta}\right\}. \qquad (\gamma) \end{aligned}\right\} \text{ (a) } k_2 = \frac{1}{r^2\sin\theta}\left\{\frac{\partial\left(rE_\phi\sin\theta\right)}{\partial\theta} - \frac{\partial\left(rE_\theta\right)}{\partial\phi}\right\}, \quad (\alpha) \Bigg\} \tag{10}$$

The boundary conditions (5) now are

$$\left.\begin{aligned} E_\theta^{(\mathrm{I})} = E_\theta^{(\mathrm{II})}, \quad E_\phi^{(\mathrm{I})} = E_\phi^{(\mathrm{II})}, \\ H_\theta^{(\mathrm{I})} = H_\theta^{(\mathrm{II})}, \quad H_\phi^{(\mathrm{I})} = H_\phi^{(\mathrm{II})}. \end{aligned}\right\} \text{ for } r = a \tag{11}$$

The equations (10), together with the boundary conditions (11) are the basic equations of our problem.

We shall represent the solution of these equations as a superposition of two linearly independent fields ( ${}^eE, {}^eH$ ) and ( ${}^mE, {}^mH$ ) each satisfying equations (10) such that

$${}^eE_r = E_r, \quad {}^eH_r = 0, \tag{12a}$$

and

$${}^mE_r = 0, \quad {}^mH_r = H_r \tag{12b}$$

It is not difficult to see that such a representation is consistent with our equations. With $H_r = {}^eH_r = 0$, equations (10a, $\beta$ ) and (10a, $\gamma$ ) become

$$\left.\begin{aligned} k_1\,{}^eE_\theta &= \tfrac{1}{r}\tfrac{\partial}{\partial r}\left(r\,{}^eH_\phi\right), \\ k_1\,{}^eE_\phi &= -\tfrac{1}{r}\tfrac{\partial}{\partial r}\left(r\,{}^eH_\theta\right). \end{aligned}\right\} \tag{13}$$

Substituting from these relations into ( 10 b, $\beta$ ) and ( 10 b, $\gamma$ ) we obtain:
and

$$\left.\begin{aligned} \left(\tfrac{\partial^2}{\partial r^2} + k^2\right)\left(r\,{}^eH_\theta\right) &= -\tfrac{k_1}{\sin\theta}\tfrac{\partial\,{}^eE_r}{\partial\phi} \quad (\text{ b}, \beta) \\ \left(\tfrac{\partial^2}{\partial r^2} + k^2\right)\left(r\,{}^eH_\phi\right) &= +k_1\tfrac{\partial\,{}^eE_r}{\partial\theta}. \quad (\text{b}, \gamma) \end{aligned}\right\} \tag{14}$$

Equations (14), together with (10a, $\alpha$ ), constitute a system of equations for ${}^eE_r, {}^eH_\theta$, and ${}^eH_\phi$. Not all solutions of this system represent physical fields, however; only those do

---

0

- See, for example, W. Magnus and F. Oberhettinger, Formulas and Theorems for the Functions of Mathematical Physics (New York, Chelsea Publishing Company, 1954), p. 146.

The components defined here are not those employed in the Absolute Differential Calculus of Ricol and Levi-Civita. There one has two sets of different but equivalent components of a vector $A$, the contravariant components $A^i$ and the covariant components $A_i$. If $e_1, e_2$, and $e_3$ are base vectors, in general non-orthogonal and of different lengths, the contravariant components (with respect to these base vectors) may be defined as the coefficients in the representation $A = A^1e_1 + A^2e_2 + A^3e_3$; and the covariant components may be defined as the coefficients in the representation $\boldsymbol{A} = A_1\boldsymbol{e}^1 + A_2\boldsymbol{e}^2 + A_3\boldsymbol{e}^3$, where $\boldsymbol{e}^1, \boldsymbol{e}^2$, and $\boldsymbol{e}^3$ are the reciprocal vectors, i.e. vectors satisfying the relations $e^i \cdot e_k = \delta_{ik}$ where $\delta_{ik}$ is the Kronecker symbol. In the special case when $e_1, e_2$, and $e_3$ are orthogonal, so are also $e^1, e^2$, and $e^3$ and the corresponding vectors in the two sets are parallel. One can then introduce one set of natural components $A_i = \sqrt{A_iA^i}$ and these have a simple geometrical interpretation: they are the orthogonal projections of $A$ on to the three directions. In the case of spherical polar coordinates, the natural components are those given by (8).

Tensor components may be treated in a similar way.

which satisfy the subsidiary conditions $\mathrm{div}^e H = 0$. We restrict ourselves to such solutions. In spherical polar coordinates, and under our assumption that ${}^6H_r = 0$, this subsidiary condition is

$$\frac{\partial}{\partial\theta}\left(\sin\theta {}^eH_\theta\right) + \frac{\partial}{\partial\phi}\left({}^eH_\phi\right) = 0 \tag{15}$$

and ensures that the remaining equation (10b, $\alpha$ ) is satisfied. For (10b, $\alpha$ ) becomes, on substitution from (13)

$$0 = \frac{1}{{k_1}^2 r^2 \sin\theta}\frac{\partial}{\partial r}\left[\frac{\partial}{\partial\theta}\left(r\sin\theta {}^eH_\theta\right) + \frac{\partial}{\partial\phi}\left(r {}^eH_\phi\right)\right]$$

and because of (15) this is identically satisfied. Strictly similar considerations apply to the complementary case with ${}^mE_r = 0$.

The solution with vanishing radial magnetic field is called the electric wave (or transverse magnetic wave) and that with vanishing radial electric field is called the magnetic wave (or transverse electric wave). We shall now show that they may each be derived from a scalar potential ${}^e\Pi$ and ${}^m\Pi$ respectively, which are known as Debye's potentials.*

It follows, first of all from (10b, $\alpha$ ), since ${}^eH_r = 0$, that ${}^eE_\phi$ and ${}^eE_\theta$ may be represented in terms of a gradient of a scalar,

$${}^eE_\phi = \frac{1}{r\sin\theta}\frac{\partial U}{\partial\phi}, \quad {}^eE_\theta = \frac{1}{r}\frac{\partial U}{\partial\theta}. \tag{16}$$

If now we put

$$U = \frac{\partial\left(r {}^e\Pi\right)}{\partial r} \tag{17}$$

then we have from (16)

$${}^eE_\theta = \frac{1}{r}\frac{\partial^2\left(r {}^e\Pi\right)}{\partial r\partial\theta}, \quad {}^eE_\phi = \frac{1}{r\sin\theta}\frac{\partial^2\left(r {}^e\Pi\right)}{\partial r\partial\phi}. \tag{18}$$

It is seen that equation (13) may be satisfied by

$$\left.\begin{aligned} {}^eH_\phi &= k_1\frac{\partial {}^e\Pi}{\partial\theta} = \frac{k_1}{r}\frac{\partial\left(r {}^e\Pi\right)}{\partial\theta}, \\ {}^eH_\theta &= -\frac{k_1}{\sin\theta}\frac{\partial {}^e\Pi}{\partial\phi} = -\frac{k_1}{r\sin\theta}\frac{\partial\left(r {}^e\Pi\right)}{\partial\phi}. \end{aligned}\right\} \tag{19}$$

If we substitute from (19) into (10a, $\alpha$ ) we obtain

$${}^eE_r = -\frac{1}{r\sin\theta}\left\{\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial {}^e\Pi}{\partial\theta}\right) + \frac{1}{\sin\theta}\frac{\partial^2 e\Pi}{\partial\phi^2}\right\}. \tag{20}$$

Substitution from (19) and (20) into (14) gives two equations, the first of which expresses the vanishing of the $\phi$ derivative, the second the vanishing of the $\theta$ derivative of one and the same expression. These equations may, therefore, be satisfied by equating this expression to zero, and this gives

0

- If $r$ denotes the radius vector from the origin, then ${}^e\Pi r$ and ${}^m\Pi r$ are radial Hertz vectors, i.e. Hertz vectors (see § 2.2.2) which point everywhere in the radial direction. (Cf. A. Sommerfeld, contribution in P. Frank and R. v. Mises: Riemann-Weber's Differentialgleichungen der mathematischer Physik (Braunschweig, Vieweg, 2nd ed., 1935; also New York, Dover Publ., 1961 ), 2. 790; C. J. Bouwkamp and H. B. G. Casimir, Physica, 20 (1954), 539 and A. Nisbet, Proc. Roy. Soc., A, 231 (1955), 260); Physica, 21 (1955), 799.

$$\frac{1}{r}\frac{\partial^2(r^e\Pi)}{\partial r^2}+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial^e\Pi}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 e\Pi}{\partial\phi^2}+k^2 e\Pi=0 \tag{21}$$

By means of this equation, (20) may be written as

$${}^eE_r=\frac{\partial^2(r^e\Pi)}{\partial r^2}+k^2 r^e\Pi \tag{22}$$

It can be verified by substituting from (18), (19), (20), (21) and (22) into (10) that we have obtained a solution of our set of equations.

In a similar way one may consider the magnetic wave and one finds that this wave can be derived from a potential ${}^m\Pi$ which satisfies the same differential equation (21) as ${}^6\Pi$. The complete solution of our field equations is obtained by adding the two fields; this gives

$$\left.\begin{array}{lll} E_r={}^eE_r+{}^mE_r=\frac{\partial^2(r^e\Pi)}{\partial r^2} & +k^2re\Pi, & (\alpha)\\ E_\theta={}^eE_\theta+{}^mE_\theta=\frac{1}{r}\frac{\partial^2(r^e\Pi)}{\partial r\partial\theta} & +k_2\frac{1}{r\sin\theta}\frac{\partial(r^m\Pi)}{\partial\phi}, & (\beta)\\ E_\phi={}^eE_\phi+{}^mE_\phi=\frac{1}{r\sin\theta}\frac{\partial^2(r^e\Pi)}{\partial r\partial\phi} & -k_2\frac{1}{r}\frac{\partial(r^m\Pi)}{\partial\theta}, & (\gamma)\end{array}\right\} \tag{23a}$$

$$\left.\begin{array}{lll} H_r={}^mH_r+{}^eH_r= & k^2r^m\Pi+\frac{\partial^2(r^m\Pi)}{\partial r^2}, & (\alpha)\\ H_\theta={}^mH_\theta+{}^eH_\theta=-k_1\frac{1}{r\sin\theta}\frac{\partial(r^e\Pi)}{\partial\phi} & +\frac{1}{r}\frac{\partial^2(r^m\Pi)}{\partial r\partial\theta}, & (\beta)\\ H_\phi={}^mH_\phi+{}^eH_\phi=k_1\frac{1}{r}\frac{\partial(r^e\Pi)}{\partial\theta} & +\frac{1}{r\sin\theta}\frac{\partial^2(r^m\Pi)}{\partial r\partial\phi}. & (\gamma)\end{array}\right\} \tag{23b}$$

Both potentials ${}^e\Pi$ and $m\Pi$ are solutions of the differential equation (21), which is nothing but the wave equation

$$\nabla^2\Pi+k^2\Pi=0,$$

written in polar coordinates. In order that the components $E_\theta, E_\phi, H_\theta$, and $H_\phi$ shall be continuous over the spherical surface $r=a$, it is evidently sufficient that the four quantities

$$k_1r^e\Pi,\quad k_2r^m\Pi,\quad \frac{\partial}{\partial r}(r^e\Pi),\quad \frac{\partial}{\partial r}(r^m\Pi) \tag{24}$$

shall also be continuous over this surface. Thus our boundary conditions also split into independent conditions for ${}^e\Pi$ and $m\Pi$. Our diffraction problem is thus reduced to the problem of finding two mutually independent solutions of the wave equation, with prescribed boundary conditions.

(b) Series expansions for the field components We first represent the solution of the wave equations as expansions with undetermined coefficients, each term representing a particular integral. We shall then determine the coefficients by using the boundary conditions.

We seek integrals of the form

$$\Pi=R(r)\Theta(\theta)\Phi(\phi) \tag{25}$$

As may easily be verified by direct substitution into (21), the functions $R,\Theta$, and $\Phi$ must satisfy the ordinary differential equations

$$\left.\begin{array}{rl} \frac{d^2(rR)}{dr^2}+\left(k^2-\frac{\alpha}{r^2}\right)rR=0, & (a)\\ \frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta}{d\theta}\right)+\left(\alpha-\frac{\beta}{\sin^2\theta}\right)\Theta=0, & (b)\\ \frac{d^2\Phi}{d\phi^2}+\beta\Phi=0, & (c)\end{array}\right\} \tag{26}$$

where $\alpha$ and $\beta$ are integration constants.

As the field $E,H$ is a single valued function of position, $\Pi$ must likewise be single valued, and this requirement imposes certain conditions on $\Theta$ and $\Phi$.

For each of the equations in (26) one can write down the general solution. For (c) it is

$$a\cos(\sqrt{\beta}\phi) + b\sin(\sqrt{\beta}\phi).$$

The condition of single-valuedness demands that

$$\beta = m^2, \quad (m = \text{ integer }) \tag{27}$$

Hence the single-valued solution of (26c) is

$$\Phi = a_m \cos(m\phi) + b_m \sin(m\phi) \tag{28}$$

The equation (26b) is the well-known equation for spherical harmonics. A necessary and sufficient condition for a single valued solution is that

$$\alpha = l(l+1), \quad (l > |m|, \text{ integer }). \tag{29}$$

We substitute for $\beta$ from (27) in (26b) and introduce the new variable

$$\xi = \cos\theta \tag{30}$$

The equation then transforms to*

$$\frac{d}{d\xi}\left\{\left(1-\xi^2\right)\frac{d\Theta}{d\xi}\right\} + \left\{l(l+1) - \frac{m^2}{1-\xi^2}\right\}\Theta = 0 \tag{31}$$

the solutions of which are the associated Legendre functions

$$\Theta = P_l^{(m)}(\xi) = P_l^{(m)}(\cos\theta). \tag{32}$$

These functions vanish identically if $|m| > l$; for each $l$ there are therefore $2l+1$ such functions, namely those with

$$m = -l, -l+1, \ldots l-1, l.$$

To integrate the remaining equation (26a) we set

$$kr = \rho, \quad R(r) = \frac{1}{\sqrt{\rho}} Z(\rho), \tag{33}$$

and obtain the Bessel equation †

$$\frac{d^2Z}{d\rho^2} + \frac{1}{\rho}\frac{dZ}{d\rho} + \left\{1 - \frac{\left(l+\frac{1}{2}\right)^2}{\rho^2}\right\} Z = 0 \tag{34}$$

The solution of this equation is the general cylindrical function $Z = Z_{l+\frac{1}{2}}(\rho)$ of order $l+\frac{1}{2}$, so that the solution of (26a) is

$$R = \frac{1}{\sqrt{kr}} Z_{l+\frac{1}{2}}(kr). \tag{35}$$

Each cylindrical function may be expressed as a linear combination of two cylindrical functions of standard type, e.g. the Bessel functions $J_{l+\frac{1}{2}}(\rho)$ and the Nuumann functions $N_{l+\frac{1}{8}}(\rho)$. For the present purpose it is convenient to employ the functions ‡

$$\psi_l(\rho) = \sqrt{\frac{\pi\rho}{2}} J_{l+\frac{1}{2}}(\rho), \quad \chi_l(\rho) = -\sqrt{\frac{\pi\rho}{2}} N_{l+\frac{1}{2}}(\rho) \tag{36}$$

The functions $\psi_l(\rho)$ are regular in every finite domain of the $\rho$-plane, including the origin, whereas the functions $\chi_i(\rho)$ have singularities at the origin $\rho = 0$, where they

become infinite. We may, therefore, use the functions $\psi_l(\rho)$ but not the functions $\chi_l(\rho)$ for representing the wave inside the sphere.

The general integral of (26a) may then be written as

$$rR = c_l\psi_l(kr) + d_l\chi_l(kr) \tag{37}$$

In particular, with $c_l = 1, d_l = -i$ we have

$$rR = \zeta_l{}^{(1)}(kr) \tag{38}$$

where

$$\zeta_l^{(1)}(\rho) = \psi_l(\rho) - i\chi_l(\rho) = \sqrt{\frac{\pi\rho}{2}}H^{(1)}_{l+\frac{1}{2}}(\rho) \tag{39}$$

$H^{(1)}$ being one of the Hankel functions.* The Hankel functions are distinguished from other cylindrical functions by the property that they vanish at infinity in the complex plane. The one used here with the index I vanishes in the half-plane of the positive imaginary part of $\rho$ and is thus suitable for the representation of the scattered wave.

According to (25) a particular integral $\Pi_l{}^{(m)}$ is obtained on multiplying together the functions given by (28), (32), and (37); we then obtain the following general solution of the wave equation:

$$\begin{aligned} r\Pi &= r\sum_{l=0}^{\infty}\sum_{m=-l}\Pi_l{}^{(m)} \\ &= \sum_{l=0}^{\infty}\sum_{m=-l}^{l}\{c_l\psi_l(kr) + d_l\chi_l(kr)\}\left\{P_l{}^{(m)}(\cos\theta)\right\}\{a_m\cos(m\phi) + b_m\sin(m\phi)\}, \end{aligned} \tag{40}$$

where $a_m, b_m, c_l$, and $d_l$ are arbitrary constants.

We must now determine these constants in such a way as to satisfy the boundary conditions. For this to be possible one must be able to express the potentials ${}^e\Pi^{(i)}$ and $m\Pi^{(i)}$ of the incident wave in a series of the form (40). To show that this can be done, we first transform the expression (4) for the incident wave into spherical polar coordinates in accordance with (8):

$$\left.\begin{aligned} E_r^{(i)} &= e^{ik^{(\mathrm{I})}r\cos\theta}\sin\theta\cos\phi, & H_r^{(i)} &= \tfrac{ik^{(\mathrm{I})}}{k_2^{(\mathrm{I})}}e^{ik^{(\mathrm{I})}r\cos\theta}\sin\theta\sin\phi, \\ E_\theta^{(i)} &= e^{ik^{(\mathrm{I})}r\cos\theta}\cos\theta\cos\phi, & H_\theta^{(i)} &= \tfrac{ik^{(\mathrm{I})}}{k_2^{(\mathrm{I})}}e^{ik^{(\mathrm{I})}r\cos\theta}\cos\theta\sin\phi, \\ E_\phi^{(i)} &= -e^{ik^{(\mathrm{I})}r\cos\theta}\sin\phi, & H_\phi^{(i)} &= \tfrac{ik^{(\mathrm{I})}}{k_2^{(\mathrm{I})}}e^{ik^{(\mathrm{I})}r\cos\theta}\cos\phi. \end{aligned}\right\} \tag{41}$$

0

- Cf. A. Sommerfeld, Partial Differential Equations of Physics (New York, Academic Press, 1949), p. 127.
† Cf. A. Sommerfeld; loc. cit., p. 86.
‡ Several slightly different definitions of the $\psi$ and $\chi$ functions exist in the literature. (Cf. G. N. Watson, A Treatise on the Theory of Bessel Functions (Cambridge University Press, 2nd ed., 1944), p. 56; or A. Sommerfeld, loc. cit., pp. 113-114. On pp. 646-647 we summarize some of the formulae relating to the $\psi_l$ functions.

There is a similar relation involving the other Hankel function:

$$J_p - iN_p = H_p^{(2)}$$

These two formulae are analogous to the expressions for the exponential functions $e^{i\rho}$ and $e^{-i\rho}$ in terms of $\cos\rho$ and $\sin\rho$.

To determine the potentials ${}^e\Pi^{(i)}$ or ${}^m\Pi^{(i)}$ it is only necessary to use one of the equations (23); the first of them yields:

$$e^{ik^{(\mathrm{I})}r\cos\theta}\sin\theta\cos\phi = \frac{\partial^2\left(r^e\Pi\right)}{\partial r^2} + k^{(\mathrm{I})^2}re\Pi^{(i)}. \tag{42}$$

The first factor on the left-hand side of this equation may be expressed in the following differentiable series of Legendre polynomials [Bauer's formula, §9.4 (9)]:

$$e^{ik^{(1)}r\cos\theta} = \sum_{l=0}^{\infty} i^l(2l+1)\frac{\psi_l\left(k^{(1)}r\right)}{k^{(1)}r}P_l(\cos\theta). \tag{43}$$

We also have the identities

$$e^{ik^{(1)}r\cos\theta}\sin\theta \equiv -\frac{1}{ik^{(1)}r}\frac{\partial}{\partial\theta}\left(e^{ik^{(1)}r\cos\theta}\right), \tag{44}$$

$$\frac{\partial}{\partial\theta}P_l(\cos\theta) \equiv -P_l^{(1)}(\cos\theta); \quad P_0^{(1)}(\cos\theta) \equiv 0. \tag{45}$$

Using these relations, the left-hand side of (42) may be expressed in the form

$$e^{ik^{(\mathrm{I})}r\cos\theta}\sin\theta\cos\phi = \frac{1}{\left(k^{(\mathrm{I})}r\right)^2}\sum_{l=1}^{\infty} i^{l-1}(2l+1)\psi_l\left(k^{(\mathrm{I})}r\right)P_l^{(1)}(\cos\theta)\cos\phi. \tag{46}$$

Accordingly we take as a trial solution of (42) a series of a similar form

$$r^e\Pi^{(i)} = \frac{1}{k^{(\mathrm{I})}2}\sum_{l=1}^{\infty}\alpha_l\psi_l\left(k^{(\mathrm{I})}r\right)P_l^{(1)}(\cos\theta)\cos\phi. \tag{47}$$

On substituting from (46) and (47) in equation (42) and comparing coefficients we obtain the relation

$$\alpha_l\left\{k^{(\mathrm{I})^2}\psi_l\left(k^{(\mathrm{I})}r\right) + \frac{\partial^2\psi_l\left(k^{(\mathrm{I})}r\right)}{\partial r^2}\right\} = i^{l-1}(2l+1)\frac{\psi_l\left(k^{(\mathrm{I})}r\right)}{r^2}. \tag{48}$$

Now from (37) (with $c_l = 1$ and $d_l = 0$ ) it follows that

$$\psi_l\left(k^{(\mathrm{I})}r\right) = rR \tag{49}$$

is a solution of equation (26a):

---

0

- This formula is an immediate consequence of the well known relation between the Bessel, Neumann and Hankel functions:

$$J_p + iN_p = {H_p}^{(1)}$$

$$\frac{d^2\psi_l}{dr^2} + \left(k^{(\mathrm{I})^2} - \frac{\alpha}{r^2}\right)\psi_l = 0, \tag{50}$$

provided [see (29)] that $\alpha = l(l+1)$. Comparing (50) with (48), we see that

$$\alpha_l = i^{l-1}\frac{2l+1}{l(l+1)}. \tag{51}$$

The calculations relating to the magnetic potential $m\prod^{(i)}$ are similar. We thus obtain the following expression for the two potentials of the incident wave:

$$\left.\begin{array}{ll} r^e\Pi^{(i)} = \frac{1}{k^{(\mathrm{I})^2}}\sum_{l=1}^{\infty} i^{l-1}\frac{2l+1}{l(l+1)}\psi_l\left(k^{(\mathrm{I})}r\right)P_l^{(1)}(\cos\theta)\cos\phi, & \text{(a)} \\ r^m\Pi^{(i)} = \frac{1}{k^{(1)}}\sum_{l=1}^{\infty} i^l\frac{k^{(\mathrm{I})}}{k_2^{(\mathrm{I})}}\frac{2l+1}{l(l+1)}\psi_l\left(k^{(\mathrm{I})}r\right)P_l^{(1)}(\cos\theta)\sin\phi. & \text{(b)} \end{array}\right\} \tag{52}$$

We have expressed both the potentials in series of the form (40) and the unknown constants can now easily be determined.

The boundary conditions (24) written more fully are

$$\left.\begin{array}{rll} \frac{\partial}{\partial r}\left\{r\left({}^e\Pi^{(i)} + {}^e\Pi^{(s)}\right)\right\}_{r=a} &= \frac{\partial}{\partial r}\left\{r^e\Pi^{(w)}\right\}_{r=a}, & \text{(a)} \\ \frac{\partial}{\partial r}\left\{r\left({}^m\Pi^{(i)} + {}^m\Pi^{(s)}\right)\right\}_{r=a} &= \frac{\partial}{\partial r}\left\{r^m\Pi^{(w)}\right\}_{r=a}, & \text{(b)} \\ k_1{}^{(1)}\left\{r\left(\Pi^{(i)} + {}^e\Pi^{(s)}\right)\right\}_{r=a} &= k_1{}^{(1)}\left\{r^e\Pi^{(w)}\right\}_{r=a}, & \text{(c)} \\ k_2{}^{(1)}\left\{r\left({}^m\Pi^{(i)} + {}^m\Pi^{(s)}\right)\right\}_{r=a} &= k_2{}^{(1)}\left\{r^m\Pi^{(w)}\right\}_{r=a}. & \text{(d)} \end{array}\right\} \tag{53}$$

According to (52), these equations can only be satisfied if only those terms occur in the expansions (40) for the unknown potentials $\Pi^{(s)}$ and $\Pi^{(w)}$ for which $m = 1$ and if, moreover for the magnetic potential, and

$$a_1 = 0$$
$$b_1 = 0$$

for the electric one.

We already noted that for the representation of $\Pi^{(w)}$ only the $\psi_l$ functions are appropriate, since these remain regular at the origin, whilst the functions $\chi_i$ become infinite there. Hence we set

$$\left.\begin{array}{ll} r^e\Pi^{(w)} = \frac{1}{k^{(\mathrm{II})^2}}\sum_{l=1}^{\infty} A_l\psi_l\left(k^{(\mathrm{II})r}\right)P_l^{(1)}(\cos\theta)\cos\phi, & \text{(a)} \\ r^m\Pi^{(w)} = \frac{i}{k^{(\mathrm{II})}k_2^{(\mathrm{II})}}\sum_{l=1}^{\infty} {}^mA_l\psi_l\left(k^{(\mathrm{II})}r\right)P_l^{(1)}(\cos\theta)\sin\phi. & \text{(b)} \end{array}\right\} \tag{54}$$

We have also seen that for the scattered wave a representation in terms of the functions $\zeta_l{}^{(1)} = \psi_l - i\chi_l$ is appropriate, $\zeta_l{}^{(1)}$ being obtained from the Hankel function $H^{(1)}$ on multiplication by $\sqrt{\pi\rho/2}$ [see (39)]. For large $\rho, H^{(1)}$ behaves as $e^{i\rho}/\sqrt{\rho}$, i.e. $\zeta_l{}^{(1)}$ behaves as $e^{i\rho}$ and $R = \zeta_l{}^{(1)}\left(k^{(1)}r\right)/r$ as $e^{ik^{(1)}r}/r$. Thus at large distances from the sphere the scattered wave is spherical, with its centre at the origin $r = 0$. We therefore set

$$\left.\begin{array}{l} r^e\Pi^{(s)} = \frac{1}{k^{(\mathrm{I})^2}}\sum_{l=1}^{\infty} B_l\zeta_l{}^{(1)}\left(k^{(\mathrm{I})}r\right)P_l{}^{(1)}(\cos\theta)\cos\phi \\ r^m\Pi^{(s)} = \frac{i}{k^{(\mathrm{I})}k_2^{(1)}}\sum_{l=1}^{\infty} {}^mB_l\zeta_l{}^{(1)}\left(k^{(\mathrm{I})}r\right)P_l{}^{(1)}(\cos\theta)\sin\phi \end{array}\right\} \tag{55}$$

If we now substitute the expressions (52), (54), and (55) into the boundary conditions (53) we obtain the following linear relations between the coefficients* ${}^eA_l, {}^mA_l, {}^eB_l$, and ${}^mB_l$ :

$$\left.\begin{array}{l} {}^{e}B_l \frac{1}{k^{(\mathrm{I})}} \zeta_l^{(1)\prime}\left(k^{(\mathrm{I})}a\right) + \frac{1}{k^{(\mathrm{I})}} i^{l-1} \frac{2l+1}{l(l+1)} \psi_l'\left(k^{(\mathrm{I})}a\right) = \frac{1}{k^{(\mathrm{II})}} {}^{e}A_l \psi_l'\left(k^{(\mathrm{II})}a\right) \\ {}^{m}B_l \frac{1}{k_2^{(\mathrm{I})}} \zeta_l^{(1)\prime}\left(k^{(\mathrm{I})}a\right) + \frac{1}{k_2^{(\mathrm{I})}} i^{l-1} \frac{2l+1}{l(l+1)} \psi_l'\left(k^{(\mathrm{I})}a\right) = \frac{1}{k_2^{(\mathrm{II})}} {}^{m}A_l \psi_l'\left(k^{(\mathrm{II})}a\right) \\ {}^{e}B_l \frac{1}{k_2^{(\mathrm{I})}} \zeta_l^{(1)}\left(k^{(\mathrm{I})}a\right) + \frac{1}{k_2^{(\mathrm{I})}} i^{l-1} \frac{2l+1}{l(l+1)} \psi_l\left(k^{(\mathrm{I})}a\right) = \frac{1}{k_2^{(\mathrm{II})}} {}^{e}A_l \psi_l\left(k^{(\mathrm{II})}a\right) \\ {}^{m}B_l \frac{1}{k^{(\mathrm{I})}} \zeta_l^{(1)}\left(k^{(\mathrm{I})}a\right) + \frac{1}{k^{(\mathrm{I})}} i^{l-1} \frac{2l+1}{l(l+1)} \psi_l\left(k^{(\mathrm{I})}a\right) = \frac{1}{k^{(\mathrm{II})}} {}^{m}A_l \psi_l\left(k^{(\mathrm{II})}a\right) \end{array}\right\} \tag{56}$$

We are only interested in the coefficients ${}^{e}B_l$ and ${}^{m}B_l$ which characterize the scattered wave. These may be obtained by the elimination of ${}^{e}A_l$ and ${}^{m}A_l$; this gives

$$\left.\begin{array}{ll} {}^{e}B_l = i^{l+1} \frac{2l+1}{l(l+1)} \frac{k_2^{(\mathrm{I})} k^{(\mathrm{II})} \psi_l'\left(k^{(\mathrm{I})}a\right)\psi_l\left(k^{(\mathrm{II})}a\right) - k_2^{(\mathrm{II})} k^{(\mathrm{I})} \psi_l'\left(k^{(\mathrm{II})}a\right)\psi_l\left(k^{(\mathrm{I})}a\right)}{k_2^{(\mathrm{II})} \zeta_l^{(1)\prime}\left(k^{(\mathrm{I})}a\right)\psi_l\left(k^{(\mathrm{II})}a\right) - k_2^{(\mathrm{II})} k^{(\mathrm{I})} \psi_l'\left(k^{(\mathrm{II})}a\right)\zeta^{(1)}\left(k^{(\mathrm{I})}a\right)}, & \text{(a)} \\ {}^{m}B_l = i^{\imath+1} \frac{2l+1}{l(l+1)} \frac{k_2^{(\mathrm{I})} k^{(\mathrm{II})} \psi_l\left(k^{(\mathrm{I})}a\right)\psi_l'\left(k^{(\mathrm{II})}a\right) - k_2^{(\mathrm{II})} k^{(\mathrm{I})} \psi_l'\left(k^{(\mathrm{I})}a\right)\psi_l\left(k^{(\mathrm{II})}a\right)}{k_2^{(\mathrm{II})} \zeta_l^{(\mathrm{I})}\left(k^{(\mathrm{I})}a\right)\psi_l'\left(k^{(\mathrm{II})}a\right) - k_2^{(\mathrm{II})} k^{(\mathrm{I})} \zeta_l^{(1)}\left(k^{(\mathrm{I})}a\right)\psi_l\left(k^{(\mathrm{II})}a\right)}. & \text{(b)} \end{array}\right\} \tag{b}$$

Finally, the components of the field vectors of the scattered wave are obtained on substituting from (55) in (23). This gives

$$\left.\begin{array}{l} E_r^{(s)} = \frac{1}{k^{(\mathrm{I}^2)}} \frac{\cos\phi}{r^2} \sum_{l=1}^{\infty} l(l+1) {}^{e}B_l \zeta_l^{(1)}\left(k^{(1)}r\right) P_l^{(1)}(\cos\theta) \\ E_\theta^{(s)} = -\frac{1}{k^{(1)}} \frac{\cos\phi}{r} \sum_{l=1}^{\infty} \left\{ {}^{e}B_l \zeta_l^{(1)\prime}\left(k^{(1)}r\right) P_l^{(1)\prime}(\cos\theta)\sin\theta \right. \\ \qquad \left. -i^{m}B_l \zeta_l^{(1)}\left(k^{(1)}r\right) P_l^{(1)}(\cos\theta) \frac{1}{\sin\theta} \right\} \\ E_\phi^{(s)} = -\frac{1}{k^{(1)}} \frac{\sin\phi}{r} \sum_{l=1}^{\infty} \left\{ {}^{e}B_l \zeta_l^{(1)\prime}\left(k^{(1)}r\right) P_l^{(1)}(\cos\theta) \frac{1}{\sin\theta} \right. \\ \qquad \left. -i^{m}B_l \zeta_l^{(1)}\left(k^{(1)}r\right) P_l^{(1)\prime}(\cos\theta)\sin\theta \right\} \\ H_r^{(s)} = \frac{i}{k^{(1)}k_2^{(1)}} \frac{\sin\phi}{r^2} \sum_{l=1}^{\infty} l(l+1) {}^{m}B_l \zeta_l^{(1)}\left(k^{(1)}r\right) P_l^{(1)}(\cos\theta) \\ H_\theta^{(s)} = -\frac{1}{k_2^{(1)}} \frac{\sin\phi}{r} \sum_{l=1}^{\infty} \left\{ {}^{e}B_l \zeta_l^{(1)}\left(k^{(1)}r\right) P_l^{(1)}(\cos\theta) \frac{1}{\sin\theta} \right. \\ \qquad \left. +i^{m}B_l \zeta_l^{(1)\prime}\left(k^{(1)}r\right) P_l^{(1)\prime}(\cos\theta)\sin\theta \right\} \\ H_\phi^{(s)} = \frac{1}{k_2^{(1)}} \frac{\cos\phi}{r} \sum_{l=1}^{\infty} \left\{ {}^{e}B_l \zeta_l^{(1)}\left(k^{(1)}r\right) P_l^{(1)\prime}(\cos\theta)\sin\theta \right. \\ \qquad \left. +i^{m}B_l \zeta_l^{(1)\prime}\left(k^{(1)}r\right) P_l^{(1)}(\cos\theta) \frac{1}{\sin\theta} \right\} \end{array}\right\} \tag{58}$$

This completes the formal solution of our boundary value problem. We shall not enter into the questions of existence and convergence of the solution.

It will be helpful to recall the meaning of the various constants. Since the medium surrounding the sphere is assumed to be a non-conductor, $\sigma^{(\mathrm{I})} = 0$. If we write $\sigma$ in place of $\sigma^{(\mathrm{II})}$ for the conductivity of the sphere, we have from (2)

$$\left.\begin{array}{lll} k_1{}^{(\mathrm{I})} &= \frac{i\omega}{c}\varepsilon^{(\mathrm{I})} = i\frac{2\pi}{\lambda_0}\varepsilon^{(\mathrm{I})}, \quad k_2{}^{(\mathrm{I})} = \frac{i\omega}{c} = i\frac{2\pi}{\lambda_0} & \\ k^{(\mathrm{I})} &= \sqrt{-k_1^{(\mathrm{I})}k_2^{(\mathrm{I})}} = \frac{2\pi}{\lambda_0}\sqrt{\varepsilon^{(\mathrm{I})}} = \frac{2\pi}{\lambda^{(\mathrm{I})}} & \\ k_1^{(\mathrm{II})} &= \frac{i\omega}{c}\left(\varepsilon^{(\mathrm{II})} + i\frac{4\pi\sigma}{\omega}\right) = i\frac{2\pi}{\lambda_0}\left(\varepsilon^{(\mathrm{II})} + i\frac{4\pi\sigma}{\omega}\right), \quad k_2^{(\mathrm{II})} = \frac{i\omega}{c} = i\frac{2\pi}{\lambda_0}, & \text{(c)} \\ k^{(\mathrm{II})} &= \sqrt{-k_1^{(\mathrm{II})}k_2^{(\mathrm{II})}} = \frac{2\pi}{\lambda_0}\sqrt{\varepsilon^{(\mathrm{II})} + \frac{4\pi\sigma}{\omega}}, & \text{(d)} \end{array}\right\} \tag{c}$$

0

- The addition of a prime to the functions $\psi_l, \zeta_l$, and $P_l{}^{(1)}$ denotes differentiation with respect to their arguments.

where $\lambda_0$ is the wavelength of the light in the vacuum and $\lambda^{(1)}$ is the wavelength in the medium surrounding the sphere.

For the purpose of later discussion it is also convenient to introduce the complex refractive index of the sphere relative to the surrounding medium. Denoting this index by $\hat{n}$ we have

$$\hat{n}^2 = \frac{\hat{n}^{(\mathrm{II})^2}}{n^{(\mathrm{I})^2}} = \frac{k^{(\mathrm{II})^2}}{k^{(\mathrm{I})^2}} = \frac{\varepsilon^{(\mathrm{II})}}{\varepsilon^{(\mathrm{I})}} + i\frac{4\pi\sigma}{\omega\varepsilon^{(\mathrm{I})}} = \frac{k_1^{(\mathrm{II})}}{k_1^{(\mathrm{I})}} \tag{60}$$

We also introduce a dimensionless parameter $q$ defined as

$$q = \frac{2\pi}{\lambda^{(\mathrm{I})}}a \tag{61}$$

i.e. $q$ is $2\pi$ times the ratio of the radius of the sphere to the wavelength of the light in the outer medium. Then, using also the relation

$$\frac{k^{(\mathrm{II})}k_2^{(\mathrm{I})}}{k^{(\mathrm{I})}k_2^{(\mathrm{II})}} = \hat{n}$$

we may express the coefficients (57) in the form

$$\left.\begin{aligned} {}^{e}B_l &= i^{l+1}\tfrac{2l+1}{l(l+1)}\tfrac{\hat{n}\psi_l'(q)\psi_l(\hat{n}q)-\psi_l(q)\psi_l'(\hat{n}q)}{\hat{n}\zeta_l{}^{(1)}(q)\psi_l(\hat{n}q)-\zeta_l{}^{(1)}(q)\psi_l'(\hat{n}q)}, \quad \text{(a)} \\ {}^{m}B_l &= i^{l+1}\tfrac{2l+1}{l(l+1)}\tfrac{\hat{n}\psi_l(q)\psi_l'(\hat{n}q)-\psi_l'(q)\psi_l(\hat{n}q)}{\hat{n}\zeta_l{}^{(1)}(q)\psi_l'(\hat{n}q)-\zeta_l{}^{(1)\prime}(q)\psi_l(\hat{n}q)}. \quad \text{(b)} \end{aligned}\right\} \tag{62}$$

These formulae take a particularly simple form when either the dielectric constant or the conductivity of the sphere is high, and at the same time the radius of the sphere is not too small. In this case $|\hat{n}| \gg 1, |\hat{n}q| \gg 1$ and (62) reduce to

$$\left.\begin{aligned} {}^{e}B_l &= i^{l+1}\tfrac{2l+1}{l(l+1)}\tfrac{\psi_l'(q)}{\zeta_l^{(1)\prime}(q)}, \quad \text{(a)} \\ {}^{m}B_l &= i^{l+1}\tfrac{2l+1}{l(l+1)}\tfrac{\psi_l(q)}{\zeta_l^{(1)}(q)}. \quad \text{(b)} \end{aligned}\right\} \tag{63}$$

This approximation is of little interest for optics, but it is of importance in connection with radio waves. It is also of historical interest as the early theories were concerned with this limiting case.*

(c) Summary of formulae relating to the associated Legendre functions and to the cylindrical functions

For convenience of further discussion we summarize here some formulae relating to the spherical harmonics and to the cylindrical functions.

**Associated Legendre functions**

The Legendre polynomials are the polynomials (in $\cos\theta$ )

$$P_l(\cos\theta) = \sum_{m=0}^{[l/2]}(-1)^m\frac{(2l-2m)!}{2^l m!(l-m)!(l-2m)!}(\cos\theta)^{l-2m}, \tag{64}$$

and the associated Legendre functions of the first kind are defined by the formula*

$$P_l^{(m)}(\cos\theta) = (\sin\theta)^m\frac{d^m P_l(\cos\theta)}{d(\cos\theta)^m} \tag{65}$$

0

• Cf. K. Schwarzschild, Mnch. Akad., Math-phys. Kl., 31 (1901), 293.

We shall also need the relations

$$\left.\begin{aligned} P_l^{(1)}(\cos\theta) &= \tfrac{l}{\sin\theta}\left\{P_{l-1}(\cos\theta) - \cos\theta P_l(\cos\theta)\right\} \\ P_l^{(1)}(\cos\theta) &= \tfrac{\cos\theta}{\sin^2\theta}\left\{P_l^{(1)}(\cos\theta) - l(l+1)\tfrac{P_l(\cos\theta)}{\sin\theta}\right\} \end{aligned}\right\} \tag{66}$$

For large $l$ one has the asymptotic approximation

$$P_l(\cos\theta) \sim \sqrt{\frac{2}{l\pi\sin\theta}}\sin\left[\left(l+\frac{1}{2}\right)\theta + \frac{\pi}{4}\right]. \tag{67}$$

**Cylindrical functions**

I. For small values of the argument $x$, we have, for $\psi_l(x)$ the series expansion

$$\psi_l(x) = \frac{x^{l+1}}{1.3\ldots(2l+1)} f_l(x), \tag{68}$$

where

$$f_l(x) = 1 - \frac{2}{2l+3}\left(\frac{x}{2}\right)^2 + \ldots \tag{69}$$

For $\zeta_l{}^{(1)}(x)$ we have the expansion

$$\zeta_l^{(1)}(x) = -i\frac{1.3\cdots(2l-1)}{x^l} e^{ix}\left\{h_l(x) - ixg_l(x)\right\}, \tag{70}$$

where $h_l(x)$ and $g_l(x)$ are power series of which the first term is unity and the second term is quadratic in $x$. The functions $\psi_l'(x)$ and $\zeta_l^{(1)\prime}(x)$ may be expressed in an analogous form:

$$\psi_l'(x) = \frac{(l+1)x^l}{1.3\cdots(2l+1)} f_l^+(x), \tag{71}$$

$$\zeta_l^{(1)\prime(x)} = il\frac{1.3\cdots(2l-1)}{x^{l+1}} e^{ix}\left\{h_l^+(x) - ig^+(x)\right\}, \tag{72}$$

where ${f_l}^+(x), {h_l}^+(x)$, and ${g_l}^+(x)$ are power series of the same kind as before.
II. For large values of the argument $x$, provided that $l$ is small compared with $|x|$, the following asymptotic formulae may be used:

$$\psi_l(x) \sim \frac{1}{2}\left\{i^{l+1}e^{-ix} + (-i)^{l+1}e^{ix}\right\}, \tag{73}$$

$$\zeta_l^{(1)}(x) \sim (-i)^{l+1}e^{ix}, \tag{74}$$

and

$$\psi_l'(x) \sim \frac{1}{2}\left\{i^l e^{-ix} + (-i)^l e^{ix}\right\}, \tag{75}$$

$$\zeta_l^{(1)\prime}(x) \sim (-i)^l e^{ix}. \tag{76}$$

0

- A slightly different definition is sometimes used, differing from the present one by a multiplicative factor $(-1)^m$.

For real values of $x$ the functions $\psi_l(x)$ and $\psi_l'(x)$ are themselves real:

$$\psi_l(x) \sim \sin\left(x - \frac{l\pi}{2}\right), \tag{77}$$

$$\psi_l'(x) \sim \cos\left(x - \frac{l\pi}{2}\right). \tag{78}$$

### 13.5.2 Some consequences of Mie's formulae

(a) The partial waves It is seen from (58) that the amplitude of the radial components ${E_r}^{(s)}$ and ${H_r}^{(s)}$ of the scattered wave fall off as the inverse square of the distance from the scattering centre, whereas the amplitudes of the other components fall off more slowly, as the inverse of this distance. At sufficiently great distances ( $r \gg \lambda$ ), in the radiation zone or wave zone, the radial components may, therefore, be neglected in comparison with the tangential components, i.e. in this region the wave is transverse.

The formulae show that the scattered wave is composed of contributions involving spherical harmonics of different orders. These contributions may be called partial waves and their strength is determined by the absolute values of the complex coefficients ${}^6B_l$ and ${}^mB_l$. These coefficients depend on the nature of the two media, and on the ratio of the radius of the sphere to the wavelength of the incident light.

Each partial wave consists of an electrical part with amplitude ${}^eB_l$ and a magnetic part with amplitude ${}^mB_l$. The magnetic lines of

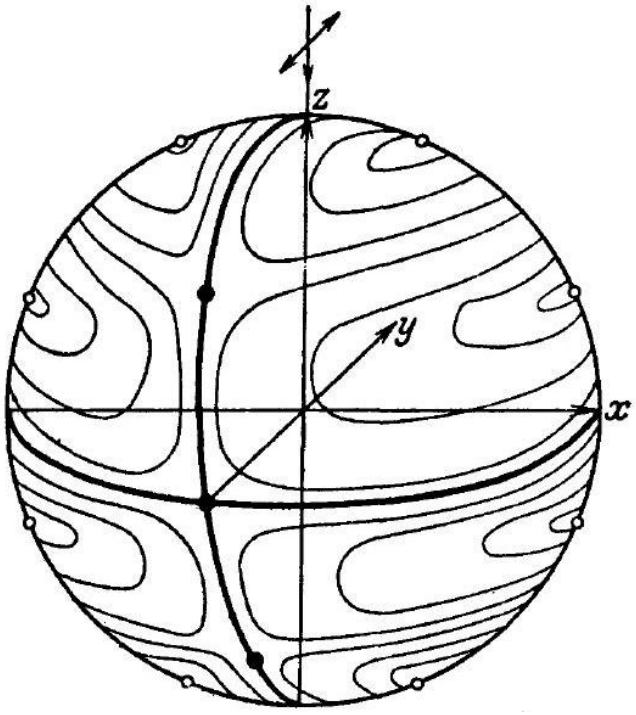



Fig. 13.8. Lines of magnetic force for the fourth electric partial wave.

force of the electric partial wave and the electric lines of force of the magnetic partial wave lie wholly on concentric spherical surfaces, since for the former ${H_r}^{(s)} = 0$ and for the latter ${E_r}^{(s)} = 0$.

Let us consider a typical partial wave, for instance the $l$ th electric wave. We see that the corresponding components ${E_\theta}^{(s)}$ and ${H_\phi}^{(s)}$ vanish at points where either $\cos\phi$ or ${P_l}^{(1)}(\cos\theta)\sin\theta$

is zero; likewise ${E_\phi}^{(s)}$ and ${H_\theta}^{(s)}$ vanish where either $\sin\phi$ or $P_l^{(1)}(\cos\theta)/\sin\theta$ is zero. Now within the interval $0 \leqslant \theta \leqslant \pi$, the function $P_l^{(1)\prime}(\cos\theta)$ vanishes $l$ times and the function $P_l^{(1)}(\cos\theta)/\sin\theta$ vanishes $(l-1)$ times, but differs from zero when $\theta = 0$ or $\pi$. It follows that for $\phi = \pm\pi/2$ all the field-components vanish $2l$ times, i.e. $(\ 4l-2\ )$ times in all. As the magnetic lines of force must be closed curves, and as we saw they lie wholly on spherical surfaces concentric with the origin it follows that each of the $2l$ zero points on the circle $\phi = 0$ or $\pi$ are the centres of families of closed magnetic loops, whereas the $2(l-1)$ zero points lying on the circle $\phi = \pm\pi/2$ are neutral points. The lines of force avoid these neutral points just as two families of rectangular hyperbolae having common asymptotes avoid their common centre.

Fig. 13.8 shows the magnetic lines of force for the fourth electrical partial wave. The two sets of points may be clearly distinguished; the first in the $xz$-section (the plane of the diagram) and the second in the $yz$-section.

Fig. 13.9(a) shows the projections on to the $yz$-plane of the magnetic lines of force situated on one of the two hemispheres on either side of the $yz$-plane, and belonging to the first four electric partial waves. In the radiation zone the electric lines of force are

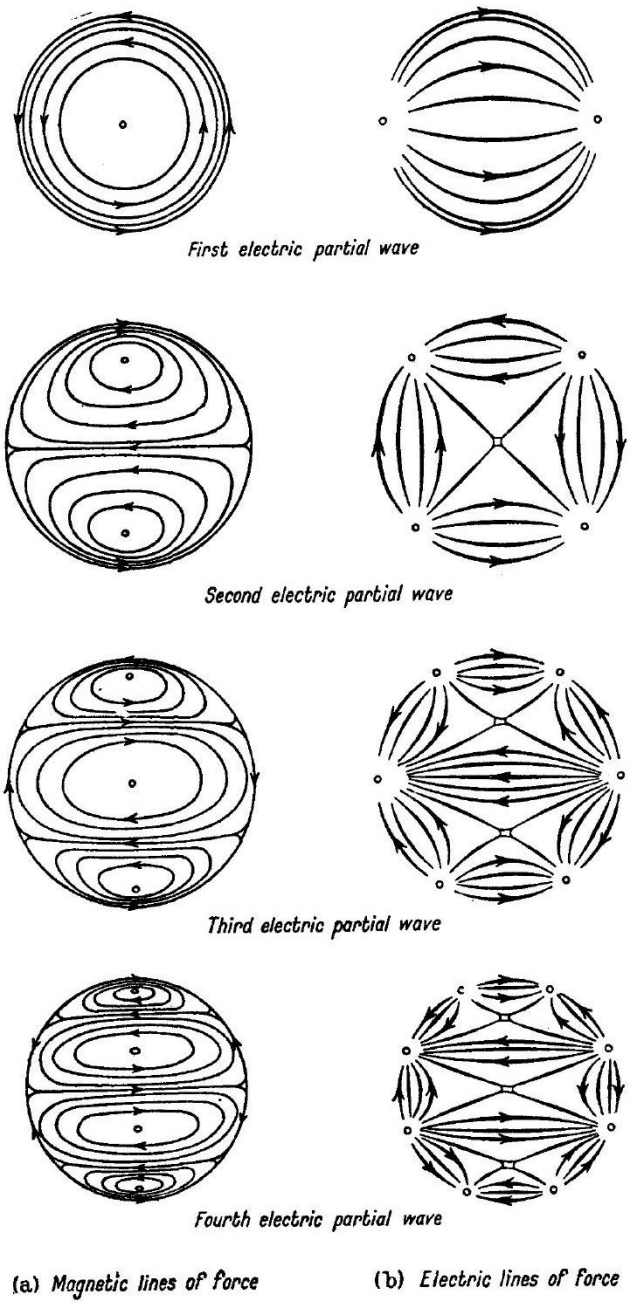



Fig. 13.9. Lines of force of the first four electric partial waves.
(After G. Mie, Ann. d. Physik (4), 25 (1908), 377.)

orthogonal to the magnetic lines, since, according to (58), one has, for each partial wave (electric or magnetic),

$$E_\theta{}^{(s)} H_\theta{}^{(s)} + E_\phi{}^{(s)} H_\phi{}^{(s)} = 0. \tag{79}$$

Fig. 13.9(b) shows these lines in a similar projection on the $xz$-plane.

Similar results hold for the magnetic partial waves, except that $\cos\phi$ and $\sin\phi$ are now interchanged. The corresponding projections of the electric lines of force of the magnetic partial waves that lie on a sphere may be obtained merely by rotating the figures through an angle of 90° about the $z$-axis.

(b) Limiting cases We shall now investigate the relative contributions of the various partial waves. The general case ( $\hat{n}$ and $q$ arbitrary) is not amenable to simple analytical treatment

and we, therefore, consider in detail only two limiting cases, namely when the radius of the sphere is large compared with the wavelength ($q \gg 1$), and when it is much smaller than a wavelength ($q \ll 1$).

I: $q \gg 1$. This is the case for which our solution must be expected to give substantially the same results as the Huygens-Kirchhoff diffraction theory, or even (as $q \to \infty$ ) the same results as geometrical optics.

If we restrict ourselves to orders $l$ that are very much less than $q$ and $|\hat{n}q|$ then we may use the asymptotic approximations (73)-(76). We have

$$\frac{\psi_l(\hat{n}q)}{\psi_l'(\hat{n}q)} \sim \frac{i^{l+1}e^{-inq} + (-i)^{l+1}e^{i\hat{n}q}}{i^l e^{-i\hat{n}q} + (-i)^l e^{i\hat{n}q}} = \frac{\cos\left[\hat{n}q - (l+1)\frac{\pi}{2}\right]}{\cos\left[\hat{n}q - l\frac{\pi}{2}\right]} = \tan\left[\hat{n}q - l\frac{\pi}{2}\right], \tag{80}$$

and the coefficients (62) becomes

$$\left.\begin{aligned} {}^eB_l &= (-1)^{l+1}\tfrac{2l+1}{l(l+1)}e^{-iq}\tfrac{\sin\left[q-l\frac{\pi}{2}\right]-\hat{n}\cos\left[q-l\frac{\pi}{2}\right]\tan\left[\hat{n}q-l\frac{\pi}{2}\right]}{1-i\hat{n}\tan\left[\hat{n}q-l\frac{\pi}{2}\right]} \\ {}^mB_l &= (-1)^{l+1}\tfrac{2l+1}{l(l+1)}e^{-iq}\tfrac{\hat{n}\sin\left[q-l\frac{\pi}{2}\right]-\cos\left[q-l\frac{\pi}{2}\right]\tan\left[\hat{n}q-l\frac{\pi}{2}\right]}{\hat{n}-i\tan\left[\hat{n}q-l\frac{\pi}{2}\right]} \end{aligned}\right\} \tag{81}$$

We note that these coefficients are rapidly oscillating functions of both $q$ and $l$, so that a small change in $q$ or $l$ may give rise to a large change in ${}^eB_l$ and ${}^mB_l$. We also observe that ${}^mB_{l+1}$ and ${}^eB_l$ are of the same order of magnitude, i.e. the amplitude of the electric partial wave of order $l$ is of the same order of magnitude as that of the magnetic partial wave of next higher order.

Since the formulae (81) were derived on the assumption that $l$ is much smaller than $q$, one cannot determine from these approximations the number of partial waves that make an appreciable contribution to the scattered field. Debye* derived asymptotic approximations valid for all orders and showed that the amplitudes of the partial waves fall off rapidly to zero as soon as $l + \frac{1}{2}$ exceeds $q$, so that only the first $q$ terms need be included.

When either the dielectric constant or the conductivity of the sphere is very large ($|\hat{n}| \to \infty$), the expressions for ${}^eB_l$ and ${}^mB_l$ reduce to

$$\left.\begin{aligned} {}^eB_l &= i(-1)^{l+1}\tfrac{2l+1}{l(l+1)}e^{-iq}\cos\left(q - l\tfrac{\pi}{2}\right), \\ {}^mB_l &= (-1)^{l+1}\tfrac{2l+1}{l(l+1)}e^{-iq}\sin\left(q - l\tfrac{\pi}{2}\right), \end{aligned}\right\} \tag{82}$$

derived most simply from (63) with the help of the asymptotic approximations (73)-(76).

The approximations of this section ($q \gg 1$) may be applied in connection with the theory of the rainbow*; the parameter $q$, determined by the size of the raindrops, is of the order $10^4$.

II. $q \ll 1$. This case is of great practical importance in connection with microscopic and submicroscopic particles in colloidal solutions. We may now use the power series expansions (68)-(72) for the cylindrical functions and obtain (restricting ourselves to terms of the leading order),

$$\left.\begin{aligned} {}^eB_l &\sim i^l \tfrac{q^{2l+1}}{l^2[1.3.5...(2l-1)]^2}\tfrac{\hat{n}^2-1}{\hat{n}^2+\frac{l+1}{l}} \\ {}^mB_l &\sim i^l \tfrac{q^{2l+3}}{l(l+1)(2l+1)(2l+3)[1.3.5...(2l-1)]^2}(\hat{n}^2-1) \end{aligned}\right\} \tag{83}$$

When either the conductivity or the dielectric constant is very large, we have from (63) †

0

- P. Debye, Ann. d. Physik (4), 30 (1909), 118; Mat. Ann., 67 (1909), 535; Sitzungsb. Mnch. Akad. Wiss, Math. Phys. Kl., 5 Abh. (1910).

$$\left.\begin{aligned} {}^{e}B_l &= i^l \tfrac{q^{2l+1}}{l^2[1.3.5...(2l-1)]^2} \\ {}^{m}B_l &= -i^l \tfrac{q^{2l+1}}{l(l+1)[1.3.5...(2l-1)]^2}. \end{aligned}\right\} \tag{84}$$

As long as the conductivity remains finite [formulae (83)], ${}^{e}B_{l+1}$ and ${}^{m}B_l$ are proportional to the same power of $q$, i.e. the amplitude of the $(l+1)$ th electric partial wave is of the same order of magnitude as the amplitude of the $l$ th magnetic partial wave, just as was the case when $q \gg 1$. If the radius of the sphere is so small that $q^2$ may be neglected in comparison with unity, only the first electric partial wave need be considered and its strength (and phase) is given by the complex amplitude

$${}^{e}B_1 = iq^3 \frac{\hat{n}^2-1}{\hat{n}^2+2} = i\left(\frac{2\pi a}{\lambda^{(1)}}\right)^3 \frac{\hat{n}^2-1}{\hat{n}^2+2} \tag{85}$$

Since ${}^{e}B_l$ is complex there is a phase difference between the incident primary field and the scattered secondary field.

On the other hand, in the limiting case $|\hat{n}| \to \infty$, the amplitude of the $l$ th electric partial wave and the amplitude of the $l$ th magnetic partial wave are of the same order of magnitude, as is seen from (84).

Let us now consider, for $|\hat{n}|$ finite, the field of the first electric partial wave, at large distances from the sphere ($r \gg \lambda$). We use in (58) the asymptotic approximations

$$\zeta_1^{(1)}(x) = -e^{ix}, \quad \zeta_1^{(1)'}(x) = -ie^{ix} \tag{86}$$

and the relations

$$P_1(\cos\theta) = \cos\theta, \quad P_1{}^{(1)}(\cos\theta) = \sin\theta, \quad P_1{}^{(1)\prime}(\cos\theta)\sin\theta = -\cos\theta \tag{87}$$

where the prime denotes differentiation with respect to $\cos\theta$, and obtain,

$$\left.\begin{aligned} E_\theta^{(s)} &= -\tfrac{i}{k^{(\mathrm{I})}}\cos\phi\cos\theta {}^{e}B_1 \tfrac{e^{ik^{(\mathrm{I})}r}}{r} = \left(\tfrac{2\pi}{\lambda^{(\mathrm{I})}}\right)^2 a^3 \tfrac{\hat{n}^2-1}{\hat{n}^2+2}\cos\phi\cos\theta \tfrac{e^{ik^{(\mathrm{I})}r}}{r}, \\ E_\phi^{(s)} &= \tfrac{i}{k^{(\mathrm{I})}}\sin\phi {}^{e}B_1 \tfrac{e^{ik^{(\mathrm{I})}r}}{r} = -\left(\tfrac{2\pi}{\lambda^{(\mathrm{I})}}\right)^2 a^3 \tfrac{\hat{n}^2-1}{\hat{n}^2+2}\sin\phi \tfrac{e^{ik^{(\mathrm{I})}r}}{r}, \\ H_\theta^{(s)} &= \tfrac{1}{k_2^{(\mathrm{I})}}\sin\phi {}^{e}B_1 \tfrac{e^{ik^{(\mathrm{I})}r}}{r} = \left(\tfrac{2\pi}{\lambda^{(\mathrm{I})}}\right)^2 \sqrt{\varepsilon^{(\mathrm{I})}} a^3 \tfrac{\hat{n}^2-1}{\hat{n}^2+2}\sin\phi \tfrac{e^{ik^{(\mathrm{I})}r}}{r}, \\ H_\phi^{(s)} &= \tfrac{1}{k_2^{(\mathrm{I})}}\cos\phi\cos\theta {}^{e}B_1 \tfrac{e^{ik^{(\mathrm{I})}r}}{r} = \left(\tfrac{2\pi}{\lambda^{(\mathrm{I})}}\right)^2 \sqrt{\varepsilon^{(\mathrm{I})}} a^3 \tfrac{\hat{n}^2-1}{\hat{n}^2+2}\cos\phi\cos\theta \tfrac{e^{ik^{(\mathrm{I})}r}}{r}. \end{aligned}\right\} \tag{88}$$

It is easy to see that equations (88) are identical with those for the radiation zone ($r \geqslant \lambda$) of an electric dipole at $O$, oscillating parallel to the $x$-axis, i.e. parallel to the electric vector of the primary incident field, and of moment $p = p_0 e^{-i\omega t}$ where

$$p_0 = a^3 \left|\frac{\hat{n}^2-1}{\hat{n}^2+2}\right| \tag{89}$$

To show this we assume first that the outer medium is vacuum $\left(\varepsilon^{(1)}=1\right)$ and we return to equations (53) and (54) of §2.2. According to these equations, the radiation field in vacuum of a linear electric dipole of moment $p_0 e^{-i\omega t}$ is given by

---

0

- Cf. B. van der Pol and H. Bremmer, Phil. Mag., (7) 24 (1937), 857; H. Bucerius, Optik, 1 (1946), 188.
† A direct transition from (83) to (84) is not simple; one has to take into account that as $|\hat{n}| \to \infty, q \to 0$ in such a way that $q^2|\hat{n}|^2$ approaches a finite value.

$$\left.\begin{aligned} E &= -\left(\tfrac{\omega}{c}\right)^2 \tfrac{1}{r^3} r \wedge (r \wedge p_0) \\ H &= \left(\tfrac{\omega}{c}\right)^2 \tfrac{1}{r^2} (r \wedge p_0), \end{aligned}\right\} \tag{90}$$

where the time harmonic factor $e^{-i\omega t}$ has been omitted and $r$ is the position vector of the point of observation with respect to the origin. From (90) it follows that if $p_0$ is the direction of the $x$-axis, the components of $E$ and $H$ are given by

$$\left.\begin{aligned} E_x &= \left(\tfrac{\omega}{c}\right)^2 p_0 \left(\sin^2\theta \sin^2\phi + \cos^2\theta\right) \tfrac{e^{ikr}}{r}, \\ E_y &= -\left(\tfrac{\omega}{c}\right)^2 p_0 \left(\sin^2\theta \sin\phi \cos\phi\right) \tfrac{e^{ikr}}{r}, \\ E_z &= -\left(\tfrac{\omega}{c}\right)^2 p_0 \sin\theta \cos\theta \cos\phi \tfrac{e^{ikr}}{r}, \\ H_x &= 0, \\ H_y &= \left(\tfrac{\omega}{c}\right)^2 p_0 \cos\theta \tfrac{e^{ikr}}{r}, \\ H_z &= -\left(\tfrac{\omega}{c}\right)^2 p_0 \sin\theta \sin\phi \tfrac{e^{ikr}}{r}. \end{aligned}\right\} \tag{91}$$

If now the transformation (8) to spherical polar coordinates is applied, this transforms into (88) (with $\varepsilon^{(I)} = 1$ ) if the expression (89) is substituted for $p_0$. In the more general case, when $\varepsilon^{(1)} \neq 1$, we must consider in place of (91) the corresponding expressions for the radiation field of a dipole in a dielectric. These expressions, which may be obtained in a similar way to (91), transform into (88) when transformation to spherical polar coordinates is made. We note that the refractive index of the sphere enters the expression (89) for the equivalent dipole moment only in the combination $(\hat{n}^2 - 1)/(\hat{n}^2 + 2)$; for dielectric substances ( $n$ real) we have already encountered this expression, in the theory of molecular refractivity (cf. § 2.3 (17)).

Since, according to (88) the amplitude is inversely proportional to the square of the wavelength, the intensity of the scattered light is inversely proportional to the fourth power of the wavelength. In such a case one speaks of Rayleigh scattering.*

The first magnetic partial wave may be similarly described in terms of a vibrating magnetic dipole. The partial waves of higher orders may be considered to be due to vibrating multipoles, but we shall not investigate this here.

(c) Intensity and polarization of the scattered light Let us now return to the general case and examine briefly the intensity and the polarization of the scattered light. As we are only interested in relative values of the intensity we may take as a measure of the intensity the square of the real amplitude of the electric vector. We consider only the distant field ( $r \gg \lambda$ ) and may therefore replace the functions $\zeta_l^{(1)}$ and $\zeta_l^{(1)}$ in (58) by their asymptotic approximations. We also set

$$\left.\begin{aligned} I_l^{(s)} &= \tfrac{\lambda^{(1)2}}{4\pi^2 r^2} \left| \sum_{l=1}^{\infty} (-i)^l \left( {}^e B_l P_l^{(1)}(\cos\theta) \sin\theta - {}^m B_l \tfrac{P_l^{(1)}(\cos\theta)}{\sin\theta} \right) \right|^2, \\ I_\perp{}^{(s)} &= \tfrac{\lambda^{(1)2}}{4\pi^2 r^2} \left| \sum_{l=1}^{\infty} (-i)^l \left( {}^e B_l \tfrac{P_l^{(1)}(\cos\theta)}{\sin\theta} - {}^m B_l P_l^{(1)}(\cos\theta) \sin\theta \right) \right|^2. \end{aligned}\right\} \tag{92}$$

Then

$$\left|E_\theta^{(s)}\right|^2 = I_\sharp^{(s)} \cos^2\phi, \quad \left|E_\phi^{(s)}\right|^2 = I_\perp^{(s)} \sin^2\phi \tag{93}$$

We define the plane of observation as the plane that contains the direction of propagation of the incident light and the direction $(\theta, \phi)$ of observation. According to (4) and (7), $\phi$ represents the angle between this plane and the direction of vibration of the electric vector of the incident wave. Since according to (93) either $E_\theta{}^{(s)}$ or $E_\phi{}^{(s)}$ vanishes when $\phi = 0$ or $\phi = \pi/2$, the scattered light is linearly polarized when the plane of observation is parallel or perpendicular to the primary vibrations. For any other direction $(\theta, \phi)$ the light is in general elliptically polarized, since the ratio $E_\theta{}^{(s)}/E_\phi{}^{(s)}$ is complex. In the special case of Rayleigh scattering,

represented by (88), the ratio $E_\theta^{(s)}/E_\phi^{(s)}$ is, however, always real, so that the scattered light is then linearly polarized for all directions of observation.

In practice one is usually concerned with the scattering of natural light. As in § 1.5 the appropriate formulae may be obtained from those for polarized light by averaging

over all directions of polarization.* Denoting by a bar this average, we now have, in place of (93), since $\overline{\cos^2\phi} = \overline{\sin^2\phi} = \frac{1}{2}$,

$$\overline{|E_\theta^{(s)}|^2} = \frac{1}{2} I_\parallel{}^{(s)}, \quad \overline{|E_\phi^{(s)}|^2} = \frac{1}{2} I_\perp. \tag{94}$$

In general neither $I_1{}^{(s)}$ nor $I_\perp{}^{(s)}$ is zero, so that the scattered light is partially polarized. In analogy with § 1.5 (42) we may define the degree of polarization $P$ of the scattered light as

$$P = \left| \frac{I_\perp{}^{(s)} - I_n{}^{(s)}}{I_\perp{}^{(s)} + I_n{}^{(s)}} \right|. \tag{95}$$

The unpolarized proportion of the scattered light then is

$$I_\perp{}^{(s)} + I_1{}^{(s)}\big)\,(1-P) = 2I_1{}^{(s)} \text{ when } \left.\begin{array}{ll} I_1{}^{(s)} & < I_\perp{}^{(s)}, \\ & = 2I_\perp{}^{(s)} \text{ when } I_1{}^{(s)} > I_\perp{}^{(s)}. \end{array}\right\} \tag{96}$$

The dependence of the intensity and polarization of the scattered light on the direction of scattering and on the physical parameters ( $\lambda, a, \hat{n}$ ) has been investigated on the basis of Mie's theory by many writers † and we cannot do more here than summarize briefly some of the main results obtained.

In Fig. 13.10 and 13.11 the intensity and the unpolarized proportion of the scattered light are shown as functions of the angle $\theta$ of observation for dielectric and metallic spheres of various sizes. The length of the radius vector of the outer curves is proportional to the intensity $I^{(s)} = I_1{}^{(s)} + I_\perp{}^{(s)}$ and that of the inner curves, unless stated otherwise, is proportional to $I_\perp{}^{(s)}$. The units are arbitrary and different in each figure. From these polar diagrams as well as from other published calculations the following general conclusions may be drawn:

Excluding the case when the conductivity or the dielectric constant is very large (in which case most of the incident light is radiated in the backward direction, i.e. "reflected"), the polar diagrams, in the limit of vanishingly small spheres ( $a \to 0$ ), are symmetrical about the plane through the centre of the sphere, at right angles to the direction of propagation of the incident light. There is an intensity maximum in the forward direction ( $\theta = 0$ ) and in the reverse direction ( $\theta = 180°$ ), and there is a minimum in the plane of symmetry ($\theta = 90°$). As the radius of the sphere is increased there is a departure from symmetry, more light being scattered in the

0

- As already pointed out on p. 98, Rayleigh showed, in celebrated papers concerning the blue colour of the sky, that scattering of this type is also caused by spontaneous fluctuations in the density of a homogeneous medium. [Lord Rayleigh, Phil. Mag., (4), XII (1871), 274, 447; (5), XLVII (1899), 375. Reprinted in his Scientific Papers, 1 (1899), 87, 104; 4 (1903), 397.]

0

- Alternatively we may use the result of § 10.8 .2 , according to which a wave of natural light may be regarded as composed of two incoherent waves of equal amplitudes, propagated in the same direction and polarized in directions at right angles to each other. The scattering of each wave may be determined separately and the total intensity is obtained by summing the intensities of the two component waves. This leads again to (94).

  † In addition to the paper by Mre already quoted we may refer to the following:
  R. Gans, Ann. d. Physik (4), 76 (1925), 29; H. Senftleben and E. Benedict, Ann. d. Physik (4), 60

forward direction than in the opposite direction. This phenomenon is often called the Mie effect. As the radius is increased still further practically all the scattered light appears around the forward direction $\theta = 0$; likewise for a conducting sphere there is a greater concentration of light in this direction. When the radius of the sphere is very large compared to the wavelength most of the incident light is, however, reflected, as follows from geometrical optics.

The dependence of the intensity of the scattered light on the radius of the sphere is illustrated in Table XIX. The Mife effect may be clearly seen from the comparison of

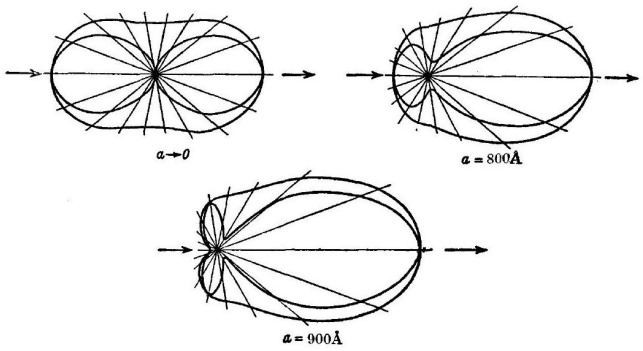



Fig. 13.10. Polar diagrams for the scattering of linearly polarized light by a spherical gold particle ( $\lambda_{0}=5500 \AA, n\hat{}\{\}\{(\mathrm{I})\}=1 \cdot 33, \hat{n}\hat{}\{\}\{(\mathrm{II})\}=0.57+2 \cdot 45 i$ ).

(After G. Mie, Ann. d. Physik, (4) 25 (1908), 429.)

the first and third row. The table indicates a very rapid increase in intensity with increasing size of the sphere; for true comparison the values in the table must be multiplied by the factor $\lambda^{(1)^2}/4\pi^2a^2 = 1/q^2$.

Table XIX

The normalized intensity $4\pi^2a^2\left(I_\perp{}^{(s)} + I_n{}^{(s)}\right)/\lambda^{(\mathrm{I}^2)}$ of light scattered by dielectric spheres of refractive index $n = 1 \cdot 25$, as function of the parameter

$q = 2\pi a/\lambda^{(1)}$

(Compiled from calculations of H. Blumer, Z. f. Phys., 38 (1926), 304)

| $\theta$ | $q = 0.01$ | $q = 0.1$ | $q = 0.5$ | $q = 1$ | $q = 2$ | $q = 5$ | $q = 8$ |
|---|---|---|---|---|---|---|---|
| 0 | $5 \cdot 0 \cdot 10^{-14}$ | $5 \cdot 0 \cdot 10^{-8}$ | $1 \cdot 2.10^{-3}$ | $2 \cdot 3 \cdot 10^{-1}$ | $4 \cdot 3$ | $9 \cdot 8 \cdot 10^{2}$ | $7 \cdot 5 \cdot 10^{3}$ |
| 90° | $2 \cdot 5 \cdot 10^{-14}$ | $2 \cdot 5 \cdot 10^{-8}$ | $5 \cdot 0 \cdot 10^{-4}$ | $3 \cdot 6 \cdot 10^{-2}$ | $2 \cdot 5 \cdot 10^{-1}$ | 2.7 | $7 \cdot 1$ |
| 180° | $5 \cdot 0 \cdot 10^{-14}$ | $4 \cdot 9 \cdot 10^{-8}$ | $7 \cdot 8 \cdot 10^{-4}$ | $1 \cdot 9 \cdot 10^{-3}$ | $2 \cdot 0 \cdot 10^{-2}$ | 1.3 | 0.9 |

When $q$ exceeds unity, i.e. when the diameter $2a$ of the sphere is greater than $\lambda^{(\mathrm{I})}/\pi$ there appear a series of maxima and minima, which at first are distributed irregularly. The

(1919), 297; H. Blumer, Z. f. Phys., 32 (1925), 119; 38 (1926), 304, 920 ; 39 (1926), 195; C. Schaln, Uppsala Astr. Obs. Ann., 1, No. 2 (1939); ibid., 1, No. 9 (1945); G. R. Paranjpe, Y. G. Naik, and P. B. Vaidya, Proc. Indian Acad., A, 9 (1939), 333, 352; H. Holl, Optik, 1 (1946), 213; ibid., 4 (1948/49), 173. A very full survey of results obtained by many writers will be found in a paper by H. C. van de Hulst, Rech. Astr. Observ. Utrecht, 11, Pt. I (1946) and in his book Light Scattering by Small Particles (New York, John Wiley and Sons; London, Chapman and Hall, 1957). See also G. Oster, Chem. Rev., 43 (1948), 319.

Tables of scattering functions for spherical particles, National Bureau of Standards (Washington, D.C. 1949), Applied Mathematics Series, 4; R. O. Gumprecht and C. M. Sliepcevich, LightScattering Functions for Spherical Particles (Ann Arbor, University of Michigan Press, 1951).

appearance of a number of maxima and minima when $q$ is large is in agreement with the Huygens-Kirchhoff theory.

Concerning polarization of the scattered light the results again are different depending on whether or not $|\hat{n}|$ is large.

For very small spheres which are highly conducting ( $\sigma \to \infty$ ) or have very high dielectric constant ($\varepsilon \to \infty$) the polarization is largest when $\theta = 60°$ (the Thomson

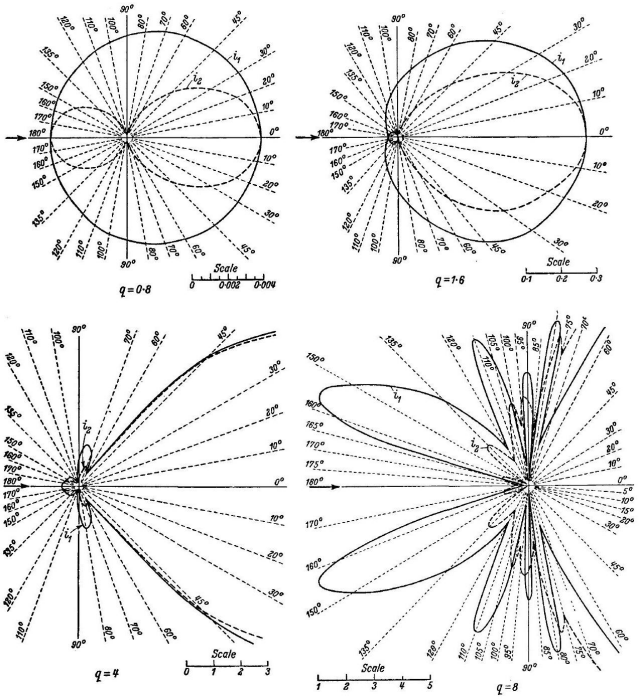



Fig. 13.11. Polar diagrams for the scattering of linearly polarized light by a dielectric sphere of refractive index $n=1 \cdot 25$ . $i_{1}=q\hat{}\{2\} I_{\perp}$, $i_{2}=q\hat{}\{2\} I_{|}$.

(After H. Blumer, Z. f. Phys., 32 (1925), 119.)

angle). With increasing radius the maximum is displaced in the direction of increasing $\theta$.

The dependence of the polarization on the angle $\theta$ of observation, for spheres of finite conductivity and finite dielectric constant, is illustrated in two typical cases in Fig. 13.10 and 13.11. When the radius of the sphere is very small ( $q \to 0$ ), the polarization diagram is, like the intensity diagram, symmetrical about the $xy$-plane

and has a maximum for $\theta = 90°$, where the polarization is complete. In this case (Rayleigh scattering), the degree of polarization may be represented by a single analytical expression, obtained on substituting from (88) and (93) into (95); this gives

$$P(\theta) = \frac{\sin^2\theta}{1 + \cos^2\theta} \tag{97}$$

This formula was derived by Rayleigh, in a different way.

As the radius of the sphere is increased, up to about $a = \lambda^{(I)}/\pi$, the maximum is displaced; in the majority of the cases that have been investigated the displacement is in the direction of larger $\theta$ for dielectric spheres and in the direction of smaller $\theta$ for absorbing spheres. When the radius of the sphere is increased still further, there appears an irregular sequence of polarization maxima.

In the direction $\theta = 90°$ the light is, for $q < 1$, almost completely polarized, with its electric vector perpendicular to the plane of observation; for larger values of $q$ this is no longer the case and the behaviour becomes irregular.

So far we have confined our attention to monochromatic light. One often deals with scattering of polychromatic light and we must therefore also consider the effects arising from the presence of components of different wavelengths. We note that the wavelength enters our formulae only through the parameter $q$ and through the refractive index $\hat{n}$. In a sufficiently small range of wavelengths, $\hat{n}$ is practically independent of the wavelength if in (60) the term that contains the conductivity $\sigma$ is small compared with the other term, i.e. for a poorly conducting sphere. On the other hand, in the limit of infinitely high conductivity, $\hat{n}$ does not enter at all. In these cases the intensities of the spectral components depend on $a/\lambda^{(\mathrm{I})}$ only. The effect of changing the wavelength is thus substantially equivalent to the effect that arises from changing the radius of the sphere by an appropriate amount. Since for different wavelengths the polarization maxima occur at different angles of observations, complicated colour changes are seen when observations on scattered light are made through a polarizing prism. This effect is called polychroism. The dependence of polarization of the scattered light on wavelength-known as dispersion of polarization-affords a very precise test of the theory.*

### 13.5.3 Total scattering and extinction

(a) Some general considerations It is of considerable practical interest to determine the total amount of light that is scattered or absorbed by the sphere. This may be calculated by evaluating the Poynting vector and integrating it over all directions. With the help of the orthogonality relations that exist between the associated Legendre functions it is possible to express the integrals in terms of the coefficients ${}^{e}B_l$ and ${}^{m}B_l$. These calculations, which are somewhat lengthy, are carried out in full in the paper by Mie. †

The total energy lost from the incident wave, i.e. the sum of the scattered and the absorbed energy, may be determined in an alternative way, from certain general considerations that apply to an obstacle of any shape whatsoever. These considerations show that there is a close connection between the energy loss and the amplitude of the scattered wave in the forward direction ( $\theta = 0$ ). From this

result, which we shall now establish, and from Mie's formula for the scattered wave, the total energy lost by scattering and absorption from a sphere may then easily be determined.

Consider a plane monochromatic wave incident on an obstacle of arbitrary form, embedded in a dielectric medium. The field at any point in the medium surrounding the obstacle may again be represented as the sum of the incident field and the scattered field

$$\boldsymbol{E} = \boldsymbol{E}^{(i)} + \boldsymbol{E}^{(s)}, \quad \boldsymbol{H} = \boldsymbol{H}^{(i)} + \boldsymbol{H}^{(s)}. \tag{98}$$

As usually, we omit a time factor $\exp(-i\omega t)$. The time averaged energy flow is represented by the averaged Poynting vector, which, according to (98) and the formula (56) of § 1.4, is given by

$$\langle \boldsymbol{S} \rangle = \left\langle \boldsymbol{S}^{(i)} \right\rangle + \left\langle \boldsymbol{S}^{(s)} \right\rangle + \langle \boldsymbol{S}' \rangle \,, \tag{99}$$

where ( $\mathscr{R}$ denoting the real part)

0

- The dispersion of polarization and polychroism was investigated by M. A. Schirmann Ann. d. Physik, (4), 59 (1919), 493.
  † G. Min, loc. cit., pp. 432-436.

$$\left\langle S^{(i)}\right\rangle = \frac{c}{8\pi}\mathscr{R}\left(E^{(i)} \wedge H^{(i)\star}\right) \tag{100a}$$

$$\left\langle S^{(s)}\right\rangle = \frac{c}{8\pi}\mathscr{R}\left(E^{(s)} \wedge H^{(s)\star}\right) \tag{100b}$$

$$\langle S'\rangle = \frac{c}{8\pi}\mathscr{R}\left(E^{(i)} \wedge H^{(s)\star} + E^{(s)} \wedge H^{(i)\star}\right) \tag{100c}$$

Consider the averaged outward flow of energy through the surface of a large sphere of radius $R$, centred at some point in the region occupied by the obstacle. The net flow per second is represented by the integral of the radial component $\langle \boldsymbol{S}\rangle_r$ of $\langle \boldsymbol{S}\rangle$, taken over the sphere, and is evidently zero when the obstacle is dielectric. If, however, the obstacle is a conductor, some of the incident energy is absorbed by it and the net outward flow through the surface of the sphere is equal in magnitude to the rate at which absorption takes place. Let $\mathscr{W}^{(a)}$ be the rate at which the energy is being absorbed by the obstacle. Then, from (99)

$$-\mathscr{W}^{(a)} = \mathscr{W}^{(i)} + \mathscr{W}^{(s)} + \mathscr{W}', \tag{101}$$

where $\mathscr{W}^{(i)}, \mathscr{W}^{(s)}$, and $\mathscr{W}'$ are the integrals of the radial components $\left\langle S^{(i)}\right\rangle_r, \left\langle S^{(s)}\right\rangle_r$, and $\langle \boldsymbol{S}'\rangle_r$ over the surface of the sphere. Now since the medium that surrounds the obstacle is assumed to be non-conducting, $\mathscr{W}^{(i)} = 0$, so that

$$\mathscr{W}^{(a)} + \mathscr{W}^{(s)} = -\mathscr{W}' = -\frac{c}{8\pi}\mathscr{R}\iint_S \left(E^{(i)} \wedge H^{(s)\star} + E^{(s)} \wedge H^{(i)\star}\right).ndS \tag{102}$$

$S$ denoting the large sphere and $n$ the unit outward normal. Thus the expression on the right of (102) represents the rate at which energy is dissipated by heat and scattering.

Let $n_0$ be the unit vector in the direction in which the incident wave is propagated, so that

$$E^{(i)} = ee^{ik^{(1)}(n_0\cdot r)} \quad , \quad H^{(i)} = he^{ik^{(1)}(n_0\cdot r)} \tag{103}$$

We assume that this wave is linearly polarized, so that $e$ and $h$ may be assumed to be real constant vectors. At a large distance from the obstacle the scattered wave may be assumed to be spherical:

$$E^{(s)} = a(n)\frac{e^{ik^{(1)}r}}{r}, \quad H^{(s)} = b(n)\frac{e^{ik^{(1)}r}}{r} \tag{104}$$

The vectors $\boldsymbol{a}(\boldsymbol{n})$ and $\boldsymbol{b}(\boldsymbol{n})$ characterize the strength of the radiation scattered in the direction $n$. Since the incident and scattered wave obey Maxwell's equations we have (cf. eq. (4), (5) of § 1.4)

$$\left.\begin{aligned} h &= \sqrt{\varepsilon^{(\mathrm{I})}}n_0 \wedge e, & b &= \sqrt{\varepsilon^{(\mathrm{I})}}n \wedge a, \\ n_0 \cdot e &= n_0 \cdot h = 0, & n \cdot a &= n \cdot b = 0, \end{aligned}\right\} \tag{105}$$

where $\varepsilon^{(1)}$ is the dielectric constant of the outer medium, assumed to be non-magnetic ($\mu = 1$). From these relations it follows that on the surface of the large sphere $S$,

$$\left.\begin{aligned} \left(E^{(i)} \wedge H^{(s)\star}\right)\cdot n &= \sqrt{\varepsilon^{(\mathrm{I})}}e \cdot a^{\star}e^{ik^{(\mathrm{I})}R(n_0\cdot n)}\frac{e^{-ik^{(\mathrm{I})}R}}{R} \\ \left(E^{(s)} \wedge H^{(i)\star}\right)\cdot n &= \sqrt{\varepsilon^{(\mathrm{I})}}\left[(n\cdot n_0)(a\cdot e) - (n\cdot e)(n_0\cdot a)\right]e^{-ik^{(\mathrm{I})}R(n_0\, n)}\frac{e^{ik^{(\mathrm{I})}R}}{R}. \end{aligned}\right\} \tag{106}$$

We substitute these expressions in (102). To evaluate the resulting integral we use the mathematical lemma,* that when $R$ is large, and $f$ is an arbitrary function of $n$,

$$\frac{1}{R}\iint_S f(n)e^{-ik^{(1)}R(n_0.n)}dS \sim \frac{2\pi i}{k^{(1)}}\left[f(n_0)\,e^{-ik^{(1)}R} - f(-n_0)\,e^{ik^{(1)}R}\right] \tag{107}$$

We then obtain

$$\left.\begin{aligned}\iint_S \left(E^{(i)} \wedge H^{(s)\star}\right)\cdot n dS &\sim -\tfrac{2\pi i}{k^{(\mathrm{I})}}\sqrt{\varepsilon^{(\mathrm{I})}}\left[e\cdot a^{\star}(n_0) - e\cdot a^{\star}(-n_0)\,e^{-2ik^{(\mathrm{I})}R}\right],\\ \iint_S \left(E^{(s)} \wedge H^{(i)\star}\right)\cdot n dS &\sim \tfrac{2\pi i}{k^{(\mathrm{I})}}\sqrt{\varepsilon^{(\mathrm{I})}}\left[e\cdot a(n_0) + e\cdot a(-n_0)\,e^{2ik^{(\mathrm{I})}R}\right],\end{aligned}\right\} \tag{108}$$

and the relation (102) becomes

$$\mathscr{W}^{(s)} + \mathscr{W}^{(a)} = \frac{c\sqrt{\varepsilon^{(\bar{I})}}}{2k^{(\bar{I})}}\mathscr{I}\left[e\cdot a(n_0)\right] \tag{109}$$

where $\mathscr{I}$ denotes the imaginary part.
The relation (109) implies that with incident light that is linearly polarized, the rate at which the energy is dissipated is proportional to a certain amplitude component of the scattered wave; the amplitude is that which corresponds to forward scattering ( $\boldsymbol{n} = \boldsymbol{n}_0$ ) and the component is in the direction of the electric vector of the incident wave.

The ratio $Q$ between the rate of dissipation of energy $\left(\mathscr{W}^{(a)} + \mathscr{W}^{(s)}\right)$ and the rate at which the energy is incident on unit cross-sectional area of the obstacle $\left(\left|\left\langle \boldsymbol{S}^{(i)}\right\rangle\right|\right)$ is called the extinction cross-section of the obstacle. From (100a), (103), and (105) it follows that $\left|\left\langle S^{(i)}\right\rangle\right| = c\sqrt{\varepsilon^{(1)}}e^2/8\pi$ so that we have, according to (109),

$$Q = \frac{\mathscr{W}^{(s)} + \mathscr{W}^{(a)}}{\left|\langle S^{(i)}\rangle\right|} = 2\lambda^{(\mathrm{I})}\mathscr{I}\left(\frac{e\cdot a(n_0)}{e^2}\right). \tag{110}$$

This formula is due to van de Hulst. †

One may define the scattering cross-section $Q^{(s)}$ and the absorption cross-section $Q^{(a)}$ of the obstacle in a similar way:

$$Q^{(s)} = \frac{\mathscr{W}^{(s)}}{\left|\langle S^{(i)}\rangle\right|}, \quad Q^{(a)} = \frac{\mathscr{W}^{(a)}}{\left|\langle S^{(i)}\rangle\right|}, \tag{111}$$

and evidently $Q = Q^{(s)} + Q^{(a)}$. For a non-absorbing obstacle $Q^{(a)} = 0$ and the extinction cross-section is then equal to the scattering cross-section.

Before applying the relation (110) to a spherical obstacle, let us consider the value of $Q$ for an obstacle that does not transmit an appreciable fraction of the incident light. We also assume that the linear dimensions of the obstacle are large compared to the wavelength. In this case the Huygens-Kirchhoff theory applies and the

0

• This lemma may be proved, after an obvious change of variables, by the application of the principle of stationary phase (cf. Appendix III, p. 752). See also D. S. Jones, Proc. Camb. Phil. Soc., 48 (1952), 736.
† H. C. van de Hulst, Physica, 15 (1949), 740. The method of proof employed here is due to D. S. Jones, Phil. Mag., 46 (1955), 957. There is an analogous theorem for atomic collisions (cf. E. Feenberg, Phys. Rev., 40 (1932), 48; M. Lax, Phys. Rev., (2), 78 (1950), 306).

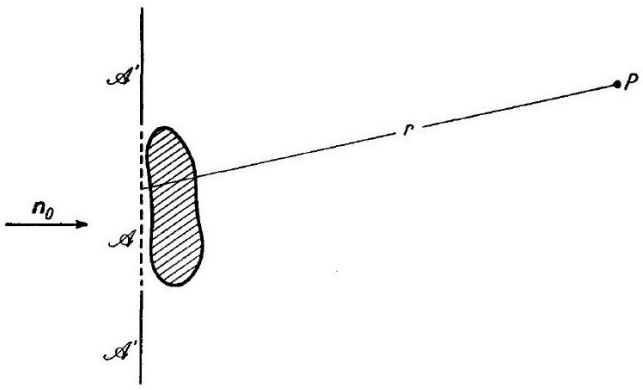



Fig. 13.12. Illustrating theorem (114), relating to the extinction cross-section of a large obstacle.

main contribution to the forward scattering arises from Fraunhofer diffraction. Let $\mathscr{A}$ be the "shadow region" and $\mathscr{A}'$ the unobstructed region of a plane wave-front of a linearly polarized wave incident upon the obstacle (Fig. 13.12), and consider the scattered field $E^{(s)}$ at a point $P$ at a large distance from the obstacle. According to the Huygens-Fresnel principle and the principle of Babinet (§ 8.3 (21))

$$E^{(s)} = \frac{i}{\lambda^{(\mathrm{I})}} e \iint_{\mathscr{A}} \frac{e^{ik^{(\mathrm{I})}r}}{r} dS, \tag{112}$$

if the angle of diffraction is small. If $P$ is very far away from the obstacle in the direction of propagation of the incident wave (forward direction), $r$ may be taken as constant and (112) gives

$$E^{(s)}(n_0) = \frac{i}{\lambda^{(1)}} De \frac{e^{ik^{(1)}r}}{r}, \tag{113}$$

where $D$ is the geometrical cross-section of the obstacle (the area $\mathscr{A}$ ). Hence in this case the vector $a(n_0)$ in (104) is given by $iDe/\lambda^{(1)}$, and (110) shows that

$$Q = 2D. \tag{114}$$

Thus the extinction cross-section of a large opaque obstacle is equal to twice its geometrical cross-section. This result appears somewhat paradoxical at first sight, as one might have expected that with a large obstacle the geometrical optics approximation would apply, and in this approximation the extinction cross-section is equal to $D$. The
explanation of this apparent contradiction* is, that no matter how large the obstacle is, and no matter how far away from it the field is considered, there is always a narrow region-the neighbourhood of the edge of the geometrical shadow-where the geometrical optics approximation does not hold. In addition to the light intercepted by the obstacle (lost by reflection and absorption), with cross-section $D$, there is an additional contribution to the extinction, arising from the neighbourhood of the edge of the shadow and this contribution is evidently also equal to $D$. In order to verify the relation (114) by experiment one must collect the light over a sufficiently wide area and far enough away from the obstacle. †
Let us now apply the general formula (110) to a spherical obstacle. According to (58) the amplitude component of the scattered wave in the direction of the electric vector of the incident wave ( $\phi = 0$ ), for forward scattering ( $\theta = 0$ ), is given by

$$\left(E_\theta^{(s)}\right)_{\theta=\phi=0} = -\frac{1}{k^{(\mathrm{I})}}\frac{1}{r}\sum_{l=1}^{\infty}\left\{{}^{e}B_l\zeta_l^{(1)'}\left(k^{(\mathrm{I})}r\right)\left[P_l^{(1)'}(\cos\theta)\sin\theta\right]_{\theta=0}\right.$$
$$\left.-i^m B_l\zeta_l^{(1)}\left(k^{(1)}r\right)\left[P_l^{(1)}(\cos\theta)\frac{1}{\sin\theta}\right]_{\theta=0}\right\} \tag{115}$$

The two terms involving the associated Legendre functions can easily be evaluated from the expansion ‡

$$P_l^{(m)}(x) = \frac{1}{2^m m!}\frac{(l+m)!}{(l-m)!}\left(1-x^2\right)^{m/2}\left\{1+c_1\left(\frac{1-x}{2}\right)+c_2\left(\frac{1-x}{2}\right)^2+\ldots\right\}, \tag{116}$$

where $c_1, c_2, \ldots$ depend on $l$ and $m$ only. From (116) and from the derivative of this expression we find that

$$\left(\frac{P_l^{(1)}(\cos\theta)}{\sin\theta}\right)_{\theta=0} = \frac{1}{2}l(l+1), \quad \left(P_l^{(1)'}(\cos\theta)\sin\theta\right)_{\theta=0} = -\frac{1}{2}l(l+1). \tag{117}$$

On substituting from (117) in (115) and using the asymptotic approximations (74) and (76) for $\zeta_l{}^{(1)}$ and $\zeta_l{}^{(1)}$, we obtain

$$\left(E_\theta^{(s)}\right)_{\theta=\phi=0} = \frac{1}{2k^{(\mathrm{I})}}\frac{e^{ik^{(\mathrm{I})}r}}{r}\sum_{l=1}^{\infty}(-i)^l l(l+1)\left[{}^{e}B_l + {}^{m}B_l\right]. \tag{118}$$

Remembering that the incident field was taken to be of unit amplitude ( $e^2 = 1$ ), the required quantity $e.a\,(n_0)\,/e^2$ is the multiplier of $e^{ik^{(1)}}/r$ in this expression; on substituting it in (110) and using the identity $\mathscr{I}(z) \equiv \mathscr{R}(-iz)$, which holds for any

complex number $z$, we finally obtain the following expressions for the extinction crosssection of a sphere, in terms of the coefficients ${}^{e}B_l$ and ${}^{m}B_l$ which are given by (62):

$$Q = \frac{\lambda^{(I)2}}{2\pi}\mathscr{R}\sum_{l=1}^{\infty}(-i)^{l+1}l(l+1)\left[{}^{e}B_l + {}^{m}B_l\right] \tag{119}$$

where $\mathscr{R}$ denotes the real part.

(b) Computational results We shall now summarize the main computational results relating to total scattering, total absorption, and the extinction of a sphere.

We saw in § 13.5.2 that when the sphere is very much smaller than the wavelength (Rayleigh scattering), only the first electric partial wave needs to be taken into account. The amplitude

0

- There is an analogous apparent paradox in quantum mechanical scattering problems, first noted by H. S. W. Massey and C. B. O. Mohr, Proc. Roy. Soc., A, 141 (1933), 434.

† It follows from a careful analysis leading to lemma (107), that the portion of $S$ which contributes appreciably to $Q$ subtends at the centre of $S$ a solid angle of order $(kR)^{-\beta}$, where $1 > \beta > 4/5$ [cf. D. S. Jones, Proc. Camb. Phil. Soc., 48 (1952), 736]. For a fuller discussion of the extinction cross-section of a large obstacle and the role played by the geometrical shadow see D. Sinclair, J. Opt. Soc. Amer., 37 (1947), 475 and L. Brillouin, J. Appl. Phys., 20 (1949), 1110.

‡ Cf. W. Magnus and F. Oberhettinger, Formulas and Theorems for the Functions of Mathematical Physics (New York, Chelsea Publishing Company, 1954), p. 54. The formula given by these authors differs from (116) by a multiplication factor ( -1)$^m$. This difference arises from their use of a slightly different definition of the associated Legendre polynomials (cf. footnote on p. 646).

of the scattered wave is then proportional to $1/\lambda^{(1)^2}$, so that the total scattering is inversely proportional to the fourth power of the wavelength. If we take into account the higher terms, which depend on the radius and the material constants, the total scattering becomes a very complicated function of the wavelength and shows selective properties.* In the case of gold, for example, even a very small sphere gives a maximum near $\lambda = 5500$Å (see Fig. 13.13).

Such maxima may be interpreted as a kind of resonance phenomena. Let us suppose that the sphere is not under the influence of the field of the incident beam of light, but that it performs a free electromagnetic oscillation. The frequency and the damping constant of this free vibration is obtained from the theory, by omitting in equations (56) those terms which do not contain the coefficients ${}^eB_l$ and ${}^mB_l$. The resulting equations are linear and homogeneous and have non-trivial solutions only if a consistency condition is satisfied. Each solution of the consistency equation corresponds to a damped proper vibration, and the frequency of these vibrations coincides very closely with the frequencies for which certain scattered partial waves have maxima of intensity. †

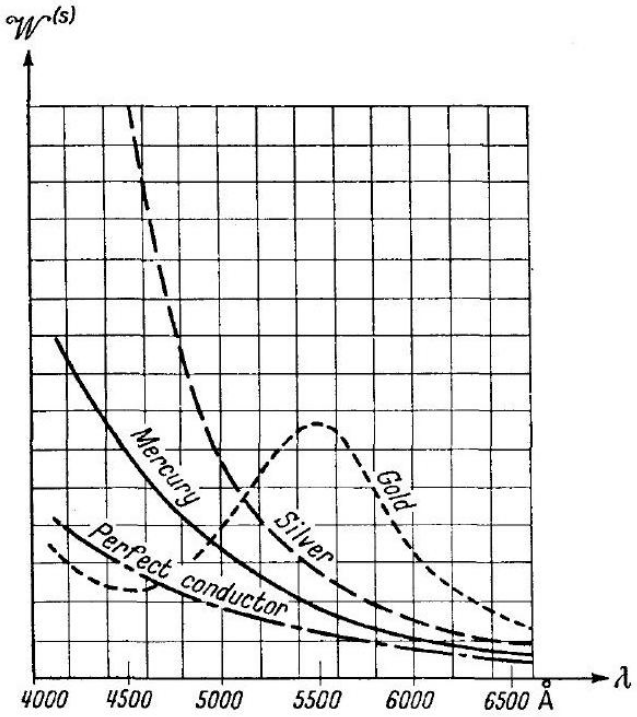



Fig. 13.13. Total scattering from very small particles ( $a \to 0$ ) as function of the wavelength.

(After R. Feick, Ann.d. Physik, (4) 77 (1925), 582.)

Calculations relating to scattering and extinction from spheres of finite radii, for selected values of the refractive index, have been carried out by many workers. The majority of the

0

• An interesting example of selective scattering was provided by the phenomena observed in September 1950, when over a large part of Europe the sun (and the moon also) appeared deep blue. Spectrographic measurements of the "blue" and normal suns gave the extinction curve of the layer causing the phenomena, and it was concluded that the blue colour was due to selective scattering by smoke, probably consisting of oil droplets, remarkably uniform in size, which had been carried by winds in the upper atmosphere from forest fires burning in Alberta (cf. R. Wilson, Mon. Not. Roy. Astr. Soc., 111 (1951), 478).
† Optical resonance was observed by R. W. Wood (Phil. Mag., (6), 3 (1902), 396), from granular films and fogs of the alkali metals.

The proper vibrations of a sphere were studied by P. Debye, Ann. d. Physik, (4), 30 (1909), 73.

calculations concern dielectric spheres ( $n$ real) and spheres that are slightly absorbing. In Fig. 13.14 a typical curve, due to B. Goldberg, is shown for spheres of refractive index $n = 1 \cdot 33$.* This is the refractive index of water, and the results are of interest in connection with the transmission of light by mist, clouds, and fogs, in connection with the theory of the Wilson cloud chamber, etc. It is seen that the curve has a series of maxima and minima and that with increased radius the extinction cross-section tends to twice the geometrical cross-section, in agreement with equation (114). The curve has also a fine structure, i.e. small subsidiary maxima and minima. Naturally, these small fluctuations are smoothed out if the scattering is caused by many spherical particles which are not of exactly the same size.

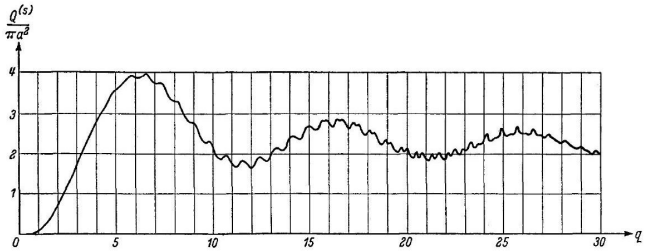



Fig. 13.14. The scattering cross-section of dielectric spheres of refractive index $n = 1.33$ as function of the parameter $q=2 \pi a / \lambda$.

(After B. Goldberg, J. Opt. Soc. Amer., 43 (1953), 1221.)

The extinction curves for dielectric spheres of other refractive indices exhibit similar behaviour. It may be shown that if $n$ is not too different from unity all the curves have a first maximum for a value of $q$ given by †$2q(n-1) \sim 4$, where $Q$ may be as large as $4\pi a^2$.

For perfectly reflecting spheres ‡$(n \to \infty)$ the first maximum of the extinction curve is found to be at $q = 1.2$ where $Q = 2.29\pi a^2$ and the first minimum at $q = 1.6$ where $Q = 2 \cdot 12\pi a^2$. Thereafter there are slight oscillations and the curve approaches $2\pi a^2$ as $q \to \infty$.

The calculations relating to absorbing spheres are much more laborious and only a few special cases have been studied in detail. In Fig. 13.15 curves relating to scattering,

absorption, and extinction by small spheres of iron are shown. For larger spheres, asymptotic formulae due to Jobst,* based on Mie's theory and on Debye's asymptotic expansions of the cylinder functions may be used for calculation. Weakly

0

- This case was also investigated by H. Holl, Optik, 4 (1948), 173 and by H. G. Houghton and W. R. Chalker, J. Opt. Soc. Amer., 39 (1949), 955. Similar curves relating to other values of the refractive index have been published by many writers, for example: M. D. Barnes and V. K. La Mer, J. Col. Sci., 1 (1946), 79; M. D. Barnes, A. S. Kenyon, E. M. Zaiser, and V. K. la Mer, ibid., 2 (1947), 349; J. L. Greenstein (Harvard circ., 1937, No. 422); H. C. van de Hulst, Rech. Astr. Observ. Utrecht, 11, Pt. 1 (1946), 43-51 and his book Light Scattering by Small Particles (New York, J. Wiley and Sons; London, Chapman and Hall, 1957), Chapter 13; R. Pendorf, J. Opt. Soc. Amer., 46 (1956), 1001.
  † The quantity $2q(n-1)$ represents the phase shift suffered by a ray of light that traverses the sphere along a diameter. This is a useful parameter, and if $n$ is not too different from unity the extinction curves plotted against it are very similar to each other.
  ‡ F. W. P. Götz, Astr. Nachr., 255 (1935), 63 and J. L. Greenstein (loc. cit.).

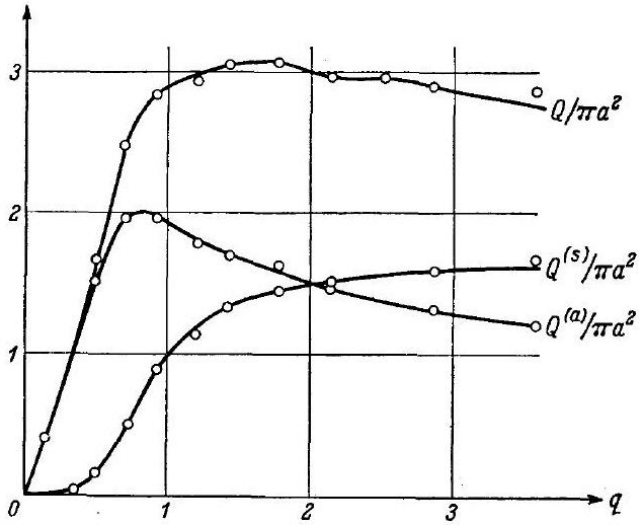



Fig. 13.15. The absorption cross-section ( $Q^{(a)}$ ), the scattering cross-section ( $Q^{(s)}$ ), and the extinction cross-section ( $Q$ ) for iron spheres of various radii. $\hat{n} = 1.27 + 1.37i$, $\lambda^{(1)} = 4200$Å.
(Based on computations of C. Schaln, Uppsala Astr. Observ. Ann., 1, No. 9 (1945).)

absorbing spheres have been studied by van de Hulst, † and we show results of his calculations in Fig. 13.16. In this latter case the general behaviour of the extinction curves is seen to be similar to those of dielectric spheres, but even a very

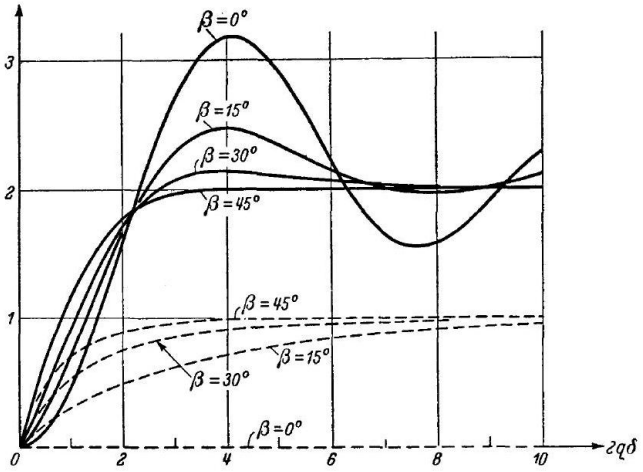



Fig. 13.16. The extinction curves $Q / \pi a^2$ (full lines) and the absorption curves $Q^{(a)}/\pi a^2$ (interrupted lines) for weakly absorbing spheres of refractive index $\hat{n} = (1+\delta)+i\delta \tan\beta$ where $\delta$ is real and small compared to unity.
(After H. C. van de Hulst, Rech. Astr. Obs. Utrecht, 11, Part 2 (1949), 28.)

small conductivity is sufficient to smooth out the small undulations completely. As the conductivity is increased still further, the first minimum disappears altogether

0

• G. Jobst, Ann. d. Physik, (4), '76 (1925), 863.
† H. C. van de Hulst, Rech. Ast. Observ. Utrecht, 11, Part 2 (1949), 27.

and the extinction curve rises asymptotically from the origin to 2 ; the absorption curves rise asymptotically from the origin to half this value.

Mir's theory may be tested experimentally by means of observations of light scattered either by a single spherical particle, or by many particles (cloudy media, colloidal solutions). Such tests may be carried out with relative ease when the particles are large, but are rather troublesome when the diameter of each particle is of the order of a wavelength or smaller. La Mer and collaborators* succeeded in testing the theory from measurements of the angular distribution of scattered light as well as the total scattering from sulphur sols in water, of particle diameter 3000Å to 5000Å. Light of vacuum wavelengths in the range from 2850Å to 10,000Å was used and a fair agreement with the predictions of Mix's theory was found. In some cases even the minute fluctuations (fine structure) of the extinction curve (see Fig. 13.14) were observed.

The scattering of light by particles of shapes other than spherical has been considered by some authors, but in general the analytical properties of the corresponding wave functions are much more complicated, so that rigorous solutions are of limited practical value †. Gans ‡ and other workers discussed the scattering of electromagnetic waves by ellipsoids with dimensions small compared to the wavelength; a rigorous solution for an ellipsoid of arbitrary size has been published by Möqlich.§ The scattering from long circular conducting cylinders was studied as early as 1905 by Seitz || and Ionatowsky 9 and the formulae obtained by them are similar to those of Mie relating to the sphere. Scattering by long circular dielectric cylinders and highly reflecting cylinders was investigated by Schaeffer and Grossmann.**

## 9.18 Optics of Crystals

### 9.18.1 14.1. The Dielectric Tensor of an Anisotropic Medium

It will be remembered that our optical theory is based on two quite separate foundations; on the one hand on Maxwell's equations § 1.1 (1) and (2), on the other hand on the material equations which in the case of an isotropic medium were given by the formulae § 1.1 (9)-(11). In dealing with crystals we must generalize these latter equations so as to take account of anisotropy. In the greater part of this chapter we assume that the medium is homogeneous, non-conducting ( $\sigma = 0$ ), and magnetically isotropic,* but allow electrical anisotropy, i.e. we consider substances whose electrical excitations depend on the direction of the electric field. In general the vector $D$ will then no longer be in the direction of the vector $E$. In place of equation § 1.1 (10) we assume the relation between $D$ and $E$ to have the simplest form which can account for anisotropic behaviour, namely one in which each component of $D$ is linearly related to the components of $\boldsymbol{E}$ :

0

- M. D. Barnes and V. K. La Mer, J. Col. Sci., 1 (1946), 79; M. D. Barnes, A. S. Kenyon, E. M. Zaiser, and V. K. La Mer, ibid., 2 (1947), 349; I. Johnson and V. K. la Mer, J. Amer. Chem. Soc., 69 (1947), 1184.
  † Approximate methods were developed by several authors. See, for example, R. W. Hart and E. W. Montroll, J. Appl. Phys., 22 (1951), 376; E. W. Montroll and R. W. Hart, ibid., 22 (1951), 1278; E. W. Montroll and J. M. Greenberq, Phys. Rev., 86 (1952), 889.

See also the review article by C. J. Bouwkamp, Rep. Progr. Phys. (London, Physical Society), 17 (1954), 35.
‡ R. Gans, Ann. d. Physik, (4), 37 (1912), 881; ibid. 47 (1915), 270.
§ F. Möqlice, Ann. d. Physik, (4), 83 (1927), 609.
|| W. Seitz, Ann. d. Physik, (4), 16 (1905), 746; ibid., 19 (1906), 554.
II W. v. Ignatowsky, Ann. d. Physik, (4), 18 (1905), 495.
** C. Schaeffer and F. Grossmann, Ann. d. Physik, (4) 31 (1910), 455; see also H. C. van de Hulst, Astrophys. J., 112 (1950), 1.

$$\left.\begin{aligned} D_x &= \varepsilon_{xx}E_x + \varepsilon_{xy}E_y + \varepsilon_{xz}E_z, \\ D_y &= \varepsilon_{yx}E_x + \varepsilon_{yy}E_y + \varepsilon_{yz}E_z, \\ D_z &= \varepsilon_{zx}E_x + \varepsilon_{zy}E_y + \varepsilon_{zz}E_z. \end{aligned}\right\} \tag{1}$$

The nine quantities $\varepsilon_{xx}, \varepsilon_{yy}, \ldots$ are constants of the medium, and constitute the dielectric tensor; the vector $\boldsymbol{D}$ is thus the product of this tensor with $\boldsymbol{E}$.

We shall write equation (1) in shorter form as

$$D_k = \sum_l \varepsilon_{kl} E_l, \tag{2}$$

where $k$ stands for one of the three indices $x, y$, and $z$, and $l$ stands for each of $x, y$, and $z$ in turn in the summation. The summation sign would be omitted in formal tensor notation, the occurrence of the index $l$ in two places in the product being understood as an instruction to sum over all $l$ 's. We shall, however, retain the summation sign, as this will help to avoid any ambiguities for readers unfamiliar with tensor calculus.

We assume that the expressions § 1.1 (31) for electric and magnetic energy densities retain their validity. Thus

$$w_e = \frac{1}{8\pi} E.D = \frac{1}{8\pi} \sum_{kl} E_k \varepsilon_{kl} E_l. \tag{3}$$

and

$$w_m = \frac{1}{8\pi} B \cdot H = \frac{1}{8\pi} \mu H^2 \tag{4}$$

We retain also the definition § 1.1 (38) of the Poynting vector, or the "ray-vector"

$$S = \frac{c}{4\pi}(E \wedge H) \tag{5}$$

and investigate whether these definitions are consistent with the principle of the conservation of energy.

We have, as in § 1.1.4, by multiplying the first Maxwell equation by $\boldsymbol{E}$ and the second by $H$ and using the vector identity § 1.1 (27),

$$\begin{aligned} -c\,\mathrm{div}(\boldsymbol{E} \wedge \boldsymbol{H}) &= \boldsymbol{E} \cdot \dot{\boldsymbol{D}} + \boldsymbol{B} \cdot \dot{\boldsymbol{H}} \\ &= \sum_{kl} \boldsymbol{E}_k \varepsilon_{kl} \dot{\boldsymbol{E}}_l + \frac{1}{2}\frac{d}{dt}\left(\mu H^2\right) \end{aligned} \tag{6}$$

If we divide both sides of this equation by $4\pi$, the second term on the right represents the rate of change of the magnetic energy per unit volume, but the first term does not represent the rate of change of the electric energy density unless

$$\frac{1}{4\pi} \sum_{kl} E_k \varepsilon_{kl} \dot{E}_l = \frac{dw_e}{dt} = \frac{1}{8\pi} \sum_{kl} \varepsilon_{kl} \left(E_k \dot{E}_l + E_l \dot{E}_k\right) \tag{7}$$

0

- There are also magnetic crystals, but as the effect of magnetization on optical phenomena (rapid oscillations) is small, the magnetic anisotropy may be neglected. The magnetic permeability shall, however, be retained, and will be represented by a scalar $\mu$, in order to preserve some symmetry in the formulae and to include weakly magnetic crystals; moreover, by retaining it, we facilitate the expression of the equations in systems of units in which $\mu$ is not equal to unity in free space.

that is, unless

$$\sum_{kl} \varepsilon_{kl} \left( E_k \dot{E}_l - \dot{E}_k E_l \right) = 0$$

The suffices $k$ and $l$ are dummy suffices; both run over the same values ( $x, y, z$ ). Hence the expression is not altered if we interchange $k$ and $l$ in the second term. This leads to

$$\sum_{kl} \left( \varepsilon_{kl} - \varepsilon_{lk} \right) E_k \dot{E}_l = 0$$

As this equation must hold whatever the value of the field, it follows that

$$\varepsilon_{kl} = \varepsilon_{lk}. \tag{8}$$

This means that the dielectric tensor must be symmetric; it has only six instead of nine independent components. Conversely, the condition (8) is sufficient to ensure the validity of equation (7), and we obtain the energy theorem in differential form (the "hydrodynamical continuity equation" § 1.1 (43))

$$-\operatorname{div} \boldsymbol{S} = \frac{dw}{dt} \quad , \quad (w = w_e + w_m) \tag{9}$$

The symmetry of the tensor $\varepsilon$ makes it possible to reduce the expression for the electric energy $w_e$ to a form in which only the squares of the field components, and not their products, enter. Consider in a space $x, y, z$ the surface of the second degree

$$\varepsilon_{xx} x^2 + \varepsilon_{yy} y^2 + \varepsilon_{zz} z^2 + 2\varepsilon_{yz} yz + 2\varepsilon_{xz} xz + 2\varepsilon_{xy} xy = \text{ const.} \tag{10}$$

The left-hand side of (10) must be a positive definite quadratic form, because if $x, y$, and $z$ are replaced by the components of $\boldsymbol{E}$ the expression becomes equal to $8\pi w_e$, and the energy $w_e$ must be positive for any value of the field vector. Therefore
equation (10) represents an ellipsoid. The ellipsoid can always be transformed to its principal axes; thus there exists a coordinate system fixed in the crystal such that the equation of the ellipsoid is

$$\varepsilon_x x^2 + \varepsilon_y y^2 + \varepsilon_z z^2 = \text{ constant } . \tag{11}$$

In this system of principal dielectric axes the material equations and the expression for the electrical energy take the simple forms

$$D_x = \varepsilon_x E_x, \quad D_y = \varepsilon_y E_y, \quad D_z = \varepsilon_z E_z \tag{12}$$

$$\begin{aligned} w_e &= \frac{1}{8\pi} \left( \varepsilon_x E_x^2 + \varepsilon_y E_y^2 + \varepsilon_z E_z^2 \right) \\ &= \frac{1}{8\pi} \left( \frac{D_x^2}{\varepsilon_x} + \frac{D_y^2}{\varepsilon_y} + \frac{D_z^2}{\varepsilon_z} \right) \end{aligned} \tag{13}$$

$\varepsilon_x, \varepsilon_y, \varepsilon_z$ are called the principal dielectric constants (or principal permittivities). It may be seen immediately from these formulae that $\boldsymbol{D}$ and $\boldsymbol{E}$ will have different directions, unless $\boldsymbol{E}$ coincides in direction with one of the principal axes, or the principal dielectric constants are all equal; in the latter case ( $\varepsilon_x = \varepsilon_y = \varepsilon_z$ ) the ellipsoid degenerates into a sphere.

A note must be added here on the effect of dispersion. Just as, in the case of isotropic substances, the dielectric constant is not a constant of the material but depends on the frequency, so in an anisotropic medium the six components $\varepsilon_{kl}$ of the dielectric tensor will also vary with

frequency. As a result not only the values of the principal dielectric constants $\varepsilon_x, \varepsilon_y, \varepsilon_z$ will vary but also the directions of the principal axes. This phenomenon is known as dispersion of the axes. It can arise, however, only in crystals in which the symmetry of the structure does not determine a preferential orthogonal triplet of directions, i.e. it can be observed only in monoclinic and triclinic systems (cf. § 14.3.1).*

If we restrict ourselves to monochromatic waves we may disregard dispersion; the quantities $\varepsilon_{kl}$ are then constants depending only on the medium.

## 9.18.2 14.2. The Structure of a Monochromatic Plane Wave in an Anisotropic Medium

### 14.2.1 The phase velocity and the ray velocity

In a monochromatic plane wave of angular frequency $\omega = 2\pi\nu$ propagated with velocity $c/n$ in the direction of the unit wave-normal $s$, the vectors $\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{H}$ and $\boldsymbol{B}$ are in complex notation proportional to $\exp\left[i\omega\left(\frac{n}{c}(r.s) - t\right)\right]$. It may be mentioned at once that in addition to the phase (or wave-normal) velocity $c/n$, we shall later have to introduce a ray (or energy) velocity, since, as will be seen, in an anisotropic medium the energy is in general propagated with a different velocity and in a direction different from that of the wave normal.

In such an oscillatory field the operation $\partial/\partial t$ is always equivalent to multiplication by $-i\omega$, while the operation $\partial/\partial x$ is equivalent to multiplication by $i\omega n s_x/c$. In particular,

$$\dot{E} = -i\omega E, \quad \operatorname{curl} E = i\omega \frac{n}{c} s \wedge E \tag{1}$$

Maxwell's equations, for a region which does not contain currents, viz.

$$\operatorname{curl} H - \frac{1}{c}\dot{D} = 0, \quad \operatorname{curl} E + \frac{1}{c}\dot{B} = 0, \tag{2}$$

become

$$ns \wedge H = -D, \quad ns \wedge E = \mu H, \tag{3}$$

where the relation $\boldsymbol{B} = \mu \boldsymbol{H}$ has been used. Eliminating $\boldsymbol{H}$ between equations (3) and using a well-known vector identity, we obtain

$$D = -\frac{n^2}{\mu} s \wedge (s \wedge E) = \frac{n^2}{\mu}[E - s(s.E)] = \frac{n^2}{\mu} E_\perp, \tag{4}$$

where $\boldsymbol{E}_\perp$ denotes the vector component of $\boldsymbol{E}$ perpendicular to $\boldsymbol{s}$ in the plane of $\boldsymbol{E}$ and $s$ (see Fig. 14.1).

We see from equation (3) that the vector $\boldsymbol{H}$ (and hence also $\boldsymbol{B}$ ) is at right angles to $\boldsymbol{E}, \boldsymbol{D}$, and $\boldsymbol{s}$, which must therefore be coplanar. Further $\boldsymbol{D}$ is seen to be orthogonal to $s$. Thus $\boldsymbol{H}$ and $\boldsymbol{D}$ are transversal to the direction of propagation $\boldsymbol{s}$, as before, but $\boldsymbol{E}$ is not. Fig. 14.1 shows the relative directions of these vectors, and in addition the unit

0

- This dispersion phenomenon is particularly conspicuous in the infra-red: ef. Th. Liebisch and H. Rubens, Sitzber. preuss. Akad. Wiss., Phys.-math. Kl. (1919), 198, 876. H. Rubens, ibid., 976

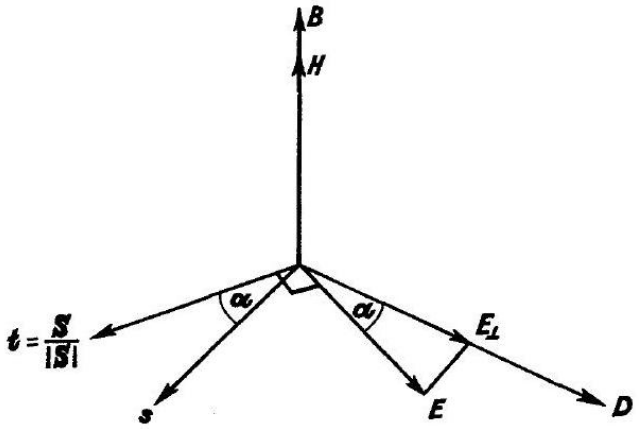



Fig. 14.1. Directions of the wave normal, of the field vectors and of the energy flow in an electrically anisotropic medium.

vector in the direction of the ray vector $S$ to be denoted $t$, which is perpendicular to $\boldsymbol{E}$ and $\boldsymbol{H}$. The angle between $\boldsymbol{E}$ and $\boldsymbol{D}$ is the same as the angle between $\boldsymbol{s}$ and $\boldsymbol{t}$, and will be denoted by $\alpha$. We see that $\boldsymbol{D}, \boldsymbol{H}$, and $s$, on the one hand, and $\boldsymbol{E}, \boldsymbol{H}$, and $\boldsymbol{t}$, on the other hand, form orthogonal vector triplets, with the common vector $\boldsymbol{H}$, rotated relatively to one another through the angle $\alpha$. An important conclusion is that in a crystal the energy is not in general propagated in the direction of the wave normal. On the other hand the theorem of equal electric and magnetic energy densities retains its validity. This follows from equations (3) since*

$$\left.\begin{aligned} w_e &= \tfrac{1}{8\pi} E \cdot D = -\tfrac{n}{8\pi} E \cdot (s \wedge H), \\ w_m &= \tfrac{1}{8\pi} B \cdot H = \tfrac{n}{8\pi} (s \wedge E) \cdot H, \end{aligned}\right\} \tag{5}$$

and by a well-known property of the scalar triple product the right-hand sides of these equations are equal to each other. Moreover, they are both equal to $n(E \wedge H).s/8\pi$, so that the total energy density $w = w_e + w_m$ is

$$w = \frac{n}{c} S.s. \tag{6}$$

We must distinguish between the phase velocity and the velocity of energy transport. The former, the phase velocity, is in the direction of the unit vector $s$ and its magnitude is

$$v_y = \frac{c}{n}. \tag{7}$$

The latter, the ray velocity, is in the same direction as the Poynting vector $S$, i.e. in the direction of the unit vector $t$. Its magnitude $v_r$ is equal to the energy that crosses in unit time an area perpendicular to the flow direction, divided by the energy per unit volume. In accordance with the energy theorem § 14.1 (9) this is given by

$$v_r = \frac{S}{w}. \tag{8}$$

From (6), (7), and (8)

0

- As we are now dealing with quadratic functions of the field variables, it is the real field vectors and not the associated complex vectors that enter equations (5) and (6) (ef. § 1.3, p. 17).

$$v_p = v_r t.s = v_r \cos\alpha. \tag{9}$$

i.e. the phase velocity is the projection of the ray velocity on to the direction of the wave normal.

It may be noted that the ray velocity, being derived from the Poynting vector shares with it (cf. § 1.1, p. 9-10) a certain degree of arbitrariness. It is nevertheless a useful concept although, like the phase velocity, it has no directly verifiable physical significance.

If $\boldsymbol{E}$ and $\boldsymbol{D}$ are known (as for instance when $\boldsymbol{E}$ is prescribed and $\boldsymbol{D}$ is determined from it by equation § 14.1 (1)), the refractive index $n$ and the wave normal $s$ are thereby determined. For in the first place, since $E_\perp$ is the vector component of $E$ in the direction of $D$,

$$\boldsymbol{E}_\perp = \left(\boldsymbol{E} \cdot \frac{\boldsymbol{D}}{\bar{D}}\right) \frac{\boldsymbol{D}}{\bar{D}}, \tag{10}$$

so that (4) gives

$$n^2 = \frac{\mu D}{E_\perp} = \frac{\mu D^2}{(E.D)}. \tag{11}$$

Further, since the unit vector $s$ is perpendicular to $D$ and coplanar with $D$ and $E$, it may be expressed in the form

$$s = \frac{E - E_\perp}{|E - E_\perp|} = \frac{E - \frac{(E.D)D}{D^2}}{\sqrt{E^2 - \frac{(E.D)^2}{D^2}}} = \frac{D^2 E - (E.D)D}{\sqrt{D^2\left[E^2 D^2 - (E.D)^2\right]}}. \tag{12}$$

By analogy with the refractive index $n$ we may also define a ray index or energy index $n_r$ by means of the formula

$$n_r = \frac{c}{v_r}. \tag{13}$$

By (7) and (9) we have,

$$n_r = n\cos\alpha \tag{14}$$

We shall now show that the ray index $n_r$ and the unit vector $t$ in the direction of the propagation of energy are given by formulae analogous to (11) and (12). We have from (14), (11), and the relation $E.D = ED\cos\alpha$,

$$n_r^2 = \frac{\mu(E.D)}{E^2} \tag{15}$$

The unit vector $t$ is perpendicular to $\boldsymbol{E}$ and coplanar with $\boldsymbol{E}$ and $\boldsymbol{D}$ and so must (apart perhaps from the sign) be given by the formulae that results in interchanging $\boldsymbol{E}$ and $\boldsymbol{D}$ in (12). Hence

$$-t = \frac{E^2 D - (E \cdot D)E}{\sqrt{E^2\left[E^2 D^2 - (E \cdot D)^2\right]}}. \tag{16}$$

The negative sign on the left ensures that $s$ and $t$ point towards the same side of $E$ and $D$ as in Fig. 14.1.

Both (12) and (16) reduce to 0/0 when $\boldsymbol{E}$ and $\boldsymbol{D}$ coincide in direction, i.e. when $\boldsymbol{E}$ is in the direction of one of the principal axes of the crystal. This is to be expected since the directions of $s$ and $t$ are then undetermined apart from the fact that they must be perpendicular to $\boldsymbol{E}$.

We may also express the magnitude of the Poynting vector in terms of $\boldsymbol{E}$ and $\boldsymbol{D}$. According to (8), (13), and (15), remembering also that for a plane wave $w = 2w_e$

$$S = v_r w = \frac{c}{n_r}\frac{\boldsymbol{E}.\boldsymbol{D}}{4\pi} = \frac{c}{4\pi\sqrt{\mu}}E\sqrt{\boldsymbol{E}.\boldsymbol{D}} \tag{17}$$

a formula that, in the case of an isotropic medium, is seen to be in agreement with equations (8) and (9) of § 1.4.

### 14.2.2 Fresnel's formulae for the propagation of light in crystals

The formulae derived in § 14.2.1 are consequences of Maxwell's equations alone and therefore independent of the properties of the medium. We shall now combine these with the material equation § 14.1 (1).

We shall use a system of coordinate axes coincident with the principal dielectric axes. The relations § 14.1 (1) then reduce to the simpler form § 14.1 (12), and substitution for $D$ into (4) gives

$$\mu\varepsilon_k \boldsymbol{E}_k = n^2\left[E_k - s_k(\boldsymbol{E}.\boldsymbol{s})\right] \quad (k = x, y, z) \tag{18}$$

Equations (18), being three homogeneous linear equations in $E_x, E_y$, and $E_z$, can be satisfied by non-zero values of these components only if the associated determinant vanishes. This implies that a certain relation must be satisfied by the refractive index $n$, the vector $s\,(s_x, s_y, s_z)$ and the principal dielectric constants $\varepsilon_x, \varepsilon_y, \varepsilon_z$. This relation may be derived by writing (18) in the form

$$E_k = \frac{n^2 s_k(\boldsymbol{E}\cdot s)}{n^2 - \mu\varepsilon_k} \tag{19}$$

multiplying it by $s_k$ and adding the resulting three equations; dividing the expression which then results by the common factor $\boldsymbol{E}.\boldsymbol{s}$, we obtain

$$\frac{s_x^2}{n^2 - \mu\varepsilon_x} + \frac{s_y^2}{n^2 - \mu\varepsilon_y} + \frac{s_z^2}{n^2 - \mu\varepsilon_z} = \frac{1}{n^2}. \tag{20}$$

This formula may be expressed in a slightly different form. We multiply both sides of (20) by $n^2$ and subtract ${s_x}^2 + {s_y}^2 + {s_z}^2 = 1$. Next we multiply the resulting expression by $-n^2$ and find that

$$\frac{s_x^2}{\frac{1}{n^2} - \frac{1}{\mu\varepsilon_x}} + \frac{s_y^2}{\frac{1}{n^2} - \frac{1}{\mu\varepsilon_y}} + \frac{s_z^2}{\frac{1}{n^2} - \frac{1}{\mu\varepsilon_z}} = 0. \tag{21}$$

We define three principal velocities of propagation by the formulae*

$$v_x = \frac{c}{\sqrt{\mu\varepsilon_x}}, \quad v_y = \frac{c}{\sqrt{\mu\varepsilon_y}}, \quad v_z = \frac{c}{\sqrt{\mu\varepsilon_z}}. \tag{22}$$

When the expression (7) is used for the phase velocity $v_p$, (19) and (21) take the form

$$E_k = \frac{v_k^2}{v_k^2 - v_p^2} s_k(E.s) \quad (k = x, y, z), \tag{23}$$

$$\frac{s_x^2}{v_p^2 - v_x^2} + \frac{s_y^2}{v_p^2 - v_y^2} + \frac{s_z^2}{v_p^2 - v_z^2} = 0. \tag{24}$$

Equations (20), (21), and (24) are equivalent forms of Fresnel's equation of wave normals. This is a quadratic equation in ${v_p}^2$, as can be seen by multiplying (24) by the product of the denominators. Thus to every direction $s$ there correspond two phase velocities $v_p$. (The two values $\pm v_p$ corresponding to any value ${v_p}^2$ are counted as one, since the negative value evidently belongs to the opposite direction of propagation $-s$.) With each of the two values of $v_p$, the equations (23) may then be solved for the ratios $E_x : E_y : E_z$; the corresponding ratios involving the $D$ vector may be subsequently obtained from § 14.1 (12). Since these ratios are real, the E and D fields are linearly polarized. Thus we have the important result that the structure of an anisotropic medium permits two monochromatic plane waves with two different linear polarizations and two different velocities to propagate in any given direction. It will be shown later that the two directions of the electric displacement vector $D$ corresponding to a given direction of propagation $s$ are perpendicular to each other.

We now show that there is an analogous formula for the ray velocity $v_r$. This is most easily done by showing first that there is a relation analogous to (4) in which the roles of $\boldsymbol{D}$ and $\boldsymbol{E}$ and of $\boldsymbol{s}$ and $\boldsymbol{t}$ are interchanged. It is convenient to introduce a vector $D_\perp$ defined as the vector component of $D$ perpendicular to $t$, in the plane of $\boldsymbol{D}$ and $\boldsymbol{t}$. It is evidently given by

$$D_\perp = D - t(D \cdot t) \tag{25}$$

Since the electric vector is also perpendicular to $t$ and is coplanar with $D$ and $t$ (see Fig. 14.1), $D_\perp$ is parallel to $E$ and so may also be expressed in the form

$$D_\perp = \left(D \cdot \frac{E}{E}\right)\frac{E}{E} = \frac{n_r^2}{\mu}E \tag{26}$$

where (15) has been used. From (25) and (26) it follows that

$$E = \frac{\mu}{n_r^2}[D - t(D.t)] = \frac{\mu}{n_r^2}D_\perp. \tag{27}$$

This equation is analogous to (4) and may be formally obtained from it by interchanging the roles of $E$ and $D, n$ and $1/n_r, \mu$ and $1/\mu$ and $s$ and - $t$. Quite generally the following rule of duality follows from the basic equations:

Let the variables be arranged in two rows as follows:

$$\left.\begin{array}{llllllllllll} \boldsymbol{E}, & \boldsymbol{D}, & \boldsymbol{s}, & \boldsymbol{t}, & \boldsymbol{c}, & \boldsymbol{\mu} & v_p, & n, & \varepsilon_x, & \varepsilon_y, & \varepsilon_z, & v_x, & v_y, & v_z, \\ \boldsymbol{D}, & \boldsymbol{E}, & -\boldsymbol{t}, & -\boldsymbol{s}, & \frac{1}{c}, & \frac{1}{\mu}, & \frac{1}{v_r}, & \frac{1}{n_r}, & \frac{1}{\varepsilon_x}, & \frac{1}{\varepsilon_y}, & \frac{1}{\varepsilon_z}, & \frac{1}{v_x}, & \frac{1}{v_y}, & \frac{1}{v_z} \end{array}\right\} \tag{28}$$

If in any relation which holds between the quantities in one row each quantity is replaced by the corresponding quantity in the other row another valid relation is obtained.

Applying this rule to Fresnel's equation of wave normals (24) we immediately obtain the required ray equation

$$\frac{t_x^2}{\frac{1}{v_r^2} - \frac{1}{v_x^2}} + \frac{t_y^2}{\frac{1}{v_r^2} - \frac{1}{v_y^2}} + \frac{t_z^2}{\frac{1}{v_r^2} - \frac{1}{v_z^2}} = 0. \tag{29}$$

We may, of course, express this equation also in forms analogous to (20) and (21). Like (24) this is again a quadratic equation and gives two possible ray velocities $v_r$ for each ray direction $t\,(t_x, t_y, t_z)$. The corresponding direction of $D$ may be obtained on solving with the appropriate value $v_r$ the equations dual to (23), namely

0

- Note that $v_x, v_y, v_z$ are not components of a vector and are defined only with reference to the principal axes.

$$D_k = -\frac{v_r^2}{v_k^2 - v_r^2} t_k (D \cdot t). \quad (k = x, y, z). \tag{30}$$

The directions of the two $E$ vectors (which as we saw are orthogonal to $t$ ) may then be obtained by using § 1.4.1 (12).

As a rule only one of the vectors $s$ or $t$ is given and it is, therefore, desirable to derive relations from which the other may be directly calculated. We have from Fig. 14.1

$$\text{E. } s = E_\perp \tan\alpha, \qquad \text{D. } t = -D \sin\alpha \tag{31}$$

But from (4), $D = n^2 E_\perp / \mu$. Hence

$$D \cdot t = -\frac{n^2}{\mu} E_\perp \sin\alpha = -\frac{n^2}{\mu} E \cdot s \cos\alpha = -\frac{1}{\mu} \frac{c^2}{v_p v_r} E \cdot s \tag{32}$$

where the relations (9) and (7) have been used. Substitution from (32) into (30) gives

$$D_k = \varepsilon_k E_k = \frac{1}{\mu} \frac{c^2 v_r}{v_p \left(v_k^2 - v_r^2\right)} t_k (E.s) \tag{33}$$

Comparing (33) with (23) and remembering that $\mu \varepsilon_k {v_k}^2 = c^2$, we obtain

$$\frac{v_p s_k}{v_k^2 - v_p^2} = \frac{v_r t_k}{v_k^2 - v_r^2} \tag{34}$$

Solving for $t_k$ we have

$$t_k = \frac{v_p}{v_r} \frac{v_k^2 - v_r^2}{v_k^2 - v_p^2} s_k \tag{35}$$

so that

$$v_r t_k - v_p s_k = v_p s_k \frac{v_r^2 - v_p^2}{v_p^2 - v_k^2} \tag{36}$$

Squaring and adding the three equations (36) and using the relation § 14.2 (9), namely $s.t = v_p / v_r$, we obtain

$$v_r^2 - v_p^2 = v_p^2 \left(v_r^2 - v_p^2\right)^2 \left[ \left( \frac{s_x}{v_p^2 - v_x^2} \right)^2 + \left( \frac{s_y}{v_p^2 - v_y^2} \right)^2 + \left( \frac{s_z}{v_p^2 - v_z^2} \right)^2 \right] \tag{37}$$

Hence we may define

$$g^2 \equiv v_p^2 \left(v_r^2 - v_p^2\right) = \frac{1}{\left( \frac{s_x}{v_p^2 - v_x^2} \right)^2 + \left( \frac{s_y}{v_p^2 - v_y^2} \right)^2 + \left( \frac{s_z^2}{v_p^2 - v_z^2} \right)^2} \tag{38}$$

This relation expresses $v_r$ in terms of $s$, since $v_p$ is already known in terms of $s$ from Fresnel's equation (24). With $v_r$ so determined, equation (35) then gives the unit ray vector $t$ as a function of $s$. Using the expression $g$, equation (35) may be written as

$$t_k = \frac{s_k}{v_p v_r} \left( v_p^2 + \frac{g^2}{v_p^2 - v_k^2} \right) \quad (k = x, y, z) \tag{39}$$

Since to every $s$ there correspond in general two phase velocities $v_p$, there are in general two ray directions* for each wave-normal direction. There are, however, two singular directions in certain crystals (biaxial crystals-cf. § 14.3.1) due to vanishing of the denominators in (39),

to each of which there corresponds an infinite number of rays; there are also two singular ray directions to each of which there corresponds an infinite number of wave normal directions. These special cases give rise to an interesting phenomenon (conical refraction), which will be investigated in §14.3.4.

### 14.2.3 Geometrical constructions for determining the velocities of propagation and the directions of vibration

Many results concerning the phase and the ray velocities and the directions of vibration may be illustrated by means of certain geometrical constructions.
(a) The ellipsoid of wave normals. By equations § 14.1 (13) the components of the vector $D$ at a given energy density $w = 2w_e$ satisfy the relation

$$\frac{D_x^2}{\varepsilon_x} + \frac{D_y^2}{\varepsilon_y} + \frac{D_z^2}{\varepsilon_z} = C \quad (C = 8\pi w = E.D) \tag{40}$$

Let us write $x, y$, and $z$ in place of $D_x/\sqrt{C}, D_y/\sqrt{C}, D_z/\sqrt{C}$, and consider these as Cartesian coordinates in space. Then

$$\frac{x^2}{\varepsilon_x} + \frac{y^2}{\varepsilon_y} + \frac{z^2}{\varepsilon_z} = 1 \tag{41}$$

This equation represents an ellipsoid, the semi-axes of which are equal to the square roots of the principal dielectric constants and coincide in directions with the principal dielectric axes. We call this ellipsoid the ellipsoid of wave normals in preference to the widely used but rather vague term "optical indicatrix" (also known as the index ellipsoid, or the reciprocal ellipsoid).

With the aid of the ellipsoid of wave normals we can find the two phase velocities

$v_p$ and the two directions of vibrations $\boldsymbol{D}$ which belong to a given wave-normal direction $s$ as follows: We draw a plane through the origin at right angles to $s$. The curve of intersection of this plane with the ellipsoid is an ellipse; the principal semi-axes of this ellipse are proportional to the reciprocals $1/v_p$ of the phase velocities, and their directions coincide with the corresponding directions of vibrations of the vector $\boldsymbol{D}$ (Fig. 14.2).

To establish this result consider the two equations that specify the ellipse:

0

- The position of the rays that correspond to a given wave normal is investigated in detail for the case of a biaxial crystal in M. Born, Optik (Berlin, Springer, 1933), pp. 235-237.

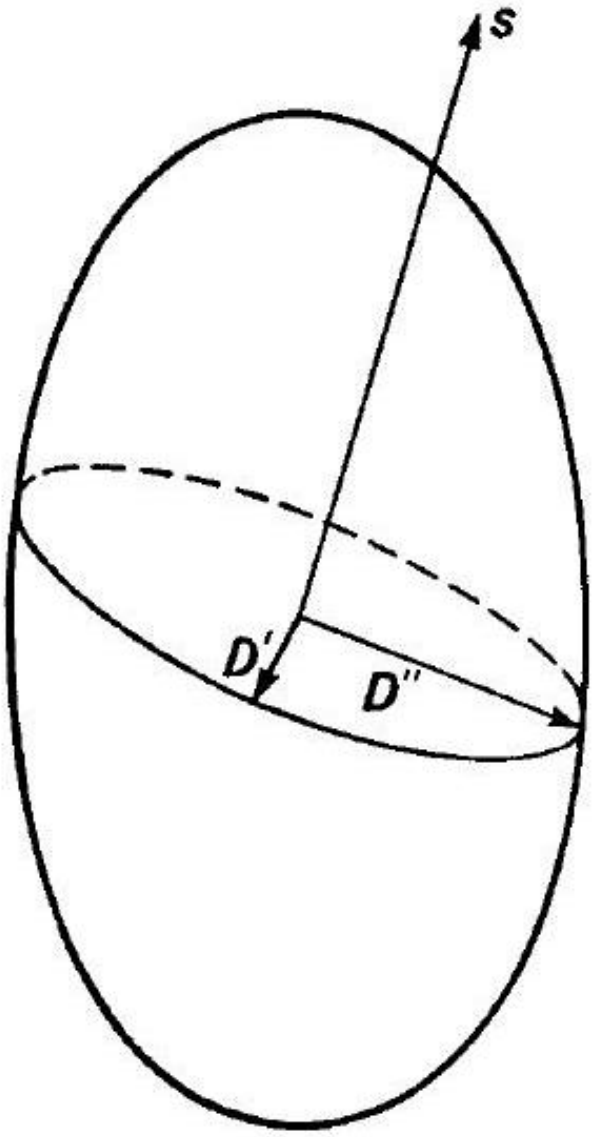



Fig. 14.2. The ellipsoid of wave normals. Construction of the directions of vibrations of the $\boldsymbol{D}$ vectors belonging to a wave normal $s$.

$$xs_x + ys_y + zs_z = 0 \tag{42}$$

$$\frac{x^2}{\varepsilon_x} + \frac{y^2}{\varepsilon_y} + \frac{z^2}{\varepsilon_z} = 1 \tag{43}$$

Since the principal axes of an ellipse are, by definition, its shortest and longest diameters, we can determine them by finding the extrema of

$$r^2 = x^2 + y^2 + z^2, \tag{44}$$

subject to the conditions (42) and (43). This we do by means of Lagrange's method of undetermined multipliers.* We introduce two multipliers $2\lambda_1$ and $\lambda_2$ and construct the function

$$F = x^2 + y^2 + z^2 + 2\lambda_1 (xs_x + ys_y + zs_z) + \lambda_2 \left( \frac{x^2}{\varepsilon_x} + \frac{y^2}{\varepsilon_y} + \frac{z^2}{\varepsilon_z} - 1 \right). \tag{45}$$

Our problem is then equivalent to finding the extremum of $F$ subject to no subsidiary conditions. The necessary conditions for the extremum of $F$ is that its derivatives with respect to $x, y$, and $z$ should vanish, i.e. that

$$x + \lambda_1 s_x + \frac{\lambda_2 x}{\varepsilon_x} = 0, \quad y + \lambda_1 s_y + \frac{\lambda_2 y}{\varepsilon_y} = 0, \quad z + \lambda_1 s_z + \frac{\lambda_2 z}{\varepsilon_z} = 0. \tag{46}$$

Multiplying these equations by $x, y$, and $z$ respectively and adding, we obtain, because of (42) and (43):

$$r^2 + \lambda_2 = 0. \tag{47}$$

Next we multiply the equations (46) by $s_x, s_y$, and $s_z$ and add, and again use (42). We then obtain

$$\lambda_1 + \lambda_2 \left( \frac{xs_x}{\varepsilon_x} + \frac{ys_y}{\varepsilon_y} + \frac{zs_z}{\varepsilon_z} \right) = 0. \tag{48}$$

Substitution for $\lambda_1$ and $\lambda_2$ from (47) and (48) into (46) gives

$$x \left( 1 - \frac{r^2}{\varepsilon_x} \right) + s_x r^2 \left( \frac{xs_x}{\varepsilon_x} + \frac{ys_y}{\varepsilon_y} + \frac{zs_z}{\varepsilon_z} \right) = 0, \tag{49}$$

with two similar equations. With a given $s$, these are three homogeneous equations for $x, y$, and $z$. They are compatible only if the associated determinant vanishes; a

condition that gives an algebraic equation for $r^2$. Now it is immediately seen that the equations (49) differ only in notation from equations (18). For if we replace $x$ by $D_x/\sqrt{C}, x/\varepsilon_x$ by $E_x/\sqrt{C}$ and $r^2$ by $D^2/C = D^2/E.D = n^2/\mu$ (in accordance with (11)), (49) becomes

$$\mu D_x = n^2 \left[ E_x - s_x (\boldsymbol{E} \cdot \boldsymbol{s}) \right] \tag{50}$$

which, together with the two similar equations, is identical with (18).
Thus we find that the roots of the determinantal equation for $n = c/v_p$ (which as we saw is of the second degree) are proportional to the lengths $r$ of the semi-axes of the elliptical section at right angles to $s$, and, moreover, that the two possible directions $x : y : z$ of the vector $D$ coincide with the directions of these axes. Since the axes of an ellipse are perpendicular to each other, we obtain the important result that the directions of vibrations of the two vectors $\boldsymbol{D}$ corresponding to a given direction of propagations are perpendicular to each other. In what follows we shall denote the two $\boldsymbol{D}$ directions that correspond to a particular wave-normal direction $s$ by $D'$ and $D''$; thus $s, D'$, and $D''$ form an orthogonal triplet.

In the special case when the direction of propagation coincides with one of the principal axes of the ellipsoid of wave normals, say the $x$ axis, the extrema of $r$ are, according to our construction, equal to the lengths of the other two semi-axes, i.e. to $\sqrt{\varepsilon_y}$ and $\sqrt{\varepsilon_z}$. But we saw that the extrema of $r$ are also equal to $n/\sqrt{\mu} = c/v_p\sqrt{\mu}$. Hence the phase velocities of waves

0

- For a full account of this method see, for example, R. Courant, Differential and Integral Calculus, Vol. II (London, Blackie & Son, Ltd., 1942), pp. 188-199.

which are propagated in the direction of the principal dielectric axis $x$ are equal to $c/\sqrt{\mu\varepsilon_y}$ and $c/\sqrt{\mu\varepsilon_z}$, i.e. to the principal velocities of propagation $v_y$ and $v_z$ introduced formally

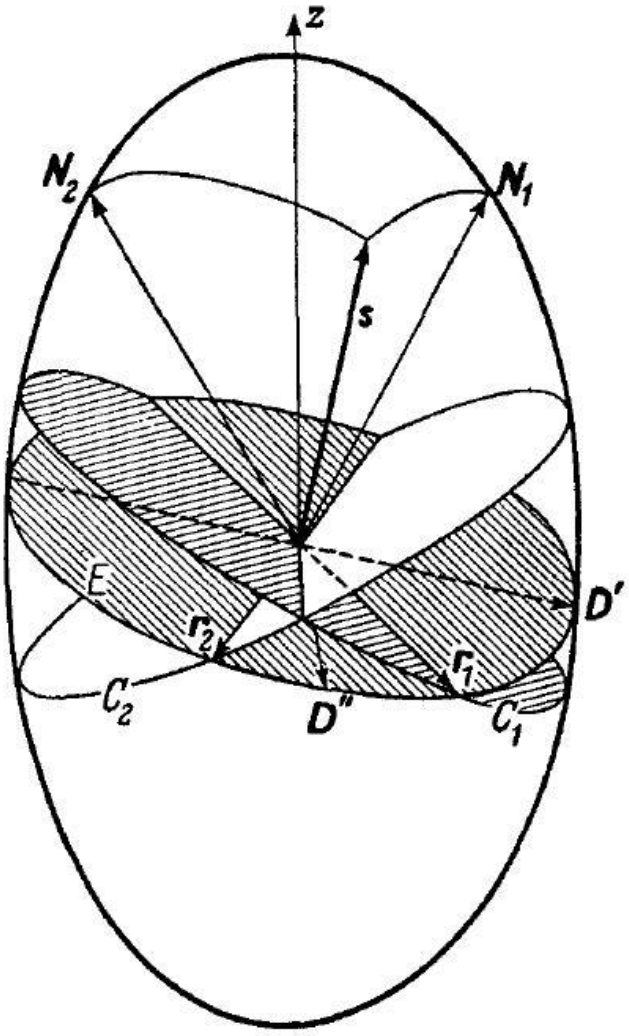



Fig. 14.3. A construction for determining the planes of vibrations ( \$s, D\hat{}\{\}\{\prime\}\$ ) and ( $s, D''$ ).

by (22). A corresponding result holds, of course, for propagation in the directions of the other two axes.

There is another construction by which the directions of vibrations can be determined. It is known that an ellipsoid has two circular sections $C_1$ and $C_2$ passing through the centre, and that the normals $N_1$ and $N_2$ to these sections are coplanar with the longest and shortest principal axes ( $z$ and $x$ ) of the ellipsoid. These two directions $N_1$ and $N_2$ are called the optic axes,* and will be considered more fully later (§ 14.3.3). Since the sections $C_1$ and $C_2$ are circular (also of the same radius), the directions $N_1$, $N_2$ have the property that there is only one velocity of propagation along them: $D$ can then take any direction perpendicular to $s$. Let $E$ be the elliptical section through the centre, at right angles to an arbitrary unit normal $s$. This plane intersects the circles $C_1, C_2$ in two radial vectors $r_1, r_2$ which are of equal length and must therefore make equal angles with the principal axes of $E$ (see Figs. 14.3 and 14.4). The required directions of vibration are therefore the bisectors of the directions $r_1, r_2$. But $r_1$ is perpendicular to $N_1$ and $s$, and is, therefore, perpendicular to the plane containing $N_1$ and $s$; similarly $r_2$ is perpendicular to the plane containing $N_2$ and $s$. If these planes intersect the

ellipse $E$ in vectors $r_1', r_2'$ the principal axes of the ellipse

must also bisect the directions of $\boldsymbol{r_1'}, \boldsymbol{r_2'}$. It follows that the planes of vibration of the electric displacement, i.e. the planes containing $s$ and $D'$ or $D''$, are the internal or external bisectors of the angle between the planes ( $N_1, s$ ) and ( $N_2, s$ ).* This construction becomes indeterminate when $s$ coincides in direction with either $N_1$ or $N_2$, as we should expect.

(b) The ray ellipsoid. Rays may be treated in the same manner as wave normals if, in accordance with the duality rule (28), one starts from the ray ellipsoid

$$\varepsilon_x x^2 + \varepsilon_y y^2 + \varepsilon_z z^2 = 1. \tag{51}$$

In particular, the central section of this ellipsoid perpendicular to a ray direction $t$ is an ellipse with semi-axes of lengths proportional to the two corresponding ray

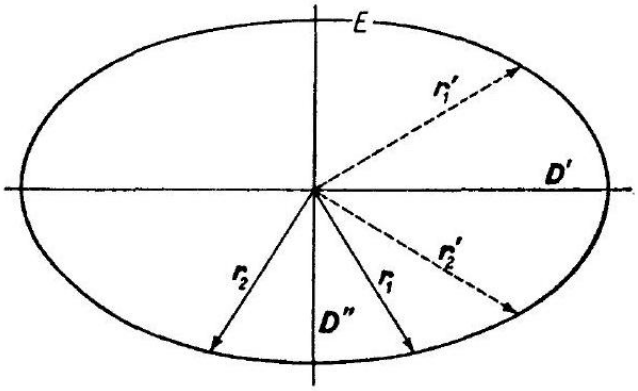



Fig. 14.4. The plane $E$ of Fig. 14.3.

velocities $v_r$, oriented in the two permissible directions $\boldsymbol{E'}$ and $\boldsymbol{E''}$ of the electric vector. Thus $\boldsymbol{t}, \boldsymbol{E'}$, and $\boldsymbol{E''}$ form an orthogonal triplet of vectors.

(c) The normal surface and the ray surface. Imagine that with a fixed point $O$ inside a crystal as origin two vectors are plotted in the same direction $s$ and of lengths that are proportional to the two corresponding phase velocities. As $s$ takes on all possible directions the end points give rise to a surface of two shells known as the wave-normal surface or, for short, the normal surface.

Similarly, the end points of position vectors plotted from a fixed origin in all directions $t$, and of lengths that are proportional to the corresponding ray velocities will generate a two-sheeted surface, called the ray surface.

These two surfaces are more complicated than the ellipsoids that we just considered. The ray surface is of the fourth degree, and the surface of normals of the sixth degree, † as may be verified from the formulae (24) and (29). There exists an important relation between these two surfaces which we shall now derive.

We have seen that if $\boldsymbol{E}$ or $\boldsymbol{D}$ is known, the directions of $\boldsymbol{s}$ and $\boldsymbol{t}$ as well as the corresponding velocities $v_p$ and $v_r$, and consequently the corresponding points ( $P$ and $P'$ in Fig. 14.5) on the two surfaces are thereby determined. Let $r$ and $r'$ be the vectors representing these points:

$$r = v_r t, \quad r' = v_p s. \tag{52}$$

0

• More precisely the optic axes of wave normals. The corresponding sections of the ray ellipsoid (see next page) define the optic ray axes.

0

We shall show that a small variation in $\boldsymbol{E}$ or $\boldsymbol{D}$ produces a change in the vector $r$ at right angles to $r'$.

We start from the equation (27)

$$\frac{1}{\mu}E = \frac{1}{n_r^2}[D - t(D.t)]. \tag{53}$$

Substitution for $\boldsymbol{t}$ from the first equation (52) gives, if we also set $n_r = c/v_r$ :

$$\frac{c^2}{\mu}E = r^2 D - r(D.r). \tag{54}$$

Suppose now that $\boldsymbol{E}$ is changed by a small amount $\delta\boldsymbol{E}$. If $\delta\boldsymbol{D}$ and $\delta\boldsymbol{r}$ are the corresponding changes in $\boldsymbol{D}$ and $\boldsymbol{r}$ respectively, we have, according to (54),

$$\frac{c^2}{\mu}\delta E = 2(r.\delta r)D + r^2\delta D - \delta r(D.r) - r(\delta r.D) - r(r.\delta D) \tag{55}$$

If we multiply both sides of this equation scalarly by $\boldsymbol{D}$ and use the relation

$$D.\delta E = \varepsilon_x E_x \delta E_x + \varepsilon_v E_y \delta E_y + \varepsilon_z E_z \delta E_z = E.\delta D, \tag{56}$$

we obtain

$$\frac{c^2}{\mu}E.\delta D = \delta D.\left[r^2 D - r(D.r)\right] + 2\delta r.\left[rD^2 - D(D.r)\right] \tag{57}$$

The terms having $\delta D$ as a factor cancel because of (54) and the term multiplying $\delta r$ may be written as $2\delta r.[(D \wedge r) \wedge D]$. Hence, since $r = v_r t$,

$$\delta r.[(D \wedge t) \wedge D] = 0 \tag{58}$$

Now the vector $D \wedge t$ is perpendicular to both $D$ and $t$ and so $(D \wedge t) \wedge D$ is in the plane of $D$ and $t$ and at right angles to $D$; it is therefore parallel to $s$ (see § 14.1). Hence

$$s.\delta r = 0 \tag{59}$$

i.e. $\delta\boldsymbol{r}$ is perpendicular to $\boldsymbol{s}$, thus proving our statement. This result implies that the tangent plane of the ray surface is always perpendicular to the corresponding wave normal. Fig. 14.5 illustrates this relation in plane section. Since the perpendicular distance from the origin to this plane is equal to $v_r t.s = v_r \cos\alpha = v_p$ by (9), it follows that the normal surface is the pedal surface of the ray surface and conversely the ray surface is the envelope of planes drawn through points on the normal surface, at right angles to the radius vectors from the origin to these points. If we know the form of either of these surfaces, the other may be determined from this relationship.

We may interpret this result in more physical terms. Consider, not a single wave, but a group formed by plane waves of the same frequency with slightly different directions of propagation. The wave normals $s$ of the component waves fill a solid angle around a

---

• Throughout this chapter $(a, b)$ denotes the plane containing the vectors $a$ and $b$.

† The normal surface and the ray surface cannot be expected to have equations of the same degree, since they are not duals of each other. In order to construct the dual of the normal surface one would have to plot vectors of lengths $1/v_r$ (not $v_r$ ), in accordance with the duality rule (28).

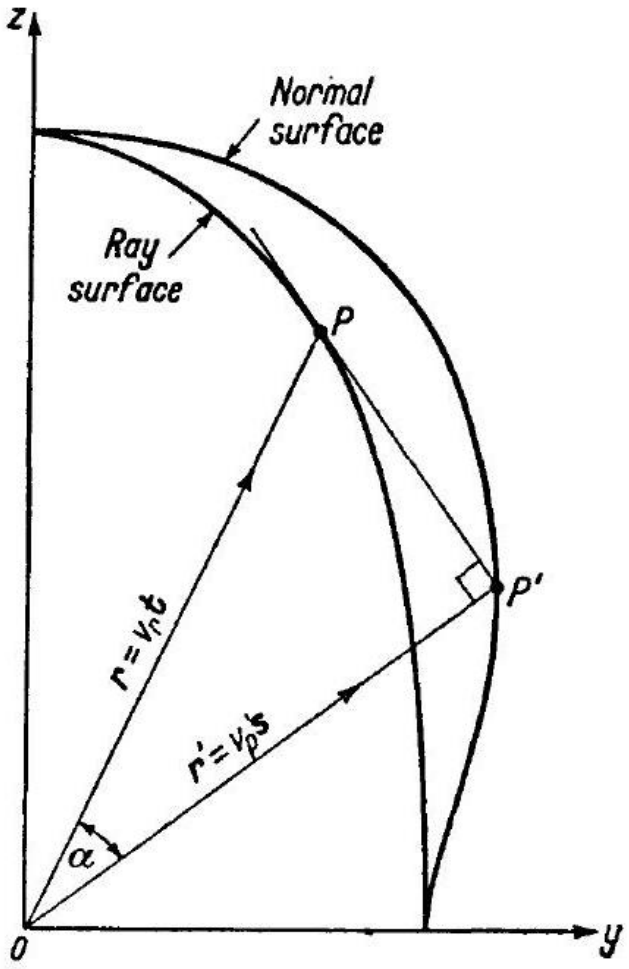



Fig. 14.5. The relation between the normal surface and the ray surface.

"mean wave normal" $s_0$, and we assume that the amplitudes of only those waves are appreciable whose normals are close to $s_0$. Suppose that at time $t = 0$ all the waves are in phase at a point $O$; the disturbance is then a maximum at this point. Let us now examine how this maximum is propagated.

Consider all the wave-fronts which pass through $O$ at time $t = 0$. After a unit time a wave-front $W$, propagated with velocity $v_p$ in the direction $s$, will have reached a position $W'$ such that the foot of the perpendicular dropped on to it from $O$ has the position vector $v_p s$; thus $W'$ is a plane perpendicular to the appropriate radius vector of the normal surface. The amplitude of the group will be largest in a region where the waves reinforce each other, i.e. where this plane intersects the planes with neighbouring wave normals. But this is precisely the region around the envelope of these planes, i.e. in the neighbourhood of the corresponding point $v_r t$ on the ray surface. This consideration confirms that the energy carried by the group is propagated with velocity $v_r$ in the direction of the unit vector $t$.

## 9.18.3 14.3. Optical Properties of Uniaxial and Biaxial Crystals

### 14.3.1 The optical classification of crystals

Transparent crystals fall into three distinct groups as regards their optical properties:
Group I. Crystals in which three crystallographically-equivalent, mutually-orthogonal directions may be chosen. These are crystals of the so-called cubic system. The equivalent directions

evidently coincide with the principal dielectric axes and one has $\varepsilon_x = \varepsilon_y = \varepsilon_z (= \varepsilon$ say ); then $\boldsymbol{D} = \varepsilon \boldsymbol{E}$ and the crystal is optically isotropic and equivalent to an amorphous body.

Group II. Crystals not belonging to group I in which two or more crystallographicallyequivalent directions may be chosen in one plane. These are crystals of the trigonal, tetragonal, and hexagonal systems, the plane containing the equivalent directions being perpendicular to the axis of threefold, fourfold, or six-fold symmetry. One dielectric principal axis must coincide with this distinguished direction, whilst for the other two one may choose any orthogonal line pair perpendicular to it. If the distinguished direction is taken as the $z$ axis one then has $\varepsilon_x = \varepsilon_y \neq \varepsilon_z$. Such crystals are said to be optically uniaxial.

Group III. Crystals in which no two crystallographically-equivalent directions may be chosen. These are the crystals belonging to the so-called orthorhombic, monoclinic and triclinic systems. Here $\varepsilon_x \neq \varepsilon_y \neq \varepsilon_z$ and the directions of the dielectric axes may or may not be determined by symmetry (see Table XX) and may therefore be wavelength dependent. Crystals of this group are said to be optically biaxial.

That all crystals fall into these three types as regards their optical properties, may be clearly seen by considering one of the associated ellipsoids, e.g. the ellipsoid of wave normals. Evidently the ellipsoid must be unchanged by the symmetry operations that leave the crystal structure unaltered.* Now there are only two steps in the degeneration of an ellipsoid into a sphere: an ellipsoid has either (a) all axes of

unequal lengths or (b) two axes equal and one unequal (spheroid, i.e. ellipsoid of revolution), or (c) all axes equal (sphere); these correspond to the three groups [in the order (III), (II), and (I)] which we just considered. The terms uniaxial and biaxial refer to the number of optic axes which the ellipsoid has, i.e. the number of diameters with the property that a plane section at right angles to them through the centre of the ellipsoid is a circle. A general ellipsoid has two such diameters (biaxial crystals), a spheroid has one (uniaxial crystals), and a sphere an infinity (isotropic crystals).

Table XX gives a survey of all the possible cases. Principal dielectric axes which may be colour-dependent are shown by two thin lines at a small angle to each other (signifying positions for two wavelengths), fixed axes by thick lines, and freely rotatable (or indeterminate) axes by broken lines ending on a circle or a sphere.

Table XX

| Crystal system | Dielectric axes | Ellipsoid of wave normals | Optical classification |
|---|---|---|---|
| Triclinic | CCC | General ellipsoid | Biaxial |
| Monoclinic | CCF | " " | " |
| Onthophombic | FFF | " " | " |
| Trigonal Tetragonal Hexagonal | FRR | Spheroid | Uniaxial |
| Cubic | RRR | Sphere | Isotropic |
| $C$ = Axis with colour dispersion<br>$F$ = Axis fixed in direction<br>$R$ = Freely rotatable, or indeterminate, axis | | | |

0

- For instance, crystals of the monoclinic class are characterized by either a twofold axis parallel to, or a mirror plane perpendicular to, one of the crystal axes, or both. In either case it is clear that the ellipsoid must have one of its axes parallel to this crystal axis, in order to remain unchanged by the symmetry operation.

Of the many accounts of crystal classes and symmetry operations, we may cite C. W. Bunn, Chemical Crystallography (Oxford, Clarendon Press, 1945), Chapter 2.

### 14.3.2 Light propagation in uniaxial crystals

We start from Fresnel's equation of wave normals §14.2 (24) and write it in the form

$$s_x^2\left(v_p^2-v_y^2\right)\left(v_p^2-v_z^2\right)+s_y^2\left(v_p^2-v_z^2\right)\left(v_p^2-v_x^2\right)+s_z^2\left(v_p^2-v_x^2\right)\left(v_p^2-v_y^2\right)=0. \tag{1}$$

For an optically uniaxial crystal with the optic axis in the $z$ direction, $v_x=v_y$. Writing* $v_o$ in place of this common velocity and $v_e$ in place of $v_z$, (1) reduces to

$$\left(v_p^2-v_o^2\right)\left[\left(s_x^2+s_y^2\right)\left(v_p^2-v_e^2\right)+s_z^2\left(v_p^2-v_o^2\right)\right]=0. \tag{2}$$

Let $\vartheta$ denote the angle which the wave normal $s$ makes with the $z$ axis; then

$$s_x^2+s_y^2=\sin^2\vartheta, \quad s_z^2=\cos^2\vartheta,$$

and (2) becomes

$$\left(v_p^2-v_o^2\right)\left[\left(v_p^2-v_e^2\right)\sin^2\vartheta+\left(v_p^2-v_o^2\right)\cos^2\vartheta\right]=0. \tag{3}$$

The two roots of this equation ( $v_p'$ and $v_p''$ say) are given by
and

$$\left.\begin{aligned} v_p'^2 &= v_o^2 \\ v_p''^2 &= v_o^2\cos^2\vartheta+v_e^2\sin^2\vartheta. \end{aligned}\right\} \tag{4}$$

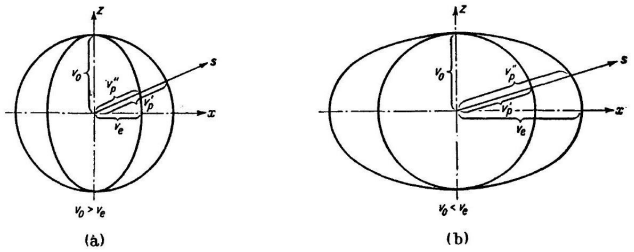



Fig. 14.6. The normal surfaces of a uniaxial crystal.
(a) Positive uniaxial crystal and (b) negative uniaxial crystal.

Equations (4) show that the two shells of the normal surface are a sphere of radius $v_p'=v_0$ and a surface of revolution (of the fourth order), an ovaloid. Thus one of the two waves that corresponds to any particular wave-normal direction is an ordinary wave, with a velocity independent of the direction of propagation, the other an

0

- Suffixes $o$ and $e$ stand here for ordinary and extraordinary, a terminology which will be

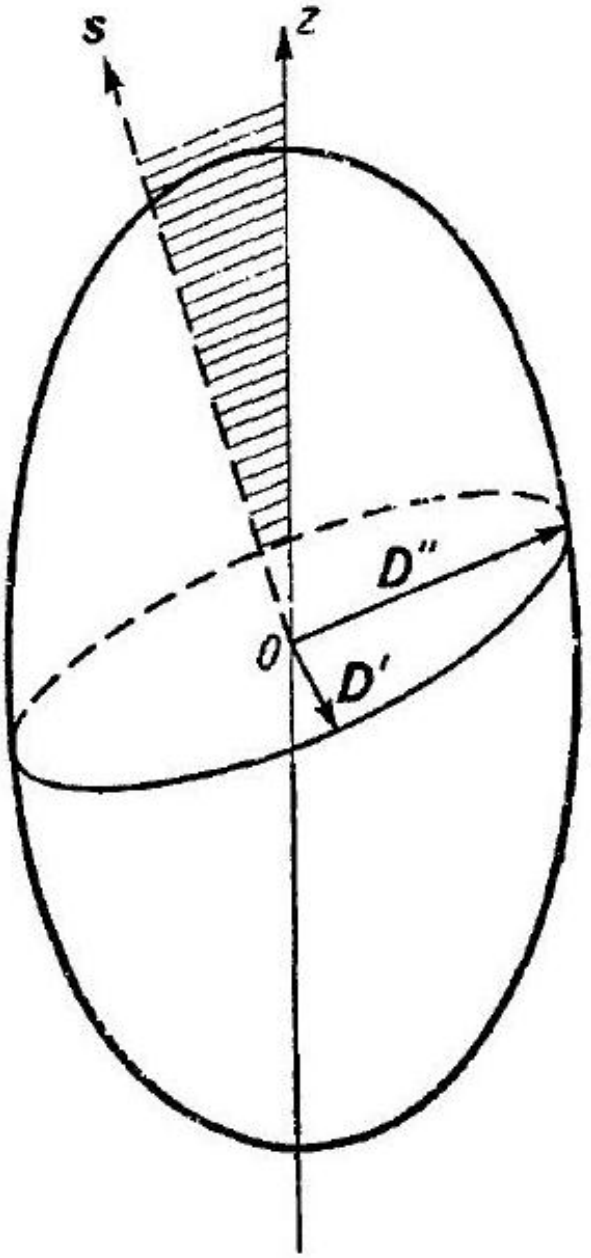



Fig. 14.7. The directions of vibrations in a uniaxial crystal.

extraordinary wave with velocity depending on the angle between the direction of the wave normal and the optic axis. The two velocities are only equal when $\vartheta = 0$, i.e. when the wave normal is in the direction of the optic axis.

When $v_o > v_e$ (see Fig. 14.6(a)), the ordinary wave travels faster than the extraordinary wave (except for $\vartheta = 0$ when they are equal); such a crystal is said to be a positive uniaxial crystal (e.g. quartz). If $v_o < v_e$ (see Fig. 14.6(b)), the ordinary wave travels more slowly than the extraordinary wave, and we speak of a negative uniaxial crystal (e.g. felspar).

The directions of vibration may conveniently be found with the help of the ellipsoid of wave normals, which now has two of its principal axes equal. The plane containing the wave normal $s$ and the optic axis $OZ$ is called the principal plane (shaded in Fig. 14.7). The ellipsoid is symmetrical about this plane. It follows that the elliptical section through $O$ by the plane perpendicular to $s$ is symmetrical about the principal plane, and therefore the principal axes of the ellipse are perpendicular and parallel to the principal plane, as shown in

Fig. 14.7. The semi-axis perpendicular to the principal plane is equal to the radius of the equatorial circle of the spheroid, i.e. it is inversely proportional to the velocity $v_0$ of the ordinary wave. We see that the vector $\boldsymbol{D}$ of the ordinary wave ( $D'$ in Fig. 14.7) vibrates at right angles to the principal plane, the vector of the extraordinary wave ( $D''$ ) is in this plane.

Optical phenomena in uniaxial crystals played a considerable part in the history of optics in connection with the question whether a "light vector" vibrates at right angles or parallel to the plane of polarization (defined as the plane of incidence for reflection from a plane air-dielectric interface at an angle of incidence at which any incident wave becomes linearly polarized, i.e. the ( $H, s$ ) plane in the language of electromagnetic theory; cf. p. 28 and p. 43). Today it is no longer relevant to discuss this question in detail,* since we know that there is no single physical entity that can be identified with a "light vector".

### 14.3.3 Light propagation in biaxial crystals

We shall now investigate the main consequences of the basic equation (1) in the general case of a biaxial crystal. It will help us to visualize the normal surface, if we consider first the sections of this surface by the three coordinate planes $x = 0, y = 0,$

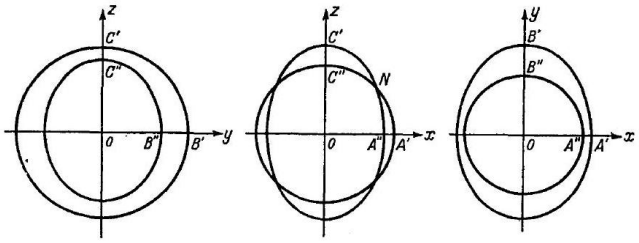



Fig. 14.8. Sections of the normal surface of a biaxial crystal.

and $z = 0$ of our reference system (principal dielectric axes). In order to fix the form of the intersection curves uniquely the three axes will be labelled so that

$$\varepsilon_x < \varepsilon_y < \varepsilon_z, \quad (v_x > v_y > v_z)\,. \tag{5}$$

If we set $s_x = 0$ in equation (1), the equation breaks into two factors, giving

$$\left.\begin{aligned} v_p'^2 &= v_x^2, \\ v_p''^2 &= v_z^2 s_y^2 + v_y^2 s_z^2. \end{aligned}\right\} \tag{6}$$

We set $v_p s_y = y, v_p s_z = z$; then ${v_p}^2 = y^2 + z^2$ and the equations become

$$y^2 + z^2 = {v_x}^2, \quad \left(y^2 + z^2\right)^2 = {v_z}^2 y^2 + {v_y}^2 z^2. \tag{6a}$$

Thus the section of the normal surface by the coordinate plane $x = 0$ is a circle and an oval. The section by each of the other two coordinate planes likewise consists of a circle and an oval, the only difference being in the relative positions of the two curves. With the choice of axes indicated by (5), the circle lies completely outside the oval in the $yz$-plane and lies completely inside it on the $xy$-plane; in the $zx$-plane the circle and the oval intersect in four points (Fig. 14.8). A quarter of the normal surface is shown in perspective in Fig. 14.9 where each of the curves $A'B'C'N$ and $A''B''C''N$ is spanned by a smooth surface. In general two surfaces intersect in a curve, but in this case the two surfaces have only four points in common, the point $N$ and the corresponding points in the other quadrants. The two lines joining the origin

with each of these points are the two optic axes of wave normals. That there are no other such points and, therefore, no other optic axes of wave normals, follows

from the geometrical theorem relating to the number of central circular sections of an ellipsoid, mentioned on p. 675. But we will now confirm this by direct calculations and establish at the same time an inequality that we shall need later.

Let

$$v_x^2 = v_y^2 + q_x, \quad v_z^2 = v_y^2 - q_z, \quad v_y^2 = v_y^2 + q, \tag{7}$$

where $q_x$ and $q_z$ are positive because of (5). With this substitution equation (1) becomes

$$s_x{}^2 q\,(q+q_z) + s_y{}^2\,(q+q_z)\,(q-q_x) + s_z{}^2\,(q-q_x)\,q = 0, \tag{8a}$$

or

$$q^2 + \left[s_x^2 q_z + s_y^2\,(q_z - q_x) - s_z^2 q_x\right] q - s_y^2 q_x q_z = 0. \tag{8b}$$

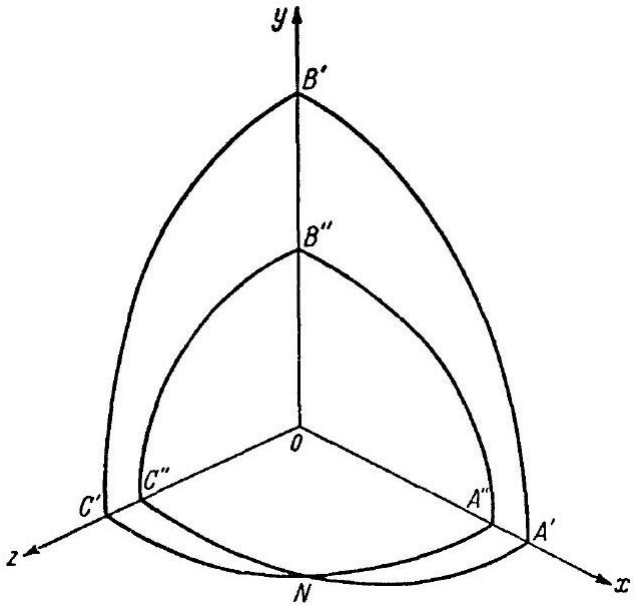



Fig. 14.9. The normal surface of a biaxial crystal.

Since the constant term $-s_y{}^2 q_x q_z$ cannot be positive, the roots of this equation must be real. If we denote them by $q'$ and $q''$, then

$$q'q'' = -s_y^2 q_x q_z \leqslant 0.$$

Hence $q'$ and $q''$ must have opposite signs. Let $q' \geqslant 0$ and $q'' \leqslant 0$.
If $q > q_x$ or $q < -q_z$ the terms on the left-hand side of (8a) would be positive. Hence $q$ must lie in the range $-q_z \leqslant q \leqslant q_x$ and so

$$-q_z \leqslant q'' \leqslant 0 \leqslant q' \leqslant q_x. \tag{9}$$

0

- For an account of the historical background see E. T. Whittaker, History of the Theories of Aether and Electricity, Vol. I: The Classical Theories (London, T. Nelson, 1951), p. 116.

The two roots $q'$ and $q''$ can therefore only be equal if they both are zero; and for this to be the case, one must, according to (8b) have simultaneously

$$s_y{}^2 = 0, \quad s_x{}^2 q_z = s_z{}^2 q_x. \tag{10}$$

Thus we obtain two directions $s$ for each of which the corresponding two velocities are equal, confirming that there are two optical axes of wave normals. As already noted earlier, these axes lie in the $xz$-plane. If $\beta$ is the angle which one of the axes makes with the $z$ direction, then $s_x = \sin\beta, s_z = \cos\beta$ where, according to (10),

$$\tan\beta = \frac{s_x}{s_z} = \pm\sqrt{\frac{q_x}{q_z}} = \pm\sqrt{\frac{v_x^2 - v_y^2}{v_y^2 - v_z^2}}; \tag{11}$$

thus the optic axes are situated symmetrically with respect to the $z$-axis.

In terms of the velocities, the inequalities (9) become

$$v_z^2 \leqslant v_p''^2 \leqslant v_y^2 \leqslant v_p'^2 \leqslant v_x^2 \tag{12}$$

a relation that we shall need later. We see that the two phase velocities are real whatever the direction of the wave normal; this conclusion is, of course, also evident from the geometrical construction for the phase velocities by means of the ellipsoid of wave normals.

The expression for the two phase velocities corresponding to a given wave-normal direction $s$ takes a very simple form if $s$ is specified in terms of the angles $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ which it makes with the two optic axes of wave normals. Since the direction cosines of the optic axes are $\pm\sin\beta, 0, \cos\beta$, the angles $\vartheta_1$ and $\vartheta_2$ are given by the equations

$$\left.\begin{aligned} \cos\vartheta_1 &= s_x\sin\beta + s_z\cos\beta, \\ \cos\vartheta_2 &= -s_x\sin\beta + s_z\cos\beta. \end{aligned}\right\} \tag{13}$$

In terms of $s_x$ and $s_z$, the roots of (8b) are

$$q = -\frac{1}{2}P \pm \frac{1}{2}\sqrt{\Delta}, \tag{14}$$

where, if we also use the identity $s_x{}^2 + s_y{}^2 + s_z{}^2 = 1$,

$$P = s_x^2 q_x - s_z^2 q_z + q_z - q_x \tag{15}$$

$$\begin{aligned} \Delta &= P^2 + 4s_y^2 q_x q_z \\ &= (q_x + q_z)^2 - 2(q_x + q_z)\left(s_x^2 q_x + s_z^2 q_z\right) + \left(s_z^2 q_z - s_x^2 q_x\right)^2 \end{aligned} \tag{16}$$

Now from (13) and (11),

$$\cos\vartheta_1\cos\vartheta_2 = \frac{q_z s_z^2 - q_x s_x^2}{q_x + q_z}, \quad \cos^2\vartheta_1 + \cos^2\vartheta_2 = 2\frac{q_x s_x^2 + q_z s_z^2}{q_x + q_z}. \tag{17}$$

Hence in terms of $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$, (15) and (16) become

$$P = q_z - q_x - (q_z + q_x)\cos\vartheta_1\cos\vartheta_2 \tag{15a}$$

$$\Delta = [(q_z + q_x)\sin\vartheta_1\sin\vartheta_2]^2 \tag{16a}$$

and we finally obtain, on substituting from (15a) and (16a) into (14), and replacing $q, q_x$, and $q_z$ by the expressions (7):

$$v_p^2 = \frac{1}{2}\left[v_x^2 + v_z^2 + \left(v_x^2 - v_z^2\right)\cos\left(\vartheta_1 \pm \vartheta_2\right)\right] \tag{18}$$

Although $v_y$ does not appear in this equation, it is contained implicitly in $\vartheta_1$ and $\vartheta_2$, since these angles depend on $\beta$ and this is a function of all three principal velocities.

In the special case of a uniaxial crystal $\vartheta_1$ and $\vartheta_2$ are equal; it may readily be verified that equation (18) then correctly reduces to (4).

Similar analysis applies to the rays. Starting from the ray equation § 14.2 (29) we find that the section of the ray surface with the coordinate plane $x = 0$ consists of the two curves

$$v_r'^2 = v_x^2, \quad \frac{1}{v_r''^2} = \frac{t_y^2}{v_z^2} + \frac{t_z^2}{v_y^2}. \tag{19}$$

If we set $v_r t_y = y, v_r t_z = z$ then $v_r^2 = y^2 + z^2$ and (19) becomes

$$y^2 + z^2 = v_x^2, \quad \frac{y^2}{v_z^2} + \frac{z^2}{v_y^2} = 1 \tag{19a}$$

i.e. the section of the ray surface by the plane $x = 0$ is a circle and an ellipse (not a more general type of an oval as in (6a)). Each of the sections by the other two coordinate planes is likewise a circle and an ellipse. Because of the inequalities (5), the circle encloses the ellipse in the $yz$-plane, whilst in the $xy$-plane the ellipse encloses the circle; and in the $xz$-plane the circle and the ellipse intersect in four points. These points specify the position of the optic ray axes $R_1, R_2$; their directions are given by the dual of (10):

$$t_y^2 = 0, \quad t_x^2\left(\frac{1}{v_y^2} - \frac{1}{v_z^2}\right) = t_z^2\left(\frac{1}{v_x^2} - \frac{1}{v_y^2}\right). \tag{20}$$

The angle $\gamma$ between either of these axes and the $z$ axis is therefore given by

$$\tan\gamma = \frac{t_x}{t_z} = \pm\frac{v_z}{v_x}\sqrt{\frac{v_x^2 - v_y^2}{v_y^2 - v_z^2}} = \pm\frac{v_z}{v_x}\tan\beta. \tag{21}$$

If $v_z < v_x$ the optic ray axis makes a smaller angle with the $z$ axis than the axis of normals.

### 14.3.4 Refraction in crystals

(a) Double refraction. Consider a plane wave incident from vacuum on a plane surface $\Sigma$ of an anisotropic medium. It will give rise to a transmitted and reflected field. We shall briefly consider the nature of the transmitted field, using substantially the same argument as in the case of an isotropic medium (§ 1.5.1). We shall confine our attention to finding the direction of propagation of the disturbance within the crystal; expressions for the amplitude ratios (corresponding to the Fresnel formulae) will not be investigated.*

Let $s$ be the unit wave normal of the incident wave and $s'$ that of the transmitted wave. We shall see shortly that in general two waves are transmitted so that there are two possible values of $s'$. The field vectors of the incident wave and of the transmitted waves are functions of $(t - r.s/c)$ and $(t - r.s'/v')$ respectively. The continuity of the field across the boundary demands that for any point $r$ on the plane $\Sigma$ and for all times $t$,

$$t - \frac{r \cdot s}{c} = t - \frac{r \cdot s'}{v'}$$

i.e.

$$r \cdot \left( \frac{s'}{v'} - \frac{s}{c} \right) = 0. \tag{22}$$

Hence the vector $s'/v' - s/c$ must be perpendicular to the boundary.
The permissible wave normals $s'$ may be determined as follows: With any point $O$ on $\Sigma$ as origin we plot vectors in all directions $s'$, each of the length $1/v'$, where $v'$ is the phase velocity corresponding to each $s'$ in accordance with Fresnel equation § 14.2 (24). The locus of the end points is a two-sheeted surface which differs from the normal surface in that each radius vector is of length $1/v'$ instead of $v'$. We call this surface the inverse surface of wave normals; it is the dual of the ray surface and, therefore, like the ray surface itself, is of the fourth degree. Since the required vector $s'/v'$ must be such that $s'/v' - s/c$ is perpendicular to $\Sigma$, its end point $Q'$ must be on the normal to $\Sigma$ through the end point $P$ of the vector $s/c$. In general, the

normal to $\Sigma$ cuts the inverse surface in four points, two of which lie on the same side of the boundary as the crystal. Hence there are two such points ( $Q'$ and $Q''$ in Fig. 14.10), and therefore two possible wave-normal directions, so that in general each incident wave will give rise to two refracted waves; to each of these waves there corresponds a ray direction and a ray velocity describing the propagation of the energy within the crystal.* This is the phenomenon of birefringence or double refraction. It is illustrated by the well-known effect that two images are observed when a small object is viewed through a slab of calcite.

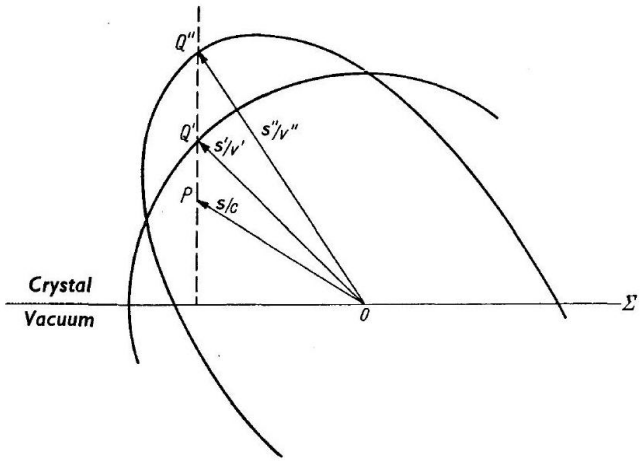



Fig. 14.10. Double refraction: construction for permissible wave normals.

Taking the boundary plane $\Sigma$ as the plane $z = 0$, (22) becomes

$$\frac{xs_x + ys_y}{c} = \frac{xs'_x + ys'_y}{v'} = \frac{xs''_x + ys''_y}{v''}, \tag{23}$$

which must be satisfied for all values of $x$ and $y$. This implies, firstly, that $s'_x/s_x = s'_y/s_y$ and $s''_x/s_x = s''_y/s_y$, so that the two refracted rays lie in the plane of incidence. Further, if $\theta_i, \theta'_t$, and $\theta''_t$ are the angles which the incident and the two transmitted waves make with the axis, we have from (23)

$$\frac{\sin\theta_i}{\sin\theta'_t} = \frac{c}{v'}, \quad \frac{\sin\theta_i}{\sin\theta''_t} = \frac{c}{v''}. \tag{24}$$

0

• These are discussed, for example, in G. Szivessy, Hdb. d. Phys., Vol. 20 (Berlin, Springer, 1928), 715.

Thus each of the transmitted waves is seen to obey the same law of refraction as in the case of isotropic media. However, the velocity $v$ now depends on $\theta_t$, so that the determination of the direction of propagation in the crystal is more complicated. In a uniaxial crystal one sheet of the inverse wave-normal surface is a sphere, so

that the phase velocity of one of the transmitted waves is then independent of $\theta_t$, this being an ordinary wave.

In the special case of normal incidence ( $\theta_i = 0$ ) one has $\theta'_t = \theta''_t = 0$, so that the two wave-normals in the crystal coincide and are in the direction of the normal to $\Sigma$. Another special case, and one of a great theoretical interest, is that of a wave propagated in the direction of one of the optic axes of a biaxial crystal. It gives rise to a phenomenon known as conical refraction, which we shall now consider.

(b) Conical refraction. We mentioned earlier that when $s$ coincides with one of the optic axes of the wave normals in a biaxial crystal, the relation between $\boldsymbol{s}$ and $\boldsymbol{t}$ has

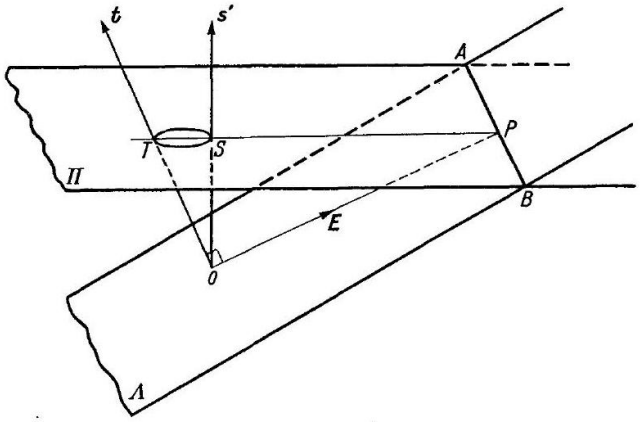



Fig. 14.11. Illustrating the position of rays that correspond to an optic axis of wave normals in a biaxial crystal.

a singularity. We must investigate the nature of this singularity before discussing the refraction phenomena associated with propagation in this special direction.

We have shown in § 14.2.3 that the electric displacement vectors $D$ associated with a wave-normal direction $s$ are parallel to the principal axes of the elliptical section through the ellipsoid of wave normals at right angles to $s$. Now in the special case when the vector $s$ is in the direction of the optic axis of wave normals, the section is circular, so that all directions of $D$ at right angles to $s$ are then permissible; consequently an infinity of directions of the electric vector $\boldsymbol{E}$ (calculated for each $D$ from § 14.1 (1)) and an infinity of ray directions $t$ (determined as in Fig. 14.1) are possible in this case. We shall now show that all these $\boldsymbol{t}$ vectors lie on the surface of a cone.

Let $s'\,(s'_x, 0, s'_z)$ be the unit vector along one of the optic axes of wave normals, referred to the principal dielectric axes, subject to the inequality (5). Then $s'_x$ and $s'_z$ are related by (11)

0

- The origin of the double refraction may also be illustrated by generalizing Huygens' construction (§ 3.3.3) to anisotropic media. A proper formulation of this approach is, however, by no means as simple as is usually given in textbooks, since Huygens' construction operates with wavelets proceeding from point sources, and not with independent plane waves, the laws of propagation of which are as a rule taken over without justification. Some of the difficulties inherent in this approach were discussed by M. G. Lam in his Leons sur la thorie mathmatique de l'lasticit des corps solides (Paris, Gauthier-Villars, 2nd ed., 1866), and V. Volterra, Acta Math., 16 (1892), 153.

and the permissible $D$ vectors, being perpendicular to $s'$, satisfy the relation,

$$s'_x D_x + s'_z D_z = 0$$

In terms of the components of the corresponding $\boldsymbol{E}$ vectors, this relation may be written as

$$s'_x \varepsilon_x E_x + s'_z \varepsilon_z E_z = 0. \tag{25}$$

In Fig. 14.11 let Π be any plane perpendicular to $s'$ and let the line in which the electric vector $\boldsymbol{E}$ is localized cut it at $\boldsymbol{P}$. Since according to (25) all the $\boldsymbol{E}$ vectors must lie in a plane Λ perpendicular to the vector with components $(s'_x\varepsilon_x, 0, s'_z\varepsilon_z)$, all the possible points $P$ must lie on the straight line $AB$ in which the planes Π and Λ intersect. Now the ray vector $t$ is coplanar with $E$ and $s'$ and perpendicular to $E$. Let $t$ and $s'$ cut the plane Π in the points $T$ and $S$ respectively. Then, by similar triangles,

$$TS.SP = OS^2 = \text{ constant } . \tag{26}$$

Since the locus of $P$ is the straight line $AB$, the locus of $T$ is the inverse of $AB$, that is*, a circle passing through the centre of inversion $S$ and with tangent at $S$ parallel to $AB$. Hence there is an infinity of rays corresponding to the optic axis of wave normals, and these rays form a surface of a cone. This cone is not a circular one, since the centre of the circle is not the foot of the perpendicular from $O$ to the plane Π.

If $E, s'$, and $t$ all are in the $xz$-plane, the direction of $t$ must be the direction $(s'_x\varepsilon_x, 0, s'_z\varepsilon_z)$ to which $E$ is always perpendicular. If this direction makes an angle $\phi$ with the $z$ axis, the aperture angle $\chi$ of the cone in this plane is given by

$$\tan\chi = \tan(\beta - \phi) = \frac{\frac{s'_x}{s'_z} - \frac{s'_x\varepsilon_x}{s'_z\varepsilon_z}}{1 + \frac{\varepsilon_x s'_x}{\varepsilon_z {s'_z}^2}} = \frac{1}{{v_y}^2}\sqrt{({v_x}^2 - {v_y}^2)\,({v_y}^2 - {v_z}^2)} \tag{27}$$

where (11) has been used. Usually $({v_x}^2 - {v_y}^2)/{v_y}^2 \ll 1, ({v_y}^2 - {v_z}^2)/{v_y}^2 \ll 1$, so that the cone is very nearly circular and of angle $\chi$.

We may show in a strictly similar manner that there is an infinity of wave normals corresponding to the optic ray axis, and these wave normals form the surface of a cone. The aperture angle $\psi$ of this cone is given by the dual of (27), i.e. by

$$\tan\psi = \frac{1}{v_x v_z}\sqrt{(v_x^2 - v_z^2)\left(v_y^2 - v_z^2\right)} = \frac{v_y^2}{v_x v_z}\tan\chi \tag{28}$$

The situation may conveniently be illustrated in terms of the normal surface and the ray surface, using the property, derived in § 14.2.3, that the normal surface is the pedal surface of the ray surface. The intersection of these surfaces with the $xz$ plane is shown in Fig. 14.12. The surface of normals intersects this plane in a circle of radius $v'_p = v_y$, and in an oval with the polar radius $v''_p$, whilst the ray surface intersects it in the same circle $v'_r = v_y$ and an ellipse $v''_r$. If the circle and the oval intersect at $N$, the line $ON$ is in the direction of the optic axis of wave normals, and the plane through $N$ perpendicular to $ON$ must touch the ray surface at all points in which the cone of permitted ray directions cuts this surface. The ray surface thus has the unusual property that certain tangent planes touch it at an infinity of points. †

The ray optic axis is represented by $OR$, where $R$ is the point of intersection of the two sheets of the ray surface; the two sheets intersect in such a way that there is an infinite number of tangent planes at $R$, with their normals lying on a cone. The normals from $O$ to these planes form the cone of wave normal directions corresponding to the ray direction $OR$. The aperture angles $\chi$ and $\psi$ of the two cones are also shown

0

in the figure. For reasons that will become evident shortly, the cone belonging to $N$, i.e. formed by rays such as $OA$, is called the cone of internal conical refraction; the cone belonging to $R$, i.e. formed by wave normals such as $OB$, is called the cone of external conical refraction.

Consider now a plate of a biaxial crystal, e.g. aragonite, cut so that its two parallel faces are perpendicular to the optic axis of wave normals. If this be illuminated by a narrow beam of monochromatic light incident normally on one of the faces of the plate, the energy will spread out in the plate in a hollow cone, the cone of internal conical refraction, and on emerging at the other side it will form a hollow cylinder, as shown in Fig. 14.13. Thus we should expect to see on a screen parallel to the crystal

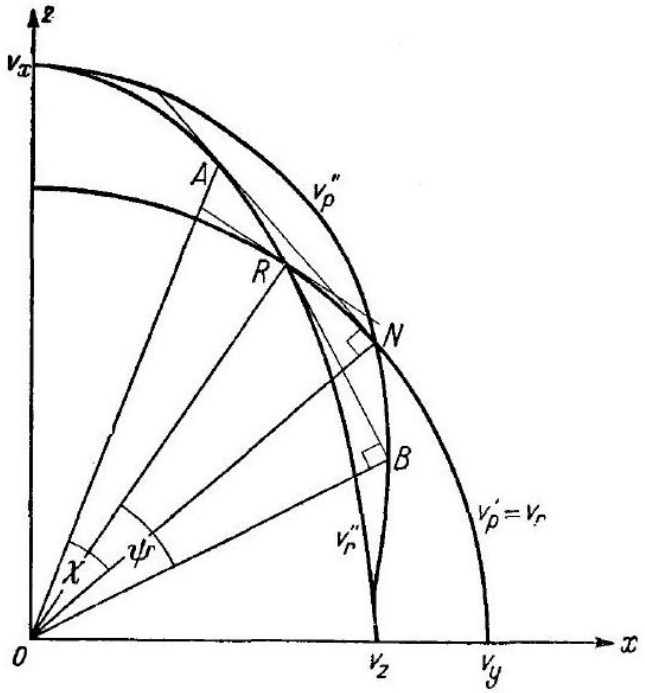



Fig. 14.12. Conical refraction: construction of the cones.

• See, for example, D. M. Y. Sommerville, Analytical Conics (London, Bell and Sons, 1941), p. 92.
† The properties of this surface are discussed in detail in G. Salmon, Analytical Geometry of Three Dimensions. Revised by R. A. P. Rogers, Vol. 2 (London, Longmans, Green & Co., 1915, 5th ed.), Chapter IV.

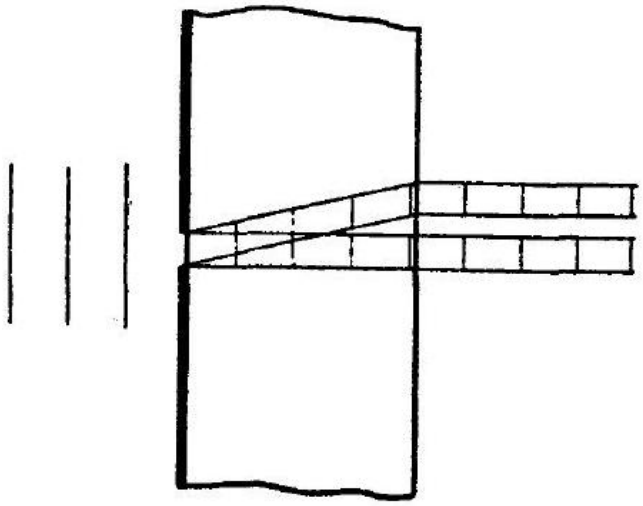


Fig. 14.13. Internal conical refraction.

face, a bright circular ring. This remarkable phenomenon was predicted in 1832 by Sir William Rowan Hamilton, and confirmed a year later by Lloyd, who investigated it in aragonite on Hamilton's instigation. The success of this experiment represented one of the most striking confirmations of Fresnel's wave theory of light and contributed greatly to its general acceptance (see Historical Introduction, p. xxiv).

In practice the demonstration of the phenomenon of conical refraction is less simple than we have just indicated, for naturally it is impossible to obtain an accurately parallel beam of monochromatic light. In an experiment one always has to use beams of finite angular aperture, and in this case, as PoggendorfF* and Haidinger † were the first to show, one observes two bright circles separated by a fine dark circle, as illustrated in Fig. 14.14. In Lloyd's first experiments this structure escaped observation, as the apertures limiting the width of his beam were too large and the two bright circles were therefore blurred into one. It remained unexplained for a long time after its discovery, until Voigt ‡ gave it an interpretation which may be summarized as follows:

0

- J. C. Poggendorff, Pogg. Ann., 48 (1839), 461.
  † W. v. Haidinger, Wiener Ber. (2), 16 (1854), 129 ; Pogg. Ann. (4), 96 (1855), 486.
  ‡ W. Voigt, Phys. Z. 6 (1905), 672, 818.



Fig. 14.14. Light distribution arising from conical refraction.

We have to consider the propagation of waves whose normals are slightly inclined to the optic axis. Each of the wave normals will give rise to two rays inside the crystal, and we should expect that the directions of these rays will differ only slightly from the directions of the generators of the cone of internal conical refraction. In order to find how the transmitted rays are distributed, we must consider part of the ray surface near the circle of contact with the tangent plane $AN$ in Fig. 14.12. This part

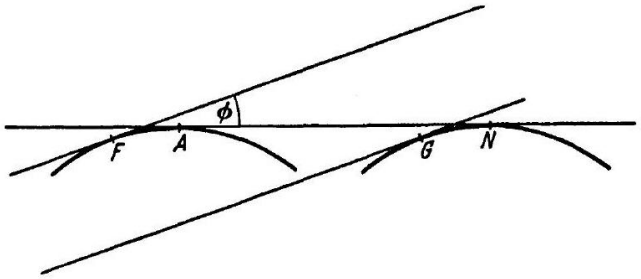



Fig. 14.15. Illustrating the position of rays belonging to wave normals that are inclined at small angles to an optic axis of wave normals.

of the surface may be likened to part of an inflated inner tube of a motor-car tyre, and the tangent plane to a flat board lying on the tyre. Fig. 14.15 shows the section with the $xz$-plane. Now the points on the ray surface which represent the directions of the two rays corresponding to the wave normal direction $s$ are the points of contact of the two tangent planes to the ray surface perpendicular to $s$ as shown in Fig. 14.5. When the wave normal $ON$ is slightly displaced from the optic axis, the tangent plane splits into two parallel planes, one of which rolls over the surface so that its point of contact moves away from the centre of the circle of contact to $F$, while the second (which cannot be represented in our model because it would have to cut through the inner tube) rolls so that its point of contact moves towards the centre to $G$. Fig. 14.15 shows this for a displacement of the wave normal in the $xz$-plane, but the same will happen for a displacement in any direction.

From these remarks it is seen that all incident rays with wave normals inclined at a small angle $\phi$ to the optic axis will give rise to pairs of rays inclined at angles $\frac{1}{2}\chi + a\phi$

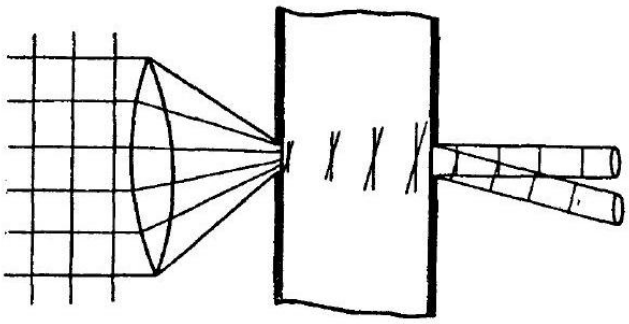


Fig. 14.16. External conical refraction.

and $\frac{1}{2}\chi - a\phi$ to the central axis of the cone of inner conical refraction, where $a$ is some constant. Thus all the energy associated with the angular range $\phi$ and $\phi + d\phi$ in the incident beam will appear in two cones of angular semi-apertures $\frac{1}{2}\chi \pm a\phi$ and angular separation $ad\phi$. But the corresponding energy in the incident beam is proportional to $\phi d\phi$, up to the maximum value of $\phi$. It follows that, at angles $\frac{1}{2}\chi \pm a\phi$, the intensity in the cone of rays is proportional to $\phi$, and in particular that it is zero at $\phi = 0$. Thus we must expect two bright circles to appear, with a dark circle between them, as is observed.

External conical refraction demonstrates the fact established earlier on that there is a whole cone of wave-normal directions corresponding to a given ray direction. This phenomenon is observed with a crystal plate cut so that its faces are perpendicular to the optic ray axis. Small apertures are placed on each face exactly opposite to each other, and one of them is illuminated with convergent light, as shown in Fig. 14.16.

Only those rays will reach the second aperture which have their directions very close to the direction of the optic ray axis, so that the waves reaching the second aperture all have their normals near the cone of the outer conical refraction. A cone of light will therefore emerge from the crystal. The angular aperture of this cone will be greater than the true angle $\psi$ of the outer conical refraction, because of refraction from emergence from the crystal. Again, two concentric circles of light are observed on a screen parallel to the crystal face, and the explanation of this double circle is similar to that given for the case of internal conical refraction.

## 9.18.4 14.4. Measurements in Crystal Optics

In this section we shall describe briefly methods for determining the character of a crystal (i.e. whether uniaxial or biaxial), the position of its optic axes and the values of its principal refractive indices. The optic axes may be located, as we shall see, from observation of interference fringes on crystal plates; the structure of the pattern indicates clearly the intersection of the optic axes with the faces of the plate. The principal refractive indices may be determined with the help of crystal prisms, from measurements of the angles of deviation or of total reflection.

As a preliminary we must discuss the production and analysis of polarized light.

### 14.4.1 The Nicol prism

One of the most commonly used instruments for the production of linearly polarized light is the Nicol prism.* It consists of a natural rhomb of calcite (uniaxial crystal) which is cut into two equal parts along a diagonal plane (represented by $AC$ in Fig. 14.17), and with the two

parts cemented together with Canada balsam. The rhomb is about three times as long as it is wide, with the angles at $B$ and $D$ of its principal section equal to 71°; the end faces $AD$ and $BC$ are ground off to reduce this angle to 68°.

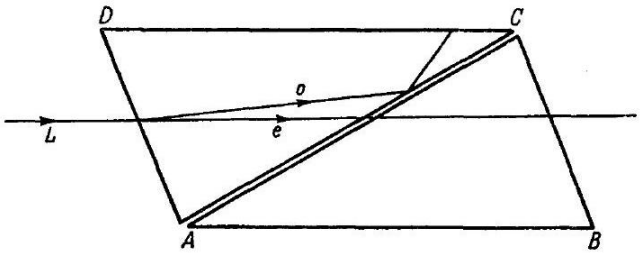



Fig. 14.17. The Nicol prism.

A ray of light incident in the direction $L$ parallel to the long edge is split into an ordinary ray and an extraordinary ray. For the first ray the Canada balsam is of lower, for the second of higher, optical density (calcite crystal: $n_0 = 1.66, n_e = 1.49$; Canada balsam $n = 1.53$ ), and it may easily be verified from the formulae of § 1.5.4 that on the Canada balsam interfaces conditions for total reflection are satisfied with respect to the ordinary ray; this ray is totally reflected towards the face $DC$ which is blackened and so absorbs it. The extraordinary ray passes through the prism with practically no lateral displacement and is linearly polarized with its $\boldsymbol{D}$ vector in the principal section (see § 14.3.2). Thus the Nicol prism produces linearly polarized light, whose direction of vibration is known.

If the incident ray is inclined to the edge of the rhomb the Nicol prism will still act as a polarizer, provided that for the ordinary ray the angle of incidence on the Canada balsam is not smaller than the critical angle; this limits the angle of the cone of incident rays in air for which the prism is effective to about* 30°.

Fairly pure linearly polarized light may also be obtained by passing natural light through a sheet of an absorbing crystalline material whose absorption coefficients for the two directions of vibrations are appreciably different. We shall consider such "polaroid sheets" in § 14.6.3.

An arrangement such as a Nicol prism, which produces linearly polarized light from light of other states of polarization, is called a polarizer. Such an arrangement may also be used as an analyzer, i.e. as a detector of linearly polarized light and of its direction of vibration. To detect linearly polarized light with a Nicol prism we only have to rotate the prism about its longitudinal axis and note whether there is a position when no light is passed through it. If there is such a position the light is linearly polarized and the direction of vibration of its $D$ vector is perpendicular to the principal section.

### 14.4.2 Compensators

Crystalline material may also be used to investigate elliptically polarized light, i.e. the directions of its axes and the ratio of their lengths. For this purpose a crystal plate of suitable material and thickness is employed, and by means of it a path difference is introduced between vibrations in two mutually orthogonal directions. In particular the path difference may be made such as to convert elliptically polarized light into light that is linearly polarized; from the analysis of the linearly polarized light the required information about the elliptically polarized light is then obtained. Such a device is called a compensator, as its function is to compensate

0

- W. Nicol, Edinburgh New Philos. Journ., 6 (1829), 83.

a phase difference. †
(a) The quarter-wave plate. Consider a plane parallel crystal plate of thickness $h$. Let the $z$ axis be normal to the plate and the $x$ and $y$ axes in the directions of the corresponding $D$-vibrations. We may assume that the plate is turned round the normal until the $x$ and $y$ axes are parallel to the principal axes of the vibrational ellipse of the incident light. The components of the $\boldsymbol{D}$ vector of the incident light are then represented by

$$\left.\begin{aligned} D_x^{(i)} &= a\cos\omega t \\ D_y^{(i)} &= b\sin\omega t \end{aligned}\right\} \tag{1}$$

On traversing the plate the two components will suffer different changes in phase because the velocities of the two rays are different. Neglecting reflection losses, the components of the $\boldsymbol{D}$ vector of the light emerging from the plate are given by

$$\left.\begin{aligned} D_x^{(t)} &= a\cos(\omega t+\delta') \\ D_y^{(t)} &= b\sin(\omega t+\delta'') \end{aligned}\right\} \tag{2}$$

where

$$\delta' = \frac{2\pi}{\lambda}n'h, \quad \delta'' = \frac{2\pi}{\lambda}n''h, \tag{3}$$

$\lambda$ being the vacuum wavelength.* Hence the phase difference introduced by the plate is

$$\delta''-\delta' = \frac{2\pi}{\lambda}(n''-n')\,h. \tag{4}$$

In particular, if the emergent light is to be linearly polarized, one must have $\delta''-\delta' = \pm\frac{1}{2}\pi$, or more generally $\delta''-\delta' = (2m+1)\pi/2$, where $m$ is any integer, so that the plate must be of thickness

$$h = \left|\frac{2m+1}{n''-n'}\right|\frac{\lambda}{4}. \tag{5}$$

With this compensator the direction of the linear vibration of the transmitted light is given by

$$\frac{D_y^{(t)}}{D_x^{(t)}} = \pm\frac{b}{a}. \tag{6}$$

A compensator which introduces a phase difference $|\delta''-\delta'| = \pi/2$, i.e. one for which the difference in the two optical thicknesses is a quarter of a wavelength, is called a quarter-wave plate †. It may conveniently be made of a foil of mica (biaxial crystal), split to the thickness $h = \lambda/4\,|n''-n'|$.

0

- There are several modifications of the Nicol prism which can be used over a larger angular field. For the description of these see, for example, L. C. Martin, An Introduction to Applied Optics, Vol. 1 (London, Pitman, 1930), p. 204, or R. W. Wood, Physical Optics (New York, Macmillan, 3rd ed., 1934), pp. 337-8.
† For a detailed survey of compensators see H. G. Jerrard, J. Opt. Soc. Amer., 38 (1948), 35. A survey of methods available for the analysis of polarized light is given in a paper by M. Richartz and Hsien-Y Hs, ibid., 39 (1949), 136.

A systematic procedure for the theoretical analysis of fully or partially polarized light is afforded by the coherency matrices or by the Stokes parameters, discussed in § 10.8.

Elliptically polarized light may be analyzed with it in the following way: the light is passed through a quarter-wave plate and then through a Nicol prism and both are rotated independently until the field seen through the Nicol prism becomes completely dark. This is the position where the axes of the mica plate are parallel to the axes of the vibrational ellipse of the incident light, and where, according to (6), the Nicol prism is set so as to extinguish linearly polarized light with its $D$ vector inclined at an angle $\tan^{-1} b/a$ to the $x$ axis. Thus the orientation of the axes of the ellipse and their ratio are found.

If the incident light is circularly polarized, then $b = +a$ or $b = -a$ and the $D$ vector of the transmitted light is linearly polarized in a direction which makes an angle of 45° or 135° respectively with the axis $OX$. The former corresponds to lefthanded, the latter to right-handed polarization.

Since the thickness of the quarter-wave plate depends on $\lambda$, exact compensation is only possible for monochromatic light of one particular wavelength. In order to achieve compensation for light of any given wavelength a wedge or a combination of wedges must be used in place of a single parallel-sided plate. We shall next consider some compensators of this type.

(b) Babinet's compensator. The compensator due to Babinet ‡ allows the realization of all phase differences (including zero). It consists of two wedges of quartz (positive uniaxial crystal), with equal acute angles. The wedges are placed against each other

as shown in Fig. 14.18, and can be displaced along their plane of contact, thus forming a parallel plate of variable thickness. In one wedge the optic axis is parallel, in the other at right angles, to the edge.

Let $n_o$ and $n_e$ be the ordinary and extraordinary refractive indices of quartz and $h_1$ and $h_2$ the thickness of the two wedges at some particular point. On passing through both wedges; the phase difference between the two rays is

$$\delta = \frac{2\pi}{\lambda} (n_e - n_o) (h_1 - h_2) . \tag{7}$$

The two contributions enter here with opposite signs, since the ray whose vector vibrates at right angles to the principal axis (the ordinary ray) is the faster one, and consequently the component vibrating in a direction parallel to the edge will be ahead of the other component in one wedge and lag behind it in the other. Now $h_1$ and $h_2$ and consequently also $\delta$ change continuously as the point of incidence moves across the plate, and $\delta$ is zero in the middle. In consequence there will be a series of lines along which the transmitted light is linearly polarized and can

0

• In contrast to § 1.3 we write $\lambda$ rather than $\lambda_0$ here, as suffix zero refers throughout this section to the ordinary ray.

† G. Airx, Trans. Camb. Phil. Soc., 4 (1833), 313.

‡ J. Babinet, C.R. Acad. Sci., Paris, 29 (1849), 514. J. Jamin, Ann. Chim. (Phys.), (3), 29 (1850), 274.

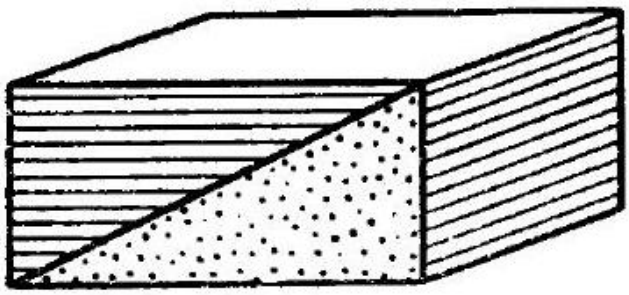


Fig. 14.18. Babinet's compensator.

thus be extinguished by a suitably oriented Nicol prism.

Taking the $x$ axis parallel to the edge, the elliptically polarized incident light may be represented by

$$\left.\begin{aligned} D_x{}^{(i)} &= a_1 \cos(\omega t + \delta_o') \\ D_y{}^{(i)} &= a_2 \cos(\omega t + \delta_o'') \end{aligned}\right\} \tag{8}$$

The light transmitted by the compensator is represented by

$$\left.\begin{aligned} D_x^{(t)} &= a_1 \cos(\omega t + \delta') \\ D_y^{(t)} &= a_2 \cos(\omega t + \delta'') \end{aligned}\right\} \tag{9}$$

where

$$\left.\begin{aligned} \delta' &= \delta_0' + \tfrac{2\pi}{\lambda} n_o (h_1 - h_2), \\ \delta'' &= \delta_0'' + \tfrac{2\pi}{\lambda} n_e (h_1 - h_2). \end{aligned}\right\} \tag{10}$$

For the emergent light to be linearly polarized one must have $\delta'' - \delta' = m\pi (m = 0, \pm 1, \pm 2, ..)$, i.e.

$$\delta_0'' - \delta_0' = -\frac{2\pi}{\lambda}(n_e - n_o)(h_1 - h_2) + m\pi. \tag{11}$$

The direction of the linear vibrations is given by

$$\frac{D_y^{(t)}}{D_x^{(t)}} = \pm \frac{a_2}{a_1}. \tag{12}$$

Suppose first that a Nicol prism is placed in front of the compensator so that the compensator receives linearly polarized light. Then $\delta_0'' = \delta_0'$, and if the compensator is followed by an analyzer that is crossed with the polarizing Nicol prism, dark bands appear running parallel to the edge through points for which the right-hand side of
(11) is a multiple of $\pi$. These dark bands determine the zero position. If next elliptically polarized light is examined by means of the compensator and the analyzing prism alone, the displacement of the dark bands from the zero position immediately determines the phase difference $\delta_0'' - \delta_0'$ of the incident light; and the amplitude ratio of the components of $D^{(i)}$ parallel and perpendicular to the edge can be determined from the orientation of the analyzer, using (12). From these data the position of the principal axes of the vibrational ellipse and their ratio can be found by application of the formulae of § 1.4.2.

(c) Soleil's compensator. For some purposes it is necessary to produce a phase difference (positive, negative, or zero) which is constant over the whole field of view. This may be achieved by a compensator due to

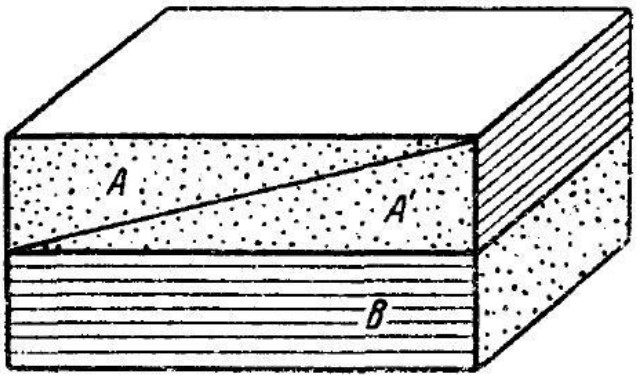



Fig. 14.19. Soleil's compensator.

Soleil.* It contains two quartz wedges $A$ and $A'$ which form, as in Babinet's compensator, a plane parallel plate, but with the difference that the optic axes in both wedges are now parallel to the edges. The lower wedge is cemented to a plane parallel quartz plate $B$ whose optic axis is at right angles to the edge (see Fig. 14.19). The effective path difference which this compensator introduces between the two rays is evidently $h_B - (h_A + h'_A)$. In the zero position, where this difference vanishes, the whole field of view can be obscured by a suitably oriented analyzer. The effective path difference can be altered by shifting the upper wedge, but for each position the path difference remains constant over the whole field. The analysis of elliptically polarized light is similar to that with Babinet's compensator.
(d) Berek's compensator. A compensator useful in biological microscopy (e.g. for measuring path differences in threads of birefringent material) is due to Berek. † It consists of a slab of a uniaxial crystalline medium with its optic axes perpendicular to the faces of the slab. Phase differences between the ordinary and extraordinary rays are introduced by tilting the slab. It is mounted so that it can be turned by a graduated wheel which may be calibrated to show directly the path differences introduced.

### 14.4.3 Interference with crystal plates

As remarked earlier, the directions of optic axes may be determined by means of interference phenomena on crystal plates. The beautiful interference effects which are then observed, and which give a striking demonstration of the agreement between theory and experiment in crystal optics, deserve to be studied in their own rights.

Consider first a beam of linearly polarized light emerging from a polarizer and incident normally on to a plane parallel crystal plate of thickness $h$. On entering the plate, each ray is divided into two rays with different velocities of propagation, and with their $\boldsymbol{D}$ vectors vibrating in two mutually orthogonal directions at right angles to the direction of the plate normal. They emerge from the plate with a certain phase difference $\delta$. If an analyzing Nicol prism is placed behind the plate, components of the two vibrations in a certain direction are then singled out and may be brought to interference in the focal plane of a lens placed behind the analyzer.

In Fig. 14.20, the plane of the drawing is parallel to the plate. $D'$ and $D''$ represent the two mutually orthogonal directions of vibrations in the crystal, and $OP$ and $OA$ are the directions of vibrations that are passed by the polarizer and the analyzer respectively. Let $\phi$ be the angle

0

- H. Soleit, C.R. Acad. Sci., Paris, 21 (1845), 426; 24 (1847), 973 ; 26 (1848), 162. J. DuboscQ and H. Soleii, ibid., 31 (1850), 248.
† M. Berek, Zbl. Miner. Geol. Palont. (1913), 388, 427, 464, 580.

that $OP$ makes with $D'$ and $\chi$ the angle between $OA$ and $OP$. The amplitude of the light incident on the plate is represented by the vector $OE$ (parallel to $OP$ ); its components in the directions of $D'$ and $D''$ are

$$OB = E\cos\phi, \quad OC = E\sin\phi. \tag{13}$$

The analyzer transmits only the components parallel to $OA$ which, as may be seen from the figure and from (13), have amplitudes

$$OF = E\cos\phi\cos(\phi-\chi), \quad OG = E\sin\phi\sin(\phi-\chi). \tag{14}$$

On leaving the plate, the two components differ in phase by the amount

$$\delta = \frac{2\pi}{\lambda}(n''-n')\,h. \tag{15}$$

According to § 7.2 (15), the intensity obtained from the interference of two monochromatic waves with phase difference $\delta$ is given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta,$$

where $I_1$ and $I_2$ are the intensities (squared amplitudes) of the two waves. With the amplitudes given by (14) we have

$$I = E^2\left\{\cos^2\chi - \sin 2\phi\sin 2(\phi-\chi)\sin^2\frac{\delta}{2}\right\}, \tag{16}$$

where the identity $\cos\delta = 1-2\sin^2\frac{\delta}{2}$ has been used.

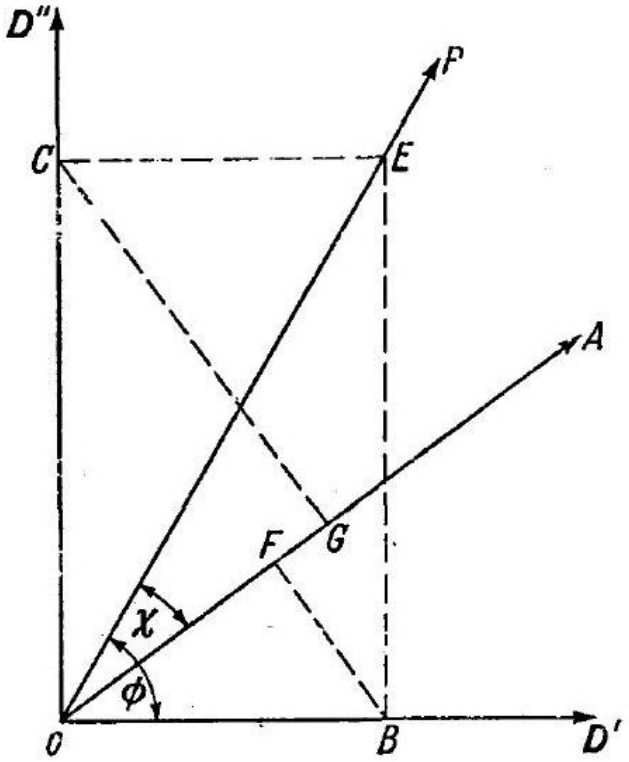



Fig. 14.20. Construction of the vibration components transmitted by a polarizer and an analyzer.

If the plate were removed ( $\delta = 0$ ), the intensity would be $I = E^2 \cos^2 \chi$; thus the second term in (16) represents the effect of the crystal plate.

We now consider two important special cases:

(a) Analyzer and polarizer parallel ( $\chi = 0$ ). In this case (16) reduces to

$$I_1 = E^2 \left(1 - \sin^2 2\phi \sin^2 \frac{\delta}{2}\right). \tag{17}$$

There is maximum transmission when

$$\phi = 0, \quad \frac{\pi}{2}, \quad \pi, \ldots, \tag{18}$$

i.e. when the direction of vibrations passed by the analyzer coincides with one of the directions of vibrations in the plate. The positions (18) are separated by positions of minima of transmission, given by $\sin 2\phi = \pm 1$, i.e. by

$$\phi = \frac{\pi}{4}, \quad \frac{3\pi}{4}, \quad \frac{5\pi}{4}, \ldots, \tag{19}$$

the minima being

$$I_{1\,\min} = E^2 \left(1 - \sin^2 \frac{\delta}{2}\right) = E^2 \cos^2 \frac{\delta}{2} \tag{20}$$

The minima are not completely dark unless $\delta$ is an odd multiple of $\pi$, i.e. unless the plate thickness has one of a number of permissible values, depending on the wavelength of the light.

(b) Analyzer and polarizer perpendicular ( $\chi = \pi/2$ ). In this case (16) gives

$$I_\perp = E^2 \sin^2 2\phi \sin^2 \frac{\delta}{2} \tag{21}$$

Comparison with (17) shows that the interference phenomena are now complementary. There are minima of complete darkness when $\phi$ has one of the values given by (18) and relative maxima for the intermediate positions (19) whose values are

$$I_{\perp \max} = E^2 \sin^2 \frac{\delta}{2} \tag{22}$$

These phenomena may be used to produce colours that are exactly complementary; one only has to pass a parallel beam of white light through two Nicou prisms separated by a crystal plate, with the prisms being first parallel ( $\chi = 0$ ) and then crossed ($\chi = \pi/2$). In order to obtain a uniform field it is essential to have the beam sufficiently well collimated to ensure that no significant phase differences are introduced.

If the light incident on the first polarizer originates in an extended incoherent source placed in the focal plane of a lens, each point in the source will give rise to an intensity distribution in the conjugate point, independently of all the other source points. In the image plane one then observes a light distribution which can be described by curves of equal intensity, forming a so-called interference figure of the crystal plate. To every point in the image plane there corresponds a direction of parallel rays entering and leaving the crystal and we must, therefore, consider the variation of the phase difference $\delta$ with this direction. It will be sufficient to consider only the case when the polarizer and analyzer are crossed (i.e. $\chi = \pi/2$ ); for when they are parallel the pattern is complementary, whilst other cases give less marked interference figures.

The intensity corresponding to a given direction of incidence depends on $\phi$ and $\delta$ and it is useful to consider the effects of varying each of these quantities separately. The curves along which $\phi$ is constant are called isogyres (or brushes), those along which $\delta$ is constant are called

isochromates. The isogyres depend on the orientation of the optic axes in the plate and are independent of the thickness of the plate and the wavelength, unless there is dispersion of the axes. The isochromates depend on the direction of the wave normals and on the thickness of the plate, and are so named because, if white light is used, they are lines of equal colour. The curves of these families along which the intensity is zero are called principal isogyres and principal isochromates and are according to (21) given by $\sin 2\phi = 0$ and $\sin \frac{1}{2}\delta = 0$ respectively. On these curves the state of polarization of the light is the same as before the passage through the crystal. This is so because on the principal isogyres the direction of vibrations passed by the analyzer coincides with one of the directions of vibrations in the crystal, and on the principal isochromates the phase difference between the two emerging beams is an integral multiple of $2\pi$. The two systems of curves are superimposed, but may be studied separately.

Before investigating the form of these curves we must consider how the phase difference $\delta$ depends on the angle of incidence. Let $SA, AB', AB''$ represent the wave normals to the incident and the two refracted waves at $A$ and let $\theta_1, \theta_2', \theta_2''$ be the angles of incidence and the two angles of refraction respectively (see Fig. 14.21).

Further let $\lambda$ be the wavelength in the first medium (air) and $\lambda' = \lambda/n', \lambda'' = \lambda/n''$ the wavelengths of the two refracted waves. The rays will emerge from the plate parallel to each other and to the incident wave normal (see Fig. 14.21) and with a phase difference

$$\delta = 2\pi \left[ \frac{AB''}{\lambda''} + \frac{B''C}{\lambda} - \frac{AB'}{\lambda'} \right], \tag{23}$$

where

$$AB' = \frac{h}{\cos \theta_2''}, \quad AB'' = \frac{h}{\cos \theta_2''}, \tag{24}$$

and

$$B''C = B''B' \sin \theta_1 = h \sin \theta_1 \left( \tan \theta_2' - \tan \theta_2'' \right). \tag{25}$$

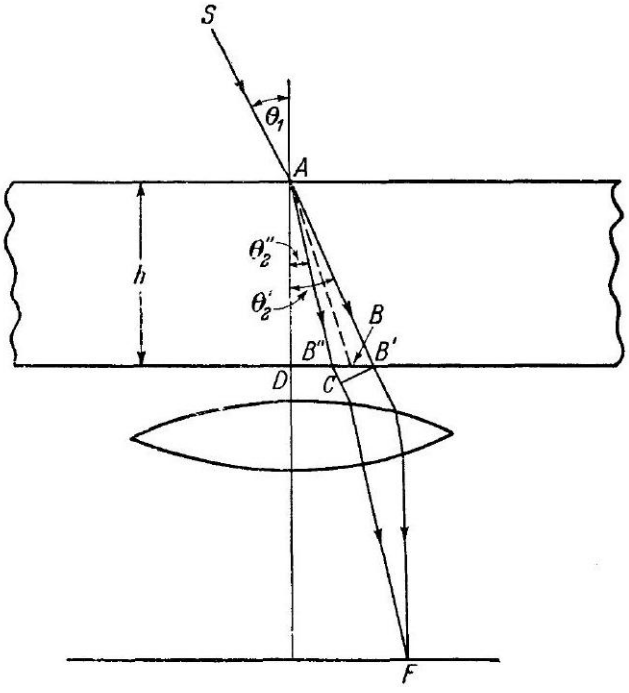



Fig. 14.21. Determination of the phase difference between two waves transmitted by a crystal plate.

Substituting from (24) and (25) into (23), we obtain

$$\delta = 2\pi h \left[ \frac{1}{\cos\theta_2''} \left( \frac{1}{\lambda''} - \frac{\sin\theta_1 \sin\theta_2''}{\lambda} \right) - \frac{1}{\cos\theta_2'} \left( \frac{1}{\lambda'} - \frac{\sin\theta_1 \sin\theta_2'}{\lambda} \right) \right]. \tag{26}$$

Making use of the law of refraction we may replace $\sin\theta_1/\lambda$ by $\sin\theta_2''/\lambda''$ in the first bracket and by $\sin\theta_2'/\lambda'$ in the second bracket, giving

$$\delta = 2\pi h \left[ \frac{\cos\theta_2''}{\lambda''} - \frac{\cos\theta_2'}{\lambda'} \right] = \frac{2\pi h}{\lambda} \left( n'' \cos\theta_2'' - n' \cos\theta_2' \right). \tag{27}$$

As the difference $n'' - n'$ is always small compared with $n'$ and $n''$, it is permissible to use in place of (27) an approximate expression. We have, to first order,

$$\begin{aligned} n'' \cos\theta_2'' - n' \cos\theta_2' &= (n'' - n') \frac{d}{dn} (n \cos\theta_2) \\ &= (n'' - n') \left[ \cos\theta_2 - n \sin\theta_2 \frac{d\theta_2}{dn} \right], \end{aligned} \tag{28}$$

where $n$ is an average value of $n'$ and $n''$ and $\theta_2$ the corresponding average of $\theta_2'$ and $\theta_2''$. We also have, on differentiation of the law of refraction $\sin\theta_1 = n \sin\theta_2$, keeping $\theta_1$ fixed,

$$0 = \sin\theta_2 + n \cos\theta_2 \frac{d\theta_2}{dn} \tag{29}$$

Hence (28) may be written as

$$n'' \cos\theta_2'' - n' \cos\theta_2' = \frac{1}{\cos\theta_2} (n'' - n'), \tag{30}$$

and (27) becomes, on substituting from (30),

$$\delta = \frac{2\pi h}{\lambda \cos\theta_2} (n'' - n') \tag{31}$$

The quantity $h/\cos\theta_2$ represents the mean geometrical path of the two rays in the plate, and this multiplied by $n'' - n'$ gives the corresponding optical path difference.

Returning to the case of an extended source, one has to consider the transmission of waves with different directions of propagation. It will be assumed that the directions make small angles with the plate normal. Let us represent each of the incident waves by its wave normal through the fixed point $A$ in Fig. 14.21. The points $F$ in which the waves are brought to a focus by the lens are in a one-to-one correspondence with the points $B$ where the transmitted wave normals $AB$ (the average of $AB'$ and $AB''$ ) strike the lower face of the plate. Since the inclination of $AB$ to the plate normal $AD$ is small, the points $F$ will form a slightly distorted image-a projectionof the points $B$. Hence the form of the isochromates is essentially given by the loci of points $B$ for which $\delta$ is constant. In particular, for the principal isochromates this constant is an integral multiple of $2\pi$. If we wish to survey the effect of varying the plate thickness, we only have to shift the plane containing the points $B$ parallel to itself.
It follows that all the isochromates may be surveyed by constructing around some point $A$ the surfaces of constant phase difference $\delta(h, \theta_2) =$ constant, called also isochromatic surfaces, and finding their intersection with the planes $h =$ constant. In determining these surfaces it must be remembered that the refractive indices $n'$ and $n''$ entering the expression (31) for $\delta$ are also functions of $\theta_2$.

We shall consider separately the form of the interference figures obtained from uniaxial and biaxial crystal plates. It will be convenient to specify the points $B$ by the polar distance

$$\rho = AB = \frac{h}{\cos\theta_2} \tag{32}$$

and by the angle $\vartheta$ (or angles $\vartheta_1$ and $\vartheta_2$ ) which $AB$ makes with the directions of the optic axis (or axes) of wave normals of the crystalline medium (Fig. 14.22).

### 14.4.4 Interference figures from uniaxial crystal plates

In a uniaxial crystal the phase velocities corresponding to a wave-normal direction that makes an angle $\vartheta$ with the optic axis are, according to § 14.3 (4), related by

$$v_p'^2 - v_p''^2 = \left(v_o^2 - v_e^2\right)\sin^2\vartheta. \tag{33}$$

Since $v_p = c/n$, and similarly for the other velocities, (33) gives

$$\frac{1}{n'^2} - \frac{1}{n'^2} = \left(\frac{1}{n_o^2} - \frac{1}{n_e^2}\right)\sin^2\vartheta. \tag{34}$$

The difference between these refractive indices is usually small compared with their values, so that (34) may be written in the approximate form

$$n'' - n' = (n_e - n_o)\sin^2\vartheta. \tag{35}$$

Substituting from this relation into (31) and using (32) we obtain

$$\delta = \frac{2\pi\rho}{\lambda}(n_o - n_o)\sin^2\vartheta. \tag{36}$$

Hence the surfaces of constant phase difference are given by

$$\rho\sin^2\vartheta = C, \quad (C = \text{ constant }). \tag{37}$$

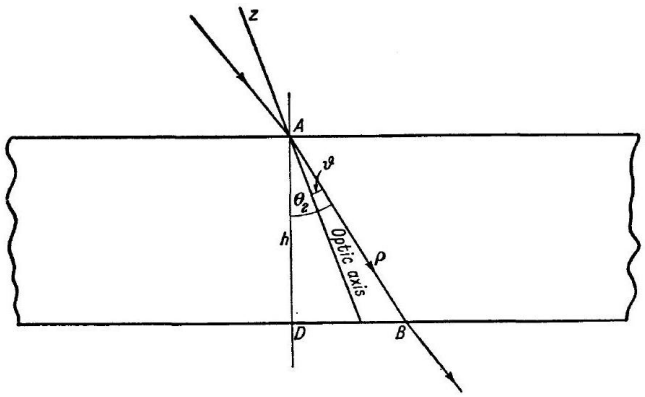



Fig. 14.22. Illustrating the theory of interference figures from crystal plates.

To visualize the form of these surfaces we take Cartesian axes with the $z$ axis along the optic axis. Then

$$\begin{aligned} \rho^2 &= x^2 + y^2 + z^2 \\ \rho^2\sin^2\vartheta &= x^2 + y^2 \end{aligned} \tag{38}$$

and, according to (37), the surfaces of constant phase difference are given by

$$\left(x^2+y^2\right)^2 = C^2\left(x^2+y^2+z^2\right). \tag{39}$$

These surfaces can be generated by rotating the curves

$$x^4 = C^2\left(x^2+z^2\right), \tag{40}$$

of which a typical one is shown in Fig. 14.23, around the $z$ axis. At large distance from the origin ( $z^2 \gg x^2$ ) the curve approaches asymptotically the parabolae

$$x^2 = \pm Cz. \tag{41}$$

In the neighbourhood of the $x$ axis

$$C^2 = \frac{x^4}{x^2+z^2} = \frac{x^2}{1+\frac{z^2}{x^2}} = x^2\left(1-\frac{z^2}{x^2}+\ldots\right) \tag{42}$$

so that the curve there is approximated by the hyperbola

$$x^2 - z^2 = C^2. \tag{43}$$

The parabolae and the hyperbola are shown as dotted curves in Fig. 14.23.
We can now determine all the isochromates by simply taking the sections of the surface (39) with planes at different distances $h$ from the origin, these planes representing the exit face of the crystal plate. It is evident from Fig. 14.23 that isochromates of very different forms can arise, depending on the orientation of this face of the crystal

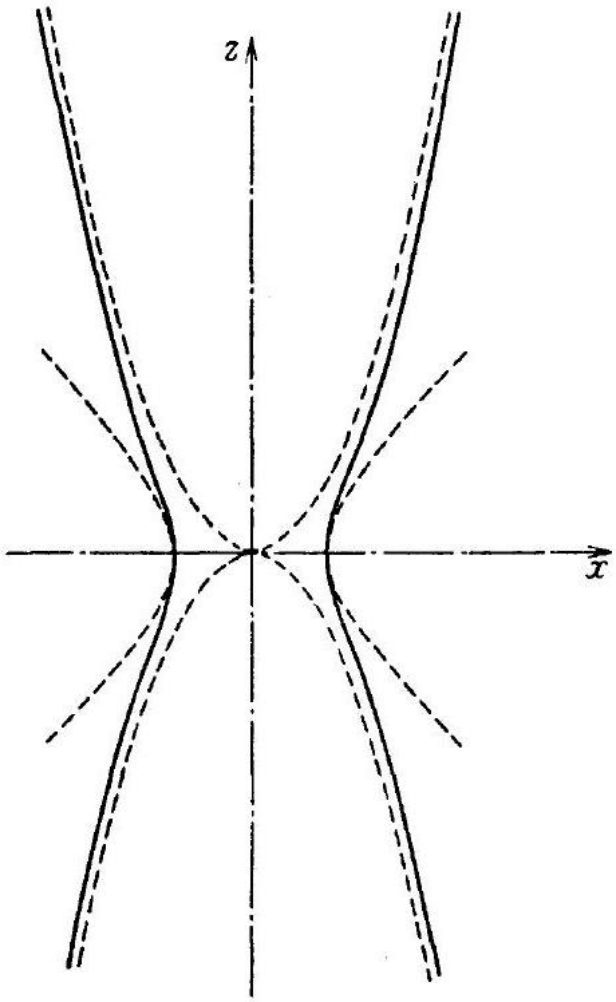



Fig. 14.23. Meridional curve of a surface of constant phase difference for an optically uniaxial crystal.

relative to the optic axis. If the face of the plate is perpendicular to the optic axis, the isochromates are evidently circular; if the normal to the face makes a small angle with the optic axis, the curves are closed, approximating to ellipses; but if the normal makes a large angle with the optic axis the curves approximate to hyperbolae.

The principal isogyres (curves on which $\sin 2\phi = 0$ ), appear as a dark blurred cross whose arms are parallel to the directions of the polarizer and the analyzer, and whose centre corresponds to a wavenormal direction parallel to the optic axis. This follows from the fact that for any given direction of the wave normal in the crystal the directions of vibration are parallel to, and perpendicular to, the principal plane containing the wave normal and the optic axis. Thus there will be darkness at all points of the field of view for which the principal plane is parallel to the direction of vibration passed by either the polarizing or the analyzing Nicol prism.

Fig. 14.24 shows a typical interference figure from a uniaxial crystal; the principal isogyres and isochromates can be clearly seen.

Interference figures from plane parallel crystal plates have an important practical application in connection with polariscopes; these are instruments used for the detection of small proportions of polarized light in a beam which is largely unpolarized. An example is Savart's plate.* It consists of two quartz plates, each cut with its optic axis at 45° to the plate normal, and cemented together so that the planes containing the plate normals and the optic axes are perpendicular to each other. Thus the direction of ordinary vibrations in one plate is the direction of extraordinary vibration in the other, and the phase differences have opposite signs; they cancel each

other exactly when the wave passes normally through the plates. This arrangement ensures that the phase difference varies very little with wavelength for nearly normal incidence, so that white light fringes may be obtained. If Savart's plate is placed between crossed Nicol prisms with its principal planes at 45° to the directions of the Nicol prisms, the interference pattern consists of light and black, almost straight, fringes. If the polarizing Nicol prism is removed, and partially polarized light passed through the system, the polarized component will give rise to fringes that are superimposed on a uniform background produced by the unpolarized

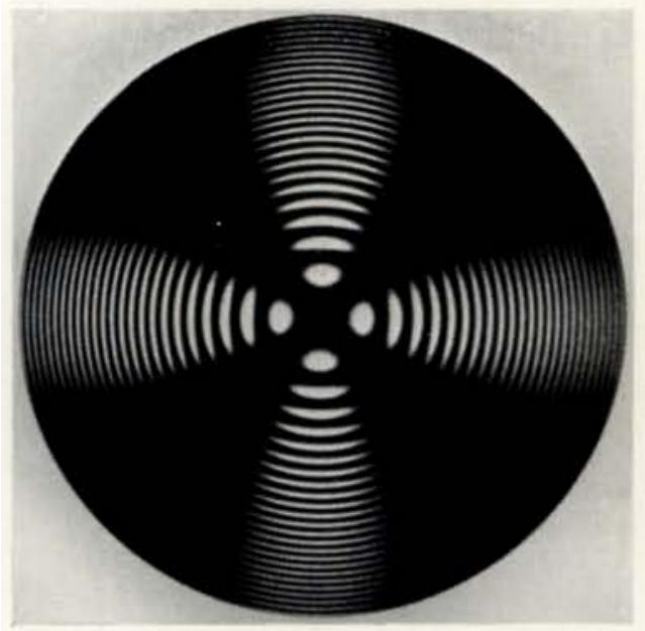


Fig. 14.24. Interference figure from fluorspar; cut perpendicular to the optic axis, between crossed Nicol prisms.

0

- Cf. Ann. d. Physik, 49 (1840), 292.

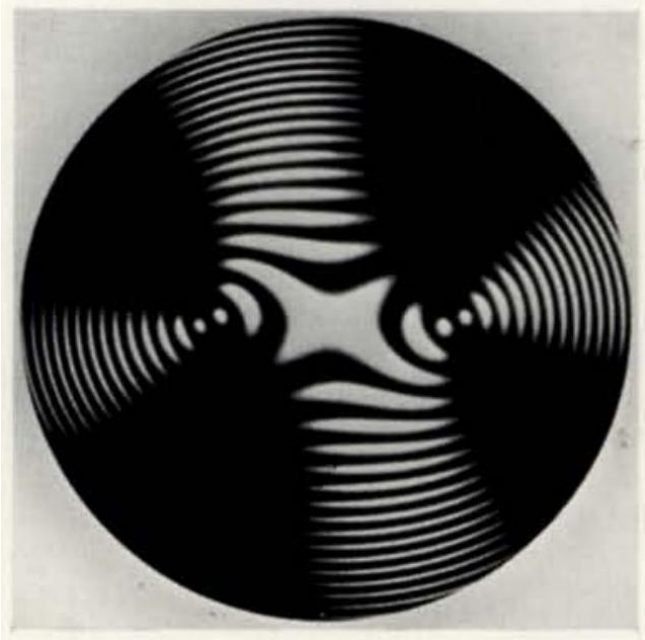


Fig. 14.26. Interference figure from Brazil topaz.

component, but the fringes can be detected even when the proportion of the polarized light is very low. The fringes have the greatest contrast when the plane of vibration of the incident light is at 45° to the principal planes of the Savart plate.

### 14.4.5 Interference figures from biaxial crystal plates

For a plane parallel biaxial crystal plate we have, in place of (33), the more general relation

$$v_p'^2 - v_p''^2 = \left(v_x^2 - v_z^2\right) \sin\vartheta_1 \sin\vartheta_2, \tag{44}$$

which follows from § 14.3 (18). Here $\vartheta_1$ and $\vartheta_2$ are the angles which the wave normal direction $AB$ makes with the two optic axes of wave normals. Since $v_p = c/n, v_x = c/n_x$, etc., (44) gives

$$\left(\frac{1}{n'^2} - \frac{1}{n'^2}\right) = \left(\frac{1}{n_x^2} - \frac{1}{n_z^2}\right) \sin\vartheta_1 \sin\vartheta_2 \tag{45}$$

or approximately, since the differences between the refractive indices are small compared with their values,

$$n'' - n' = (n_z - n_x) \sin\vartheta_1 \sin\vartheta_2. \tag{46}$$

Substituting from (46) into (31) and again setting $h/\cos\vartheta_2 = \rho$, we obtain for the phase difference $\delta$ the expression

$$\delta = \frac{2\pi\rho}{\lambda} (n_z - n_x) \sin\vartheta_1 \sin\vartheta_2. \tag{47}$$

We see that the surfaces of constant phase difference are now given by

$$\rho \sin\vartheta_1 \sin\vartheta_2 = C. \quad (C = \text{ constant }). \tag{48}$$

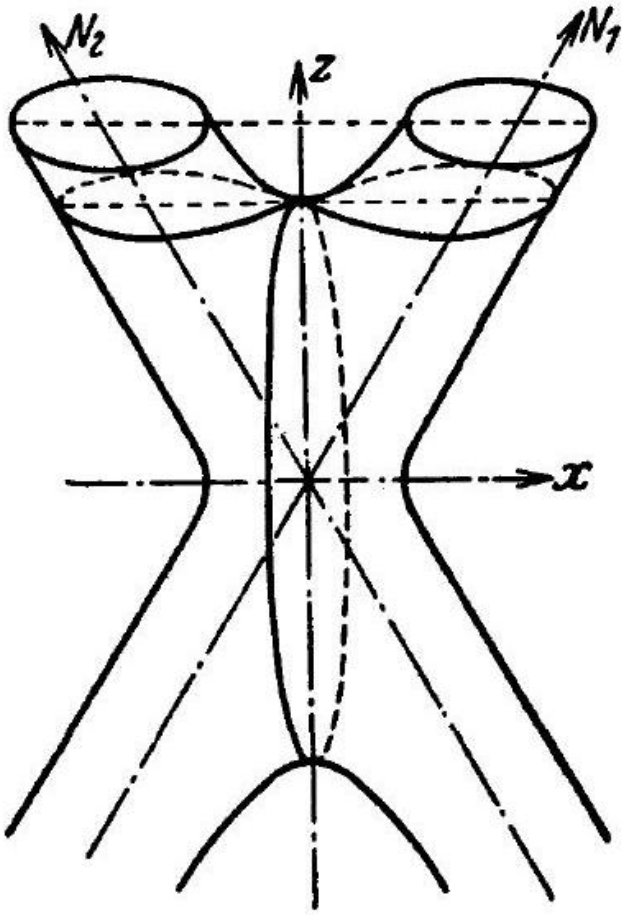



Fig. 14.25. Surface of constant phase difference for an optically biaxial crystal.

In the direction of each optic axis ($\vartheta_1 = 0$ or $\vartheta_2 = 0$) $\rho$ tends to infinity, so that the surfaces approximate asymptotically to cylinders surrounding the optic axes. When $\vartheta_1$ is small, $\vartheta_2$ is approximately equal to the angle $2\beta$ between the two optic axes, and (48) then becomes

$$\rho \sin \vartheta_1 = \frac{C}{\sin 2\beta} \tag{49}$$

But $\rho \sin \vartheta_1$ is the distance of a point on the surface from the optic axis $\vartheta_1 = 0$; similarly for $\rho \sin \vartheta_2$. Hence the "asymptotic cylinders" are circular. The general form of a surface of constant phase difference is shown in Fig. 14.25. It is evident that near the optic axes the isochromates are closed curves, approximating to ellipses, surrounding the two points in the focal plane that correspond to the optic axes.

The principal isogyres are obtained, as before, by finding all directions of the wave normal such that the directions of vibrations in the crystal coincide with the directions passed by the Nicol prisms. We may use the construction of § 14.2.3 (a), where it was shown that the planes of vibration bisect the angles between the planes ( $N_1, s$ ) and $(N_2, s)$, $N_1$ and $N_2$ being the directions of the optic axes. Thus if the crystal has its $z$ axis vertical and the directions of vibrations which are transmitted by the Nicol prisms are parallel to the $x$ and $y$ axes, the principal isogyres will lie in the $xz$-plane and the $yz$-plane, and the interference figure will thus show a dark cross with one pair of arms passing through the points corresponding to the optic axes.

More generally, if the crystal is in any other orientation relative to the directions of the Nicol prisms, the principal isogyres will have the form of rectangular hyperbolae with asymptotes in the directions of vibrations transmitted by the Nicol prisms, and with arms passing through the

points corresponding to the optic axes. If, with both the Nicol prisms fixed, the crystal plate is rotated in its plane, the pattern of isogyres will change, but the isochromates will, apart from rotation, remain the same; for the isochromates are defined by conditions which do not depend on the directions of the Nicol prisms. A typical interference figure from biaxial crystal plate is shown in Fig. 14.26.

### 14.4.6 Location of optic axes and determination of the principal refractive indices of a crystalline medium

Since the isochromates form closed curves around the optic axis (or axes), the observation of interference figures affords a ready method of determining whether a crystal is uniaxial or biaxial, and of locating the axes. The interference figures may be observed in a microscope fitted with two Nicol prisms (called polarizing microscope), either by removing the eyepiece and focusing the eye on the back focal plane of the objective lens (thus reproducing the conditions of Fig. 14.21) or by inserting an additional lens in the body of the microscope so that the back focal plane of the objective can be observed through the eyepiece. The second method has the advantage that a larger image of the interference figure is seen, and measurements may be made on the interference figure by the use of a calibrated scale in the eyepiece. Thus the angle between the optic axes of a biaxial crystal may be measured (naturally one must take into account the fact that the light is refracted as it leaves the crystal). Even very small crystal fragments, such as may occur in thin mineralogical sections, are sufficient for the location of the optic axes and measurements of their inclination by these means.

The principal refractive indices, $n_x, n_y, n_z$ of a crystal may be determined either from measurements of the angle of deviation or of total reflection in a prism, or by immersion in a series of liquids of graded refractive index.

The prism method is more convenient for uniaxial crystals than for biaxial crystals. A prism is cut with its refracting edge parallel to the optic axis of wave normal. Then the ordinary and extraordinary waves have their $\boldsymbol{D}$ vectors respectively perpendicular and parallel to this edge. The two refractive indices can be found from the deviations of the two rays which emerge from the prism when an unpolarized beam is incident on one of its faces. By means of a Nicol prism one can distinguish between the ordinary and extraordinary rays.

The immersion method depends on the fact that a transparent body is not visible when immersed in a liquid of equal refractive index. Since a crystal has two refractive indices for any given direction of propagation, it will be visible in any liquid if it is observed with unpolarized light. However, if polarized light is used with its $D$ vector in one of the directions of vibrations in the crystal, the crystal will be invisible in a liquid of the appropriate refractive index, $n'$ or $n''$. If the directions of the principal dielectric axes are known, the crystal can be oriented so that light travels parallel to each axis in turn, and $n'$ and $n''$ can be made equal to pairs of $n_x, n_y$, and $n_z$ in turn.

If the directions of the principal dielectric axes are unknown, one can obtain a fair estimate of the refractive indices by immersing a large number of crystals, of the type under consideration, with random orientations, in a series of liquids of graded refractive indices. Each crystal will become invisible at two different values of
the refractive index, for two directions of vibrations of the incident light. If these two refractive indices are $n'$ and $n''$, we have, by the inequalities § 14.3 (12)

$$n_x \leqslant n' \leqslant n_y \leqslant n'' \leqslant n_z \tag{50}$$

Thus $n_x$ equals the lower limit of the values of $n', n_z$ equals the upper limit of the values of $n''$, and $n_y$ is equal both to the upper limit of $n'$ and to the lower limit of $n''$; these limits should coincide if a sufficient number of measurements have been taken.

If the crystals are uniaxial, then every crystal will give $n_o$ for one of its refractive indices, and the other will range between $n_o$ and $n_e$.

## 9.18.5 14.5. Stress Birefringence and Form Birefringence

### 14.5.1 Stress birefringence

When a transparent isotropic material is subject to mechanical stresses it may become optically anisotropic. This phenomenon, known as stress birefringence or the photo-elastic effect was first noted by Brewster* and finds useful practical applications. We shall only briefly indicate how optical methods may be used to obtain information about the state of stress in an initially isotropic material. As a preliminary we must consider the relations that exist between the elastic and the optical constants of matter.

The state of stress and the state of strain in an elastic solid body are characterized by second-order tensors, the stress tensor $P_{kl}$ and the strain tensor $r_{kl}$, the components of which are linearly related to each other. These two tensors are always symmetric, but their principal axes are generally different from those of the dielectric tensor, which, as we saw in § 14.1, determines the optical properties of the body.

When a stress is applied to the body, the dielectric tensor is modified, and it may be assumed in the first approximation that the changes in the components of the dielectric tensor are linearly related to the six stress components, and hence also to the six strain components. Thus we are led to the introduction of two new sets of coefficients, the stress-optical constants and the strain-optical constants which characterize these relationships.

If we refer coordinates to the principal dielectric axes of the unstressed material, the ellipsoid of wave normals has the equation

$$\frac{x^2}{\varepsilon_x} + \frac{y^2}{\varepsilon_y} + \frac{z^2}{\varepsilon_z} = 1 \tag{1}$$

On applying a stress whose components are $P_{xx}, P_{xy}, \ldots$, this ellipsoid is changed into another one whose equation may be written as

$$a_{xx}x^2 + a_{yy}y^2 + a_{zz}z^2 + a_{yz}yz + a_{zx}zx + a_{xy}xy = 1. \tag{2}$$

By our assumptions, each coefficient $a_{kl}$ differs from the corresponding coefficient in (1) by a linear function in the $P'$ s. We thus have six relations, of which two typical ones are

$$\left.\begin{aligned} a_{xx} - \tfrac{1}{\varepsilon_x} &= q_{11}P_{xx} + q_{12}P_{yy} + q_{13}P_{zz} + q_{14}P_{yz} + q_{15}P_{zx} + q_{16}P_{xy}, \\ a_{yz} &= q_{41}P_{xx} + q_{42}P_{yy} + q_{43}P_{zz} + q_{44}P_{yz} + q_{45}P_{zx} + q_{46}P_{xy}. \end{aligned}\right\} \tag{3}$$

In this notation each numeral 1-6 in the subscripts refers to a pair of axes, thus: $1 = xx, 2 = yy, 3 = zz, 4 = yz, 5 = zx, 6 = xy$.

There is a similar set of equations relating the coefficients $a_{kl}$ to the strain components. The optical effect of the stress may also be described in terms of the deformation of the ray ellipsoid, and thus two further sets of linear equations with 36 coeffcients are obtained. These coefficients are related to those of the ellipsoid of wave normals, since the two ellipsoids always have their principal axes in the same directions, and the semi-axes of one are reciprocals of the other.

0

- D. Brewster, Phil. Trans. (1815), 60; (1816), 156. Trans. Roy. Soc, Ed., 8 (1818), 369.

The relations (3) take a simpler form if there are symmetry elements present in the structure. For crystals of the cubic system the three principal axes $x, y$, and $z$ are equivalent, and in consequence the following relations hold between the stress-optical coefficients*

$$\left.\begin{array}{c} q_{11} = q_{22} = q_{33} \\ q_{12} = q_{21} = q_{23} = q_{32} = q_{31} = q_{13} \\ q_{44} = q_{55} = q_{66} \end{array}\right\} \tag{4}$$

all the remaining coefficients being zero.

For isotropic materials, the relations (3) must remain unaltered for any change of axes. This is only possible if the stress-optical constants satisfy the conditions for cubic symmetry, and in addition the relation

$$2q_{44} = q_{11} - q_{12} \tag{5}$$

holds. Thus in this case there are only two independent constants. Since all systems of axes are now equivalent, we may use any set of axes and, in particular, the principal axes of the stress tensor; then $P_{yz} = P_{zy} = P_{xy} = 0$ and we have in place of (3) the simpler relations

$$\left.\begin{array}{c} a_{xx} - \frac{1}{\varepsilon} = q_{11}P_{xx} + q_{12}P_{yy} + q_{12}P_{zz}, \\ a_{yy} - \frac{1}{\varepsilon} = q_{12}P_{xx} + q_{11}P_{yy} + q_{12}P_{zz}, \\ a_{zz} - \frac{1}{\varepsilon} = q_{12}P_{xx} + q_{12}P_{yy} + q_{11}P_{zz}, \\ a_{yz} = a_{zx} = a_{xy} = 0. \end{array}\right\} \tag{6}$$

Thus the principal axes of the stress tensor and those of the ellipsoid of wave normals are the same in this case, as one would expect from symmetry considerations.

Although cubic crystals such as rocksalt are optically isotropic when unstrained, they nevertheless behave differently, when strained, from truly isotropic materials like glass. The effect of stress may be conveniently observed by viewing the body between crossed Nicol prisms (or other polarizing devices such as polaroid sheetscf. § 14.6.3). Consider a sheet of material of thickness $h$ and let light be incident normally on it. Suppose that the sheet is stressed so that two of the principal axes, say $x$ and $y$, of the stress tensor and, therefore, of the dielectric tensor, lie in the plane

of the sheet and make angles $\phi$ and $\phi + \pi/2$ with the directions of the polarizer and the analyzer, as in § 14.4.3. The section of the ellipsoid of normals by the $xy$-plane is the ellipse

$$a_{xx}x^2 + a_{yy}y^2 = 1 \tag{7}$$

where $a_{xx}$ and $a_{yy}$ are given by (6). The refractive indices $n'$ and $n''$ for the two waves propagated in the sheet are given by

$$n' = \frac{1}{\sqrt{a_{xx}}}, \quad n'' = \frac{1}{\sqrt{a_{yy}}} \tag{8}$$

Hence

0

- The relations which exist between the stress-optical coefficients of each of the crystal system are discussed by F. Pockels, Ann. d. Physik, 37 (1889), 158. Also his Lehrbuch der Kristalloptik, (Leipzig, 1906), pp. 469-74. They are summarized in G. Szivessy, Handbuch der Physik, Vol. 21 (Berlin, Springer, 1929), 840.

$$\frac{1}{n'^2} - \frac{1}{n''^2} = a_{xx} - a_{yy}$$
$$= (q_{11} - q_{12})(P_{xx} - P_{yy}) \tag{9}$$

Now $n'$ and $n''$ will in practice usually differ only slightly from $n_x$, so that, approximately

$$n'' - n' = \frac{1}{2} n_x^3 (q_{11} - q_{12})(P_{xx} - P_{yy}). \tag{10}$$

Substitution into § 14.4 (31) gives the phase difference $\delta$ between the two waves emerging from the sheet:

$$\delta = \frac{\pi h}{\lambda} n_x^3 (q_{11} - q_{12})(P_{xx} - P_{yy}). \tag{11}$$

Thus the phase difference is proportional to $P_{xx} - P_{yy}$, which represents twice the shearing stress over planes inclined at 45° to the $x$ and $y$ directions. There is in this case only one relevant stress-optical constant, namely $q_{11} - q_{12}$.

It follows that if a stressed sheet of glass or transparent plastic is observed between crossed Nicol prisms, bright and dark fringes will be seen, and these fringes are contours of equal shearing stress. Such a "stress pattern" is shown in Fig. 14.27. The fringes are seen at maximum intensity in any given region of the pattern only if the principal axes of the stress system in this region are at 45° to the directions of vibrations transmitted by the Nicol prisms. When the principal axes of the stress are parallel to the directions transmitted by the Nicol prisms, the fringes disappear and the field of view becomes black; thus the directions of the axes of the stress system can be determined by rotating the crossed Nicol prisms, while the magnitude of the shearing stress can be obtained from the order of the fringes. This method is used to investigate stresses in engineering structures; a model of the structure is made from a suitable plastic material and the effect of the stresses is directly observed in the manner just described. In this way lengthy calculations are often avoided.*

### 14.5.2 Form birefringence

The birefringent properties of crystals may be explained in terms of the anisotropic electrical properties of molecules of which the crystals are composed. Birefringence may, however, arise from anisotropy on a scale much larger than molecular, namely when there is an ordered arrangement of similar particles of optically isotropic material whose size is large compared with the dimensions of molecules, but small compared with the wavelength of light. We then speak of form birefringence.

From optical measurements information may often be obtained about the submicroscopic particles that give rise to form birefringence. We shall explain the principle of the method by considering the somewhat idealized case of a regular assembly of particles that have the form of thin parallel plates. Let $t_1$ be the thickness of each plate and $t_2$ the widths of the spaces between them (Fig. 14.28). Further let $\varepsilon_1$ be the dielectric constant of each plate and $\varepsilon_2$ the dielectric constant of the medium in which they are immersed.

Suppose that a plane monochromatic wave is incident on the assembly and assume first that its electric vector is perpendicular to the plates. If the linear dimensions of the faces of the

0

- For a full account of this method see for instance E. G. Coker and L. N. G. Filon, A Treatise on Photoelasticity (Cambridge University Press, 1931), or M. M. Frocht, Photoelasticity, Vol. I (1941), Vol. II (1948), (New York, John Wiley).

plates are assumed to be large but the thicknesses $t_1$ and $t_2$ small compared to the wavelength, the field in the plates and in the spaces may be considered

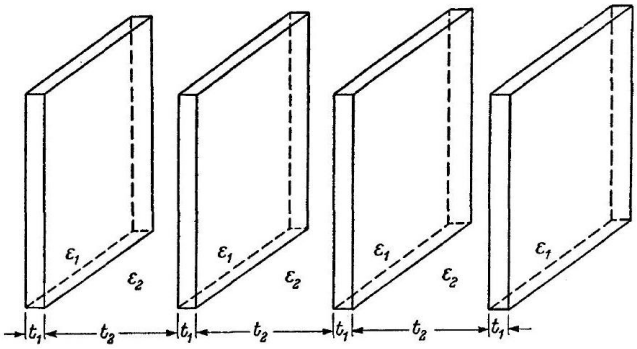



Fig. 14.28. A regular assembly of thin parallel plates.

to be uniform. Further, according to § 1.1.3, the normal component of the electric displacement must be continuous across a surface at which the properties of the medium change abruptly. Hence the electric displacement must have the same value $D$ inside the plates and in the spaces. If $E_1$ and $E_2$ are the corresponding electric fields,

$$E_1 = \frac{D}{\varepsilon_1}, \quad E_2 = \frac{D}{\varepsilon_2}, \tag{12}$$

and the mean field $\boldsymbol{E}$ averaged over the total volume is

$$E = \frac{t_1 \frac{D}{\varepsilon_1} + t_2 \frac{D}{\varepsilon_2}}{t_1 + t_2}. \tag{13}$$

The effective dielectric constant $\varepsilon_\perp$ is, therefore,

$$\varepsilon_\perp = \frac{D}{E} = \frac{(t_1 + t_2)\,\varepsilon_1 \varepsilon_2}{t_1 \varepsilon_2 + t_2 \varepsilon_1} = \frac{\varepsilon_1 \varepsilon_2}{f_1 \varepsilon_2 + f_2 \varepsilon_1}, \tag{14}$$

where $f_1 = t_1/(t_1 + t_2)\,, f_2 = t_2/(t_1 + t_2) = 1 - f_1$ are the fractions of the total volume occupied by the plates and by the surrounding medium respectively.

Suppose next that the incident field has its electric vector parallel to the plates. According to § 1.1.3, the tangential component of the electric vector is continuous across a discontinuity surface, so that in this case the electric field will have the same value $\boldsymbol{E}$ inside the plates and in the spaces. The electric displacements in the two regions are

$$D_1 = \varepsilon_1 E, \quad D_2 = \varepsilon_2 E, \tag{15}$$

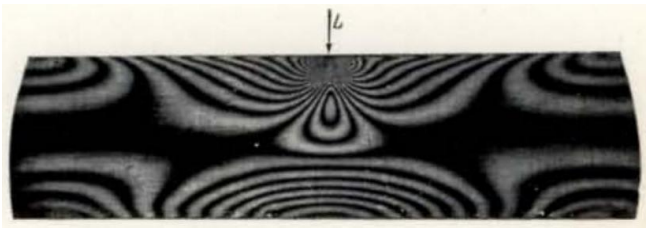



Fig. 14.27. The stress pattern of a beam under the action of a concentrated load. The position of the load is indicated by $L$; the ends (not shown) are clamped.

so that the mean electric displacement $D$ is

$$D = \frac{t_1\varepsilon_1 E + t_2\varepsilon_2 E}{t_1 + t_2} \tag{16}$$

Hence the effective dielectric constant is now given by

$$\varepsilon_1 = \frac{D}{E} = \frac{t_1\varepsilon_1 + t_2\varepsilon_2}{t_1 + t_2} = f_1\varepsilon_1 + f_2\varepsilon_2 \tag{17}$$

Since the effective dielectric constant is the same for all directions parallel to the plates, but different for directions normal to the plates, the assembly behaves as a uniaxial crystal with its optic axis perpendicular to the plane of the plates. The difference $\varepsilon_1 - \varepsilon_\perp$ is always positive, since according to (14) and (17),

$$\varepsilon_1 - \varepsilon_\perp = \frac{f_1 f_2 \left(\varepsilon_1 - \varepsilon_2\right)^2}{f_1\varepsilon_2 + f_2\varepsilon_1} \geqslant 0 \tag{18}$$

The ordinary wave has its electric vector perpendicular to the optic axis, i.e. parallel to the plane of the plates. Equation (18) implies that the assembly always behaves like a negative uniaxial crystal (cf. § 14.3.2). In terms of refractive indices the last equation may be written as

$$n_o^2 - n_o^2 = -\frac{f_1 f_2 \left(n_1^2 - n_2^2\right)^2}{f_1 n_2^2 + f_2 n_1^2} \tag{19}$$

For assemblies of particles of less idealized forms the calculations are naturally more complicated.*

A case of considerable practical interest is that of an assembly of parallel and similar thin cylindrical rods. It was shown by Wiener † that if the rods occupy a small fraction of the total volume ( $f_1 \ll 1$ ), one has, in place of (19),

$$n_e^2 - n_o^2 = \frac{f_1 f_2 \left(n_1^2 - n_2^2\right)^2}{\left(1 + f_1\right) n_2^2 + f_2 n_1^2} \tag{20}$$

Such an assembly therefore behaves as a positive uniaxial crystal, with its optic axis parallel to the axes of the rods.

Observations on form birefringence are useful in biological microscopy. The sign of the observed difference indicates whether the shape of the particles is nearer to that of a rod or a plate, and if $n_1$ and $n_2$ are known, it may be possible to estimate from equation (19) or (20) the fraction of the volume occupied by the particles. To distinguish between form birefringence and intrinsic birefringence in the material of the particles, the refractive index $n_2$ of the medium is varied; the form birefringence will disappear when $n_2 = n_1$, while intrinsic birefringence will be unaffected by variation of $n_2$. If both forms of birefringence are present, a graph of $|{n_e}^2 - {n_0}^2|$ plotted against $n_2$ will show a minimum, but not a zero, at $n_2 = n_1$.

0

- For a general treatment see O. Wiener, Abh. Schs. Ges. Akad. Wiss., Math.-Phys. Kl. No. 6. 32 (1912), 575. Formulae relating to ellipsoidal particles are also given by W. L. Bragg and A. B, Pippard, Acta Cryst., 6 (1953), 865.
  † Loc. cit. p. 581. The principal dielectric constants of a rectangular array of parallel cylinders were also calculated by Lord Rayleigh, Phil. Mag., (5), 34 (1892), 481, and equation (20) is in agreement with his results even for values of $f_1$ that are not very small compared to unity, provided that the difference between the refractive indices $n_1$ and $n_2$ is small.

The ordering of the particles that give rise to form birefringence may be permanent or semi-permanent as for example in tobacco-mosaic virus crystals,* or it may be a temporary ordering of similar particles suspended in a liquid. A suspension of similar particles in a liquid appears optically isotropic if the particles are randomly oriented, as will usually be the case if the liquid is stationary; if, however, the liquid is made to flow, there will be a tendency for the particles to align themselves in a particular direction, and the assembly will then behave as a crystal. This effect may be observed by placing the suspension between two coaxial cylinders, one rotating and the other fixed, and by viewing the liquid between crossed Nicol prisms, with light propagated in a direction parallel to the axes of the cylinders. † This phenomenon, first observed by Maxwell, ‡ is of assistance in investigations of flow of liquids past obstacles, giving information about the direction and magnitude of the velocity gradient.

## 9.18.6 14.6. Absorbing Crystals

### 14.6.1 Light propagation in an absorbing anisotropic medium

The crystalline media which we have so far considered have been characterized, with regard to their optical properties, by the dielectric tensor $\varepsilon_{kl}$. To describe media which are not only anisotropic, but are also absorbing, we must in addition introduce the conductivity tensor $\sigma_{kl}$. The directions of the principal axes of the two tensors will not be the same in general and, in consequence, the theory of propagation of light in such media is rather complicated. However, the principal axes of the two tensors coincide in direction for crystals of the higher symmetry classes (of at least orthorhombic symmetry), and we shall confine our attention to crystals of this type, as they illustrate all the essential features of the general theory. We then only have to replace the real dielectric constants $\varepsilon_x, \varepsilon_y$, and $\varepsilon_z$, by complex ones, $\hat{\varepsilon}_x, \hat{\varepsilon}_y$, and $\hat{\varepsilon}_z$. We shall see that all the earlier formulae of optics of crystals are formally retained, provided that all quantities which depend on $\hat{\varepsilon}_x, \hat{\varepsilon}_y$, and $\hat{\varepsilon}_z$ are assumed to be complex.

We start from Maxwell's equations for a conducting medium, viz.

$$\operatorname{curl} H = \frac{1}{c}\dot{D} + \frac{4\pi}{c}j, \quad \operatorname{curl} E = -\frac{1}{c}\dot{B}, \tag{1}$$

and consider the propagation of a plane, damped wave. In the complex notation the vectors $\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{B}, \boldsymbol{H}$, and $\boldsymbol{j}$ are then each proportional to

$$\exp\left[i\omega\left(\frac{\hat{n}}{c}(r.s) - t\right)\right].$$

In analogy to § 14.2 (3) we obtain the equations

$$\hat{n}s \wedge H = -D + \frac{4\pi}{i\omega}j, \quad \hat{n}s \wedge E = B \tag{2}$$

Setting $\boldsymbol{B} = \mu\boldsymbol{H}$ and eliminating $\boldsymbol{H}$ between these two equations, we find

$$\mu\left[\boldsymbol{D} + \frac{4\pi i}{\omega}\boldsymbol{j}\right] = \hat{n}^2[\boldsymbol{E} - \boldsymbol{s}(\boldsymbol{s}.\boldsymbol{E})] \tag{3}$$

0

- Cf. M. H. F. Wilkins, A. R. Stokes, W. E. Seeds, and G. Oster, Nature, 166 (1950), 127-9.
  † The determination of the orientation of the particles from such measurements is discussed by P. Boeder, Z. f. Phys., 75 (1932), 258, and by J. T. Edsall in Advances in Colloidal Science, Vol. I, edited by E. O. Kraemer (New York, Interscience Publishers Inc., 1942), p. 269.
  ‡ J. C. Maxwell, Proc. Roy. Soc., 22 (1873), 46. Also his Collected Papers, Vol. 2 (Cambridge University Press, 1890), 379.

If we take the coordinate axes in the direction of the principal dielectric tensor (which by our assumption coincide with those of the conductivity tensor), we have

$$\left.\begin{array}{llllll} D_x & = \varepsilon_x E_x, & D_y & = \varepsilon_y E_y, & D_z & = \varepsilon_z E_z \\ j_x & = \sigma_x E_x, & j_y & = \sigma_y E_y, & j_z & = \sigma_z E_z \end{array}\right\} \tag{4}$$

Substituting from (4) into (3) and introducing the complex dielectric constants

$$\hat{\varepsilon}_k = \varepsilon_k + \frac{4\pi i}{\omega}\sigma_k, \quad (k = x, y, z) \tag{5}$$

(3) becomes

$$\mu\hat{\varepsilon}_k \boldsymbol{E}_k = \hat{n}^2 \left[\boldsymbol{E}_k - s_k(\boldsymbol{E}.\boldsymbol{s})\right]. \tag{6}$$

This relation is formally identical with the relation § 14.2 (18), the real constants $\varepsilon_k$ and $n$ being replaced by complex constants $\hat{\varepsilon}_k$ and $\hat{n}_k$. Re-writing (6) in the form

$$E_k = \frac{\hat{n}^2 s_k(\boldsymbol{E}\cdot\boldsymbol{s})}{\hat{n}^2 - \mu\hat{\varepsilon}_k} \tag{7}$$

we obtain by the same argument as in §14.2 (or simply by using the formal substitution $\varepsilon_k \to \hat{\varepsilon}_k, n \to \hat{n}$ ), Fresnel's equation

$$\frac{s_x^2}{\frac{1}{\hat{n}^2} - \frac{1}{\mu\hat{\varepsilon}_x}} + \frac{s_y^2}{\frac{1}{\hat{n}^2} - \frac{1}{\mu\hat{\varepsilon}_y}} + \frac{s_z^2}{\frac{1}{\hat{n}^2} - \frac{1}{\mu\hat{\varepsilon}_z}} = 0. \tag{8}$$

Introducing the complex velocities

$$\hat{v}_p = \frac{c}{\sqrt{\mu\hat{\varepsilon}}} = \frac{c}{\hat{n}}, \quad \hat{v}_x = \frac{c}{\sqrt{\mu\hat{\varepsilon}_x}} = \frac{c}{\hat{n}_x} \tag{9}$$

etc., we may write Fresnel's equation again in the form § 14.2 (24)

$$\frac{s_x^2}{\hat{v}_p^2 - \hat{v}_x^2} + \frac{s_y^2}{\hat{v}_p^2 - \hat{v}_y^2} + \frac{s_z^2}{\hat{v}_p^2 - \hat{v}_z^2} = 0. \tag{10}$$

These relations are strictly analogous to those relating to non-absorbing crystals; their physical interpretation is, however, somewhat different. From (8) or (10) we again obtain a quadratic equation for $\hat{n}^2(s)$, i.e. we find two refractive indices and two principal vibrations $D'$ and $D''$ corresponding to each direction of propagation $s$. From (7) we see that the ratios $D_x : D_y : D_z$ are complex, so that the principal vibrations are in general no longer linear, but elliptical. A further difference is that the electric displacement vectors are no longer perpendicular to the wave-normal $s$. For we have from the first equation (2), on scalar multiplication by $s$,

$$s.\boldsymbol{D} = \frac{4\pi}{i\omega}s.j = \frac{4\pi}{i\omega}\left[\frac{\sigma_x}{\varepsilon_x}s_x D_x + \frac{\sigma_y}{\varepsilon_y}s_y D_y + \frac{\sigma_z}{\varepsilon_z}s_z D_z\right] \tag{11}$$

and the right-hand side of (11) will in general not be zero. However, if the ratios $4\pi\sigma_k/\omega\varepsilon_k$ are small compared with unity, the component of $D$ in the direction of $s$ will be small compared with $D$ itself.

Further analysis is considerably simplified if the absorption is assumed to be small, i.e. if metals are excluded and only substances that are to some extent transparent are considered. We shall restrict our discussion to this case. Formally, weak absorption implies that the second power of the attenuation index $\kappa$ may be neglected in comparison with unity. We may therefore write, on using (9),

$$\left.\begin{aligned} \hat{n} &= n(1+i\kappa), \quad \hat{n}^2 = n^2(1+2i\kappa), \\ \hat{v}_p = \tfrac{c}{n(1+i\kappa)} &= v_p(1-i\kappa), \quad \hat{v_p}^2 = v_p{}^2(1-2i\kappa), \end{aligned}\right\} \tag{12}$$

with similar expression with subscripts $x, y$, and $z$, e.g.

$$\hat{n}_x = n_x\,(1+i\kappa_x)\,, \quad \hat{v}_x = c/\hat{n}_x = v_x\,(1-i\kappa_x) \qquad \text{etc.}$$

Also

$$\mu\hat{\varepsilon}_k = \hat{n}_k^2 = n_k^2\,(1+2i\kappa_k) = \mu\varepsilon_k\,(1+2i\kappa_k) \quad (k=x,y,z) \tag{13}$$

and comparison with (5) gives

$$\kappa_k = \frac{2\pi}{\omega}\frac{\sigma_k}{\varepsilon_k} \tag{14}$$

Returning to Fresnel's equation (10) we separate the real and imaginary parts. A typical term of (10) is

$$\frac{s_x^2}{\hat{v_p}^2 - \hat{v_x}^2} = \frac{s_x^2}{v_p{}^2 - v_x{}^2 - 2i\left[\kappa v_p{}^2 - \kappa_x v_x{}^2\right]} = \frac{s_x{}^2}{v_p{}^2 - v_x{}^2}\left[1 + 2i\frac{\kappa v_p{}^2 - \kappa_x v_x{}^2}{v_p{}^2 - v_x{}^2}\right], \tag{15}$$

and it follows that the real part of (10) is Fresnel's equation in the old form § 14.2 (24). The imaginary part gives the equation

$$\begin{aligned} \kappa v_p^2 \left\{ \frac{s_x^2}{\left(v_p^2 - v_x^2\right)^2} + \frac{s_y^2}{\left(v_p^2 - v_y^2\right)^2} + \frac{s_z^2}{\left(v_p^2 - v_z^2\right)^2} \right\} \\ = \frac{\kappa_x v_x^2 s_x^2}{\left(v_p^2 - v_x^2\right)^2} + \frac{\kappa_y v_y^2 s_y^2}{\left(v_p^2 - v_y^2\right)^2} + \frac{\kappa_z v_z^2 s_z^2}{\left(v_p{}^2 - v_z{}^2\right)^2} \end{aligned} \tag{16}$$

For any given wave-normal direction $\boldsymbol{s}$, Fresnel's equation in general gives two phase velocities $v_p$ as before. Some of the energy carried by the two waves is now absorbed, and (16) gives approximate values for the two attenuation indices.

We may express the formula for $\kappa$ in a different form. From (7), (4), (9), and (13),

$$\begin{aligned} D_k &= \frac{c^2}{\mu}\frac{\varepsilon_k}{\hat{\varepsilon}_k}\frac{s_k(\boldsymbol{E}\cdot\boldsymbol{s})}{\hat{v}_k^2 - \hat{v}_p^2} \\ &= -\frac{c^2}{\mu}\frac{s_k(\boldsymbol{E}\cdot\boldsymbol{s})}{v_p^2 - v_k^2}\left[1 + 2i\frac{(\kappa-\kappa_k)\,v_p^2}{v_p^2 - v_k^2}\right] \quad (k=x,y,z), \end{aligned} \tag{17}$$

where again only terms up to the first power in the attenuation indices have been retained. Now the imaginary term in (17) involves the difference in the $\kappa$ 's and may in many cases be neglected. This means that one neglects the ellipticity of the vibrations. In this approximation the directions of the two $\boldsymbol{D}$ vectors belonging to any given wave-normal $s$ are the same as for a non-absorbing crystal which has the same (real) principal dielectric constants. We then have

$$\frac{s_k^2}{\left(v_\eta^2 - v_k^2\right)^2} = \left(\frac{\mu}{c^2}\right)^2 \frac{D_k^2}{(s\cdot E)^2}, \tag{18}$$

and (16) may be written as

$$\kappa v_p^2 = \frac{\kappa_x v_x^2 D_x^2 + \kappa_y v_y^2 D_y^2 + \kappa_z v_z^2 D_z^2}{D^2} \tag{19}$$

This formula may evidently break down as $v_p$ approaches one of the principal velocities $v_k$, since the imaginary part in (17), which we neglected, involves the difference ${v_p}^2 - {v_k}^2$ in the denominator (see remarks at the end of § 14.6.3).

Since the two coefficients $\kappa'$ and $\kappa''$ belonging to a given wave-normal direction $s$ are in general different, the two waves are absorbed with different strengths. These two coefficients may be frequency-dependent and vary in different ways with the frequency, so that if white light be incident upon the crystal, the crystal will in general appear coloured, and the colour will depend on the direction of vibration of the incident light. This phenomenon is known as pleochroism; in the case of a uniaxial crystal one also speaks of dichroism; in the case of a biaxial crystal one speaks of trichroism.*

For a uniaxial crystal $[\hat{\varepsilon}_x = \hat{\varepsilon}_y = \hat{\varepsilon}_o, \hat{\varepsilon}_z = \hat{\varepsilon}_e$; also $\kappa_x = \kappa_y = \kappa_o, \kappa_z = \kappa_e]$ the relations take a simpler form. As in the case of a non-absorbing uniaxial crystal, Fresnel's equation breaks into two factors (cf. § 14.3.2), giving

$$\left.\begin{aligned} \hat{v}_p'^2 &= \hat{v}_o^2 \\ \hat{v}_p'^2 &= \hat{v}_o^2 \cos^2\vartheta + \hat{v}_o^2 \sin^2\vartheta, \end{aligned}\right\} \tag{20}$$

where, as before, $\vartheta$ is the angle which the wave normal $s$ makes with the optic axis. The first equation gives, on separating real and imaginary parts,

$$v_p' = v_o, \quad \kappa_p' = \kappa_o, \tag{21a}$$

and the second gives

$$v_p''^2 = v_o^2 \cos^2\vartheta + v_e^2 \sin^2\vartheta, \quad \kappa_p'' v_p''^2 = \kappa_o v_o^2 \cos^2\vartheta + \kappa_e v_e^2 \sin^2\vartheta, \tag{21b}$$

where again we have neglected the terms involving the second power of the attenuation indices. We see that the absorption of the ordinary wave is the same for all directions of propagation.

For a biaxial crystal the relations are much more complicated and we shall confine our attention to special cases of interest. As in § 14.3.3 we consider first those directions of propagation for which $s_x = 0$. Fresnel's equation (10) then gives, in analogy with § 14.3 (6), the equations

$$\left.\begin{aligned} \hat{v}_p'^2 &= \hat{v}_x^2 \\ \hat{v}_p'' &= \hat{v}_z^2 s_y^2 + \hat{v}_y^2 s_x^2. \end{aligned}\right\} \tag{22}$$

Separating real and imaginary parts, we obtain the two velocities and the two attenuation indices corresponding to the wave-normal direction $s\,(0, s_y, s_z)$ :

$$v_p'^2 = v_x^2, \quad \kappa_p' = \kappa_x, \tag{23a}$$

and

$$v_p''^2 = v_z^2 s_y^2 + v_y^2 s_z^2, \quad \kappa_p'' v_p''^2 = \kappa_z v_z^2 s_y^2 + \kappa_y v_y^2 s_z^2 \tag{23b}$$

Similar relations hold, of course, for directions of propagation at right angles to the $y$ and $z$ directions.

0

- This terminology derives from the fact that for a uniaxial crystal there are two characteristic colours and for a biaxial crystal there are three. Some authors, however, mean by dichroic material, any material whose absorption coefficient depends on the state of polarization of the incident light.

In general there are no real values $s_y$ and $s_z$ for which the two roots $\hat{v}_p'$ and $\hat{v}_p''$ are equal. Directions may be found for which the real phase velocities $v_p'$ and $v_p''$ are equal, but the corresponding attenuation indices ( $\kappa'$ and $\kappa''$ ) will in general be different.

Next we consider propagation in directions that are not very different from that of an optic axis of wave normals. In order to apply (19) we must determine the directions of $D'$ and $D''$. This may be done by using the result established in § 14.2.3, according to which the two planes of vibrations ( $D', s$ ) and ( $D'', s$ ) bisect the angles between the planes ( $N_1, s$ ) and ( $N_2, s$ ), $N_1$ and $N_2$ being the axes of wave normals. Let $\psi$ be the angle between the plane ( $N_1, s$ ) and the $xz$-plane (which contains both the optic axes). Since the plane ( $N_2, s$ ) is nearly parallel to the $xz$-plane, it follows from the theorem just

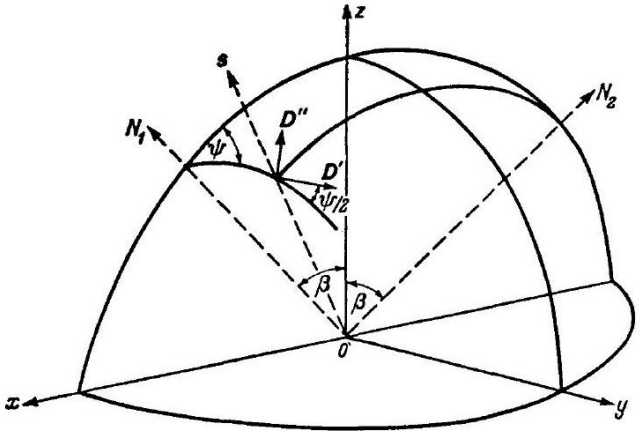



Fig. 14.29. Illustrating the theory of absorbing crystals.

quoted that the angle between $D'$ and the $xz$-plane is nearly $\psi/2$ (see Fig. 14.29). The component of $D'$ in the $xz$-plane is therefore $D' \cos(\psi/2)$. To obtain the $x$ component we must project this vector on to the $x$ axis. Since, approximately, $s$ and $N_1$ coincide, the projection angle is nearly equal to the angle between

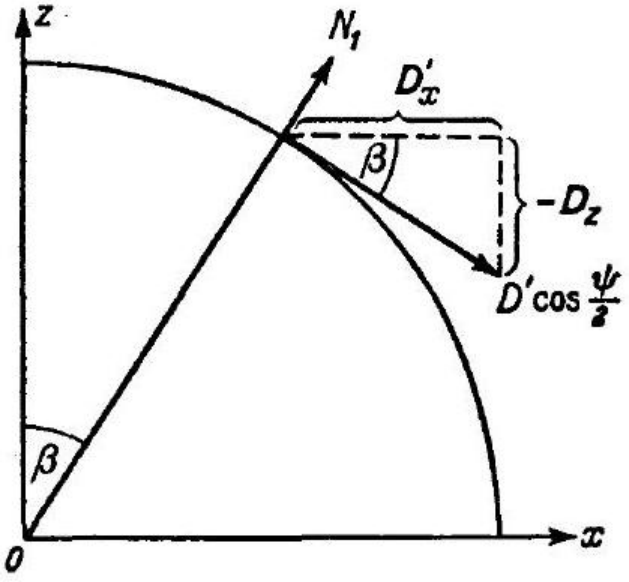



Fig. 14.30. Illustrating the theory of absorbing crystals.

the optic axis $N_1$ and the $z$ axis and consequently (see Fig. 14.30) $D_x' = D' \cos\beta \cos(\psi/2)$.

In a similar way we obtain the other components. Hence

$$\left.\begin{aligned} D'_x &= D' \cos\tfrac{\psi}{2} \cos\beta, \\ D'_y &= D' \sin\tfrac{\psi}{2}, \\ D'_z &= -D' \cos\tfrac{\psi}{2} \sin\beta. \end{aligned}\right\} \tag{24a}$$

The vector $D''$ is orthogonal to $s$ and $D'$; its components may be immediately obtained by replacing $\psi/2$ by $\psi/2 + \pi/2$ in (24a), giving

$$D''_x = -D'' \sin\frac{\psi}{2} \cos\beta, \quad D''_y = D'' \cos\frac{\psi}{2}, \quad D''_z = D'' \sin\frac{\psi}{2} \sin\beta. \tag{24b}$$

We substitute from (24a) and (24b) into (19), and also use the approximation $v'_p = v''_p = v_y\,(v_x > v_y > v_z)$, which is justified since we are restricting ourselves to directions not too far from that of the optic axis. We then obtain the required attenuation indices $\kappa'$ and $\kappa''$ :

$$\left.\begin{aligned} \kappa' v_y^2 &= \left(\kappa_x v_x^2 \cos^2\beta + \kappa_z v_z^2 \sin^2\beta\right) \cos^2\tfrac{\psi}{2} + \kappa_y v_y^2 \sin^2\tfrac{\psi}{2}, \\ \kappa'' v_y^2 &= \left(\kappa_x v_x^2 \cos^2\beta + \kappa_z v_z^2 \sin^2\beta\right) \sin^2\tfrac{\psi}{2} + \kappa_y v_y^2 \cos^2\tfrac{\psi}{2} \end{aligned}\right\} \tag{25}$$

In view of the approximation involved in (19) these formulae cannot be expected to remain valid when $s$ is in the immediate neighbourhood of the optic axis. In the limiting case when the wave normal is along the optic axis, the angle $\psi$ is undetermined; to obtain the corresponding attenuation indices we return to Fresnel's equation. For any $s$ direction in the $xz$-plane ( $s_y = 0$ ) we obtain, as in equations (23), by equating real and imaginary parts

$$\left.\begin{aligned} v'^2_p &= v_y^2, \quad \kappa'_p = \kappa_y, \\ v''_p &= v_x s_z^2 + v_z s_x^2, \quad \kappa''_p v''^2_p = \kappa_x v_x^2 s_z^2 + \kappa_z v_z^2 s_x^2. \end{aligned}\right\} \tag{26}$$

In particular, for the optic axis $s_x = \sin\beta, s_z = \cos\beta$, where $\beta$ is the angle given by § 14.3 (11), between either of the optic axes and the $z$ axis. Also $v'_p = v''_p = v_y$ and the equations on the right of (26) become, if we also write $\kappa_\perp$ in place of $\kappa'$ and $\kappa_\parallel$in place of $\kappa''$ :

$$\left.\begin{aligned} \kappa_\perp &= \kappa_y \\ \kappa_\parallel v_y^2 &= \kappa_x v_x^2 \cos^2\beta + \kappa_z v_z^2 \sin^2\beta. \end{aligned}\right\} \tag{27}$$

$\kappa_\perp$ is the attenuation index for a $D$ wave polarized at right angles to the plane of the optic axes, and $\kappa_\parallel$is the index for a $D$ wave polarized in the plane of the axes. Thus the absorption of a wave propagated in the direction of the optic axis depends on its direction of vibration.

It is convenient to express $\kappa'$ and $\kappa''$ in terms of the indices $\kappa_\parallel$and $\kappa_\perp$ and the azimuthal polarization angle $\psi$, by substituting from (27) into (25). This gives

$$\left.\begin{aligned} \kappa' &= \kappa_\parallel \cos^2\tfrac{\psi}{2} + \kappa_\perp \sin^2\tfrac{\psi}{2} = \tfrac{\kappa_\parallel + \kappa_\perp}{2} + \tfrac{\kappa_\parallel - \kappa_\perp}{2} \cos\psi, \\ \kappa'' &= \kappa_\parallel \sin^2\tfrac{\psi}{2} + \kappa_\perp \cos^2\tfrac{\psi}{2} = \tfrac{\kappa_\parallel + \kappa_\perp}{2} - \tfrac{\kappa_\parallel - \kappa_\perp}{2} \cos\psi. \end{aligned}\right\} \tag{28}$$

### 14.6.2 Interference figures from absorbing crystal plates

We shall now briefly consider interference effects with absorbing crystal plates which are cut at right angles to an optic axis of wave normals. The theory is not very different from that relating to non-absorbing crystal plates; the only important distinction is that the two interfering rays are absorbed with different strengths, and as a result the visibility of the fringes is decreased. Other conclusions, and in particular the expression for the phase difference, remain unchanged in our approximation, since the geometrical laws of propagation are the same as before.

On travelling a distance $l$ in an absorbing medium, the amplitude of a plane wave is, according to § 13.1 (19), reduced by a factor $\exp\{-\omega n\kappa l/c\}$. Hence with the same arrangement and notation as in § 14.4.3 (cf. Fig. 14.20) the amplitudes of the principal vibrations on emergence from the plate are given by [cf. § 14.4 (13)]

$$OB = Ee^{-\frac{\omega\kappa'}{v'}l}\cos\phi, \quad OC = Ee^{-\frac{\omega\kappa''}{v''}l}\sin\phi \tag{29}$$

Here $l = h/\cos\theta_2$, $h$ being the thickness of the plate and $\theta_2$ the angle which the wave normal in the plate makes with the axis, it being assumed that the two paths in the plate are the same for both waves; this is approximately so, if we restrict ourselves to wave-normal directions close to the optic axis. In the same approximation we may take $v' = v'' = v_y$ in (29). It is convenient to set

$$u = \frac{\omega l}{v_y} \sim \frac{\omega l}{v'} \sim \frac{\omega l}{v''} \tag{30}$$

With this substitution,

$$OB = Ee^{-\kappa' u}\cos\phi, \quad OC = Ee^{-\kappa'' u}\sin\phi \tag{31}$$

In place of § 14.4 (14) we now obtain for the amplitudes of the waves, after passing through the polarizer and the analyzer, the expressions (cf. Fig. 14.20)

$$OF = Ee^{-\kappa' u}\cos\phi\cos(\phi-\chi), \quad OG = Ee^{-\kappa'' u}\sin\phi\sin(\phi-\chi) \tag{32}$$

The total intensity of the light brought to interference is

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta \tag{33}$$

where, apart from unessential proportional factors, $I_1 = OF^2, I_2 = OG^2$ and the phase difference $\delta$ is calculated as before.

We now examine some special cases of interest.

(a) Uniaxial crystals. In this case we have

$$\left.\begin{array}{ll} v_x = v_y = v_o, & v_z = v_e, \\ \kappa_x = \kappa_y = \kappa_o, & \kappa_z = \kappa_e. \end{array}\right\} \tag{34}$$

The direction of vibration of the extraordinary ray is in the principal plane, i.e. in the plane containing the wave normal and the optic axis. We may, therefore, identify the angle between the $D$ vector of the extraordinary wave and the direction $OP$ of the polarizer with the angle $\phi$ in equation (32); to retain agreement with (21a) and (21b) we must, however, interchange $\kappa'$ and $\kappa''$. With the Nicol prisms crossed ( $\chi = \pi/2$ ) we have from (32)

$$OF = Ee^{-\kappa' u}\cos\phi\sin\phi, \quad OG = -Ee^{-\kappa' u}\sin\phi\cos\phi \tag{35}$$

and (33) gives

$$I = \frac{E^2}{4}\sin^2 2\phi\left\{e^{-2\kappa' u} + e^{-2\kappa'' u} - 2e^{-(\kappa'+\kappa'')u}\cos\delta\right\} \tag{36a}$$

For the optic axis, $\kappa' = \kappa''$ and $\delta = 0$, and (36a) reduces to

$$I_0 = 0 \tag{36b}$$

hence at the centre of the pattern (the point corresponding to the optic axis) there is darkness. The field of view is crossed by a dark isogyre, given by $\sin 2\phi = 0$, i.e. by $\phi = 0$ and

$\phi = \pi/2$. Thus the isogyre has the form of a cross with its arms parallel to the directions of the polarizer and the analyzer (cf. § 14.4.4). The minima and maxima of intensity are

$$\left.\begin{aligned} I_{\min} &= \tfrac{E^2}{4}\sin 2\phi\left\{e^{-2\kappa' u} + e^{-2\kappa'' u} - 2e^{-(\kappa'+\kappa'')u}\right\}, \\ I_{\max} &= \tfrac{E^2}{4}\sin 2\phi\left\{e^{-2\kappa' u} + e^{-2\kappa'' u} + 2e^{-(\kappa'+\kappa'')u}\right\}, \end{aligned}\right\} \tag{37}$$

so that the visibility $\mathscr{V}$ of the fringes is

$$\mathscr{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2e^{-(\kappa'+\kappa')u}}{e^{-2\kappa' u} + e^{-2\kappa' u}} = \frac{1}{\cosh\left\{(\kappa' - \kappa'')\, u\right\}}. \tag{38}$$

Thus the visibility is the greater the smaller is the difference between $\kappa'$ and $\kappa''$. Since for wave-normal directions near the optic axis ( $\vartheta$ small) $\kappa'$ and $\kappa''$ are nearly equal, the fringes should be clearly visible in this region, provided that sufficient light is transmitted by the plate. If $\kappa_o$ is small compared with $\kappa_e$ (e.g. for magnesium platinum cyanide) there will be relatively little absorption near the optic axis and the central fringes will be bright. If $\kappa_0$ is large compared to $\kappa_e$ (e.g. for tourmaline), the absorption will be relatively weakest near the optic axis and the central fringes will be dark. In both cases the visibility decreases from the centre towards the edge of the field. If $\kappa_o$ and $\kappa_e$ are nearly equal, the interference figure will be similar to that of a nonabsorbing crystal, apart from a decrease in visibility with increasing distance from the centre.

(b) Biaxial crystals. We again restrict ourselves to the case of crossed Nicol prisms ( $\chi = \pi/2$ ) and assume that the faces of the plate are perpendicular to one of the optic axes ( $N_1$ ). Only waves whose directions of propagation are close to this axis will be considered.

Let $N_1, N_2$, and $Q$ be the points of intersection of the two optic axes and of the wave normal with the exit face of the plate. Further let $\psi$ be the angle which the line $N_1Q$ makes with $N_1N_2$ and $\alpha$ the angle between the plane of the optic axes and the plane of vibrations transmitted by the polarizer $P$.

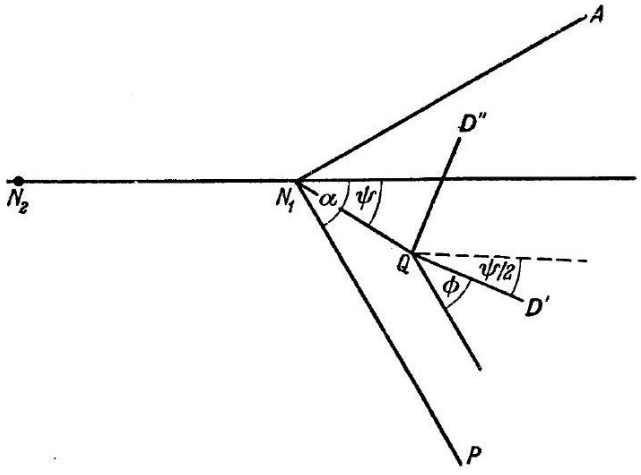



Fig. 14.31. Illustrating the theory of interference on absorbing biaxial crystal plates.

If $N_1Q \ll N_1N_2$ the angle which the direction of vibration $D'$ makes with $N_1N_2$ is approximately $\psi/2$ so that the angle $\phi$ between $D'$ and the direction $OP$ is approximately (see Fig. 14.31)

$$\phi = \alpha - \frac{\psi}{2}. \tag{39}$$

The amplitudes (32) of the waves brought to interference, therefore, are

$$OF = E\cos\left(\alpha - \frac{\psi}{2}\right)\sin\left(\alpha - \frac{\psi}{2}\right)e^{-\kappa' u}, \quad OG = -E\sin\left(\alpha - \frac{\psi}{2}\right)\cos\left(\alpha - \frac{\psi}{2}\right)e^{-\kappa' u}. \quad (40)$$

Hence the intensity is

$$I = \frac{E^2}{4}\sin^2(2\alpha - \psi)\left\{e^{-2\kappa' u} + e^{-2\kappa'' u} - 2e^{-(\kappa' + \kappa'')u}\cos\delta\right\} \quad (4\text{la})$$

For a wave propagated in the direction of the optic axis, $\psi$ is undetermined. In this special case we resolve the vibrations into components parallel and perpendicular to the plane containing the optic axes, i.e. we set $\psi = 0$ and write $\kappa_|$and $\kappa_\perp$ in place of $\kappa'$ and $\kappa''$ in (41a) (cf. (28)). Also $\delta = 0$, so that we obtain, in place of (41a),

$$I_0 = \frac{E^2}{4}\sin^2 2\alpha\left\{e^{-\kappa_n u} - e^{-\kappa_+ u}\right\}^2 \quad (41\text{b})$$

Equation (41a) shows that the intensity is zero along a line on which $\sin(2\alpha - \psi) = 0$, this being a principal isogyre. Whilst in the case of a non-absorbing crystal the isogyre passed through the point corresponding to the optic axis, now this will no longer be the case, unless, as seen from (41b), $\alpha = 0$ or $\alpha = \pi/2$, i.e. unless the plane of the polarizer is parallel or at right angles to the plane containing the two optic axes.

The dark isochromates, given by $\cos\delta = 1$, i.e. by $\delta = 0, \pm 2\pi, \pm 4\pi, \ldots$ are rings surrounding the point which corresponds to the optic axis. The visibility of the fringes, again given by (38), is appreciable only when $\kappa'$ and $\kappa''$ are nearly equal. According to (28) this is so when $\psi$ is nearly $\pi/2$ or $-\pi/2$.

Let us consider how the intensity changes with $\psi$ when $\delta$ is kept constant. If we substitute from (28), equation (41a) may be written in the form

$$I = \frac{E^2}{2}e^{-(\kappa_{11} + \kappa_\perp)u}\sin^2(2\alpha - \psi)\left\{\cosh\left[(\kappa_1 - \kappa_\perp)\, u\cos\psi\right] - \cos\delta\right\}. \quad (42)$$

The term in the brace brackets is greatest when $|\cos\psi|$ is greatest and least when $|\cos\psi|$ is least. Hence the intensity is maximum when $\psi = 0$ and $\psi = \pi$, and minimum when $\psi = \pi/2$ and $-\pi/2$. Thus in addition to the dark isogyre $\psi = 2\alpha$, the field of view is also crossed by a dark "brush" with its arms perpendicular to the plane containing the optic axes.*

### 14.6.3 Dichroic polarizers

In the preceding section we have studied the propagation of polarized light through an absorbing crystal plate. Here we shall consider the effect of such a plate on natural (unpolarized) light.

We again assume that the faces of the plate are perpendicular to the optic axis (or to one such axis if it is biaxial) of wave normals, and consider propagation in a direction close to that of the optic axis. A parallel beam of natural light may be regarded as consisting of two mutually incoherent beams of equal amplitudes, polarized in any two mutually orthogonal directions perpendicular to the direction of propagation (cf. § 10.8(56) ). Let us choose as the directions of vibration of the two partial beams the directions of the principal vibrations in the crystal. If $E$ is the amplitude of either of the partial beams on entering the plate, their amplitudes after travelling a distance $l$ in the plates are

$$E' = Ee^{-\kappa' u}, \quad E'' = Ee^{-\kappa'' u} \quad (43)$$

- From equation (44) which, by (28), may be written as

$$I = 2E^2 e^{-(\kappa_1+\kappa_\perp)u} \cosh\left\{(\kappa_1 - \kappa_\perp)\, u \cos\psi)\right\},$$

it is seen that the dark brush also appears in the absence of a polarizer and analyzer, when natural light is passed through the plate.
where as before $u = \omega l/v_y$, it being assumed that the light is quasi-monochromatic and of mean frequency $\omega$. The total intensity is then given by

$$I = I' + I'', \tag{44}$$

where

$$I' = I_0 e^{-2\kappa' u}, \quad I'' = I_0 e^{-2\kappa'' u} \tag{45}$$

and $I_0 = E^2$. For the direction of the optic axes, $\kappa'$ and $\kappa''$ must be replaced by $\kappa_1$ and $\kappa_\perp$.

We see that after the light has travelled a distance $l$ in the medium, the amplitudes of the two components are in the ratio $\exp\{-(\kappa' - \kappa'')\, u\}$, so that the light has become

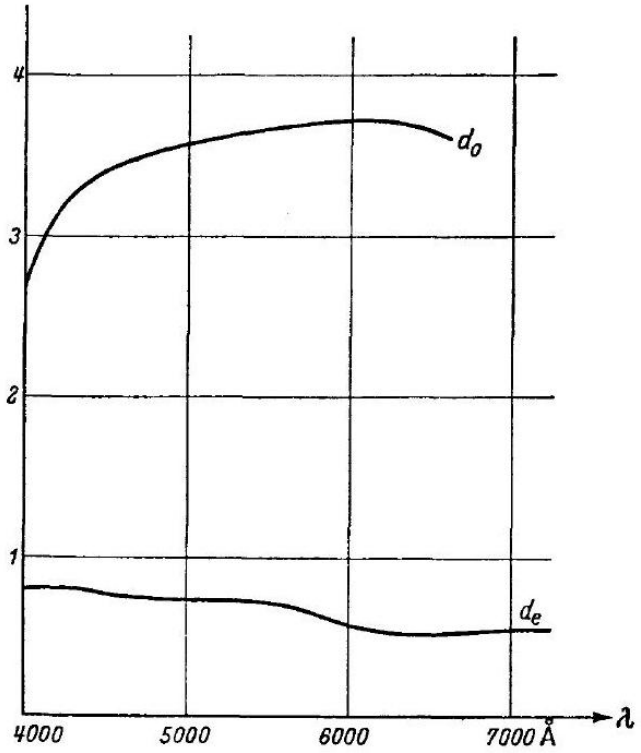



Fig. 14.32. Dichroism of black tourmaline plate about 0.2 mm thick, cut \$24\hat{}\{\}\{\circ\}\$ from parallelism with the optic axis, for light incident at right angles to the optic axis.

$$d = 4\pi\,(\log_{10}\theta)\,(n\kappa//\lambda) \sim 5 \cdot 5n\kappa l/\lambda.$$

[After E. H. Land and C. D. West, contribution in Colloid Chemistry, Vol. 6, ed. J. Alexander (New York, Reinhold Publishing Corporation, 1946), 167.]
partially polarized. If the two attenuation indices are very different, a relatively thin sheet of the material will be sufficient to transform the incident unpolarized beam into one that is nearly linearly polarized; i.e. the plate acts as a polarizer. An example of a natural crystal which will act as polarizer of this type is tourmaline, which suppresses the ordinary ray much more strongly than the extraordinary ray. However, for most wavelengths it also absorbs

a considerable part of the extraordinary ray (see Fig. 14.32), so that it is not suitable for practical applications. It is possible, chiefly as the result of researches of Land (c. 1932) and his co-workers, to produce synthetic dichroic materials which act as excellent polarizers. These materials, known commercially as polaroids, are not single crystals, but are sheets of organic polymers with long-chain molecules, brought into almost complete alignment by stretching or by some other treatment,* and sometimes dyed. A polymer which is particularly

suitable for this purpose is the synthetic product polyvinyl alcohol $(-CH_2 - CHOH-)_x$. In Fig. 14.33, the dichroism of iodine on oriented polyvinyl alcohol is shown as function of the wavelength.

The ratio of the absorption coefficients of a good dichroic polarizer may be as high as 100 : 1. It may transmit about 80 per cent of light polarized in one direction and less than 1 per cent of light polarized in the direction at right angles to it. Large sheets of such polarizing materials are obtainable; machines giving pieces thirty inches wide and indefinitely long are in use in commercial production. In this respect

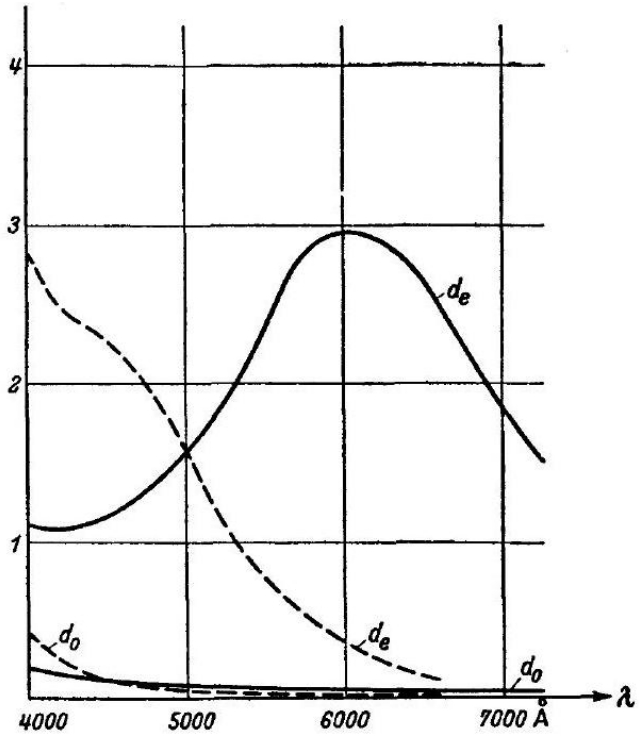



Fig. 14.33. Dichroism of iodine on oriented polyvinyl alcohol, blue (full line) and brown (broken line).
$d=4 \pi\left(\log _{10} \theta\right)(n \kappa l / \lambda) \sim 5 \cdot 5 n \kappa l / \lambda$.

[After E. H. Land and C. D. West, contribution in Colloid Chemistry, Vol. 6, ed. J. Alexander (New York, Reinhold Publishing Corporation, 1946), 177.]
these "polaroid sheets" are superior to the Nicol prisms the dimensions of which are limited by the scarcity of large pieces of calcite of good optical quality.

Finally, let us recall that in the main part of this chapter we have assumed that the ellipticity of the principal light vibrations in the crystal may be neglected. This assumption is certainly not

0

- For a fuller discussion of dichroic polarizers see the article referred to in the captions to Figs. 14.32 and 14.33 or E. Land, J. Opt. Soc. Amer., 41 (1951), 957.

always justified; it may be shown, for example, that in every absorbing crystal there exist four directions, lying in pairs near the optic axes, for which the polarization is circular. However, with weaker and weaker absorption the regions of appreciable ellipticity are more and more restricted to the neighbourhood of these special directions, which themselves tend to coincidence with the optic axis. Taking this into account, the general character of the phenomenon is retained and we, therefore, do not need to enter into a more detailed discussion. Naturally our analysis does not reveal some of the finer aspects of the optical theory of absorbing crystals; for their discussion we must refer elsewhere.*

0

- See, for example, G. Szivessy, Handbuch der Physik, Vol. 20 (Berlin, Springer, 1928), 861904

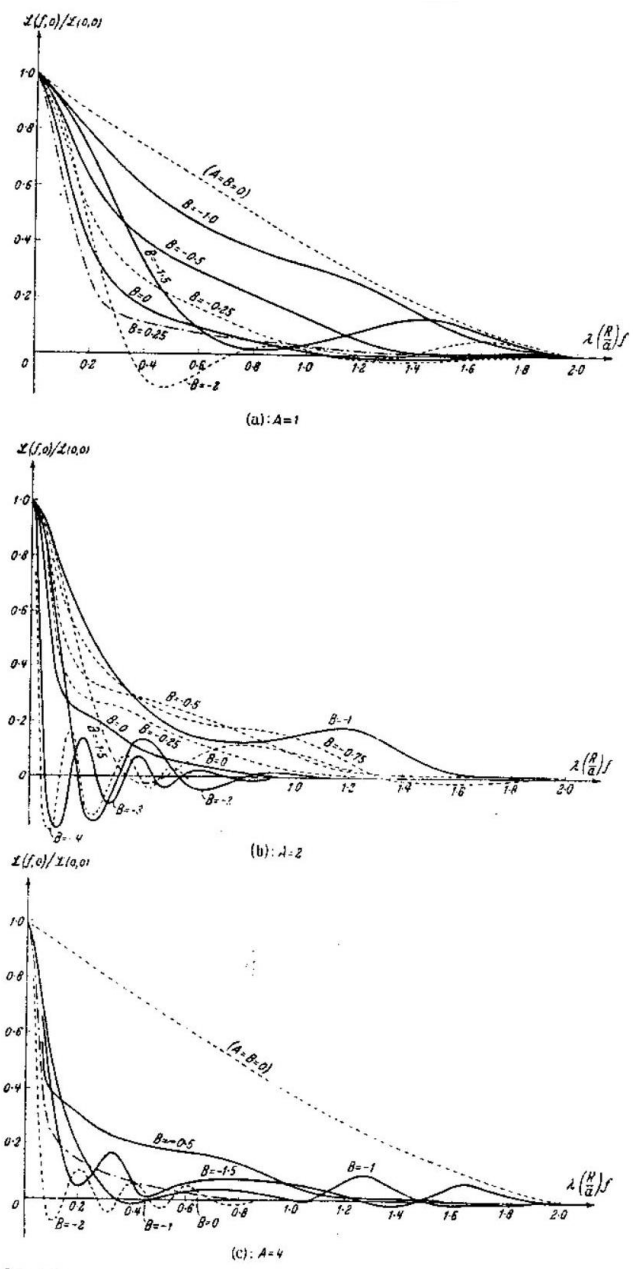



Fig. 9.15. The normalized frequency response curves for incoherent illumination, at selected focal settings of a system suffering from a small amount of primary spherical aberration $\Phi=A\left(\rho\hat{}\{4\}+B \rho\hat{}\{2\}\right) \lambda,|G|=1$. The value $B = 0$ corresponds to the paraxial focal plane and $B = -2$ corresponds to the receiving plane through the marginal focus.

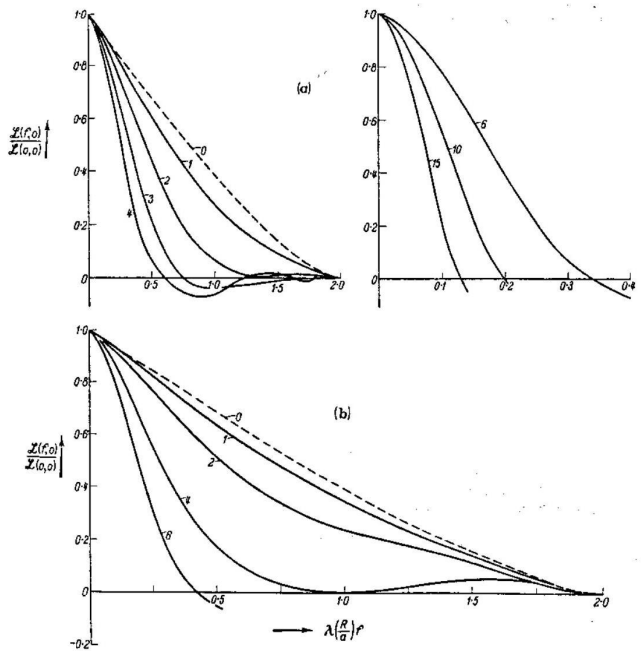


Fig. 9.16. The normalized frequency rosponse curves for incoherent illumination, in the presence of primary astigmatism $\Phi=\frac{m \lambda}{\pi} \rho\hat{}{2} \cos \hat{}{2} \theta,|G|=1$, for receiving plane midway between the tangential and sagittal focal lines. Line periodic along the meridian $\theta = \pi/4[(a)]$, and along the meridians $\theta = 0$ or $\theta = \pi/2[(b)]$. The number on each curve is the value of $m$.

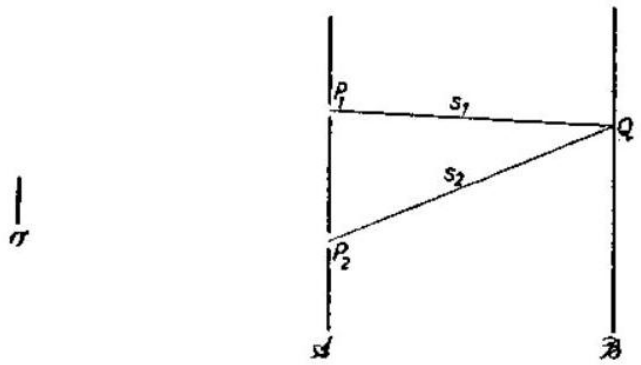


Fig. 10.1. An interference experiment with polychromatic light from an extended source $\sigma$.

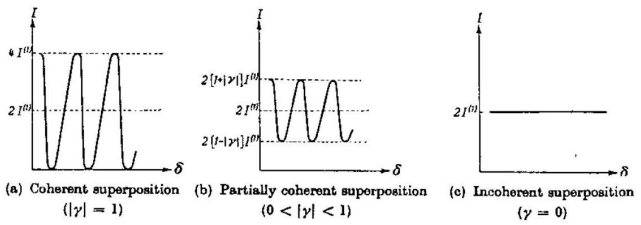



Fig. 10.2.. Intensity distribution in the interference pattern produced by two quasimonochromatic beams of equal intensity $I\hat{}{(1)}$ and with degree of coherence $|\gamma|$.

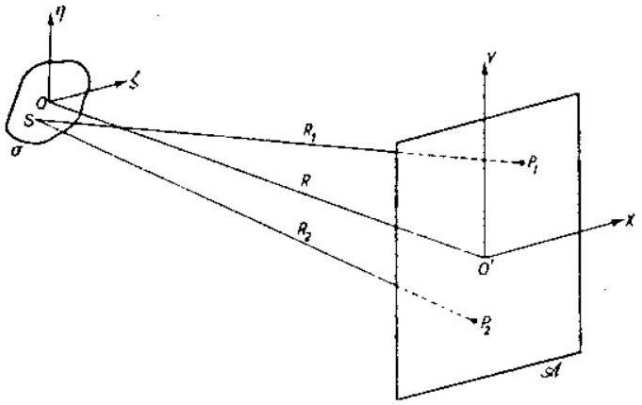



Fig. 10.3. Illustrating the van Cittert-Zernike theorem.

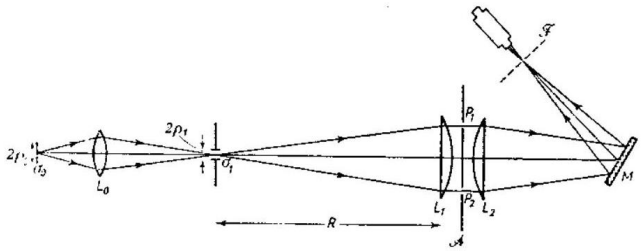



Fig. 10.4. The Diffractometer.

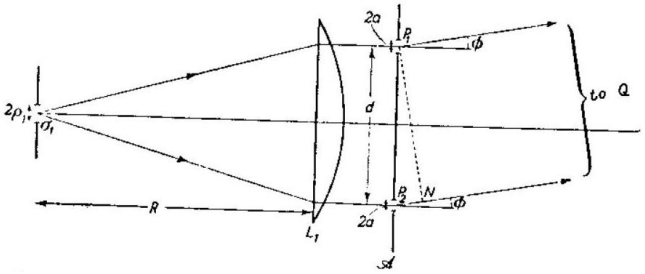



Fig. 10.5. Calculation of the intensity distribution in the focal plane of the Diffractometer.

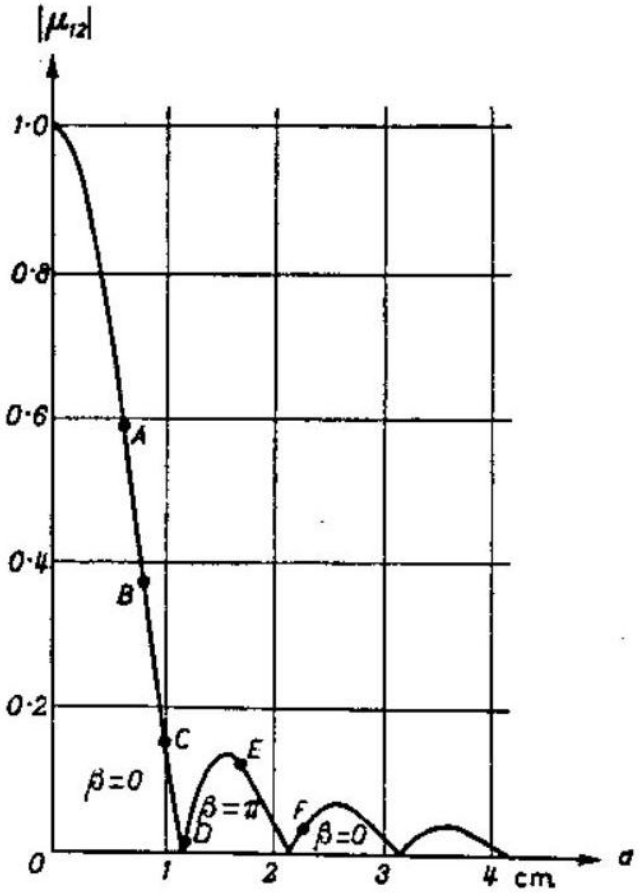



Fig. 10.7. Two beam interference with partially coherent light. The degree of coherence as function of the separation $d$ of the two illuminated apertures in the Diffractometer. ( $\rho_{1}=0.45 \times 10\hat{}\{-2\} \mathrm{~cm}$, R=152 $\mathrm{~cm}$, $\lambda=5790$ \AA; incoherent illumination of $\sigma_1$ assumed.)

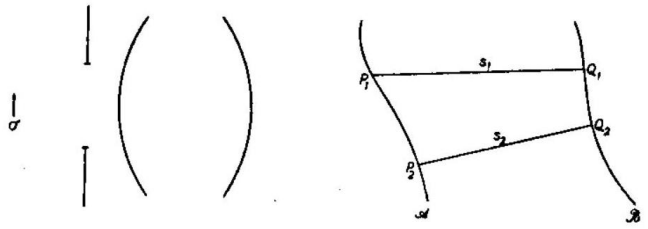



Fig. 10.8. Propagation of mutual intensity: illustrating formula (45).

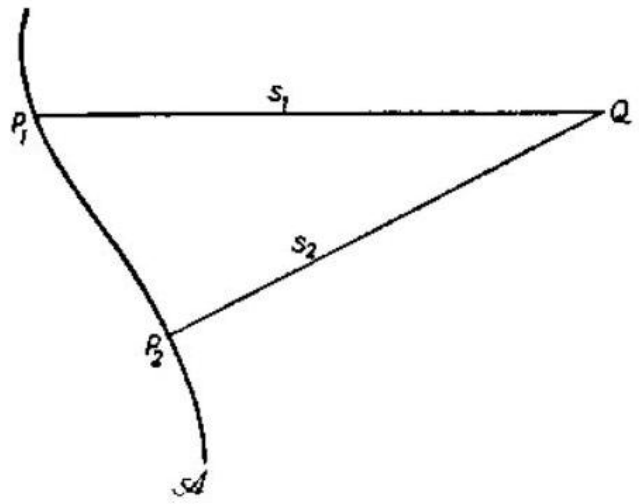



Fig. 10.9. Illustrating formula (46).

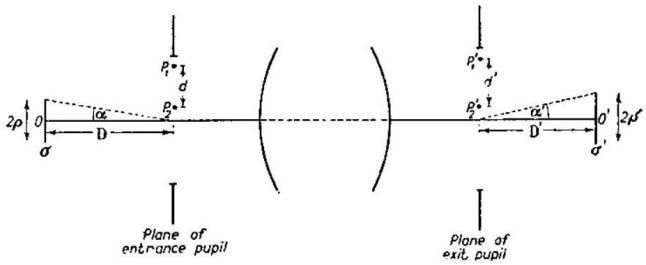



Fig. 10.10. Calculation of the degree of coherence in the image of an incoherent light source.

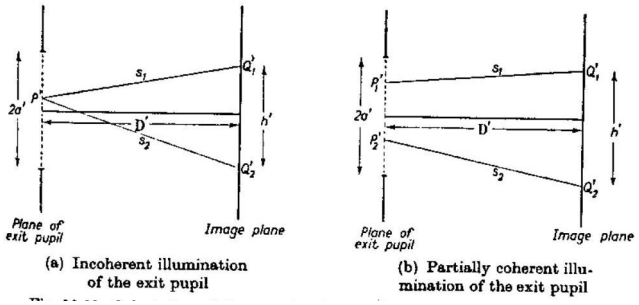



Fig. 10.11. Calculation of the complex degree of coherence in the image plane.

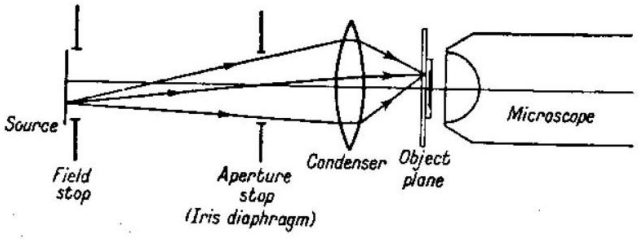



Fig. 10.12. Critical illumination.

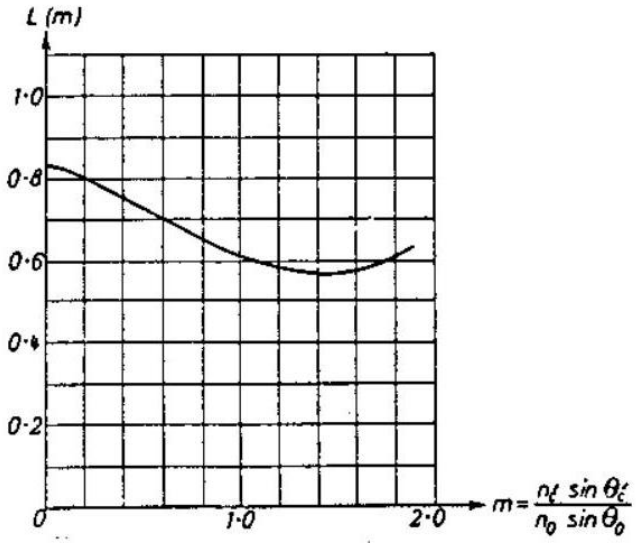



Fig. 10.13. Effect of the condenser aperture on the resolution of two pinholes of equal brightness. (After H. H. Hopkins and P. M. Barham, Proc. Phys. Soc., 63 (1950), 72.)

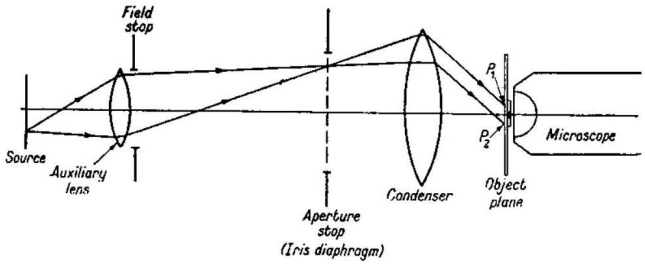



Fig. 10.14. Köhler's method of illumination.

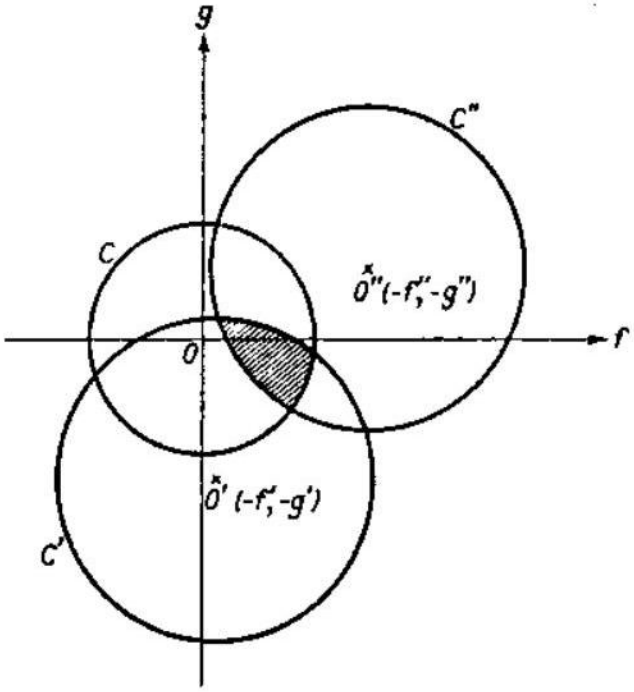



Fig. 10.15. The effective domain of integration (shown shaded) for the transmission cross-coefficient $\mathscr{F}\left(f\hat{}\{\prime}, g\hat{}\{\prime} ; f\hat{}\{\prime \prime}, g\hat{}\{\prime \prime}\right)$ of an image-forming system with a circular exit pupil of radius $a$. The object is assumed to be transilluminated by quasi-monochromatic light of mean wavelength $\bar{\lambda}_0$ via a condenser of numerical aperture $mn_0 \sin\theta_0$, where $n_0 \sin\theta_0$ is the numerical aperture of the image-forming system. $C'$ and $C''$ are circles of radius $a/\bar{\lambda}R \sim n_0 \sin\theta_0/M\bar{\lambda}_0$ centred at the points $O'(-f', -g')$ and $O''(-f'', -g'')$ respectively. $C$ is a circle of radius $mn_0 \sin\theta_0/\lambda_0$ centred on the origin. ( $R$ = radius of Gaussian reference sphere, $M$ = Gaussian magnification.)

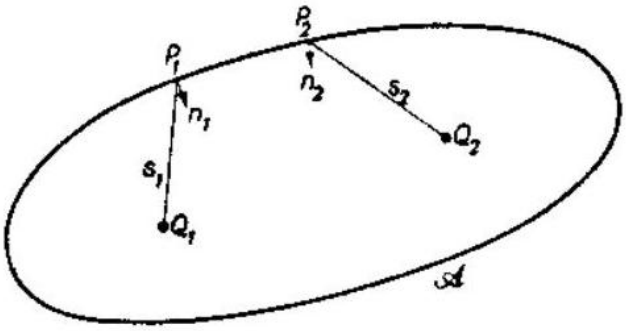



Fig. 10.16. Notation used in the rigorous formulation of the propagation law for mutual coherence.

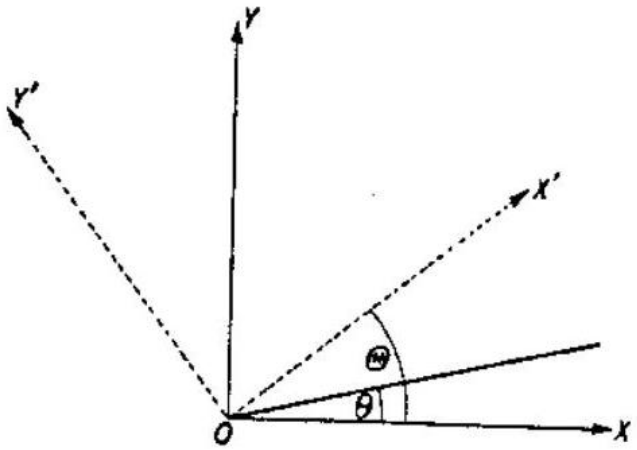



Fig. 10.17. Illustrating notation.

# Part III
# Fundamentals of Quantum Optics

## 10 Basics of Laws of Quantum Optics

(photoeffect, Compton effect, pressure of light etc. currenlty, this is writtein in the QM or QFT)

### 10.1 Photoelectric effect

(I'll write later about it in context of optics)
(see Landsberg for example)

### 10.2 Compton effect

(I'll write later about it in context of optics)
(see Landsberg for example)

### 10.3 Pressure of light

(I'll write later about it in context of optics)
(see Landsberg for example)

## 11 Basics Models of Quantum Optics

(this is written in the note on quantum optics)

### 11.1 Lasers: Basics

#### 11.1.1 Basic Model of a Laser

(тут только про излучение по идее, а не про лазеры. ибо на них очень большой раздел в следующей части есть.)
(в 1й части выписаны основные формулы лазеров??)

**простейшая модель лазера в двух словах**

(как пойму - напишу)

**переходы в одном атоме**

Приведем типичную модель лазера, которая качественно описывает его излучение.

Рассмотрим процессы взаимодействия излучения с атомом — процессы, играющие фундаментальную роль в квантовых генераторах света.

Пусть в неком объеме с зеркальными стенками находится В одном и том же состоянии $n$ фотонов и один атом в возбужденном состоянии с уровнями энергии $E_2$ и $E_1$, такими что $\dot{E}_2 - E_1 = \hbar\omega$, где $\omega$ - частота фотона.

Динамическое равновесие в системе поддерживается за счет того, что все время происходит то поглощение одного из фотонов, то испускание фотона атомом.

Когда атом находится в нижнем состоянии $E_1$, то число фотонов в системе равно $(n+1)$. За большое время $t$ атом много раз будет находиться то в первом, то во втором состоянии, и число переходов из нижнего состояния в верхнее $N_{12}$ практически (с точностью до единицы) будет равно числу обратных переходов $N_{21}$.

Обозначим как $W_{12}^{(1)}$ вероятность перехода атома в возбужденное состояние под действием одного фотона в единицу времени, а $t_1$ - время, которое атом проводит в среднем в нижнем состоянии, теперь запишем

$$N_{12} = W_{12}^{(1)}(n+1)t_1 \approx N_{21}$$

Эту формулу можно переписать следующим образом:

$$N_{12} = nW_{12}^{(1)}t_1 + w_{12}^{(1)}t_1 = N_{21}^{\text{инд}} + N_{21}^{\text{сп}}$$

Первый член в формуле (6.37) зависит от числа имеющихся в системе фотонов и поэтому он описывает индуцированные переходы $N_{21}^{\text{инд}}$, т.е. переходы, происходящие под действием внешнего излучения.

Второй член вообще не зависит от излучения (от числа фотонов) и он описывает спонтанные, т.е. самопроизвольные переходы атома из возбужденного состояния $N_{21}^{\text{с}}$.

Как следует из формулы (6.37), отношение числа индуцированных переходов к числу спонтанных определяется лишь числом фотонов

$$\frac{N_{21}^{\text{инд}}}{N_{21}^{\text{сп}}} = n.$$

Отношение числа вынужденных переходов к числу спонтанных равно, естественно, отношению вероятностей этих переходов, т.е. формула (6.38) может быть записана в виде

$$\frac{w_{21}^{\text{инд}}}{w_{21}^{\text{сп}}} = n$$

откуда следует, что

$$w_{21}^{\text{сп}} = \frac{w_{21}^{\text{инд}}}{n} = W_1^{\text{инд}}$$

где $W_1^{\text{инд}}$ - вероятность индуцированного перехода под действием одного фотона.

мы получили два весьма существенных результата. 1.

вероятность индуцированного излучения пропорциональна числу имеющихся в системе фотонов.

другими словами, чем больше фотонов, тем больше вероятность испустить еще один.

вероятность спонтанного излучения $w^{\text{сп}}$ равна вероятности индуцированного перехода под влиянием одного фотона $w_1^{\text{инд}}$.

На частном примере мы на самом деле показали справедливость более общего утверждения, которое носит название принципа детального равновесия и играет фундаментальную роль в науке.

**принцип детального равновесия**

принцип детального равновесия формулируется следующим образом: если между двумя состояниями возможен переход в одном направлении, то он возможен и в обратном направлении, и вероятности этих переходов (вернее приведенные вероятности, т.е. вероятности переходов, деленные на статистический вес конечного состояния) одинаковы.

Фактически этот принцип является аналогом классической теоремы об обратимости времени - если возможна некоторая траектория частицы, которая проходится телом в одном направлении, то та же самая траектория может быть пройдена телом и в обратном направлении, ибо при изменении знака времени все скорости переменят знак, а траектории останутся неизменными.

**случай многих атомов - (?) вывод формулы Бозе-Эйнштейна**

Теперь мы можем рассмотреть более общий случай, когда имеется в системе много таких атомов, из которых $n_1$ в равновесии число атомов, переходящих из основного состояния в возбужденное, должно быть равно числу обратных переходов за тот же временной интервал, и, если среднее число фотонов данной энергии в системе равно $\bar{n}$, то условие равновесия можно записать в виде

$$N_1 \bar{n} W_{12}^{(1)} t = N_2 \bar{n} W_{21}^{\text{инд}}\, t + N_2 W_{21}^{\text{с}} t$$

Если температура системы будет столь большая, что $E_2 - E_1 \ll\ll kT$, то число атомов в основном и возбужденном состояниях будет почти одинаково ($N_1 \approx N_2$).

Из выражения (6.40) и того, что при большой температуре $\bar{n} \gg 1$ и вторым членом в уравнении (6.39) можно пренебречь, следует, что должно выполняться следующее соотношение между вероятностями переходов:

$$W_{12}^{(1)} = W_{21}^{\text{инд}} = W_{21}^{\text{сп}}$$

Поэтому из условия (6.41) имеем, что при любой конечной температуре

$$N_1 \bar{n} = N_2 (\bar{n} + 1)$$

то есть

$$\frac{N_2}{N_1} = \frac{\bar{n}}{\bar{n} + 1}$$

Но заселенность уровней при данной температуре определяется больцмановским фактором:

$$N_2 N_1 = \frac{e^{-E_2/k_{\text{B}}T}}{e^{-E_1/k_{\text{B}}T}} = e^{-\Delta E/k_{\text{E}}T} = \exp\left(-\frac{\hbar\omega}{kT}\right)$$

и мы приходим к соотношению

$$\bar{n}\left[1 - \exp\left(-\frac{\hbar\omega}{k_{\text{B}}T}\right)\right] = \exp\left(-\frac{\hbar\omega}{k_{\text{E}}T}\right)$$

Окончательно получаем, что при равновесии число фотонов с частотой $\omega$ в одном состоянии равно

$$\bar{n} = \frac{1}{e^{\hbar\omega/k_{\text{B}}T} - 1}$$

Мы видим, что $\bar{n}$ не что иное, как планковский множитель.

Эта формула называется распределением Бозе-Эйнштейна для средних чисел заполнения $\bar{n}$, так как, как уже указывалось, этой формулой определяется среднее число частиц, находящихся в данном состоянии.

Хотя это распределение получено для фотонов, оно справедливо также для любых бозе-частиц, т.е. частиц с целым спином.

**взаимодействие излучения со средой**

Перейдем к рассмотрению взаимодействия излучения со средой.

Пусть на некую среду, содержащую $N$ атомов при температуре $T$, падает вдоль оси $z$ монохроматическое излучение интенсивности $I$ и частоты $\omega$ такое, что энергия этого излучения равна разности энергетических уровней атомов среды, т.е.

это излучение является резонансным для среды.

Пусть $N_1$ это число атомов среды, находящихся на основном состоянии, а $N_2$ - число атомов в возбужденном состоянии; естественно, число частиц на обоих уровнях равно всегда полному числу частиц $N_1 + N_2 = N$. В соответствии с полученными выше выводами в среде будут происходить как процессы поглощения, так и спонтанного и индуцированного испусканий, и уравнение энергетического баланса имеет вид

$$dI = -W_a^{(1)} I N_1 dz + W_a^{(1)} I N_2 dz + W_s N_2 dz$$

где $W_a^{(1)}, W_s$ - вероятности поглощения и спонтанного излучения одного фотона.

Если спонтанное излучение значительно менее вероятно, чем индуцированное (достаточно высока интенсивность падающего излучения), то

$$\frac{dI}{I} = W_a^{(1)} \left(N_2 - N_1\right) dz$$

и мы получаем, что интенсивность падающего излучения зависит от расстояния как

$$I = I_0 \exp\left[W_a^{(1)} \left(N_2 - N_1\right) z\right]$$

если в среде заселенность уровней соответствует равновесной, то всегда $N_2 < N_1$, при этом происходит затухание только падающей волны.

Иначе говоря, в термодинамически равновесной среде поглощение всегда оказывается большим, чем излучение.

Чтобы происходило усиление падающего излучения, в среде должна быть неравновесная заселенность уровней, т.е. $N_2 > N_1$. О такой среде говорят как о неравновесной, или как активной среде, т.е.

что среда должна находиться при отрицательной температуре, имея в виду, конечно, абсолютную температуру.

Иногда говорят также, что это среда с инверсной заселенностью уровней.

Все это одно и то же.

**общая болтавня**

Впервые явление вынужденного излучения удалось использовать практически в 1954 г. при создании мазера.

мазер это аббревиатура слов «Microwave Amplification by Stimulated Emission of Radiation»).

Первый оптический лазер был создан только в 1960 г.

Слово лазер является аббревиатурой слов «Light Amplification by Stimulated Emission of Radiation», что означает «усиление света вынужденным излучением».

Лазеры представляют собой источники очень высокой монохроматичности и направленности.

Существенным моментом всех «обычных» источников света является то, что каждый атом излучает «сам по себе», и индуцированное излучение не играет существенной роли.

А вот принципиальным моментом в работе лазеров является то, что возбужденные атомы излучают когерентно.

Эта особенность лазеров обусловлена различием физической природы спонтанного и индуцированного излучений.

При спонтанном испускании частота фотона может быть произвольной в пределах контура спектральной линии, ширина которого определяется только шириной энергетических уровней, между которыми происходит лазерный переход.

Произвольно также направление распространения фотона и фаза.

В отличие от спонтанного излучения, при вынужденном испускании фотоны полностью тождественны фотонам, воздействующим на систему возбужденных атомов.

Дело в том, что под действием пришедшей электромагнитной волны электрон, находящийся в воз- бужденном состоянии, начинает колебаться строго с частотой внешнего поля и в той же фазе.

Поэтому, если воздействующая электромагнитная волна монохроматична, то индуцированно ис- пускаемая волна будет тоже монохроматической и иметь ту же частоту и те же направления распространения и поляризации.

Особенности вынужденного испускания позволяют генерировать когерентное излучение.

Первоисточником является процесс спонтанного излучения, причем наибольшее число фотонов бу- дет испущено, конечно, на резонансной частоте, т.е.

частоте, соответствующей разности энергий между атомными уровнями.

Далее в игру вступает индуцированный процесс, вероятность ко- торого тоже максимальна на частоте перехода. Рассмотрим, каким образом можно создать активную среду.

В простейшем варианте активная среда представляет собой трехуровневую систему (см.рис. 6.4) — основной 1, возбужденный 3 и метастабильный 2.

Внешнее излучение («накачка») переводит атомы из основ- ного состояния 1 в возбужденное состояние 3.

Из этого состояния атомы частично возвращаются в основное состояние, а частично переходят в метастабильное состояние 2.

В этом состоянии они накапливаются, пока не возникнет индуцирован- ное излучение.

Чтобы свет источника «накачки» использовался эффективно, нужно, чтобы уровень, на который производится накачка (на рис. 6.4 это уровень 3) был достаточно широким, а значит, короткоживущим.

Уровень 2, наоборот, должен быть долгоживущим, чтобы спонтанное излучение с этого уровня не могло конкурировать с индуцированным.

Газоразрядная трубка «световой накачки», обвивающая рабо- чий объем лазера, создает мощный поток фотонов, возбуждаю- щих атомы рабочего вещества.

Заметим, что никакое накопление возбужденных атомов во втором возбужденном состоянии невоз- можно.

То самое излучение, которое переводит атомы из основ- ного состояния в возбужденное (состояние 3 на рис.

6.4), вызы- вает индуцированное излучение этих атомов и их ускоренный переход в основное состояние.

В процессе возбуждения часть атомов, однако, высвечивается из второго в первое возбужденное состояние и там накапливается. Выше мы показали, что в активной среде происходит усиление падающего света.

Однако чтобы превратить усилитель в генератор когерентных колебаний, необходимо создать поло- жительную обратную связь.

В лазере обратную связь обычно по- лучают, помещая активную среду между двумя параллельными зеркалами, имеющими высокий коэффициент отражения.

Два па- раллельных зеркала образуют оптическую резонансную систему, называемую открытым резонатором.

Фактически открытый ре- зонатор является разновидностью интерферометра Фабри–Перо.

Отличительной особенностью открытого резонатора является то, что для осевых пучков его добротность оказывается очень высокой; потери обусловлены только дифракцией на краях зерка- ла.

Эти потери очень малы и добротность открытого резонатора может достигать огромной величины — порядка 109.

Для всех наклонных к оси открытого резонатора световых пучков доброт- ность оказывается малой, так как через несколько отражений наклонные лучи уходят через открытую боковую поверхность. Таким образом, открытый резонатор осуществляет одновременно как частотную фильтрацию излучения (в силу высокой добротно- сти колебаний), так и пространственную фильтрацию (высокую добротность имеют строго осевые пучки).

Излучаемая волна отражается от зеркал, опять возвращается в активное вещество, вновь вызывая индуцированные переходы.

Атомы в метастабильном состоянии как бы ждут лишь того момента, когда «нужные» фотоны «приказывают» им излучать свет.

В качестве примера лазера, работающего по трехуровневой схеме, рассмотрим лазер на кристалле рубина.

Рубином называется кристалл корунда $Al_2O_3$, в котором часть ионов алюминия заменена ионами хрома ($\approx 0,05\%$). Именно хром придает корунДу красную окраску и его ионы играют основную роль в работе лазера.

КорунД - это диэлектрик с широкой запрешенной зоной между валентной зоной и зоной проводимости (см.

& 5.9).

Энергетические уровни хрома в корунде образуют две узКие пОлОсы $E_3$ и $E_4$ примесных состояний, лежащие в запрещенной зоне (рис.

6.5).

При оптической накачке атомы хрома переходят ИЗ основного состояниЯ $E_1$ в возбужденное - энергетические

полосы $E_3$ и $E_4$)длины волн этих переходов расположены в зе- леной и синей частях спектра видимого света).

Благодаря значительной ширине этих полос примерно $10 - 15\%$ лучистой энергии лампы-вспышки передается возбужденным атомам хрома

Спонтанный возврат атомов хрома на исходный уровень (он показан штриховой линией на рис.

6.5) менее вероятен, чем переход на расположенные ниже по энергии узкие метастабильные уровни $E_{2a}$ и $E_{26}$. Время жизни атома в этих состояниях составляет около $10^{-3}$ с, что очень велико по атомным масштабам.

Это способствует быстрому созданию инверсной заселенности уровней рабочего перехода $E_2 \to E_1$. Более интенсивной является линия с длиной волны $\lambda = 694,3$ нм (с уровня $E_{2a}$), именно она и усиливается при работе лазера.

Зеркалами оптического резонатора могут служить либо хорошо отполированные и покрытые отражающим слоем торцы самого кристалла либо специально изготовленные зеркала, расположенные вне кристалла.

Одно (а иногда и оба) из зеркал делают частично прозрачным.

Через него световые кванты выходят из прибора, образуя лазерный луч, как это показано на рис.

6.5. Резонатор обладает свойством не только накапливать в себе фотоны, излучаемые активной средой, но и осуществлять синхронизацию волн определенных частот.

Для синфазности излучения надо, чтобы фаза соответствующей волны частоты $\nu$ за один усилительный цикл (один проход от одного зеркала до другого и обратно) менялась на число, кратное $2\pi$. В терминах длин волн это условие соответствует тому, что на длине резонатора должно укладываться целое число полуволн.

При двойном проходе через резонатор длины $L$ это соответствует условию

$$2L = n\lambda = n\frac{c}{\nu}$$

Здесь с — скорость света в веществе, если кристалл занимает весь объем резонатора.

Из этого условия следует, что резонанс ными частотами открытого резонатора являются

$$\omega_n = \frac{\pi c}{L}n$$

В силу высокой добротности открытого резонатора, Ширина его резонансных частот $\Delta\omega_{\text{рез}}$ оКазывается мноГО меньшей естественной Ширины лазерного перехода $\Delta$ шест.

Поэтому резонатор Как бы «выбирает» из всего

испускаемого диапазона частот лишь дискретные линии, и мо- нохроматичность излучения определяется не естественной шири- ной, а добротностью резонатора с активной средой (рис.

6.6). Чем больше число активных атомов, тем, естественно, выше генерируемая лазером мощность.

Наибольшая плотность атомов

в твердом теле, однако при этом возникают трудности с отводом тепла, возможен перегрев кристалла.

Поэтому, как правило, все твердотельные лазеры работают не в непрерывном, а в импульс- ном режиме, что обусловлено быстро наступающим перегревом кристалла.

Лампы-вспышки создают длинный световой импульс, продолжающийся около миллисекунды.

Нарастание и прекраще- ние лазерного импульса происходит намного быстрее.

Поэтому за время работы лампы-вспышки успевает родиться множество быстро следующих друг за другом лазерных импульсов.

Эти импульсы могут заметно отличаться друг от друга не только по амплитуде, но и по фазе, поэтому монохроматичность излучения не очень высока, но очень велика мощность излучения, ибо в кристалле очень много активных атомов.

В рубиновом лазере длиной 20 см и диаметром 1,5 см энергия излучения составляет 1–2Дж (мощность излучения до 200 МВт).

Сфокусированный луч твердотельного лазера легко разрезает металлические листы толщиной 1–2 мм, и поэтому они достаточно широко использу- ются в технологических целях.

Под действием мощного лазер- ного импульса вначале металл плавится, а затем в его парах происходит электрический пробой — явление, также широко используемое в прикладных целях. Современные твердотельные лазеры обладают столь большой импульсной мощностью, что под действием излучения может даже происходить пробой воздуха, как это видно из приведенной фотографии—рис.

6.7.

Видно, что следующие друг за другом ко- роткие импульсы света из мощного неодимового лазера образуют в воздухе светящуюся цепочку.

Рассмотрим теперь работу газового гелий-неонового лазера. Эти лазеры работают в непрерывном режиме, их мощность до- стигает 100 мВт, они обладают высокой монохроматичностью, достаточно просты в изготовлении и нашли очень широкое распространение.

Лазер представляет собой наполненную смесью

He и Ne стеклянную трубку, в которой имеются электроды, служащие для поддержания газового разряда.

Возбуждение ато- мов происходит в результате столкновений их с электронами

газоразрядной плазмы. В гелий-неоновой смеси основным газом является ге- лий, его в 10 раз больше неона, и поэтому в основном в газовом разряде возбуж- даются атомы гелия, а ра- бочими переходами являются переходы в Ne.

Упрощенная схема возбужденных состоя- ний He и Ne показана на рис.

6.8. В чистом неоне достичь инверсной заселенности по- чти невозможно, ибо на долгоживущем уровне 5 сразу же собирается много возбуж- денных атомов, за счет этого происходит и пополнение ла

зерного нижнего уровня 1 и инверсии между уровнями 1 и 3 не возникает.

Наличие гелия резко меняет ситуацию.

Для ра- боты лазера существенны два уровня с энергиями примерно 20 эВ.

Спонтанный радиационный переход с этих уров- ней на основной происходит с малой вероятностью в силу кван- товых их характеристик (правил отбора по моменту и четности). Поэтому время жизни атома на уровнях 2?

и 3?

гелия очень велико. Однако этих правил запрета нет для электронов, и в результа- те электронных ударов на этих метастабильных уровнях накап- ливается очень много атомов гелия.

Принципиальным моментом является то, что эти уровни гелия почти совпадают с уровнями 2 и 3 неона.

Благодаря этому при столкновениях возбужденных атомов гелия с невозбужденными атомами неона интенсивно происходят безызлучательные переходы атомов гелия в основное состояние с резонансной передачей энергии атомам неона. Этот процесс возбуждения атомов неона на рис.

6.8 изображен штриховыми стрелками.

В результате такого резонансного воз- буждения концентрации атомов неона на уровнях 2 и 3 сильно возрастают, и возникает их инверсная заселенность уровней по отношению к уровням 1 и 4, что и приводит к лазерному эф- фекту.

Опустошение уровня 1 происходит за счет переходов на уровень 5, снятие возбуждения с которого обусловлено столк- новениями атомов неона со стенками трубки.

Поэтому эффективность гелий-неоновых лазеров оказывается максимальной при вполне определенном диаметре стеклянной трубки 2–5 мм.

В зависимости от длины оптического резонатора и устройства его зеркал удается выделить тот или иной лазерный переход.

Как правило, в гелий-неоновом лазере реализуется красное излучение, соответствующее переходу 3–1 с длиной волны 632,8 нм.

Газовые лазеры обладают очень вы- сокой монохроматичностью, но, как правило, если не делать специальных ограничений пучка, структура выхо- дящего из него пучка в поперечном направлении достаточно сложна, как это видно из приведенной на рис. 6.9 фотографии распределения интенсивности в пучке газового лазера.

Выходящий луч имеет в диаметре около 0,5 см;.

# Part IV
# Problems

## 12 Basic Parts of Optics

### 12.0.1 Basic problems of geometrical optics

#### O2-1.1 Focal length of jar with mercury

A jar with mercury rotates uniformly around a vertical axis with an angular velocity of $\omega = 1\text{c}^{-1}$. The surface of mercury takes a concave shape and is used as a mirror. Determine the focal length of this mirror.

#### O2-1.4 Biconvex thin lens

Find the focal length $f$ of a biconvex thin lens bounded by spherical surfaces with radii $R_1 = 25\text{mm}$ and $R_2 = 40$ mm; The refractive index of the lens glass is $n = 1.5$.

#### O2-1.5 Putting a setup in water

С помощью тонкой собирающей стеклянной линзы с показателем преломления $n = 3/2$ получено действительное изображение предмета на расстоянии 10 cm от линзы. После того как предмет и линзу погрузили в воду, не изменяя расстояния между ними, изображение получилось на расстоянии 60 см от линзы. Найти фокусное расстояние $f$ линзы, если показатель преломления воды $n' = 4/3$.

#### O2-1.7

Плоская стеклянная пластинка толщиной 3 мм рассматривается в микроскоп. Сначала микроскоп устанавливают для наблюдения верхней поверхности пластинки, а затем смещают тубус микроскопа вниз до тех пор, пока не будет отчетливо видна нижняя

#### O2-1.10

Вычислить угол наименьшего отклонения $\partial$ для призмы с очень малым преломляющим углом $A$ с учетом членов второго порядка малости (относительно Å).

#### O2-1.11

Чему равен угол наименьшего отклонения $\delta$ для линии $D$ натрия в призме с преломляющим углом 60°? Для линии $D$ показатель преломления стекла призмы $n = 1,62$.

#### O2-1.12

На плоскопараллельную стеклянную пластинку под углом $\varphi$ падает узкий пучок света шириной $a$ (рис. 265), содержащий две спектральные компоненты. Показатели преломления стекла для этих длин волн различны: $n_1$ и $n_2$. Определить минимальную толщину $h_{\min}$ пластинки, при которой свет, пройдя через пластинку, будет распространяться в виде двух отдельных пучков, каждый из которых содержит только одну спектральную компоненту.

#### O2-1.13

Для обращения изображения часто используют так называемую призму Дове (рис. 266), представляющую собой усеченную прямоугольную равнобедренную призму. Определить длину $l$ основания призмы, если ее высота $h = 2,11$ cm, а показатель преломления стекла $n = 1,41$. Призма должна обращать пучок света максимального сечения.

#### O2-1.14

На стеклянный клин перпендикулярно его грани падает тонкий луч света (рис. 267). Показатель преломления стекла
$n = 1,41$, угол $\alpha = 10°$. Сколько светлых пятен будет видно на экране, поставленном за клином?

### O2-1.15

Перед торцом стеклянного цилиндрического световода, показатель преломления которого равен $n$, на его оси расположен точечный источник света. Найти угловую апертуру a пучка света, проходящего через световод.

### O2-1.16

Цилиндрический стакан с жидкостью поставлен на монету, рассматриваемую сквозь боковую стенку стакана. Указать наименьшую возможную величину показателя преломления $n$ жидкости при котором монета не видна.

### O2-1.17

С каким углом a нужно взять трапецеидальный сосуд с водой $ABCD$ (рис. 268), чтобы сквозь сго боковую стенку не было видно предмета, положенного под сосуд? Показатель преломления воды $n = 1,33$. Дно сосуда имеет форму прямоугольника.

### O2-1.18

Человек, стоящий на берегу пруда, смотрит на камень, находящийся на дне. Глубина пруда $h = 1$ м. На каком расстоянии $h'$ от поверхности воды получится изображение камня, если луч зрения составляет с нормалью к поверхности воды угол $\varphi = 60°$? Показатель преломления воды $n = 1,33$.

### O2-1.22

Матовое стекло фотографического аппарата установлено так, что резким выходит изображение предмета, находящегося на расстоянии 5 м. До какого диаметра $D$ нужно задиафрагмировать объектив с фокусным расстоянием 20 cm, чтобы не было заметной нерезкости в изображении предметов, находящихся на 0,5 м ближе снимаемого? (Нерезкость считать незаметной, если размытость деталей не превышает 0,1 мм.)

### O2-1.23

При фотографировании на пленке из-за конечной разрешающей способности получаются резко изображенными не только те предметы, на которые сфокусирован объектив фотоаппарата, но также и предметы, находящиеся несколько ближе и несколько дальше этого расстояния. Оказалось, что при наведении объектива фотоаппарата на предмет, находящийся на расстоянии $L_0 = 10$ м, ближняя граница глубины резкости расположена на расстоянии $L_1 = 7,8$ м. Определить дальнюю границу $L_2$.

### O2-1.24

Как сместится фокус фотоаппарата, если внутрь аппарата на пути лучей (перпендикулярно к оптической оси) поместить плоскопараллельную стеклянную пластинку толщиной $d = 6$ мм с показателем преломления $n = 1,5$? (Объектив сильно задиафрагмирован.)

### O2-1.25

Доказать, что если линза находится перед глазом и движется в сторону, то наблюдателю кажется; что предмет, рассматриваемый через линзу, движется в ту же сторону, что и линза, если линза рассеивающая, и в противоположную, если линза собирающая.

При меч ание. Если собирающая линза используется как лупа (предмет помещается между фокусом и линзой), то получается прямое изображение. Если же, отодвинув собирающую линзу достаточно далеко от глаза, рассматривать через нее удаленные предметы, то получаются их обратные изображения. В этом случае при смещении линзы в сторону изображение смещается в ту же сторону.

### O2-1.26

Показать, что наименьшее расстояние между двумя оптически сопряженными относительно собирающей линзы точками равно $4f$, где $f$ - фокусное расстояние линзы.

### O2-1.27

Собирающая линза дает изображение некоторого объекта на экране. Высота изображения равна $h_1$. Оставляя неподвижным экран и объект, начинают двигать линзу к экрану и находят, что при втором четком изображении объекта высота изображения равна $h_2$. Найти действительную высоту предмета $h$.

### O2-1.28

Расстояние от лампочки до экрана $L = 50$ см. Линза, помещенная между ними, дает четкое изображение лампы на экране при двух положениях, расстояние между которыми $l = 10$ cm. Найти фокусное расстояние $f$ линзы.

### O2-1.29

У двояковыпуклой тонкой линзы серебрится одна из поверхностей. Найти фокусное расстояние $f$ полученного таким 123 образом зеркала. Радиус кривизны чистой поверхности равен $R_1$, радиус кривизнь посеребренной поверхности - $R_2$, показатель преломления материала линзы равсн $n$.

### O2-1.30

Две одинаковые плосковыпуклые тонкие линзы с показателем преломления $n$ посеребрены: одна с плоской стороны, другая с выпуклой. Найти отношснис фокусных расстояний $f_1$ и $f_2$ по.ученных сложных зеркал, если свет в обоих случаях падает с непосеребрснной стороны.

### O2-1.31

Фотографическим аппаратом, объектив которого имеет фокусное расстояние, меняющееся от 12 cm до 20 cm, требуется сфотографировать предмет, находящийся на расстоянии 15 см от объектива. Какую жинзу нужно добавить к объективу, чтобы изображение вышло резким при максимально возможном фокусном расстоянии?

### O2-1.32

Две тонкие линзы с фокусными расстояниями $f_1$ и $f_2$ находятся на расстоянии $l$ друг от друга, образуя центрированную систему. Найти фокусное расстояние $f$ этой системы, а также положения се главных плоскостей.

### O2-1.35

Фокусное расстоянис объектива зрительной трубы равно $f_1 = 60$ cm, а окуляра $-f_2 = 4$ cm. Показатель преломления стекла объектива и окуляра $n = 3/2$. Труба погружастся в воду, заполняющую ее внутреннюю часть. Каким объективом из стекла того же сорта следует заменить объсктив трубы, чтобы в нее можно было рассматривать удаленные предметы в воде? Чему будет при этом равно увеличсние трубы, если показатель преломления воды $n' = 4/3$?

### O2-1.36

Человск с нормальным зрением рассматривает удаленный предмет с помощью зрительной трубы Галилея. В качестве объектива и окуляра используются линзы с фокусными расстояниями $f_1 = 40$ cm и $f_2 = -2$ cm. При каких расстояниях $L$ между объективом и окуляром наблюдатель увидит четкое изображение предмета, если глаз может аккомодироваться от 10 cm до бесконечности?

### O2-1.37

Галилеева труба 9-кратного увеличения имеет длину 40 cm. После того как объектив и окуляр трубы заменили собирающими линзами, труба стала давать то же увеличение. Определить фокусные расстояния $f_1'$ и $f_2'$ этих линз, а также фокусные расстояния $f_1$ и $f_2$ объектива и окуляра галилеевой трубы.

## 12.0.2 Advanced problems on geometrical optics 1

### ТУЙМ-107. Зеркала (1996)

Два плоских квадратных зеркала имеют одну общую сторону и образуют двугранный прямой угол. Тонкая собирающая линза с фокусным расстоянием $f$ и диаметром $2f$ касается обоих зеркал, а её главная оптическая ось проходит через линию пересечения зеркал перпендикулярно ей. На главной оптической оси на расстоянии $1,5f$ от центра линзы находится точечный источник света $S$. Найдите положение его изображения $S'$, созданного лучами, вышедшими из линзы в сторону источника.

## Solution

Суть решения в том, что (???? мб покрутить геометрию прокатит???)

Из геометрических соображений следует, что предмет $S$ находится перед передним фокусом $F$ линзы, а задний её фокус совпадает с точкой касания зеркал $C$ (рис. 199). Интересующие нас световые лучи сначала проходят через линзу, затем отражаются от пары зеркал, после чего снова проходят через линзу в обратном направлении. Для нахождения изображения $S'$ вычислим положения промежуточных изображений, даваемых по отдельности линзой и зеркалами по мере прохождения лучей через систему.

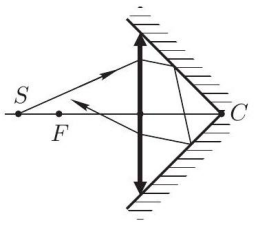



Рис. 199

Для иллюстрации рассуждений на рис. 200 показаны продолжения лучей, создающих промежуточные изображения $S_1, S_2$ и $S_3$, а на рис. 201 - вспомогательные построения для поиска изображений. Точки $O, A$ и $B$ - это соответственно оптический центр линзы и её точки касания с каждым из зеркал.

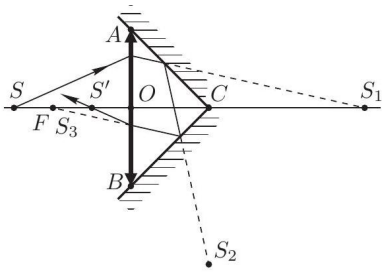



Рис. 200

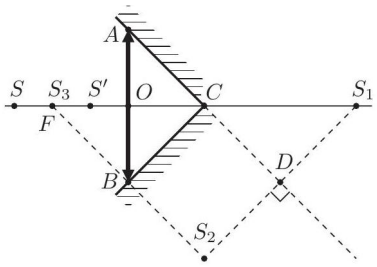



Рис. 201

• Пусть $a$ и $b$ - расстояния от линзы до предмета и до изображения соответственно, тогда по формуле тонкой линзы

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{f}$$

откуда

$$b=\frac{af}{a-f}$$

причём значение любого из расстояний $(f, a, b)$ в этой формуле может быть отрицательным: $f < 0$ означает рассеивающую линзу, $a < 0$ - мнимый предмет, $b < 0$ - мнимое изображение. Точка $S_1$ - изображение источника $S$ в линзе без учёта зеркал, поэтому из формулы (238) при $a = SO = 1,5f$ получаем $OS_1 = b = 3f$. 2. Зеркало даёт изображение, симметричное предмету относительно плоскости зеркала, причём это свойство применимо как к действительным, так и к мнимым предметам. Точка $S_2$ - изображение предмета $S_1$ в зеркале $AC$ без учёта второго зеркала, значит, построив к плоскости зеркала $AC$ перпендикуляр $S_1D$ и продолжив его на такое же расстояние, находим $S_2$.

• Рассуждая аналогично, строим точку $S_3$ - изображение предмета $S_2$ в зеркале $BC$ без учёта линзы. Последовательное применение двух отражений относительно взаимно перпендикулярных прямых эквивалентно одному отражению относительно их точки пересечения (вспомните свойство центральной симметрии графика нечётной функции), поэтому $S_3C = CS_1$, откуда

$$S_3O = S_3C - OC = CS_1 - OC = OS_1 - 2OC = 3f - 2f = f$$

то есть изображение $S_3$ совпадает с передним фокусом $F$ линзы.

• Искомая точка $S'$ является действительным изображением мнимого предмета $S_3$, поэтому из формулы (238) при $a = -S_3O = -f$ получаем $S'O = b = f/2$.

В самом начале решения был выбран луч, который сперва отражается от зеркала $AC$, а потом - от $BC$, однако существуют лучи, которые испытывают отражения в обратном порядке. В данном случае из симметрии ясно, что они пересекутся в той же точке, однако при другом расположении зеркал могли бы возникнуть два конечных изображения. Таким образом, окончательный ответ: изображение будет на расстоянии $f/2$ перед линзой.

## ТУЙМ-108. Котлован (1999)

Человек, стоя на краю высокого обрыва, смотрит на ровное плоское дно котлована шириной $L$, заполненного водой глубиной $h$ (рис. 62). Высота обрыва $H$. Размеры котлована удовлетворяют неравенствам $L \gg H \gg h$. Показатель преломления воды равен $n$. Как зависит от расстояния до обрыва видимая глубина котлована?

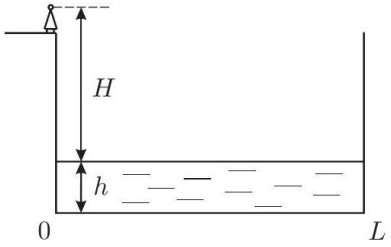


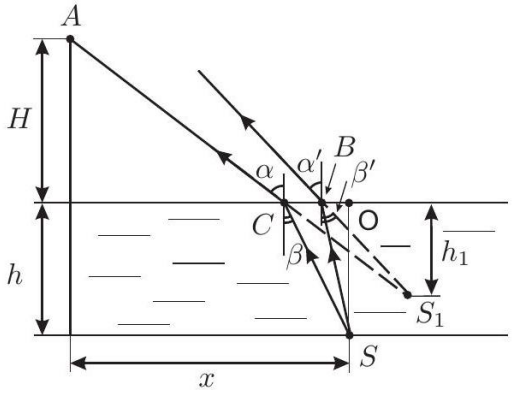

Рис. 62

### Solution

Суть решения в том, что (??? просто нужно расписать геометрию, все просто (???))

### Solution ТУЙМ-108. Котлован (1999)

Человек, находящийся в точке $A$ (рис. ниже), оценивает глубину водоёма по двум близким лучам, идущим из точек $B$ и $C$.




Ему кажется, что лучи исходят не из точки $S$, а из точки $S_1$, и поэтому видимая глубина водоёма $h_1$. Изменения направлений лучей на поверхности воды подчиняются закону Снеллиуса:

$$\frac{\sin\alpha}{\sin\beta} = n, \quad \frac{\sin\alpha'}{\sin\beta'} = n$$

где $n$ - показатель преломления воды, а углы $\alpha, \beta, \alpha'$ и $\beta'$ показаны на рис. 202. Пусть $\Delta$ - длина отрезка $BC$. Тогда из прямоугольных треугольников, показанных на рис. 202, следует, что

$$\Delta = h\,(\operatorname{tg}\beta - \operatorname{tg}\beta') = h_1\,(\operatorname{tg}\alpha - \operatorname{tg}\alpha')$$

Углы $\alpha$ и $\alpha', \beta$ и $\beta'$ мало отличаются, поэтому, если $\alpha - \alpha' = \Delta\alpha$ и $\beta - \beta' = \Delta\beta$, то

$$\operatorname{tg}\beta - \operatorname{tg}\beta' \approx \frac{\Delta\beta}{\cos^2\beta}, \quad \operatorname{tg}\alpha - \operatorname{tg}\alpha' \approx \frac{\Delta\alpha}{\cos^2\alpha}.$$

Подставляя полученные соотношения в формулу (240), находим

$$h_1 = h\frac{\cos^2\alpha}{\cos^2\beta}\cdot\frac{\Delta\beta}{\Delta\alpha}.$$

Запишем закон Снеллиуса (239) для маленьких приращений углов:

$$\cos\alpha\Delta\alpha = n\cos\beta\Delta\beta$$

откуда

$$h_1 = \frac{h}{n}\cdot\frac{\cos^3\alpha}{\cos^3\beta} = \frac{h}{n}\cdot\frac{\cos^3\alpha}{\left(1-\frac{\sin^2\alpha}{n^2}\right)^{3/2}}.$$

Для угла $\alpha$ из рис. 202 находим

$$\sin\alpha = \frac{x-OB}{\sqrt{(x-OB)^2+H^2}} \approx \frac{x}{\sqrt{x^2+H^2}}$$

так как $OB \ll x$ при малых $h$. Осталось подставить полученное выражение для $\sin\alpha$ в формулу (241) и произвести алгебраические преобразования:

$$h_1 = \frac{n^2hH^3}{\left((n^2-1)\,x^2+n^2H^2\right)^{3/2}}.$$

График этой зависимости, построенный с помощью MathCAD, представлен на рис. 203. Видно, что вдали от обрыва водоём кажется более мелким.

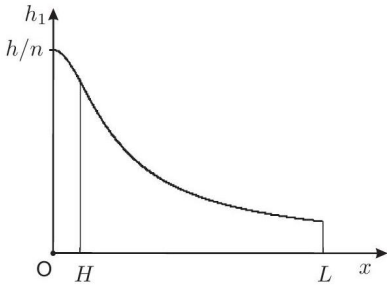



Рис. 203

## ТУЙМ-109. Световой зайчик (2000)

Человек, стоящий на расстоянии $h$ от длинной ровной стены, освещает её лучом фонарика, вращая фонарик в горизонтальной плоскости слева направо с постоянной угловой скоростью $\omega$. Учитывая конечность скорости распространения света $c$, найдите, как с точки зрения человека будут изменяться со временем положение светового зайчика на стене и скорость его движения.

### Solution

Суть решения в том, что
Если бы скорость света была бесконечно большой, то человек видел бы световое пятно в том месте, где луч фонарика пересекает стену. Из-за конечности скорости распространения света человек увидит в этом месте зайчик спустя некоторое время. В этот момент луч фонарика будет уже направлен в иную точку.
Найдём скорость зайчика с точки зрения наблюдателя:

$$v = \frac{dx}{dt_1},$$

где $v$ - проекция на ось $x$ скорости зайчика (рис. 204), который наблюдатель видит в момент времени $t_1$. Луч фонарика был направлен в эту точку в момент времени $t$. Время $t_1$ больше $t$ на величину, необходимую свету для прохождения от $A$ к $B$ и обратно (рис. 204):

$$t_1 = t + \frac{2h}{c \cdot \cos \omega t}.$$

Поскольку переменные $t_1$ и $t$ связаны соотношением (243), целесообразно представить уравнение (242) в виде

$$v = \frac{dx}{dt}\frac{dt}{dt_1} = \frac{dx}{dt}\left(\frac{dt_1}{dt}\right)^{-1}$$

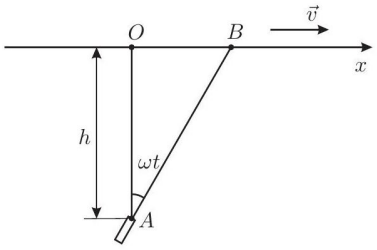



Рис. 204
Из рис. 204 видно, что

$$x = h \operatorname{tg} \omega t.$$

Продифференцировав функции (243) и (245), найдём производные, входящие в уравнение (244), и после их подстановки получим

$$v = \frac{c\omega h}{c \cdot \cos^2 \omega t + 2\omega h \sin \omega t}.$$

Формулы (245), (246) и (243) содержат ответы на поставленные в задаче вопросы. Проанализируем полученные соотношения.
Изобразим приблизительный график зависимости (243) (рис. 205).

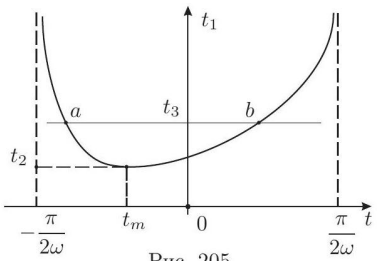



Рис. 205
Величина $t_1$ имеет минимальное значение $t_2$ в точке $t_m$, для нахождения которой следует приравнять к нулю производную функции $t_1(t)$:

$$\sin \omega t_m = \frac{\omega h}{c} - \sqrt{1 + \left(\frac{\omega h}{c}\right)^2}.$$

По формуле (245) можно найти координату зайчика $x_m$ в момент времени $t_2$:

$$x_m = -\frac{h}{\sqrt{2}}\sqrt{\sqrt{\left(\frac{c}{\omega h}\right)^2 + 1} - 1}.$$

При подстановке выражения (247) в формулу (246) обнаруживаем, что функция $v(t)$ испытывает разрыв в точке $t_m$, меняя знак. Поведение функции $v(t)$ качественно представлено графиком, изображённым на рис. 206. Сплошными линиями показан график при $\omega h < c/2$, пунктирными - при $\omega h > c/2$. Сопоставляя величины $v$ и $t_1$ при одинаковых значениях параметра $t$, можно нарисовать и приближённый график зависимости наблюдаемой скорости зайчика $v\,(t_1)$ (рис. 207).

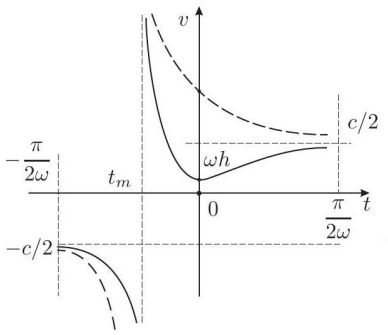



Рис. 206

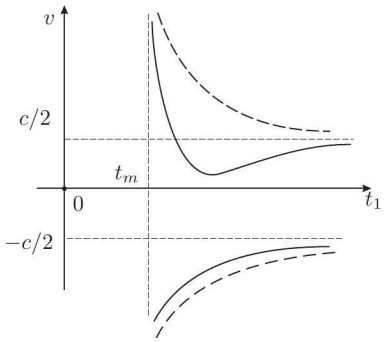



Рис. 207

О каком движении светового зайчика говорят эти графики? Раньше всего наблюдатель увидит зайчик, находящийся в точке $x_m$ в момент времени $t_2$. В последующие моменты времени $t_1 = t_3 > t_2$ наблюдатель увидит два зайчика, координаты которых соответствуют точкам $a$ и $b$ на рис. 205. Скорости этих зайчиков направлены в противоположные стороны. При $t_1 \to \infty$ проекция скорости одного стремится к $c/2$, а другого - к $-c/2$. После следующего прохождения вращающегося луча фонарика через точку $x_m$ появится ещё одна пара разбегающихся зайчиков и так далее.

Не является ли обнаруженное здесь парадоксальное свойство световых зайчиков результатом ошибки данной теоретической модели? Одним из тестов на ошибочность новых физических теорий служит принцип соответствия. Согласно этому принципу новая теория в предельных частных случаях даёт те же результаты, что и верная старая теория. Релятивистская физика даёт немало парадоксальных результатов, противоречащих «здравому смыслу». Однако она успешно выдерживает испытание принципом соответствия. Испытаем этим принципом результаты данной задачи. Посмотрим, что дают полученные формулы при $c \to \infty$. Из уравнения (242) следует, что $t_1 = t$, то есть наблюдатель видит зайчик в точке $x$ в тот же момент времени, в который через эту точку проходит луч фонарика. Из формулы (246) следует, что $v = \omega h/\cos^2 \omega t$. Это естественно согласуется с формулой (245), и в этом случае не возникает никаких пар зайчиков. Таким образом, полученные в задаче результаты согласуются с принципом соответствия.

## ТУЙМ-110. Сферическое зеркало (2001)

Солнечные лучи падают на вогнутое сферическое зеркало диаметром $D$ параллельно его оси симметрии. Радиус кривизны поверхности зеркала $R \gg D$. В фокальной плоскости зеркала перпендикулярно его оси симметрии помещён непрозрачный экран радиусом $r$. Как зависит средняя освещённость светового пятна на экране от радиуса экрана?

### Solution

Суть решения в том, что

Задача 110. Сферическое зеркало (2001 год)

На рис. 208 показан ход луча, падающего на зеркало на расстоянии $x$ от его оптической оси. Точка $O$ - центр кривизны зеркала. Фокус $F$ удалён от центра на расстояние $OF = R/2$.

Отражённый луч пересекает главную оптическую ось в точке $F_x$, для которой, как видно из рисунка, $OF_x = \frac{R}{2\cos\alpha}$. Угол $\alpha$ зависит от $x$, так как $\sin\alpha = x/R$. Найдём, на каком расстоянии $y$ от главной оптической оси луч пересекает фокальную плоскость. Это нетрудно сделать с помощью рис. 208:

$$y = \left(\frac{R}{2\cos\alpha} - \frac{R}{2}\right) \mathrm{tg}(2\alpha)$$

В соответствии с условием задачи углы $\alpha$ малы, поэтому можно использовать приближения

$$\sin\alpha \approx \operatorname{tg}\alpha \approx \alpha \quad \text{и} \quad \cos\alpha \approx 1 - \frac{\alpha^2}{2}$$

которые существенно упрощают формулу (249):

$$y \approx \frac{R}{2}\alpha^3 = \frac{x^3}{2R^2}$$

Чтобы найти среднюю освещённость экрана, следует световой поток, падающий на него, поделить на площадь пятна на экране. Какие же лучи попадают на экран? На зеркало падают лучи, проходящие мимо экрана. Для них

$$r < x < \frac{D}{2}$$

и, согласно формуле (250),

$$\frac{r^3}{2R^2} < y < \frac{D^3}{16R^2}$$

Все эти лучи попадают на экран, если он достаточно велик, то есть при

$$r > \frac{D^3}{16R^2}$$

В этом случае площадь освещённого пятна на экране составляет

$$S_1 = \pi\left(\left(\frac{D^3}{16R^2}\right)^2 - \left(\frac{r^3}{2R^2}\right)^2\right).$$

Световой поток Φ, падающий на это пятно, найдём по формуле $\Phi = NS_2$, где $N$ - мощность солнечного излучения, приходящегося на единичную площадку, перпендикулярную лучам, а $S_2$ - площадь так же ориентированной площадки, которую пересекают лучи, попадающие на зеркало. Эта площадь равна

$$S_2 = \pi\left(\left(\frac{D}{2}\right)^2 - r^2\right).$$

Таким образом, получаем формулу для средней освещённости $E$ светового пятна при условии $r > D^3/\left(16R^2\right)$:

$$E = \frac{N\left(\left(\frac{D}{2}\right)^2 - r^2\right)}{\left(\frac{D^3}{16R^2}\right)^2 - \left(\frac{r^3}{2R^2}\right)^2} = \frac{4R^4N}{\left(\frac{D}{2}\right)^4 + r^2\left(\frac{D}{2}\right)^2 + r^4}.$$

Если экран небольшой, то есть $r \leqslant D^3/\left(16R^2\right)$, то на него попадают не все лучи, отражённые зеркалом. Площадь освещённого пятна $S_1$ в этом случае равна

$$S_1 = \pi\left(r^2 - \left(\frac{r^3}{2R^2}\right)^2\right).$$

Площадь $S_2$ в этом случае принимает значение

$$S_2 = \pi\left(x_1^2 - r^2\right)$$

где $x_1$ находится из формулы (250) при подстановке в неё $y = r : x_1^3 = 2R^2r$. Зная $S_1$ и $S_2$, находим освещённость $E$ при условии $r \leqslant D^3/\left(16R^2\right)$:

$$E = N\frac{x_1^2 - r^2}{r^2 - \left(\frac{r^3}{2R^2}\right)^2} = N\frac{\left(2R^2r\right)^{2/3} - r^2}{r^2 - \left(\frac{r^3}{2R^2}\right)^2}.$$

При $r \ll R$ из полученного выражения находим

$$E \approx 2^{2/3} \cdot N\left(\frac{R}{r}\right)^{4/3}$$

то есть при $r \to 0$ освещённость неограниченно возрастает. Такое заключение получилось в рамках геометрической оптики, справедливой, когда длина волны света $\lambda \ll D$. Дифракция света приводит к тому, что в фокальной плоскости получается не точка, а пятнышко, что ограничивает величину $E$. При больших $r$, изменяющихся в пределах от $D^3/\left(16R^2\right)$ до $D/2$, то есть в $8R^2/D^2$ раз, освещённость

в соответствии с формулой (254) изменяется всего в 3 раза, что говорит о слабой зависимости $E$ от $r$. Качественно характер зависимости $E(r)$ отражает рис. 209, на котором

$$E_m = \frac{64}{3} N \left(\frac{R}{D}\right)^4.$$

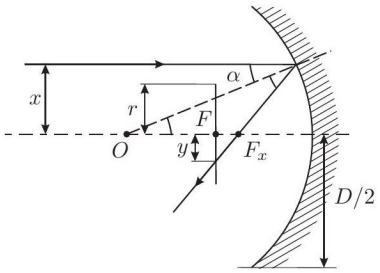



Рис. 208

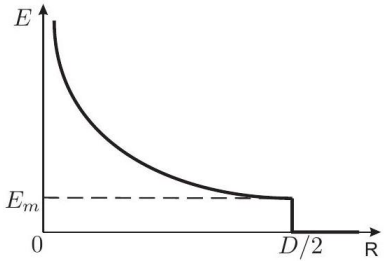


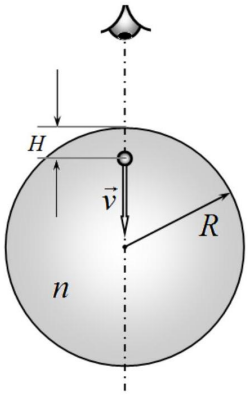

Рис. 209

## BelPho.2016.resp-11-2. Аквариум

Имеется тонкостенный прозрачный сосуд шарообразной формы радиусом $R$, заполненный прозрачной жидкостью, показатель преломления которой $n$. В сосуде равномерно опускается сверху вниз маленький шарик со скоростью $\vec{v}$ относительно сосуда. В следствие преломления света, кажущаяся глубина шарика $h$ (положение его изображения) будет отличаться от его истинной глубины $H$.

### Solution pr. Часть 1. Вид сверху.

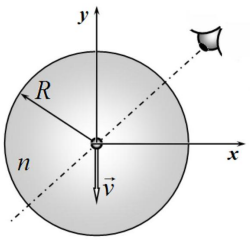
За движением шарика наблюдают сверху из точки расположенной на большом расстоянии от сосуда.




1.1 При каком положении шарика его кажущаяся глубина $h$ будет совпадать с истинной глубиной $H$, не зависимо от показателя преломления жидкости. Ответ обоснуйте.
1.2 Найдите зависимость кажущейся глубины $h$ шарика от его действительной глубины $H$ при $H < R$;
1.3 Найдите зависимость кажущейся глубины $h$ шарика от его действительной глубины $H$ при $H \geq R$;
1.4. Постройте график зависисимости, кажущейся глубины щарика $h$ от его истинной глубины $H$. Данный график постройте в относительных координатах $y = \frac{h}{R}$ от $x = \frac{H}{R}$, при двух значениях $n_1 = 1,5$ и $n_2 = 2,5$.
1.5. Найдите зависимость кажущейся скорости движения шарика от его действительной глубины $H$
1.6. Постройте графики полученных в п. 1.5 зависимостей при двух значениях $n_1 = 1,5$ и $n_2 = 2,5$. в безразмерных координатах $\chi = \frac{u}{v}$ от $x = \frac{H}{R}$;

## Solution pr. Часть 2. Вид сбоку.

Глаз наблюдателя находится сбоку на прямой, проходящей через центр шара и направленной под углом 45° к вертикали. Показатель преломления жидкости равен $n_1 = 1,5$.
2.1 Постройте вектор видимой скорости движения шарика $\vec{u}$ при наблюдении сбоку, в момент времени, когда шарик проходит центр шара. Найдите координаты этого вектора в системе отсчета, показанной на рисунке.




## Solution sol.Часть 1.

1.1 Если шарик находится в центре сосуда, то все лучи выходящие от него будут пересекать стенки сферического аквариума по нормали, следовательно, без изменения направления распространения. Поэтому при наблюдении с любой стороны при любом показателе преломления изображение шарика также будет находиться в центре шара.
1.2 Ход лучей показан на рисунке 1. Видно, что изображение А предмета **В** находится на глубине $h$ , меньшей, чем действительная глубина $H$.
Применим теорему синусов к $\triangle$OBC:

$$\frac{\text{OB}}{\sin\alpha} = \frac{\text{OC}}{\sin\left(180^\circ - \alpha - \gamma\right)} \Rightarrow$$
$$\frac{R-H}{\sin\alpha} = \frac{R}{\sin(\alpha+\gamma)}$$

Поскольку углы $\alpha, \beta$ и $\gamma$ малые, то

$$\sin\alpha \approx \alpha, \quad \sin(\alpha+\gamma) \approx \alpha+\gamma \Rightarrow$$
$$\frac{R-H}{\alpha} = \frac{R}{\alpha+\gamma} \Rightarrow \gamma = \frac{H\alpha}{R-H}.$$

Совершенно аналогично рассматривая $\triangle$ OAC, получим: $\gamma = \frac{h\beta}{R-h}$. Приравняем правые части

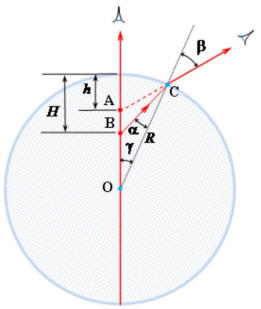



Рис. 1
выражений, полученных для $\gamma$:

$$\frac{H\alpha}{R-H} = \frac{h\beta}{R-h} \Rightarrow \frac{\alpha}{\beta} = \frac{h}{H}\cdot\frac{R-H}{R-h}.$$

Используя закон преломления света, получим: $\frac{1}{n} = \frac{\sin\alpha}{\sin\beta} \approx \frac{\alpha}{\beta} \Rightarrow \frac{1}{n} = \frac{h}{H}\cdot\frac{R-H}{R-h} \Rightarrow$

$$h = \frac{HR}{nR-(n-1)H}.$$

Введем безразмерный параметр $x = \frac{H}{R}$. В результате: $h = \frac{H}{n-(n-1)\xi}$ или

$$y = \frac{h}{R} = \frac{x}{n-(n-1)x}.$$

1.3 Ход лучей в этом случае показан на рисунке 2 . Видно, что изображение предмета находится на глубине, большей, чем действительная глубина. Ход решения в этом случае аналогичен.
Применим теорему синусов к $\triangle$BOC:

$$\frac{\mathrm{OB}}{\sin\alpha} = \frac{\mathrm{OC}}{\sin\gamma} \Rightarrow$$
$$\frac{H-R}{\sin\alpha} = \frac{R}{\sin\gamma}$$

Поскольку углы $\alpha, \beta$ и $\gamma$ малые, то

$$\sin\alpha \approx \alpha, \quad \sin\gamma \approx \gamma \Rightarrow$$

$$\frac{H-R}{\alpha} = \frac{R}{\gamma} \Rightarrow \gamma = \frac{R\alpha}{H-R}.$$

Применяя теорему синусов к $\Delta$AOC, получим:

$$\frac{\mathrm{AO}}{\sin\beta} = \frac{\mathrm{OC}}{\sin\delta} \Rightarrow \frac{h-R}{\sin\beta} = \frac{R}{\sin\delta}$$

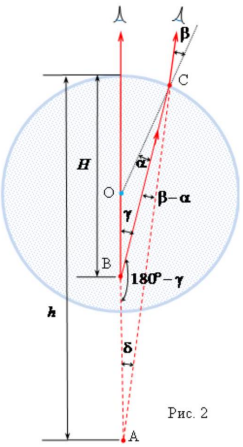


Рис. 2

На рисунке 2 видно, что в $\triangle$ AOC $\delta = 180° - (180° - \gamma) - (\beta - \alpha) = \gamma - \beta + \alpha$.
В результате: $\frac{h-R}{\sin\beta} = \frac{R}{\sin(\gamma-\beta+\alpha)} \Rightarrow \gamma - \beta + \alpha = \frac{R\beta}{h-R}$.
Подставляя выражение для $\gamma$ и используя закон преломления $\frac{1}{n} = \frac{\sin\alpha}{\sin\beta} \approx \frac{\alpha}{\beta}$, получим:
$\frac{R}{H-R}\alpha - \alpha n + \alpha = \frac{R}{h-R}\alpha n \Rightarrow h = \frac{HR}{nR-(n-1)H}$
или

$$y = \frac{h}{R} = \frac{x}{n-(n-1)x},$$

что совпадает с формулой полученной для случая, когда тело было выше центра шара.
1.4 Графики полученных зависимостей показаны на рисунке 3 . При $n = 2,5$ эта график этой зависимости терпит разрыв и устремляется к $y \to \pm\infty$. Физический смысл этого явления можно понять: шарик переходи через фокус сферической поверхности.

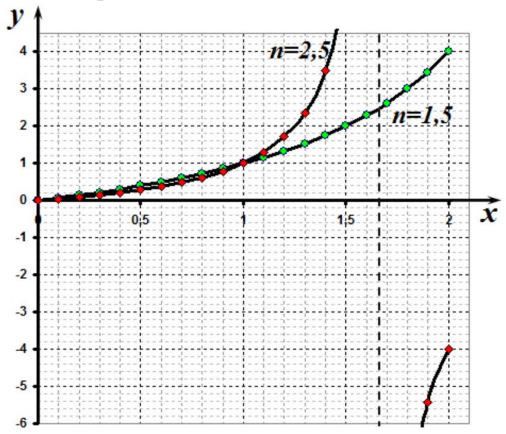


Рис. 3

1.5. Для вычисления видимой скорости движения шарика следует вычислить производную от видимой координаты $u = \frac{dh}{dt}$. Т.к. $h = \frac{HR}{nR-(n-1)H}$, то

$$u = \frac{(nR-(n-1)H)R + HR(n-1)}{(nR-(n-1)H)^2} v \Rightarrow u = \frac{nR^2}{(nR-(n-1)H)^2} \Rightarrow$$

В безразмерных координатах эта зависимость имеет вид

$$\chi = \frac{n}{(n-(n-1)x)^2}.$$

Графики этих функций показаны на рис. 4
1.6 График этих зависимостей показан на рисунке 4. Эта функция также терпит разрыв в указанной точке. Но скорость движения изображения шарика все время остается положительной.

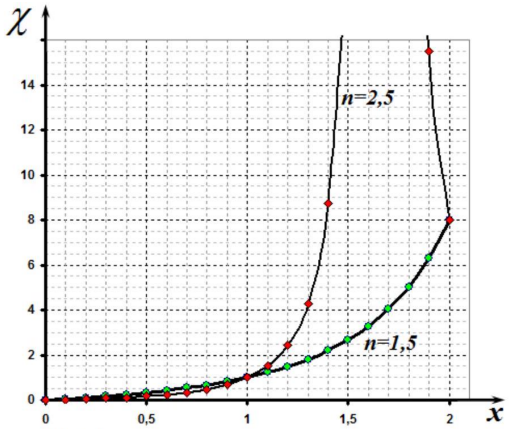



Рис. 4

## Solution sol.Часть 2.

2.1 В общем случае положение изображения зависит от точки наблюдения. Поэтому в данной части задачи нельзя пользоваться результатами, полученными в первой части задачи. Для построения изображения надо построить, по крайней мере, два луча и найти их точку пересечения.

Направим ось $X$ вдоль линии наблюдения. В этой системе шарик движется под углом 45° к осям координат. Рассмотрим положение шарика в момент времени непосредственно предшествующий его прохождению через центр, когда его координаты равны $x = a$, $y = a$, причем величина $a$ является малой. В качестве первого луча возьмем луч, попадающий на стенку в точке $x = 0$. Построение этого луча понятно из

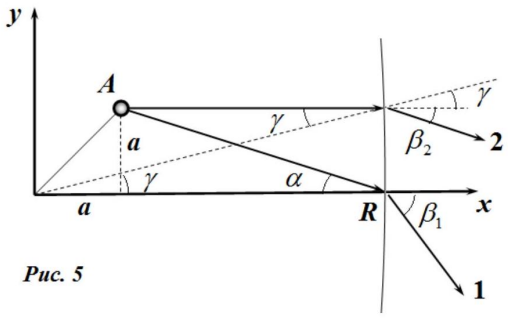



Рис. 5

рисунка 5. Полагая все углы малыми находим угол, под которым выходит преломленный луч.

$$\beta_1 = n\alpha = n\frac{a}{R-a}. \tag{1}$$

Уравнения этого луча имеет вид

$$y = -\frac{na}{R-a}(x-R) \tag{2}$$

Второй луч направим параллельно оси $X$. Для этого луча также не сложно найти уравнение:
Из закона преломления и геометрических построений следует, что угол наклона преломленного луча равен

$$\beta_2 + \gamma = n\gamma \Rightarrow \beta_2 = (n-1)\gamma \tag{3}$$

где $\gamma = \frac{a}{R}$. Тогда уравнение этого луча имеет вид

$$y = a - \frac{(n-1)a}{R}(x-R). \tag{4}$$

Изображением шарика является точка пересечения этих лучей. Координаты этой точки являются результатом совместного решения уравнений (2) и (4). Аккуратное решение этой системы уравнений приводит к следующему результату.

$$x = R - R\frac{R-a}{R+(n-1)a}$$
$$y = \frac{nRa}{R+(n-1)a} \tag{5}$$

Далее необходимо упростить данные выражения, полагая $a \ll R$. В формуле для координаты $y$ можно пренебречь $a$ в знаменателе:

$$y \approx na \tag{6}$$

Для координаты $x$ следует проделать следующие преобразования

$$x = R - R\frac{R-a}{R+(n-1)a} = R\left(1 - \frac{1-\frac{a}{R}}{1+(n-1)\frac{a}{R}}\right) \approx R\left(1 - \left(1 - \frac{a}{R} - n - 1\right)\frac{a}{R}\right)\right) \approx na \tag{7}$$

Таким образом, мы получили, что координаты изображения просто в $n$ раз больше действительных координат шарика. Следовательно, все кинематические характеристики движения изображения шарика (в т.ч. и вектор скорости) так же будут в $n = 1,5$ больше соответствующих характеристик движения самого шарика. Таким образом. Вектор скорости видимого движения совпадает по направлению с вектором $\vec{V}$, но его длина в полтора раза больше.

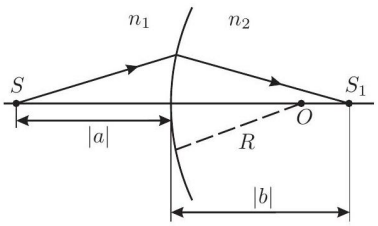
## 12.0.3 Advanced problems on geometrical optics 2

### ТУЙМ-111. Аквариум (2004)

Пучок света, проходя через пустой сферический аквариум, испытывает преломление на поверхностях сфер, разделяющих стеклянные стенки с воздухом, подобно преломлению на границах линз оптической системы. В задаче исследуются характеристики аквариума как оптической системы. Можно использовать формулу сферической поверхности:

$$\frac{n_1}{a} - \frac{n_2}{b} = \frac{n_1 - n_2}{R}.$$

Здесь $n_1$ и $n_2$ - показатели преломления первой и второй (по ходу луча) сред, разделённых этой поверхностью, а величины $a, b$ и $R$ - взятые со знаками «плюс»или «минус» расстояния от поверхности сферы до источника $S$, его изображения $S_1$ и центра сферы $O$. Знак «плюс» берётся, если расстояние отсчитывается по ходу луча, а «минус» в противоположном случае.




**1)** Найдите для узкого пучка света фокусное расстояние стенки пустого аквариума. Толщина стенки $\delta = 5$ мм, радиус аквариума $R = 10$ см, показатель преломления стекла $n = 1,6$.

**2)** Найдите фокусное расстояние всего аквариума, который отличается от рассмотренного в первом пункте лишь тем, что имеет толщину $\delta = R/2$ (толстый аквариум).

**3)** На расстоянии $2R$ от центра толстого аквариума помещают точечный источник света. На каком расстоянии $s$ от внешней поверхности аквариума наблюдатель, находящийся с противоположной от источника стороны, увидит изображение светящейся точки?

### Solution

Суть решения в том, что использовать обычные формулы (????)

Прохождение узкого пучка света через участок стенки аквариума можно рассматривать как последовательное преломление на плоско-выпуклой линзе с радиусом кривизны $R$, на плоско-параллельной пластине толщиной $\delta$ и на плоско-вогнутой линзе с радиусом кривизны $R - \delta$ (рис. 210). Прямая, соединяющая вершину рассматриваемого участка $P$ с центром аквариума $O$, служит главной оптической осью. Рассмотрим луч пучка, падающий параллельно этой оси, и найдём точку его пересечения с осью после прохождения стенки. Эта точка является фокусом.

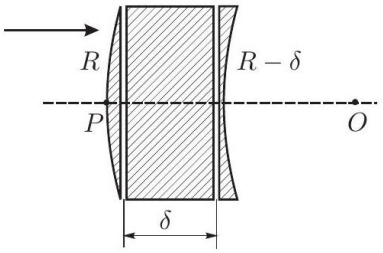



Рис. 210

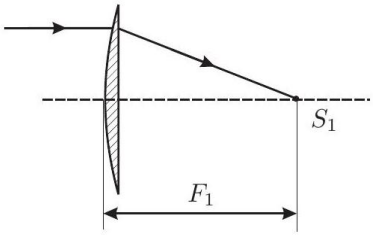



Рис. 211 После преломления на первой линзе луч идёт к точке $S_1$ (рис. 211), удаление которой от линзы

$$F_1 = \frac{R}{n-1}$$

Далее луч проходит через плоско-параллельную пластину, смещающую точку пересечения луча с осью на некоторую величину $\Delta$, которую можно найти из треугольника $ABC$ (рис. 212) по теореме синусов:

$$\frac{BC}{\sin(\alpha-\beta)} = \frac{AB}{\sin(180^\circ-\alpha)}, \quad \text{или} \quad \frac{\Delta}{\sin(\alpha-\beta)} = \frac{\delta}{\cos\beta\cdot\sin\alpha}.$$

Учитывая что рассматриваются лишь лучи, близкие к оптической оси, так что $\sin\alpha \approx \alpha, \sin\beta \approx \beta, \cos\alpha \approx \cos\beta \approx 1$, а также $\alpha/\beta \approx n$, из последнего равенства найдём

$$\Delta = \frac{\delta\sin(\alpha-\beta)}{\cos\beta\cdot\sin\alpha} \approx \delta\frac{n-1}{n}.$$

В точку $S_2$ луч не попадает, поскольку преломляется примыкающей к пластине плоско-вогнутой линзой, фокусное расстояние которой

$$F_2 = -\frac{R-\delta}{n-1}$$

Дальнейший ход луча показан на рис. 213. Луч выходит из системы так, как будто бы идёт из точки $S_3$.

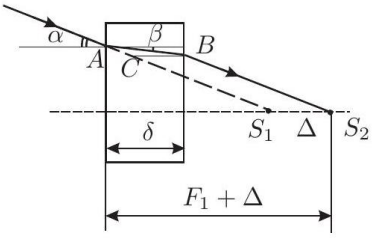



Рис. 212

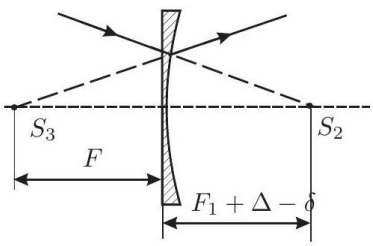


Рис. 213

Расстояние от $S_3$ до линзы, взятое со знаком «минус», является фокусным расстоянием стенки аквариума $F$. Для показанной на рис. 213 линзы $S_3$ является мнимым изображением мнимого источника $S_2$. Для нахождения $F$ применим формулу линзы:

$$\frac{1}{F} - \frac{1}{F_1 + \Delta - \delta} = \frac{1}{F_2}$$

или, с учётом формул (256), (257) и (258),

$$\frac{1}{F} = -\frac{n-1}{R-\delta} + \frac{1}{\frac{R}{n-1} + \delta\frac{n-1}{n} - \delta} \approx -\frac{n-1}{nR^2}\delta,$$

поскольку $\delta \ll R$. Итак,

$$F = \frac{n}{n-1} \cdot \frac{R^2}{\delta} = -5,3 \text{ м}$$

то есть стенка аквариума подобна рассеивающей линзе.

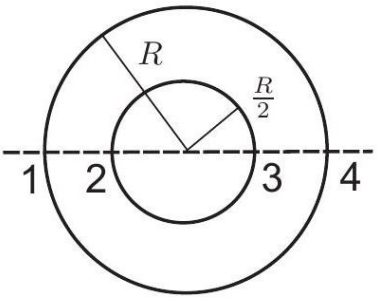



Рис. 214

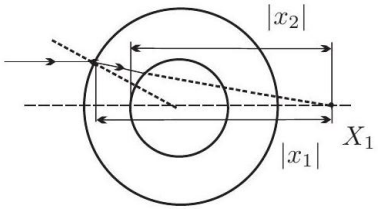



Рис. 215

Найдём фокусное расстояние толстого аквариума. Для этого нужно рассматривать преломление луча на четырёх поверхностях 1, 2, 3 и 4 (рис. 214). Нужно применить формулу сферической поверхности

$$\frac{n_1}{a} - \frac{n_2}{b} = \frac{n_1 - n_2}{R}$$

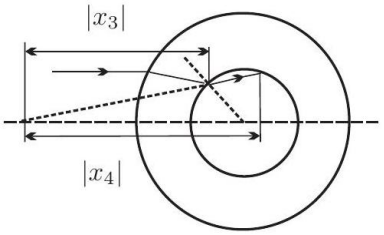



Рис. 216

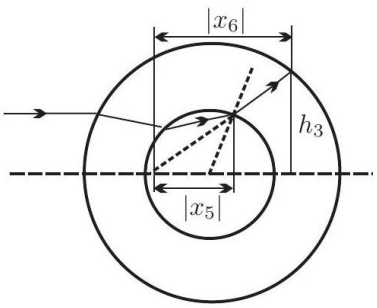


Рис. 217

Луч падает на первую поверхность параллельно главной оптической оси, поэтому $a = -\infty, b = x_1$ - расстояние от первой поверхности до точки схождения преломлённых лучей $X_1, n_1 = 1, n_2 = n$ (рис. 215). Из формулы (259) получаем

$$\frac{n}{x_1} = \frac{n-1}{R}$$

откуда

$$x_1 = \frac{n}{n-1}R$$

Найдём расстояние от точки $X_1$ до второй поверхности (рис. 215):

$$x_2 = x_1 - \frac{R}{2} = \frac{n+1}{2(n-1)}R$$

Применяем формулу (259) для второй поверхности (рис. 216), учитывая, что в этом случае $n_1 = n, n_2 = 1$:

$$\frac{n}{x_2} - \frac{1}{x_3} = \frac{n-1}{R/2}.$$

Отсюда находим положение мнимого изображения, даваемого этой поверхностью:

$$x_3 = -\frac{R}{2} \cdot \frac{n+1}{n-1}$$

Найдём расстояние от полученного изображения до третьей поверхности:

$$x_4 = -x_3 + R = \frac{3n-1}{2(n-1)}R$$

Снова применим формулу (259) для нахождения положения нового изображения (рис. 217):

$$-\frac{1}{x_4} - \frac{n}{x_5} = \frac{1-n}{-R/2}$$

откуда

$$x_5 = \frac{3n-1}{6(1-n)}R$$

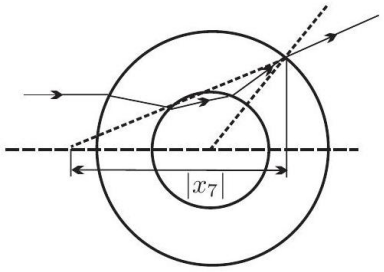



Рис. 218

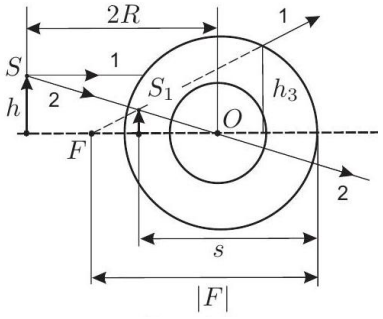


Рис. 219

Найдём расстояние от этого изображения до последней поверхности:

$$x_6 = -x_5 + \frac{R}{2} = \frac{3n-2}{3(n-1)}R$$

И снова применим формулу (259) для нахождения положения фокуса всей системы (рис. 218):

$$-\frac{n}{x_6} - \frac{1}{x_7} = \frac{n-1}{-R}$$

откуда

$$x_7 = \frac{3n-2}{2(1-n)}R.$$

Это и есть искомое фокусное расстояние

$$F = -\frac{3n-2}{2(n-1)}R \approx -23,3 \text{ cm}$$

Толстый аквариум подобен рассеивающей линзе.

Для нахождения положения $S_1$ источника $S$ рассмотрим два луча 1 и 2, на пересечении которых и находится $S_1$ (рис. 219). Из подобия треугольников с вершинами в точках $O$ и $F$ получим

$$\frac{h}{h_0} = \frac{2R}{s-R}; \quad \frac{h_3}{h_0} = \frac{|F|}{|F|-s}$$

где $h_0$ - размер изображения, $h$ - размер предмета, $h_3$ - расстояние от оптической оси до точки пересечения луча 1 с внешней поверхностью аквариума (рис. 219). При выводе соотношений предполагалось, что углы между всеми изображёнными на рис. 219 лучами и главной оптической осью малы. Из соотношений (260) получим

$$s = \frac{|F|R\left(2 + \frac{h}{h_3}\right)}{|F|\frac{h}{h_3} + 2R}.$$

Пользуясь подобием треугольников на рис. 215,216 и 217, можно получить следующие соотношения между расстояниями от оптической оси до точек пересечения лучей со сферами:

$$\frac{h}{h_1} = \left|\frac{x_1}{x_2}\right|; \quad \frac{h_1}{h_2} = \left|\frac{x_3}{x_4}\right|; \quad \frac{h_2}{h_3} = \left|\frac{x_5}{x_6}\right|.$$

Перемножая эти соотношения, получим

$$\frac{h}{h_3} = \left|\frac{x_1}{x_2}\cdot\frac{x_3}{x_4}\cdot\frac{x_5}{x_6}\right| = \frac{n}{3n-2}.$$

Осталось в формулу (261) подставить соотношение (262), а также выражение ДЛЯ $F$:

$$s = \frac{7n-4}{5n-4}R = 18$$

Задача 112. Стеклянная призма (2006 год)

Из соображений симметрии для того, чтобы луч вышел из призмы параллельно $AC$, отрезок луча $FG$ должен быть параллелен $AC$. При этом $\angle GFD = \angle EFA = \theta/2$ (рис. 220). Тогда из треугольника $AEF$ находим $\angle AEF = 180^\circ - 3\theta/2$. Поэтому для преломления в точке $E$ имеем

$$\frac{\sin(90^\circ - \theta/2)}{\sin(90^\circ - 3\theta/2)} = n, \quad \frac{\cos(\theta/2)}{\cos(3\theta/2)} = n$$

После тригонометрических расчётов находим

$\cos(3\theta/2) = \cos(\theta + \theta/2) = \cos\theta\cdot\cos(\theta/2) - \sin\theta\cdot\sin(\theta/2) = 4\cos^3(\theta/2) - 3\cos(\theta/2)$.

Подставляя полученное соотношение в закон Снеллиуса, получаем

$$\cos(\theta/2) = \sqrt{\frac{3n+1}{4n}}, \quad \text{откуда} \quad \theta = 2\arccos\left(\sqrt{\frac{3n+1}{4n}}\right) \approx 35^\circ 39'.$$

Пусть $\alpha$ - угол падения луча на призму, $\beta$ и $\beta + \Delta\beta$ - углы преломления, отвечающие показателям преломления $n$ и $n + \Delta n$ (рис. 220). Согласно закону Снеллиуса $\sin\alpha = n\sin\beta$, откуда, дифференцируя при фиксированном угле падения $\alpha$, имеем

$$n\cos\beta\cdot\Delta\beta + \Delta n\sin\beta = 0.$$

Пусть $\alpha'$ - угол выхода луча из призмы, $\beta'$ и $\beta' + \Delta\beta'$ - углы падения на границе стекло-воздух, отвечающие показателям преломления $n$ и $n + \Delta n$ (рис. 220). Тогда аналогично случаю падения можно записать

$$n \cos\beta' \cdot \Delta\beta' + \Delta n \sin\beta' = \cos\alpha' \cdot \Delta\alpha'$$

Поскольку $n \gg \Delta n$, можно положить $\alpha = \alpha'$ и $\beta = \beta'$. Однако следует отметить, что $\Delta\alpha \neq \Delta\alpha'$ и $\Delta\beta \neq \Delta\beta'$. Покажем, что $\Delta\beta = -\Delta\beta'$. Действительно, из треугольников $AEF$ и $CHG$ находим $\angle EFA = 90^\circ - \beta - \theta$ и $\angle HGC = 90^\circ - \beta' - \theta$. Далее рассмотрим треугольник $FGD$, для него $\theta = \angle GFD + \angle FGD$, а поскольку $\angle FGD = \angle HGC$ и $\angle GFD = \angle EFA$, имеем $\beta + \beta' = 180^\circ - 3\theta$. Теперь, производя дифференцирование, приходим к искомому результату: $\Delta\beta = -\Delta\beta'$. Для нахождения угла $\varepsilon = \Delta\alpha'$ между двумя выходящими из призмы лучами воспользуемся полученным выше соотношением между $\Delta\beta$ и $\Delta\beta'$, а также уравнениями (263) и (264):

$$-\Delta\beta = \frac{\Delta n \sin\beta}{n\cos\beta} = \Delta\beta' = \frac{\varepsilon\cos\alpha - \Delta n\sin\beta}{n\cos\beta}, \quad \varepsilon = 2\frac{\sin\beta}{\cos\alpha}\Delta n = 2\frac{\cos(3\theta/2)}{\sin(\theta/2)}\Delta n.$$

Пользуясь полученными ранее выражениями для $\cos(3\theta/2)$ и $\cos(\theta/2)$, приходим к ответу:

$$\varepsilon = \frac{2}{n}\sqrt{\frac{3n+1}{n-1}}\Delta n \approx 1,75'$$

Для определения линейного расстояния $y$ между двумя изображениями, видимыми в фокальной плоскости объектива, достаточно знать $f$ и $\varepsilon$:

$$y = f\,\mathrm{tg}\,\varepsilon \approx \varepsilon f = \frac{2f}{n}\sqrt{\frac{3n+1}{n-1}}\Delta n = 0,2 \text{ мм}$$

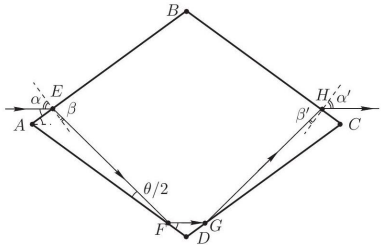



Рис. 220

## ТУЙМ-112. Стеклянная призма (2006)

Стеклянная призма, основание которой имеет ромбическую форму, находится в воздушном пространстве (рис. 64), причём у призмы $\angle BAD = \angle BCD = \theta$. Узкий луч жёлтого света падает в плоскости рисунка параллельно диагонали $AC$ на грань $AB$ призмы. Луч испытывает полное внутреннее отражение от граней $AD$ и $DC$ и выходит из призмы через грань $BC$. При решении задачи считайте, что для жёлтого света показатель пре-

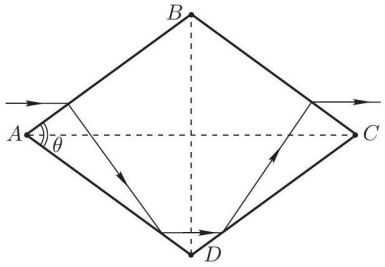



Рис. 64 ломления стекла $n = 1,60$, а для воздуха $n' = 1$. 1. При каком значении угла $\theta$ раствора призмы (как функции показателя преломления $n$) полное отклонение луча от своего первоначального направления будет нулевым? Найдите численное значение $\theta$, выраженное в градусах и минутах.

Оставим расположение призмы в пространстве и направление падающего луча прежними, но изменим характер падающего излучения. Пусть теперь падающий луч состоит из жёлтого дублета ртути. Дублет состоит из волн с длинами 579,1 нм и 577,0 нм. Показатели преломления стекла для этих длин волн равны соответственно $n = 1,60$ и $n + \Delta n$, где $\Delta n = 1,3 \cdot 10^{-4}$. Луч света, выходящий из призмы, направляется соосно в астрономический телескоп, настроенный на бесконечное расстояние.

1) Определите угловое расстояние $\varepsilon$ между двумя изображениями, видимыми в телескопе, как функцию $n$ и $\Delta n$. Найдите также численное значение $\varepsilon$.

2) Известно, что фокусное расстояние объектива телескопа $f = 0,4$ м. Найдите линейное расстояние между двумя изображениями, видимыми в фокальной плоскости объектива.

### Solution

Суть решения в том, что

## ТУЙМ-113. Холмс выбирает костюм (2007)

Шерлок Холмс, стоя перед плоским зеркалом в примерочной, обнаружил, что не видит в зеркале пола, хотя себя видит целиком (вплоть до низа ботинок) и нижний край зеркала касается пола. На какой высоте $x$ от пола находится изображение в зеркале верхнего края цилиндра (шляпы) Холмса? Край цилиндра выше уровня глаз Холмса на $h = 30$ см. Выходя из примерочной, Холмс заметил, что перестал видеть свои глаза в зеркале, когда отошёл от него в $n = 3$ раза дальше. Величину $n$ Холмс определил, подсчитав соответствующие плитки паркета на полу. Пол ровный и горизонтальный. Во время опытов Холмс стоял вертикально.

Примечание. Для произвольного угла $\alpha$ справедливы формулы $\sin 2\alpha = \frac{2\,\mathrm{tg}\,\alpha}{1+\mathrm{tg}^2\,\alpha}$, $\cos 2\alpha = \frac{1-\mathrm{tg}^2\,\alpha}{1+\mathrm{tg}^2\,\alpha}$, $\mathrm{tg}\,2\alpha = \frac{2\,\mathrm{tg}\,\alpha}{1-\mathrm{tg}^2\,\alpha}$.

### Solution

Суть решения в том, что

### Solution ТУЙМ-113. Холмс выбирает костюм (2007)

Холмс не увидел в зеркале пола, значит, луч, идущий от низа ботинок и попавший в глаз, отразился от нижнего края зеркала, а до отражения шёл горизонтально на очень малой высоте над полом. Это возможно, только если плоскость зеркала составляет некий угол $\varphi$ с вертикалью. Исходя из этого построим изображение Холмса $AB$ и цилиндра $BC$ в зеркале (рис. 221). Изображение $B'C'$ цилиндра образует угол $2\varphi$ с вертикалью, следовательно,

$$x = h\cos 2\varphi = h\cdot\frac{1-\mathrm{tg}^2\,\varphi}{1+\mathrm{tg}^2\,\varphi}.$$

Для нахождения угла $\varphi$ рассмотрим дополнительно положение, когда Холмс перестал видеть себя ниже уровня глаз (рис. 222), отойдя от зеркала на расстояние $L = nl$, где $l$ - расстояние от него до зеркала в первом случае. Во втором положении луч, идущий от глаза и попавший в глаз, отразился от нижнего края зеркала, так как в противном случае Холмс увидел бы своё изображение ниже уровня глаз. Поскольку этот луч возвращается в исходную точку, он перпендикулярен зеркалу. Пусть $H$ - высота глаз Холмса над уровнем пола, тогда из геометрических соображений в первом и втором положениях выражаем

$$\frac{H}{l} = \mathrm{tg}\,2\varphi = \frac{2\,\mathrm{tg}\,\varphi}{1-\mathrm{tg}^2\,\varphi},\quad \mathrm{tg}\,\varphi = \frac{H}{L} = \frac{H}{nl}$$

откуда

$$\mathrm{tg}^2\,\varphi = 1-\frac{2}{n}.$$

Подставив это выражение в формулу (265), находим

$$x = \frac{h}{n-1} = 15\ \text{cm}.$$

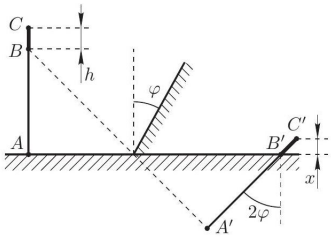



Рис. 221

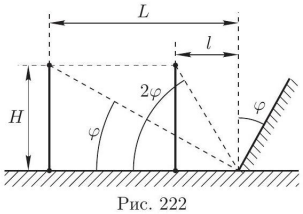



Рис. 222

## ТУЙМ-114. Без окон и дверей (2009)

Для съёмок фильма построили пустой павильон без окон и дверей. Серые бетонные стены, пол и потолок павильона отражают падающий на них свет диффузно во все стороны с коэффициентом отражения $R = 10\%$ для всех длин волн. Внутри павильона подвешен точечный источник света. Для увеличения освещённости режиссёр самолично заменил источник света на новый, имеющий в $\alpha = 2$ раза большую яркость, и перекрасил стены, пол и потолок в белый цвет, за счёт чего коэффициент отражения вырос в $\beta = 4$ раза. Во сколько раз $\gamma$ увеличилась средняя освещённость павильона?

### Solution

Суть решения в том, что

### Solution ТУЙМ-114. Без окон и дверей (2009)

Пусть $P$ - мощность излучения первого источника, $S$ - суммарная площадь стен, пола и потолка павильона, $E_1$ и $E_2$ - средние освещённости до и после произведённых изменений, тогда условия равенства количеств испущенного и поглощённого света в первом и втором случаях имеют вид

$$P = SE_1(1 - R), \quad \alpha P = SE_2(1 - \beta R)$$

откуда

$$\gamma = \frac{E_2}{E_1} = \frac{\alpha(1 - R)}{1 - \beta R} = 3$$

## ТУЙМ-115. Детектор перед зеркалом (2009)

На плоское зеркало падает под углом $\varphi$ к нормали плоская (идущая от бесконечно удалённого источника) монохроматическая световая волна с длиной волны $\lambda$. С помощью маленького детектора регистрируется сигнал, пропорциональный интенсивности света в точке наблюдения.

• Найдите частоту $\nu_1$ переменного сигнала, если зеркало неподвижно, а детектор движется со скоростью $v \ll c$, направленной под углом $\alpha$ к плоскости зеркала, но не обязательно лежащей в плоскости падения.

Какова в этом случае максимальная частота $\nu_1\,\mathrm{max}$ при всевозможных значениях $\varphi$ и $\alpha$? При каких $\varphi_1$ и $\alpha_1$ она достигается?

Найдите частоту $\nu_2$ переменного сигнала, если детектор неподвижен, а зеркало движется поступательно со скоростью $u \ll c$, лежащей в плоскости падения и перпендикулярной лучам падающей волны.

Какова в этом случае максимальная частота $\nu_2\,\mathrm{max}$ при всевозможных значениях $\varphi$? При каком $\varphi_2$ она достигается?

Эффект Доплера в данной задаче не рассматривается.

### Solution

Суть решения в том, что

### Solution ТУЙМ-115. Детектор перед зеркалом (2009)

В пространстве перед зеркалом происходит интерференция падающей и отражённой волн. Результат интерференции зависит в данном случае только от разности хода волн. Для любых двух точек, равноудалённых от зеркала, разность хода одинакова.

Пусть детектор расположен в точке $D$ на расстоянии $x$ от зеркала (рис. 223), тогда искомая разность хода $\Delta = AD - BD$, так как точки $A$ и $B$ принадлежат одному волновому фронту (свет достигает их одновременно). По условию угол падения $\angle A'AN = \varphi$, по закону отражения угол отражения $\angle DAN = \varphi$. Из геометрических соображений находим $\angle HDA = \varphi$, $\angle BAD = 2\varphi - \pi/2$

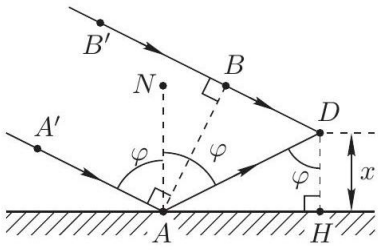


Рис. 223

$$\Delta = AD - BD = AD \cdot \left(1 - \sin\left(2\varphi - \frac{\pi}{2}\right)\right) = \frac{x}{\cos\varphi} \cdot (1 + \cos 2\varphi) = 2x\cos\varphi.$$

Разность хода изменяется на $\lambda$ при изменении $x$ на

$$d = \frac{\lambda}{2\cos\varphi}$$

Таким образом, мы нашли расстояние между соседними интерференционными максимумами или минимумами. Найдём частоту сигнала детектора:

$$\nu = \frac{1}{T} = \frac{|\dot{x}|}{d} = \frac{2|\dot{x}|\cos\varphi}{\lambda}$$

• При движении детектора $|\dot{x}| = v\sin\alpha$, откуда

$$\nu_1 = \frac{2v\sin\alpha\cos\varphi}{\lambda}$$

• Максимальная частота $\nu_1\,\max = 2v/\lambda$ достигается при $\varphi_1 = 0$ и $\alpha_1 = 90^\circ$.
При движении зеркала $|\dot{x}| = u\sin\varphi$, откуда

$$\nu_2 = \frac{2u\sin\varphi\cos\varphi}{\lambda} = \frac{u\sin 2\varphi}{\lambda}$$

• Максимальная частота $\nu_2\,\max = u/\lambda$ достигается при $\varphi_2 = 45^\circ$.

## ТУЙМ-116. Авторские очки (2010)

Автор задачи, желая разобраться в работе широко рекламируемых очков с большим количеством регулярно расположенных отверстий («дырчатых очков»), проделал в непрозрачном экране недалеко друг от друга две небольшие дырочки (расстояние между ними $l \sim 1,5$ мм, их диаметры $d \sim 0,5$ мм), расположил этот экран в непосредственной близости от глаза точно напротив зрачка и стал рассматривать через эти две дырочки яркую светящуюся точку.

В результате эксперимента было обнаружено, что изображение источника на сетчатке двоится (рис. 65), если источник расположен ближе $a_1 = 25$ см или дальше $a_2 = 42$ см от глаза, и сливается в одно пятнышко (рис. 66) при промежуточных положениях источника. Объясните наблюдаемые эффекты и определите оптическую силу $D$ необходимых автору очков. Считайте, что хрусталик глаза в процессе опыта расслаблен и его оптическая сила не изменяется.



Рис. 65



Рис. 66

### Solution

Суть решения в том, что
(я даун)

### Solution ТУЙМ-116. Авторские очки (2010)

Хрусталик является почти идеальной линзой и создаёт ровно одно изображение независимо от количества падающих лучей, однако глаз видит, регистрируя лучи, попавшие в те или иные точки сетчатки, поэтому далее нужно различать понятия «изображение» и «изображение на сетчатке».

Если источник расположен слишком близко, то его изображение получается далеко за сетчаткой, так что расстояние между его изображениями на сетчатке большое и они воспринимаются как два различных светлых пятнышка. При увеличении расстояния $a$ до источника расстояние между пятнышками

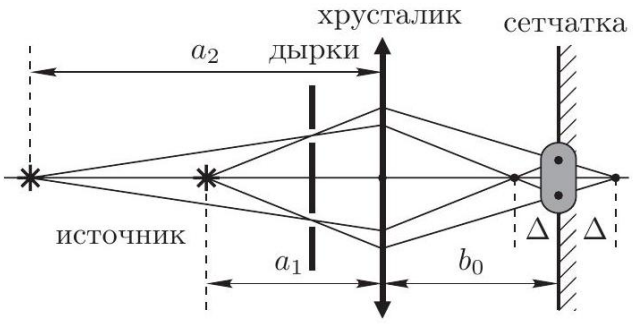



Рис. 224

уменьшается и при $a = a_1$ становится равным размеру чувствительного элемента сетчатки («пикселя»), поэтому пятнышки воспринимаются как одно удлинённое светлое пятно, хотя изображение источника получается на расстоянии $\Delta$ за сетчаткой. При увеличении расстояния $a$ в промежутке от $a_1$ до $a_2$ расстояние между пятнышками сначала уменьшается, а потом увеличивается, вновь достигая одного «пикселя» при $a = a_2$, когда изображение источника расположено примерно на расстоянии $\Delta$ перед сетчаткой. При дальнейшем увеличении расстояния $a$ пятнышки вновь разъезжаются.

Пусть $f$ - фокусное расстояние хрусталика, $b_0$ - расстояние от хрусталика до сетчатки, тогда формула линзы для двух граничных положений источника (рис. 224) имеет вид

$$\frac{1}{a_1} + \frac{1}{b_0 + \Delta} = \frac{1}{f}, \quad \frac{1}{a_2} + \frac{1}{b_0 - \Delta} = \frac{1}{f}$$

Складывая эти уравнения и учитывая $\Delta \ll b_0$, получим

$$\frac{1}{a_1} + \frac{1}{a_2} + \frac{2}{b_0} = \frac{2}{f}$$

Оптическую силу $D$ очков подбирают так, чтобы изображение бесконечно удалённого предмета попадало точно на сетчатку:

$$\frac{1}{\infty} + \frac{1}{b_0} = \frac{1}{f} + D, \quad \text{откуда} \quad D = \frac{1}{b_0} - \frac{1}{f}$$

а с учётом равенства (266) находим

$$D = -\frac{1}{2}\left(\frac{1}{a_1} + \frac{1}{a_2}\right) \approx -3,2 \text{ дптр.}$$

### ТУЙМ-117. Прямые углы в линзе (2011)

На координатной плоскости (рис. 67) ось $Ox$ совпадает с главной оптической осью тонкой линзы, а ось $Oy$ лежит в плоскости линзы. Предметом является треугольник $ABC$ с прямым углом при вершине $C$, а изображением - треугольник $A'B'C'$ с прямым углом при вершине $C'$, являющейся изображением точки $C$. Фокусное расстояние линзы $f$, а также координаты $x$ и $y$ точки $C$ известны, причём $x < 0, y \neq 0$. Найдите тангенсы углов наклона отрезков $CA$ и $CB$ к положительному направлению оси $Ox$.

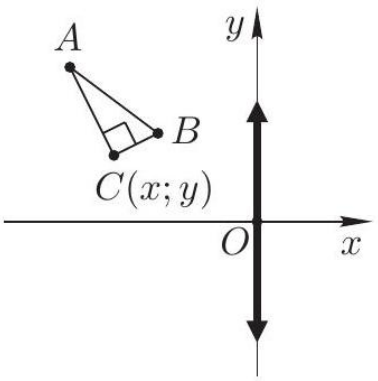


Рис. 67
Примечание. Линза может быть как собирающей ($f > 0$), так и рассеивающей ($f < 0$).

### Solution

Суть решения в том, что
(даже нет желания пробовать, я совсем все забыл, хочу ответ.)

## Solution ТУЙМ-117. Прямые углы в линзе (2011)

Пусть $x'$ и $y'$ - координаты точки $C'$, тогда из свойств тонкой линзы

$$\frac{1}{x'} + \frac{1}{-x} = \frac{1}{f} \quad \text{и} \quad \frac{x'}{x} = \frac{y'}{y}$$

находим

$$x' = \frac{fx}{f+x} \quad \text{и} \quad y' = \frac{fy}{f+x}.$$

Рассмотрим отрезок $CD$ длиной $l$, образующий угол $\alpha$ с осью $Ox$, и его изображение - отрезок $C'D'$, образующий угол $\alpha'$ с осью $Ox$. Выразим координаты точки $D$ и её изображения $D'$:

$$x_D = x + l\cos\alpha, \quad y_D = y + l\sin\alpha;$$
$$x'_D = \frac{fx_D}{f+x_D} = \frac{f(x+l\cos\alpha)}{f+x+l\cos\alpha}, \quad y'_D = \frac{fy_D}{f+x_D} = \frac{f(y+l\sin\alpha)}{f+x+l\cos\alpha)}.$$

Зная координаты двух точек отрезка, можно найти его наклон:

$$\operatorname{tg}\alpha' = \frac{y'_D - y'}{x'_D - x'} = \frac{(f+x)\operatorname{tg}\alpha - y}{f}$$

Пусть $\alpha_1$ и $\alpha_2$ - углы наклона катетов треугольника $ABC$, $\alpha'_1$ и $\alpha'_2$ - углы наклона катетов треугольника $A'B'C'$, тогда условия прямоугольности треугольников примут вид

$$\operatorname{tg}\alpha_1 \operatorname{tg}\alpha_2 = -1, \quad \operatorname{tg}\alpha'_1 \operatorname{tg}\alpha'_2 = -1$$

Подставив во второе условие выражения

$$\operatorname{tg}\alpha'_1 = \frac{(f+x)\operatorname{tg}\alpha_1 - y}{f}, \quad \operatorname{tg}\alpha'_2 = \frac{(f+x)\operatorname{tg}\alpha_2 - y}{f}, \quad \operatorname{tg}\alpha_2 = -\frac{1}{\operatorname{tg}\alpha_1},$$

ПОЛуЧИМ

$$\frac{(f+x)\operatorname{tg}\alpha_1 - y}{f} \cdot \frac{(f+x)\cdot\frac{-1}{\operatorname{tg}\alpha_1} - y}{f} = -1$$

что сводится к квадратному уравнению относительно $\operatorname{tg}\alpha_1$:

$$(f+x)y\operatorname{tg}^2\alpha_1 + \left(x^2 + 2fx - y^2\right)\operatorname{tg}\alpha_1 - (f+x)y = 0$$

Двумя корнями этого уравнения являются $\operatorname{tg}\alpha_1$ и $\operatorname{tg}\alpha_2$ :

$$\operatorname{tg}\alpha_{1,2} = \frac{y^2 - 2fx - x^2 \pm \sqrt{\left(x^2+2fx-y^2\right)^2 + 4y^2(f+x)^2}}{2y(f+x)}$$

Отметим, что формула справедлива при любых допустимых значениях параметров и что $x \neq -f$, так как иначе изображение не было бы треугольником.

## ТУЙМ-118. Отражения (2012)

Два круглых зеркала диаметром $D = 1$ м расположены друг напротив друга в почти параллельных плоскостях, угол между которыми $\varphi = 0,001$. Отрезок, соединяющий центры зеркал, имеет длину $a = 0,2$ м и перпендикулярен биссекторной плоскости двугранного угла $\varphi$ между плоскостями зеркал. Какое максимальное число $N$ отражений от зеркал может испытать луч света в такой системе? Явления, связанные с волновой природой света, в данной задаче не рассматриваются.

Примечание. Для $x \ll 1$ можно считать, что $\operatorname{tg} x \approx \sin x \approx x$, $\cos x \approx 1 - x^2/2$. При $nx \ll 1$ справедливо соотношение $(1+x)^n \approx 1 + nx$.

### Solution

Суть решения в том, что (?????)

Построить изображение в оптической системе с зеркалом можно, не только отразив лучи в зеркале, но и отразив все элементы системы в зеркале и удалив рассматриваемое зеркало.

Применим этот метод в данной задаче. Построим изображение одного из зеркал во втором и удалим второе зеркало, повторим эту операцию для новой пары зеркал и т. д. После большого числа отражений множество удалённых зеркал и их изображений будет иметь вид кольца с центром $O$ и радиусом $R \approx a/\varphi \gg D$ (рис. 225).

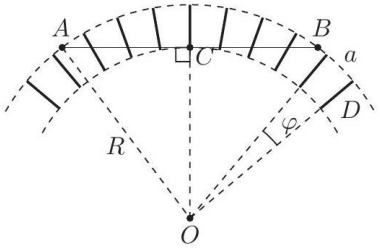



Рис. 225

Луч, идущий на рисунке прямолинейно внутри этого кольца, на самом деле отражается при каждом пересечении изображения удалённого зеркала, то есть испытывает примерно одно отражение на каждом своём отрезке длиной $a$. Наибольшее число отражений будет иметь самый длинный луч, то есть луч $AB$, концы $A$ и $B$ которого расположены вблизи одного края зеркала, а середина $C$ - вблизи противоположного, причём в зависимости от чётности числа отражений указанные точки могут оказаться как на одном, так и на обоих зеркалах, что в любом случае не влияет на ответ. Таким образом, можем найти искомое число отражений:

$$N = \frac{AB}{a} = \frac{2AC}{a} = \frac{2\sqrt{OA^2 - OC^2}}{a} = \frac{2\sqrt{R^2 - (R-D)^2}}{a} =$$
$$= \frac{2\sqrt{2RD - D^2}}{a} \approx \frac{2\sqrt{2RD}}{a} = 2\sqrt{\frac{2D}{\varphi a}} = 200.$$

## ТУЙМ-119. Прибрежные волны на море (2012 год) (????)

Под каким углом $\beta$ подойдут к прямому пологому берегу гребни (фронты) волн, если на глубине $H = 700$ м волны имели длину $\lambda = 50$ м, а их гребни образовывали с берегом угол $\alpha = 15°$? Глубина моря вблизи берега $h = 0,5$ м.

Примечание. Закон дисперсии (зависимость фазовой скорости $v$ от длины волны $\lambda$) морских волн имеет вид $v = \sqrt{\frac{g}{k}\,\mathrm{th}(kz)}$, где $g$ - ускорение свободного падения, $k = 2\pi/\lambda$ - волновое число, $z$ - глубина моря в рассматриваемой точке, а гиперболический тангенс th - это функция $\mathrm{th}\,x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Можно использовать предельные случаи закона дисперсии: на «глубокой» воде (при $z \gg \lambda$)$v \approx \sqrt{g/k}$; на «мелкой» воде (при $z \ll \lambda$)$v \approx \sqrt{gz}$.

### Solution

Суть решения в том, что (?? не нужно боятся задач про воду??? я хз вообще, что за пиздец???)
(я даун)

Показатель преломления среды обратно пропорционален скорости волны в данной среде, поэтому преломление морских волн может быть рассчитано с помощью «оптического» закона Снеллиуса, который удобно записать в виде

$$\frac{\sin\alpha_1}{v_1} = \frac{\sin\alpha_2'}{v_2}$$

где $\alpha_1$ и $\alpha_2'$ - углы падения и преломления соответственно, $v_1$ и $v_2$ - скорости волн в соответствующей среде.

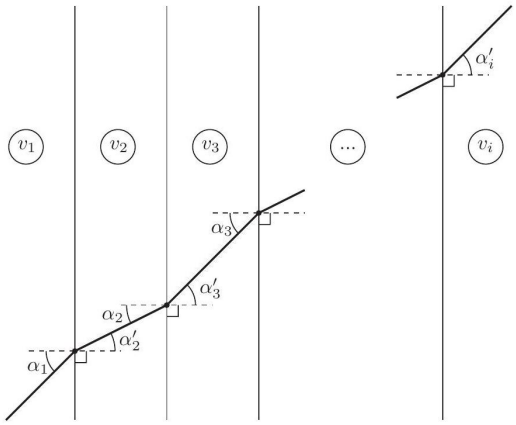



Рис. 226

Рассмотрим преломление волны при прохождении через набор произвольных однородных изотропных сред, границы раздела между которыми являются параллельными плоскостями (рис. 226). Обозначим через $v_1, v_2, \ldots, v_i$ скорости волны в каждой среде и предположим, что эффект полного внутреннего отражения отсутствует. Запишем законы преломления для каждой границы раздела сред и свойства равенства накрест лежащих углов:

$$\frac{\sin\alpha_1}{v_1} = \frac{\sin\alpha_2'}{v_2}, \quad \alpha_2' = \alpha_2, \quad \frac{\sin\alpha_2}{v_2} = \frac{\sin\alpha_3'}{v_3}, \quad \alpha_3' = \alpha_3, \quad \ldots$$

Последовательно объединяя записанные уравнения, получаем

$$\frac{\sin\alpha_1}{v_1} = \frac{\sin\alpha_2}{v_2} = \frac{\sin\alpha_3}{v_3} = \ldots = \frac{\sin\alpha_i'}{v_i}$$

откуда

$$\sin\alpha_i' = \frac{v_i}{v_1}\cdot\sin\alpha_1$$

Скорость волн зависит от расстояния до берега, поэтому море в целом является неоднородной средой, однако если разбить его на множество тонких вертикальных параллельных берегу слоёв, то каждый из них можно считать однородным. Поскольку $H \gg \lambda \gg h$, скорости волн $v_\alpha$ и $v_\beta$ вдали и вблизи берега соответственно найдём из предельных случаев закона дисперсии:

$$v_\alpha = \sqrt{\frac{g\lambda}{2\pi}}, \quad v_\beta = \sqrt{gh}$$

Применяя результат (268) и выражения (269), получаем

$$\beta = \arcsin\left(\frac{v_\beta}{v_\alpha}\sin\alpha\right) = \arcsin\left(\sqrt{\frac{2\pi h}{\lambda}}\sin\alpha\right) \approx 3,72^\circ.$$

Конечный результат сходится с экспериментальными наблюдениями, согласно которым гребни прибрежных волн почти параллельны берегу независимо от их направления в открытом море.

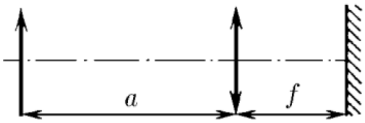
### Solution Примечание.

Отметим, что инвариант (267), записанный через скорости или показатели преломления, полезно помнить как готовый факт, так как он встречается и оказывается полезен во многих задачах.

## 12.0.4 Advanced problems on geometrical optics 3

### ФР.кр-ВШЭ?.вар2-1/5 Изображение после линзы и зеркала

За собирающей линзой с фокусным расстоянием $f$ находится плоское зеркало. Найти на каком расстоянии от линзы будет находится изображение предмета, который находится на расстоянии $a$ от линзы (стрелка на рис.). Расстояние от линзы до изображения отсчитывать слева направо (с учетом знака).



Указание: считать, что при формировании изображения свет проходит слева направо и взаимодействует с каждым оптическим элементом только один раз.

### Solution

(??? по идее все просто, нужно только не обсчитаться??? еще раз потом перерешаю ее!!)

Пусть расстояние предмета от линзы равно $x'$. По уравнению линзы: $\frac{1}{a}+\frac{1}{x'}=\frac{1}{f}$, поэтому $x'=\frac{fa}{a-f}$.

Расстояние изображения предмета в линзе до зеркала: $x''=f-\frac{fa}{a-f}$.

Расстояние от линзы до изображения после отражения от зеркала: $x'''=x'+2x$" (неважно, $x''>0$ или $x$ " $<0$). Подставим:

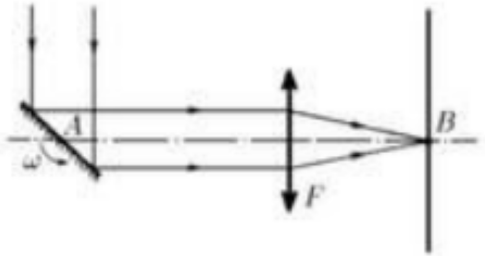
$$x'''=\frac{fa}{a-f}+2\left(f-\frac{fa}{a-f}\right)=\frac{fa+2(f(a-f)-fa)}{a-f}=\frac{f(a-2f)}{a-f}.$$

Ответ: $x^{m\prime\prime}=\frac{f(a-2f)}{a-f}$.

Главное - не усложнить задачу и не начать вторые считать взаимодействия с каждым оптическим прибором. (!? тут вроде все просто, но чет знак не тот был у меня, мб потом разберусь.)

## ФР.кр-ВШЭ?.вар1-1/5 Изображение во вращающемся зеркале

На горизонтально расположенной оптической оси $AB$ собирающей линзы с фокусным расстоянием $F$, находится плоское зеркало, которое вращается с угловой скоростью $\omega$ вокруг оси, перпендикуляриой плоскости рисунка, " проходящей через точку $A$ (см. рис.).




На зеркале вертикально падает параллельный пучок лучей, который после прохождения линзы фокусируется иа экране, расположеньом в фокальой плоскости линзы и перпендикулярном плоскости рисунка. Найти скорость светового пятна $B$ иа экране, когда оно пересекает оптическую ось.

### Solution Решение

Обозначим угол, который составляет плоскость зеркала с оттической осыо, как $\varphi$. Легко видеть, что угол, который составляет отраженый луч с горизонтальо, равеи $\frac{\pi}{2}-2\varphi$. Изображение отраженного от зеркала пучка будет иметь на экраие вертикальную координату $y(\varphi)=F\cdot\operatorname{tg}\left(\frac{\pi}{2}-2\varphi\right)$,

Соответетвено, скорость изоражения равна

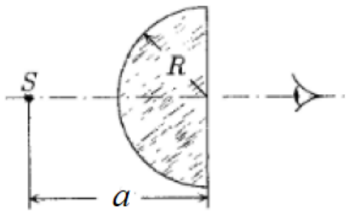
$$\mathrm{v},=\frac{dy}{dt}=F\cdot\frac{d}{dt}\operatorname{tg}\left(\frac{\pi}{2}-2\varphi\right)=F\cdot\frac{-2\dot{\varphi}}{\cos^2\left(\frac{\pi}{2}-2\varphi\right)}=F\cdot\frac{2\omega}{\cos^2\left(\frac{\pi}{2}-2\varphi\right)},$$

Т.К. $\omega=-\dot{\varphi}$ (см. рисунок). Когда изображение пересекает оптическую ось, $\varphi=\pi/4$. Соответетвенио, $v_y=2\omega F$.

Ответ: $v_{,}=2\omega F$.

## ФР.кр-ВШЭ?.вар2-2/5. О толстой линзе (!?!?!?)

Из стекла с показателем преломления $n$ изготовлена плосковыпуклая линза с радиусом кривизны передней поверхности $R$ (см. рис.). Линза находится в воздухе. Через такую линзу рассматривается точечный предмет $S$, расположенный на расстоянии $a$ от ее плоской поверхности.




а) На каком расстоянии от плоской поверхности линзы наблюдатель видит изображение (знак расстояния в ответе не важен)?
б) Слева или справа от этой плоскости находится изображение, если $a = 2R$, а $n < 2$?

### Solution

(???????? оптику учить нужно, чтобы понять это решение, не готов к нему вообще???? также тут построения есть, про которые самому нужно догадываться???)

$\Phi_2 = 0, \Phi_1 = \frac{n-1}{R} \Rightarrow \Phi_\Sigma = \Phi_1 \Rightarrow \begin{cases} O_2H_2 = -\frac{\Phi_1}{\Phi_\Sigma}\frac{d}{n} = -\frac{R}{n} \\ O_1H_1 = 0 \end{cases}$

Уравнение линзы: $\frac{1}{x} + \frac{1}{x'} = \Phi_\Sigma$, в котором $x = a - R \Rightarrow x' = \frac{R(R-a)}{a-na+Rn}$ (эта координата отсчитывается от точки $\mathrm{H}_2$).

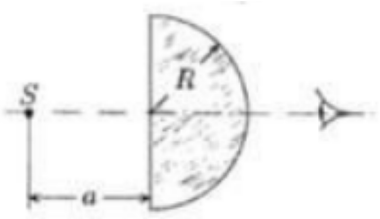
$$L = x' - \frac{R}{n} = \frac{R(R-a)}{a-na+Rn} - \frac{R}{n} = \frac{R(R-a)n - R(a-na+Rn)}{(a-na+Rn)n} = \frac{Ra}{(-a+na-Rn)n}$$

В частном случае, когда $a = 2R, n < 2 : L = \frac{2R^2}{(-2R+2nR-Rn)n} = \frac{2R}{(n-2)n} < 0$, т.е. изображение находится слева от задней поверхности линзы.

Ответ: а) $|L| = \left|\frac{Ra}{(na-a-Rn)n}\right|$. б) При $a = 2R, n < 2 : L = \frac{2R}{(n-2)n} < 0$, т.е. изображение находится слева от задней поверхности линзы.

## ФР.кр-ВШЭ?.вар1-2/5. О толстой линзе (!?!?!?)

Из стекла с показателем преловнления $n$ изготовлена плосковыпуклая линза с радиусом кривизыы залией поверхности $R$ (см. рис.). Линза находится в воздухе. Через такую лиизу рассматривается точечный предмет $\boldsymbol{S}$, расположенный на расстоянии $a$ от ее плоской поверхности.




а) На каком (по модулю) расстоянии $L$ от плоской поверхности линзы наблюдатель видит изображение?
б) Слева или справа от этой плоскости находится изображение, если $a = R/2$, а $n < 2$?

### Solution

$\Phi_1 = 0, \Phi_2 = \frac{n-1}{R} \Rightarrow \Phi_2 = \Phi_\Sigma \Rightarrow \begin{cases} O_1H_1 = \frac{\Phi_2}{\Phi_\Sigma}\frac{d}{n} = \frac{R}{n} \\ O_2H_2 = 0 \end{cases}$.

Уравнение линзы: $\frac{1}{x} + \frac{1}{x'} = \Phi_2$, в котором $x = a + \frac{R}{n} \Rightarrow x' = \frac{an+R}{n^2a-na-R}R$ (эта координата отсчитывается от точки $H_2$).

$$L = x' + R = \frac{an+R}{n^2a-na-R}R + R = \frac{an^2R}{n^2a-na-R}$$

В частном случае, когда $a = R/2, n < 2 : L = \frac{n^2R}{n^2-n-2} = \frac{n^2R}{(n-2)(n+1)} < 0$, т.е. изображение находися слева от передней поверхности лиизы.

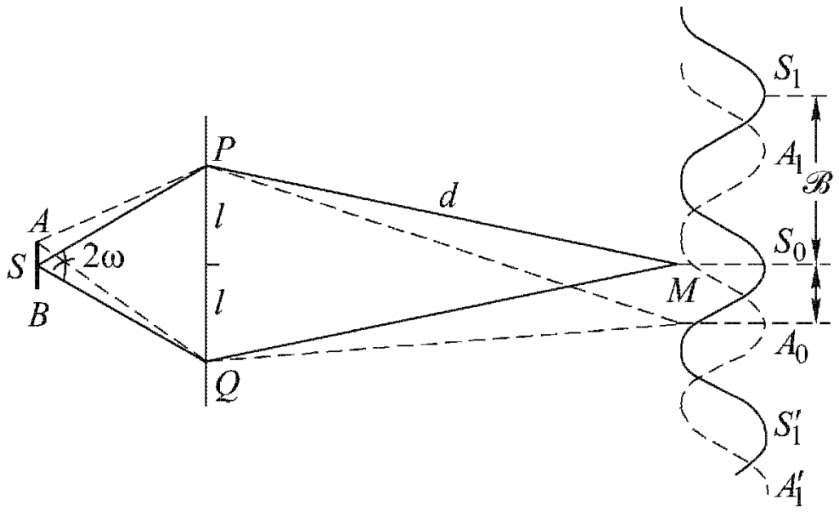
Ответ: а) $|L| = \left|\frac{an^2R}{n^2a-na-R}\right|$, б) При $a = R/2, n < 2 : L = \frac{n^2R}{n^2-n-2} = \frac{n^2R}{(n-2)(n+1)} < 0$, т.е. изображение находится слева от передией поверхности линзы.

## ФР. Размеры источника для интерференции в билинзе

Каков должен быть диаметр $D$ светящейся нити, которая используется в установке с билинзой, изготовленной из линзы с фокусным расстоянием $f = 12$ см, чтобы получить отчетливую интерференционную картину? Половинки линзы раздвинуты на расстояние $d = 1,1$ мм. Нить находится на расстоянии $a = 50$ см от билинзы. Экран находится на расстоянии $L = 480$ см от билинзы. Длина волны света $\lambda = 610$ нм.

### Solution

(??? суть??)
Для определения условия существования полос, будем иметь в виду ситуацию, описанную в книге Ландсберга, а именно, схему:



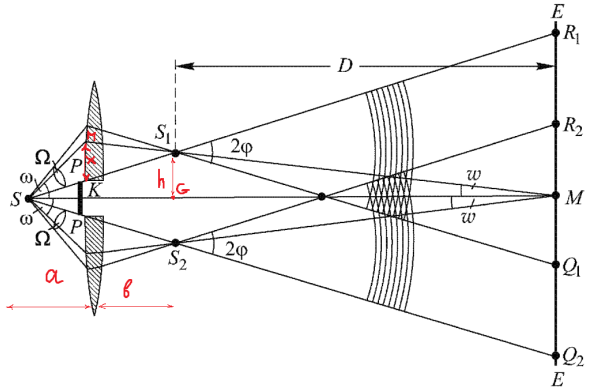

Тогда, по Ландсбергу (формула (17.1)) условие хорошего наблюдения интерференции от протяженного источника можно записать в виде

$$2b \sin\omega \leqslant \frac{\lambda}{4},$$

где $\omega$ - апертура интерференции, а $b$ - половина длины источника. Отсюда, максимальный диаметр нити

$$D = 2b = \frac{\lambda}{4 \sin\omega}.$$

Для определения апертуры, рассмотрим геометрию лучей в билинзе:




Обозначим расстояние $MP$ как $x$, тогда

$$\sin\omega = \frac{x + d/2}{\sqrt{a^2 + (x + d/2)^2}} \approx \frac{x + d/2}{a},$$

так как предположим, что $\sqrt{a^2+(x+d/2)^2} \approx a$, потому что в нашем треугольнике один катет сильно больше другого. Задача сводится к поиску $x$. Согласно формуле тонкой линзы,

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{F},$$

где $b$ - расстояние до $S_1$ от линзы. Тогда из подобия треугольников справа от билинзы имеем

$$\frac{x+d/2}{L}=\frac{h}{L-b},$$

а из подобия треугольников слева, имеем

$$\frac{d/2}{a}=\frac{h}{a+b}.$$

В итоге

$$x+d/2=\frac{d(a+b)}{2a}\frac{L}{L-b},$$

тогда $x+d/2 \approx 4.15\cdot 10^{-3}$ м, что по порядку совпадает с $d$, таким образом, наше приближение действительно применимо. Также $\sin\omega = \frac{dL(a+b)}{2a^2(L-b)}$ и

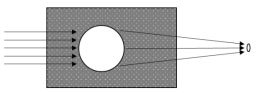
$$D=2b=\frac{\lambda}{4\sin\omega}=\frac{\lambda a^2(L-b)}{2dL(a+b)}, \qquad \text{где } \; b=\left(\frac{1}{F}-\frac{1}{a}\right)^{-1}.$$

Подставим числа, $b=(1/0.12-1/0.5)^{-1}\approx 0.158$ м,

$$D=\frac{6.1\cdot 10^{-7}0.5^2(4.8-0.158)}{2\cdot 4.8\cdot 1.1\cdot 10^{-3}(0.5+0.158)}\approx 1.02\cdot 10^{-4} \text{ м}.$$

## МФТИ-ГКЭ.2020.4А Линзирование сферической полостью

В практических задачах рентгеновской оптики было предложено использовать собирающие линзы в виде сферических полостей внутри металлических блоков. Оценить в приближении параксиальной оптики минимальный размер пятна $d$, в который можно сфокусировать плоскопараллельный пучок рентгеновского излучения с энергией кванта $E_\gamma = 14$ кэВ при помощи такой линзы, если радиус сферической полости, сделанной в алюминии, составляет $R = 0,3$ мм. Плотность алюминия $\rho = 2,7$ г/см $^3$.



### Solution

Для применения формулы фокусного расстояния нужно знать коэффициент преломления, он дается из плазменной частоты, которую можно посчитать, используя знания о типе материала. Далее, размер пятна дается из дифракционной расходимости. (???)

Для расчета фокусного расстояния линзы $F = R/2(n-1)$ необходимо, прежде всего, рассчитать коэффициент преломления по формуле

$$n=\left(1-\frac{\omega^2}{\omega^2}\right)^{1/2}$$

Поскольку в алюминии ($A=27, Z=13$) энергия связи даже самых «близких к ядру" электронов не превышает $E=Z^2\text{Ry}=2,23$ кэВ, то в формулу для расчета плазменной частоты

$$\omega_p^2=\frac{4\pi e^2 n_e}{m_e}$$

нужно подставить полную концентрацию электронов

$$n_e=\frac{Z\rho N_A}{\mu_{Al}}=\frac{13\cdot 2,7\cdot 6,02\cdot 10^{23}}{27}=7,8\cdot 10^{23}\ \text{см}^{-3}$$

При такой концентрации плазменная частота составит $\omega_p = \left(\frac{4\cdot3{,}14\cdot4{,}8^2\cdot10^{-20}\cdot7{,}8\cdot10^{23}}{9{,}1\cdot10^{-28}}\right)^{1/2} = 5{,}0\cdot10^{16}\text{с}^{-1}$, что намного меньше частоты рентгеновского излучения $\omega = \frac{E_\gamma}{\hbar} = \frac{14\cdot10^3\cdot1{,}6\cdot10^{-12}}{1{,}05\cdot10^{-27}} = 2{,}1\cdot10^{19}\text{с}^{-1}$. С учетом этого, выражение (1) можно переписать в виде

$$n = \left(1 - \frac{\omega_p^2}{\omega^2}\right)^{1/2} \approx 1 - \frac{\omega_p^2}{2\omega^2}$$

Таким образом, воздух в полости является оптически более плотной средой для рентгеновского излучения, чем алюминий, и поэтому сферическая полость будет играть роль собирающей линзы в алюминии. Для фокусного расстояния получаем:

$$F = \frac{R}{2}\cdot\frac{2\omega^2}{\omega_p^2} = R\left(\frac{\omega}{\omega_p}\right)^2$$

Угловая дифракционная расходимость $\varphi = \lambda/2R$ (??????), так что размер пятна равен в фокальной плоскости:

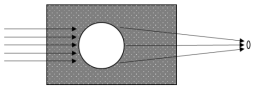
$$d = F\cdot\varphi = R\left(\frac{\omega}{\omega_p}\right)^2\cdot\frac{\lambda}{2R} = \frac{\pi c\omega}{\omega_p^2} == \frac{3{,}14\cdot3\cdot10^8\cdot2{,}1\cdot10^{19}}{25\cdot10^{32}} \approx 8\text{МКМ}$$

### Solution О применимости этой задачи

В реальности сферические аберрации приводят к тому, что такого разрешения можно добиться лишь при помощи 30 последовательно расположенных «линз». См. статью A. Snigirev, V. Kohn, I. Snigireva, B. Lengeler, "A compound refractive lens for focusing high-energy X-rays" // Nature. 1996. V.384. № 6604. P.49. Отметим, что при заданном материале линзы размер пятна, в который можно сфокусировать излучение (в использованном приближении!): 1) не зависит от размеров «линзы», 2) обратно пропорционален длине волны используемого излучения.

## МФТИ-ГКЭ.2020.4Б Линзирование многими сферическими полостями

В практических задачах рентгеновской оптики было предложено использовать собирающие линзы в виде сферических полостей внутри металлических блоков. Оценить в приближении параксиальной оптики эффективное фокусное расстояние системы таких линз, представляющую собой 30 вплотную расположенных сферических полостей с радиусом $R = 0{,}3$ мм. Энергия кванта составляет $E_\gamma = 14$ кэВ. Плотность алюминия $\rho = 2{,}7$ г/см$^3$.



### Solution

(???)

По аналогии с решением задачи 4А

$$F_1 = R\left(\frac{\omega}{\omega_p}\right)^2 = 3\cdot10^{-4}\left(\frac{2{,}1\cdot11^{19}}{5{,}0\cdot10^{16}}\right)^2 \approx 53\text{ m}.$$

Поскольку максимальный угол отклонения лучей очень мал, для составной линзы можно пользоваться параксиальным приближением, так что для ее фокусного расстояния получим

$$F_{30} = \frac{F_1}{30} \approx 1.8\text{ m}$$

# 12.0.5 Problems about photometry

## O2-1.38 Зрительная труба

The telescope with the focal length of the lens $f = 50$ cm is set to infinity. How far $\Delta l$ should the eyepiece of the tube be moved to clearly see objects at a distance of 50 m?

### Solution

(??? геом оптика тут что ли?)

## O2-1.39

На систему линз, изображенную на рис. 270, падает слева параллельный пучок света. Найти положение точки схождения этого пучка посте прохождения системы.

## O2-1.40

Найти изображение точки, которая находится на расстоянии 10 cm слева от крайней левой линзы системы, изображенной на рис. 271.

## O2-1.42

Найти положения главных плоскостей толстой линзы, имеющей форму шара радиусом $R$. Определить фокусные расстояния $f$ и $f'$ и положения фокальных точек такой линзы, когда она сделана: 1) из воды $(n_{\text{В}} = 4/3)$; 2) из стекла $(n_{\text{СТ}} = 3/2)$. При каком показателе преломления фокальные точки не выйдут наружу?

## O2-1.43

Радиус стеклянного $(n = 1,5)$ шара $R = 4$ см.
1) Найти расстояние $x'$ от центра шара до изображения предмета, который расположен в 6 см от поверхности шара.
2) Найти увеличение изображения.

## O2-1.48

Какова освещенность $E$ на горизонтальной площадке, освещаемой небесной полусферой, если считать яркость неба повсюду равномерной и равной $B$?

## O2-1.49

Какую освещенность $E$ следует создать на белом листе бумаги с коэффициентом отражения $k = 0,85$, чтобы его яркость была $B = 3 \cdot 10^4/\text{m}^2$? Можно считать, что бумага рассеивает свет по закону Ламберта.

## O2-1.50

Освещенность, получаемая при нормальном падении солнечных лучей на поверхность Земли, составляет приблизительно $E_0 = 100000$ лк. Какова освещенность $E$ изображения Солнца, даваемого свободной от аберраций линзой с диаметром $D = 5$ cm и фокусным расстоянием $f = 10$ см? Угловой диаметр Солнца $\alpha = 30'$.

## O2-1.51

Освещенность, получаемая при нормальном падении солнечных лучей на поверхность Земли, около $E = 10^5$ лк. Считая, что излучение Солнца подчиняется закону Ламберта, и пренебрегая поглощением света в атмосфере, определить яркость Солнца, если известно, что радиус земной орбиты $R = 1,5 \cdot 10^8$ км, а диаметр Солнца $D = 1,4 \cdot 10^6$ км.

## O2-1.52

Найти освещенность поверхности Земли у экватора светом, отраженным Луной в полночь в полнолуние. Считать, что Солнце является ламбертовым источником света, а Луна - ламбертовым отражателем. Яркость Солнца $B_{\text{С}} = 1,5 \cdot 10^9/\text{m}^2$, радиус Солнца $R_{\text{С}} = 7 \cdot 10^8$М, расстояние от Солнца до Земли (и Луны) $R_0 = 1,5 \cdot 10^{11}$ м, расстояние от Луны до Земли $R_1 = 3,8 \cdot 10^8$ м, видимый радиус Луны $R_{\text{Л}} = 1,7 \cdot 10^6$ м. Коэффициент отражения лунной поверхности $k = 7\%$.

### O2-1.53

Спектрограф имеет объектив коллиматора диаметром $D$ с фокусным расстоянием $f_1$ и объектив камеры того же диаметра с фокусным расстоянием $f_2$. Источник с яркостью $B$ резко отображается на входную щель спектрографа при помощи конденсора один раз с увсличением (расстояние от конденсора до щели равно L), другой раз с уменьшением. Каков должен быть диаметр конденсора $D_{\text{K}}$, чтобы в обоих его положениях освещенность на фотопластинке была одинаковой? Чему равна освещенность $E$ в этом случае, если пренебречь потерями на отражение и поглощение?

### O2-1.55

Какую мощность должна иметь лампа в осветителе микроскопа для того, чтобы можно было производить киносъемку микрообъектов в воздухе с частотой кадров $v = 10^4$ Гц? Объектив микроскопа имеет увеличение $k_1 = 20$ и апертуру $a = 0,4$, окуляр - увеличение $k_2 = 8$. Коэффициент преобразования энергии в активный для фотослоя свет в лампе накаливания $\eta = 2,5\%$. Тело свечения лампы имеет размеры $S = 2 \times 5$ мм. Чувствительность фотослоя $H = 20$ cm$^2$/ эрг (т. е. 1 эрг световой энергии на площади 20 cm$^2$ дает нормальную плотность почернения). Пропускающая способность объекта $\chi = 40\%$. Потерями света в линзах пренебречь.

### O2-1.56

Действительное изображение, сформированное собирающей линзой, рассматривается сначала непосредственно, а затем на белом экранс. Как зависит в обоих случаях яркость изображения от диаметра линзы?

### O2-1.57

Найти яркость изображения Луны, наблюдаемой в телескоп с объективом диаметром 75 мм, при увеличениях: 1) 20-кратном; 2) 25-кратном; 3) 50-кратном. Яркость Луны, видимой невооруженным глазом, принять за единицу. Диаметр зрачка глаза считать равным 3 mm.

### O2-1.58* Кажущееся уменьшение яркости в микроскопе

Как известно, яркость изображения в оптической системе не зависит от его увеличения. Почему же при наблюдении в микроскоп изображение кажется менее ярким, если применить большее увеличение?

Найти: 1) освещенность изображения в микроскопе с числовой апертурой 1 (сухая система) и увеличением $625$; 2) освещенность изображения в микроскопе с числовой апертурой 1,5 (иммерсия с $n = 1,5$) и увеличением 1500. Освещенность объекта принять за единицу. Расстояние ясного зрения равно 25 cm, диаметр зрачка глаза считать равным 2 мм. Потерями света в микроскопе пренебречь.

### ФР.кр-ВШЭ?.вар1-тренир-3/5. Освещенность в круглой комнате

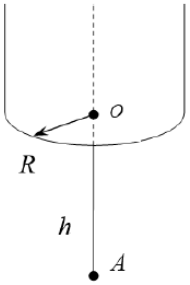
В комнате, которая имеет форму кругового цилиндра радиуса $R$, на потолке в центре круга горит плоский светильник, размеры которого пренебрежимо малы по сравнению с радиусом комнаты. Найти на каком расстоянии от потолка (от плоскости потолка по вертикали) освещенность стены комнаты (т.е. боковой поверхности цилиндра) будет максимальной, если светильник можно считать ламбертовым источником.

#### Solution

$E = LS_0 \frac{\cos^3\theta \sin\theta}{R^2} = \frac{LS_0}{R^2} \cos^3\theta \sin\theta$.
Дифференцируя по $\theta$, получаем: $\operatorname{tg}\theta_{\max} = 1/\sqrt{3} \Rightarrow h_{\max} = R/\sqrt{3}$.
Ombem: $h_{\max} = R/\sqrt{3}$.

### ФР.кр-ВШЭ?.вар2-3/5. Освещенность под полубесконечным цилиндром

У полубесконечного кругового цилиндра радиусом $R$ с открытыми основаниями светится внутренняя боковая поверхность с яркостью $L$ (см. рис.).




Найти освещенность плоскости, перпендикулярной образующей цилиндра, в т. $A$ на оси цилиндра. Точка $A$ находится на расстоянии $h$ от нижнего основания цилиндра.

### Solution

Известно для протяженных источников, что

$$E = L \int\limits_{S_i} r^{-2}\left(\theta_f\right) \cos\theta_f \cdot \cos\left(\theta_i\right) dS_i,$$

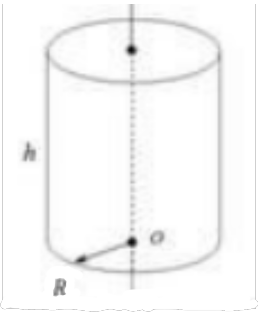
где обозначим $\theta_f = \theta$, тогда $S_i = 2\pi R^2 \operatorname{ctg}\theta \Rightarrow dS_i = 2\pi R^2 \frac{d\theta}{\sin^2\theta}, r = \frac{R}{\sin\theta}, \theta_i = \frac{\pi}{2} - \theta$

Отсюда: $E = 2\pi R^2 L \int\limits_0^{\theta_0} \left(\frac{R}{\sin\theta}\right)^{-2} \cos\theta \cdot \sin\theta \frac{d\theta}{\sin^2\theta} = 2\pi L \int\limits_0^{\theta_0} \cos\theta \cdot \sin\theta d\theta = \pi L \frac{R^2}{R^2+h^2}$

Ответ: $E = \pi L \frac{R^2}{R^2+h^2}$.

## ФР.кр-ВШЭ?.вар1-3/5. Освещенность сразу под полубесконечным цилиндром

У кругового цилиндра высотой $h$ и радиусом $R$ с открытыми основаниями светится внутренняя боковая поверхность с яркостью $L$. Найти освещенность плоскости основания шилицда (перпендикуаярной образующей циаивдра) в центре шижнего основания (т. $O$ иа рис.).




### Solution

$E = L \int r^{-2}\left(\theta_f\right) \cos\theta_f \cdot \cos\left(\theta_i\right) dS_i$, где обозначим $\theta_f = \theta$, тогда

$$S_i = 2\pi R^2 \operatorname{ctg}\theta \Rightarrow dS_i = 2\pi R^2 \frac{d\theta}{\sin^2\theta}, r = \frac{R}{\sin\theta}, \theta_i = \frac{\pi}{2} - \theta.$$

Отсюда: $E = 2\pi R^2 L \int\limits_2^2 \left(\frac{R}{\sin\theta}\right)^{-2} \cos\theta \cdot \sin\theta \frac{d\theta}{\sin^2\theta} = 2\pi L \int\limits_{h_1}^{\pi 2} \cos\theta \cdot \sin\theta d\theta = \pi L \frac{h^2}{R^2+h^2} = \frac{\pi L}{(R/h)^2+1}$

Ответ: $E = \pi L \frac{h^2}{R^2+h^2} = \frac{\pi L}{(R/h)^2+1}$

## ФР.теория Освещенность бесконечной плоскостью

Найти освещенность бесконечной плоскостью с ламбертовской яркостью (освещаемая площадка параллельна плоскости)

### Solution

$L(\theta_f) = L;\quad \theta_i(\theta_f) = \theta_f;\quad r(\theta_f) = h/\cos\theta_f, h$ - где расстояние от площадки до плоскости; $S_i = \pi h^2 \operatorname{tg}^2\theta_f \Rightarrow \frac{dS_i}{d\theta_f} = \pi h^2 2\operatorname{tg}\theta_f \frac{1}{\cos^2\theta_f} = 2\pi h^2 \frac{\sin\theta_f}{\cos^3\theta_f}$. Интеграл

$$E = \int\limits_{S_f} L(\theta_f) r^{-2}(\theta_f) \cos\theta_f \cdot \cos(\theta_i(\theta_f)) \left(\frac{dS_i}{d\theta_f}\right) d\theta_f$$

в данном случае:

$$E = \frac{L}{h^2} \int\limits_0^{\pi/2} \cos^4\theta_f 2\pi h^2 \frac{\sin\theta_f}{\cos^3\theta_f} d\theta_f = \pi L \int\limits_0^{\pi/2} \sin 2\theta_f d\theta_f = \pi L$$

### ФР.теория Освещенность полусферой

Найти освещенность полусферой с ламбертовской яркостью (освещаемая площадка параллельна плоскости сечения).

$L(\theta_f) = L;\quad \theta_i(\theta_f) = \pi/2;\quad r(\theta_f) = R;\quad S_i = \pi h^2 \operatorname{tg}^2\theta_f \Rightarrow \frac{dS_i}{d\theta_f} = \pi h^2 2\operatorname{tg}\theta_f \frac{1}{\cos^2\theta_f} = 2\pi h^2 \frac{\sin\theta_f}{\cos^3\theta_f}$. Интеграл

$$E = \int\limits_{S_f} L(\theta_f) r^{-2}(\theta_f) \cos\theta_f \cdot \cos(\theta_i(\theta_f)) \left(\frac{dS_i}{d\theta_f}\right) d\theta_f$$

в данном случае:

$$E = \frac{L}{h^2} \int\limits_0^{\pi/2} \cos^4\theta_f 2\pi h^2 \frac{\sin\theta_f}{\cos^3\theta_f} d\theta_f = \pi L \int\limits_0^{\pi/2} \sin 2\theta_f d\theta_f = \pi L$$

## 12.0.6 Problems about transmission and reflectance of light

### ФР.кр-ВШЭ?.вар1-4/5 Интенсивность преломленного пучка (!?!?)

Плоский пучок света с интенсивностью $I_0$ падает из воздуха под углом $\varphi$ на поверхности раздела со средой с показателем преломления $n$. В этом процессе коэффициент отражения падающего света по интенсивности равен $\rho$. Найти интенсивность преломленного пучка.

Указание: закон сохранения энергии в данном случае пишется для проекций на нормаль к границе (проведенной, например, от воздуха к среде) вектора Умова-Пойнтинга (средний по времени модуль этого вектора и есть интенсивность света): $S_n^{(n)} + S_n^{(n)} = S_n^{(-)}$, где индекс "$e$" соответствует падающему пучку, "$r$" - отраженному, а "$d$" - преломленному.

### Solution

Соотношение $S_n^{(n)} + S_*^{(n)} = S_n^{(f)}$ перепишем в виде

$$I^{(i)}\cos\varphi - I^{(i)}\cos\varphi = I^{(i)}\cos\psi$$

(зиак минус перед отраженым пучком означает распространение эергии в противоположную сторону от падающего пучка). Учтем, что $I^{(n)} = I_0, I^{(n)} = \rho I_9$. Тогда соотношение (1) перепишетея визе:

$$I^{(u)} = (1-\rho)\frac{\cos\varphi}{\cos\psi} I_0.$$

$$\sin\varphi = n\sin\psi \qquad \Rightarrow \qquad \sin\psi = \frac{\sin\varphi}{n} \qquad \Rightarrow \qquad \cos\psi = \sqrt{1 - \frac{\sin^2\varphi}{n^2}}.$$

Ответ: $I^{(i)} = (1-\rho)\frac{n\cos\varphi}{\sqrt{n^2 - \sin^2\varphi}} I_0$.

## ФР.кр-ВШЭ?.вар2-4/5 Интенсивность преломленного пучка (!?!?)

Плоский пучок света с интенсивностью $I_0$ падает из воздуха под некоторым углом на поверхность раздела со средой с показателем преломления $n$. В этом процессе коэффициент отражения падающего света по интенсивности равен $\rho$. Найти интенсивность преломленного пучка, если угол преломления равен $\psi$.

Указание: закон сохранения энергии в данном случае пишется для проекций на нормаль к границе (проведенной, например, от воздуха к среде) вектора Умова-Пойнтинга (средний по времени модуль этого вектора и есть интенсивность света): $S_n^{(e)} + S_n^{(r)} = S_n^{(d)}$, где индекс "$e$" соответствует падающему пучку, "$r$" - отраженному, а "$d$"- преломленному.

### Solution

Соотношение $S_n^{(e)} + S_n^{(r)} = S_n^{(d)}$ перепишем в виде

$$I^{(e)} \cos\varphi - I^{(r)} \cos\varphi = I^{(d)} \cos\psi$$

где знак минус перед отраженным пучком означает распространение энергии в противоположную сторону от падающего пучка. У нас $I^{(e)} = I_0, I^{(r)} = \rho I_0$, поэтому

$$I^{(d)} = (1-\rho)\frac{\cos\varphi}{\cos\psi} I_0.$$

$$\sin\varphi = n\sin\psi \qquad \Rightarrow \qquad \cos\varphi = \sqrt{1 - n^2\sin^2\psi}$$

$$I^{(d)} = (1-\rho)\frac{\sqrt{1-n^2\sin^2\psi}}{\cos\psi} I_0.$$

Comment: Формула $I_e\cos^2(\varphi) + I_r\cos^2(\varphi) = I_d\cos^2(\psi)$ неверная, потому что вектор Пойтинга - интенсивность. (???????)

## О2-2.1(4 мин) Сохранение энергии и формулы Френеля

Проверить с помощью формул Френеля, что поток энергии падающей волны через границу раздела сред равен сумме потоков энергии преломленной и отраженной волн через ту же границу.

(??? по идее все просто, но я забыл формулы.)

## О2-2.2 Коэффициент отражения (???)

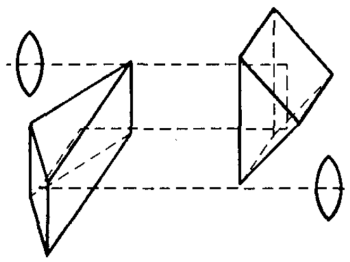
Исходя непосредственно из граничных условий для электрического и магнитного полей на границе вакуума и диэлектрика, найти коэффициент отражения $\rho$ света при нормальном падении на границу раздела. Выразить коэффициент отражения через показатель преломления диэлектрика $n$. Найти значения для отражения света от поверхности воды ($n = 1,33$) и стекла ($n = 1,5$).

## О2-2.3(2 мин) Коэффициент пропускания (???)

Найти коэффициент пропускания $T$ при нормальном падении света из воздуха на стекло с показателем преломления $n = 1,5$.

## О2-2.4 Потери в бинокле (??)

Сколько процентов светового потока теряется на отражение в призматическом бинокле?
Показатель преломления стекла призм и линз равен 1,5.
Схема бинокля дана на рис. 274.



### О2-2.5. Формулы Френеля с магнитной проницаемостью

При выводе формул Френеля предполагается, что магнитная проницаемость среды $\mu = 1$. Как изменятся формулы Френеля, если не вводить этого предположения?

(?? ну теорию переучить нужно, что сказать???)

### О2-2.7 Азимут преломленной волны (???)

Угол между плоскостью колебаний поляризованного света и плоскостью падения называется азимутом колебания. Найти азимут преломленной волны $\delta$ и азимут отраженной волны $\rho$, если азимут падающей волны $\alpha$, а угол падения $\varphi$.

(??? по идее простой же должна быть она?)

### О2-2.8(3 мин) Отражение от стекла

1) Найти угол полной поляризации для света, отраженного от стекла с показателем преломления $n = 1,5$. 2) Найти степень поляризации преломленного света $\Delta = \frac{\mathcal{I}_\perp - \mathcal{I}_\parallel}{\mathcal{I}_\perp + \mathcal{I}_\parallel}$ при падении света под тем же углом к плоскости, перпендикулярной плоскости падения, $\mathcal{O}_\parallel$−поляризованного в плоскости падения.) Падающий свет естественный.

### О2-2.9 Максимальная поляризация (??)

Под каким углом нужно отразить луч от кристалла каменной соли ($n = 1,544$), чтобы получить максимальную поляризацию отраженного луча? Падающий свет - естественный.

### О2-2.10 Поляризация после отражения от стекла

Найти степень поляризации света, отраженного от поверхности стекла под углами $0°, 45°, 56°51'$ и $90°$ (показатель преломления стекла $n = 1,53$). Падающий свет естественный.

### О2-2.14 Коэффициент отражения при обратном падении

Свет падает из среды $I$ на среду 2 под углом  и преломляется под углом $\psi$. Доказать, что коэффициент отражения не изменится, если свет будет падать из среды 2 на среду I под углом $\psi$.

### О2-2.15* Изменение амплитуды в отражении при наличии диэлектрика (???)

На плоскую поверхность отражающей среды нанесен слой прозрачного диэлектрика толщиной $l$ с показателем преломления $n$. Падающая волна поляризована либо в плоскости падения, либо перпендикулярно к ней. Найти связь между комплексными амплитудами $E$ и $R$ падающей и отраженной волн, если коэффициенты Френеля на верхней и нижней границах слоя (для прямого прохождения света) равны соответственно $r_1$ и $r_2$.

(?? что это за либо так, либо так?? мб изменю условие задачи.)

### О2-2.16 Зеркало с слоем диэлектрика не меняет амплитуду

Пользуясь формулами Френеля, показать, что при отражении плоской электромагнитной волны от идеального зеркала, покрытого сверху слоем прозрачного диэлектрика, амплитуда отраженной волны равна амплитуде падающей. Изменяется лишь фаза волны, как это и должно быть согласно закону сохранения энергии.

### О2-2.17 Прохождение через две пластины

Имеются две параллельные полупрозрачные плоскости. Коэффициенты отражения и пропускания первой из них равны $\rho_1$ и $\tau_1$, а второй - $\rho_2$ и $\tau_2$ соответственно. Степень монохроматичности падающего света невелика, так что интерференции не происходит, а имеет место сложение интенсивностей света. Найти коэффициенты отражения $\rho$ и пропускания $\tau$ для совокупности обеих плоскостей.

### О2-2.18 Прохождение через $m$ пластин

Имеется $m$ параллельных полупрозрачных плоскостей. Коэффициенты отражения и пропускания каждой из них равны $\rho$ и $\tau$. Найти коэффициент отражения $\rho_m$ и коэффициент пропускания $\tau_m$ всей системы $m$ плоскостей. Степень монохроматичности падающего света невелика, так что интерференции не происходит, а имеет место сложение интенсивностей света.

## О2-2.19 Стопа Столетова

Стопа Столетова состоит из плоскопараллельных стеклянных пластинок с показателем преломления $n = 1,5$. На нее под углом Брюстера падает свет, поляризованный в плоскости падения. Начертить график для коэффициентов отражения и пропускания стопы в зависимости от числа $N$ пластинок.

## ФР. Параметры света от пластины с помощью формул Френеля (????!!)

Естественный свет падает под углом Брюстера на поверхность стекла. Определить с помощью формул Френеля:
а) коэффициент отражения $\rho$ (отношение отраженной интенсивности к падающей);
б) степень поляризации преломленного света (отношение разностей интенсивностей в перпеникулярных направлениях к их сумме).

### Solution

(а) Из уравнений Френеля $\begin{cases} I'_\perp = I_\perp \frac{\sin^2(\theta_1-\theta_2)}{\sin^2(\theta_1+\theta_2)} \text{ - по углу Брюстера} \\ I'_\parallel = 0 \end{cases}$

$$I'_\perp = I_\perp \sin^2(\theta_1 - \theta_2) = \frac{1}{2} I \left(\sin\theta_1 \cos\theta_2 - \cos\theta_1 \sin\theta_2\right)^2$$

$$\operatorname{tg}\theta_1 = n, \qquad \sin\theta_1 = \frac{n}{\sqrt{n^2+1}}, \qquad \cos\theta_1 = \frac{1}{\sqrt{n^2+1}}, \qquad \sin\theta_2 = \cos\theta_1$$

$$I'_\perp = \frac{I}{2}\left(\frac{n^2-1}{n^2+1}\right)^2$$

Таким образом, при $n = 1,5$ коэффициент отражения $\rho = \frac{I'_\perp}{I} = \frac{1}{2}\left(\frac{n^2-1}{n^2+1}\right)^2 = 0,074$.

(б) Для преломленного света по углу Брюстера

$$I''_\perp = I_\perp - I'_\perp = \frac{1}{2} I \left(1 - \left(\frac{n^2-1}{n^2+1}\right)^2\right) = \frac{1}{2} I \frac{4n^2}{\left(n^2+1\right)^2}$$

$$I''_\parallel = \frac{I}{2}$$

Степень поляризации преломленного света:

$$P := \frac{I''_\parallel - I''_\perp}{I''_\parallel + I''_\perp} = \frac{\left(n^2+1\right)^2 - 4n^2}{\left(n^2+1\right)^2 + 4n^2} = \frac{\left(n^2-1\right)^2}{2\left(n^2+1\right)^2 - \left(n^2-1\right)^2} = \frac{\rho}{1-\rho}$$

$$P(\rho = 0,074) = 0,08.$$

## О2-2.20 Поляризация преломленного луча

Найти степень поляризации преломленного луча по выходе его из стеклянной пластинки с показателем преломления $n = 1,5$ при углах падения $20, 45, 60$ и $80°$. Падающий свет естественный.

## О2-2.21 Отступления от формул Френеля (???)

На опыте наблюдаются отступления от формул Френеля для отражения света от прозрачных изотропных сред, которые в основном сводятся к двум: 1) не существует угла полной поляризации, при котором свет с электрическим вектором, лежащим в плоскости падения, не отражается совсем; 2) при отражении линейно поляризованного света, плоскость колебаний которого не совпадает с плоскостью падения и не перпендикулярна к ней, получается эллиптическая поляризация, особенно отчетливо заметная в окрестности угла Брюстера (т. е. угла падения $\varphi = \operatorname{arctg} n$). Показать, что каждое из этих отступлений является следствием другого.

## О2-2.22 угол полной поляризации при магнитных проницаемостях (???)

Будет ли существовать угол полной поляризации, если магнитные проницаемости $\mu_1$ и $\mu_2$ граничащих сред отличны от единицы?

### O2-2.23 Нулевая отражательная способность (????)

Показать, что отражательная способность среды, для которой $\varepsilon = \mu$ ($\varepsilon$ - диэлектрическая, $\mu$ - магнитная проницаемости), равна нулю.

### O2-2.24

Рассчитать преломляющий угол параллелепипеда Френеля, сделанного из стекла с показателем преломления $n = 1,7$.

### O2-2.25

Азимут колебаний падающей линейно поляризованной волны равен $+45°$. Какая получится эллиптическая поляризация отраженного света на границе стекло-воздух, правая или левая?

### O2-2.26

При каком угле падения $\varphi$ разность фаз $\delta$ между компонентой отраженной волны, поляризованной в плоскости падения, и компонентой, поляризованной перпендикулярно к ней, достигает максимума при полном внутреннем отражении, если падающая волна линейно поляризована? Чему равен этот максимум?

### O2-2.27

Каким показателем преломления $n$ должно обладать вещество, чтобы при помощи однократного полного внутреннего отражения на границе его с воздухом можно было превращать линейно поляризованный свет в поляризованный по кругу? Азимут колебаний падающего света равен $45°$.

### O2-2.28 Изменение поляризаций отражением (???)

Падающий свет поляризован линейно с азимутом колебаний $+45°$. Можно ли путем однократного отражения превратить его в свет, поляризованный по правому кругу?

(нужно шаблон про это заготовить...)

### O2-2.29 Изменение поляризаций отражением (???)

Каким должен быть минимальный показатель преломления параллелепипеда Френеля, чтобы при азимуте колебаний падающего света в $+45°$ выходящий свет был поляризован по правому кругу?

### O2-2.30 Минимум коэффициента отражения

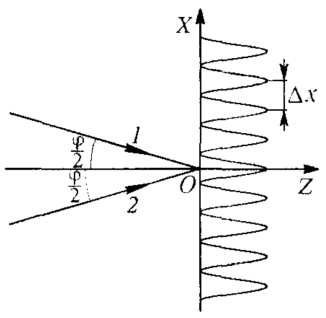
Каким должен быть показатель преломления среды, чтобы коэффициент отражения естественного света имел минимум при угле падения между 0 и 90°?

(?? какая вообще формула для него? хз...)

## 12.0.7 Problems about interference of monochromatic light

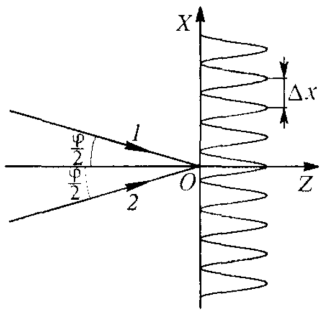
### O2-3.1* Обычная интерференция (???)

Направления распространения двух плоских волн одной и той же длины $\lambda$ составляют угол $\varphi/2$ с нормалью к плоскости экрана, на котором наблюдаются интерференционные полосы (рис. 276). Показать, что при малых $\varphi$ расстояние $\Delta x$ между соседними интерференционными полосами $\Delta x \approx \lambda/\varphi$.



## O2-3.2

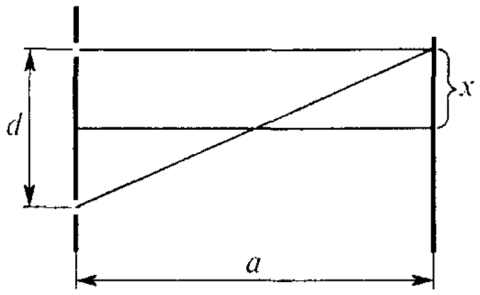
Направления распространения двух плоских волн одной и той же длины $\lambda$ составляют угол $\varphi/2$ с нормалью к плоскости экрана, на котором наблюдаются интерференционные полосы (рис. 276).




при малых $\varphi$ расстояние $\Delta x$ между соседними интерференционными полосами $\Delta x \approx \lambda/\varphi$. Как изменится это выражение, если интерферирующие лучи падают на экран не симметрично относительно нормали?

## O2-3.3 Длина волны в опыте Юнга по расстоянию первого максимума

Найти длину волны $\lambda$ монохроматического излучения, если в опыте Юнга расстояние первого интерференционного максимума от центральной полосы $x = 0,05$ cm. Данные установки: $a = 5$ m, $d = 0,5$ cm.




## O2-3.4

На пути одного луча в интерференционной установке Юнга стоит трубка длиной $l = 2$ cm с плоскопараллельными стеклянными основаниями и наблюдается интерференционная картина, когда эта трубка наполнена воздухом. Затем трубка наполняется хлором и при этом наблюдается смещение интерференционной картины на $N = 20$ полос. Вся установка помещена в термостат, поддерживающий постоянную температуру. Наблюдения проводятся со светом линии $D$ натрия ($\lambda = 5890$ Å). Принимая показатель преломления воздуха $n = 1,000276$, вычислить показатель преломления хлора. В какую сторону смещаются полосы интерференции при наполнении сосуда хлором?

## O2-3.8 Biprisma

Найти число полос интерференции $N$, получающихся с помощью бипризмы, если показатель преломления ее $n$, преломляющий угол $\alpha$, длина волны источника $\lambda$. Расстояние от источника света до бипризмы равно $a$, а расстояние от бипризмы до экрана равно $b$.

## O2-3.9

Полосы интерференции получаются с помощью бипризмы Френеля с малым преломляющим углом и щелевого источника света, параллельного ребру бипризмы. Интерференционные полосы наблюдаются на экране, расположенном перпендикулярно к оси установки. Нулевая полоса получается в центре экрана - на оси (точнее, в плоскости симметрии установки. Расстояние от источника до бипризмы равно $a$, от бипризмы до экрана - b. В какую сторону и на какую величину $x$ сместится нулевая интерференционная полоса, если щелевой источник света немного сместить в направлении, перпендикулярном к оси оптической системы, на величину $h$?

### O2-3.10

От двух когерентных источников света $S_1$ и $S_2$ (рис. 279) получена система интерференционных полос на экране $AB$, удаленном от источников на расстояние $a = 2$ м. Во сколько раз изменится ширина интерфсренционных полос, если между источниками и экраном поместить собирающую линзу с фокусным расстоянием $f = 25$ cm? Рассмотреть два случая: 1) расстояние линзы от источников равно $2f; 2)$ источники $S_1$ и $S_2$ находятся в фокальной плоскости линзы.

### O2-3.11

Из линзы с фокусным расстоянием $f = 50$ cm вырезана центральная часть ширина a, как показано на рис. 280. Обе половины линзы сдвинуты до соприкосновения. По одну сторону линзы помещен точечный источник монохроматического света ($\lambda = 6000$ Å). С противоположной стороны линзы помещен экран, на котором наблюдаются полосы интерференции. Расстояние между соседними светлыми полосами $\Delta x = 0,5$m Экрана вдоль оптической оси. Найти $a$.

### O2-3.12

Плоская монохроматическая волна длины $\lambda$ падает перпендикулярно к ребру клина из стекла с углом при вершине $\alpha \ll 1$ (рис. 281). Показатель преломления стекла $n$, угол падения волны $\varphi_1$. Найти расстояние $\Delta x$ между соседними максимумами интерференционных полос на экране, расположенном перпендикулярно к отраженному свету.

### O2-3.13

Свет с длиной волны $\lambda = 6000$ Å падает на тонкую мыльную пленку. Угол падения $\varphi = 30°$. В отраженном свете на пленке наблюдаются интерференционные полосы. Расстояние между соседними 134 полосами $\Delta x = 4$ мм. Показатель преломления мыльной пленки $n = 1,33$. Вычислить угол a между поверхностями пленки.

### O2-3.14

При какой толшине пленки исчезают интерференционные полосы при освещении ее светом с длиной волны $\lambda = 6 \cdot 10^{-5}$ cm′? Показатель преломления пленки $n = 1,5$.

### O2-3.15

В оптических приборах потери света при прохождении через лрибор происходят главным образом вследствие отражения света от поверхностей оптических деталей. Для увеличения поверхностной прозрачности стекла его поверхность покрывают тонкой пленкой, показатель преломления которой меньше показателя преломления стекла. Каковы должны быть толщина пленки и се показатель преломления, чтобы отражательная способность стекла обратилась в ноль?

### O2-3.16

В очень тонкой клиновидной пластинке в отраженном свете при нормальном падении наблюдаются интерференционные полосы. Расстояние между соседними темными полосами $\Delta x = 5$ мм. Зная, что длина световой волны $\lambda = 5800$ Å, а показатель преломления пластинки $n = 1,5$, найти угол $\alpha$ между гранями пластинки.

### O2-3.17

С помощью воздушного клина с углом при вершине $\alpha$ наблюдаются полосы равной толщины в отраженном монохроматическом свете. Свет падает на клин нормально. Найти распределение освещенности $E$ в интерференционной картине на поверхности клина. Считать интенсивности световых пучков, отраженных от обеих поверхностей клина, одинаковыми и равными $\mathcal{I}_0$.

### O2-3.18

Спутник Земли, поднимаясь над горизонтом, излучает радиоволны длины $\lambda = 10$ см. Микроволновый детектор расположен на берегу озера на высоте $h = 1$ м над уровнем воды (рис. 282). Рассматривая поверхность воды как идеальный проводник, определить, при каком угле a спутника над горизонтом детектор зарегистрирует 1-й и 2-й максимумы интенсивности сигнала. Рассмотреть случаи горизонтальной и всртикальной поляризации.

## O2-3.19

Радиоизлучение космического источника длины волны $\lambda$, имсющего угловой размер $\psi$, принимается горизонтальным вибратором, служащим антенной. Вибратор расположен на отвесном берегу на высоте $h$ над уровнем моря. Рассматривая поверхность воды как плоское зеркало, определить, как будет меняться интенсивность 135 принимаемого сигнала в зависимости от угла $\alpha$ возвышения источника над горизонтом. При каких значениях углового размера источника интенсивность принимаемого сигнала не будет зависеть от $\alpha$? Для простоты расчет провести для малых значений а и .

## O2-3.23

На поверхности жидкости с показателем преломления $n_1$ плавает очень тонкая линза с показателем преломления $n < n_1$, геометрические размеры которой показаны на рис. 284. Рассчитать, какая картина будет видна в отраженном монохроматическом свете с длиной волны $\lambda$, если смотреть на линзу сверху.

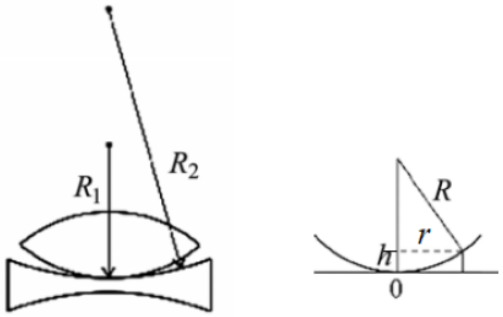
## O2-3.24 Расстояние между 20-м и 21-м кольцами Ньютона (???)

Найти расстояние $\Delta l$ между двадцатым и двадцать первым светлыми кольцами Ньютона, если расстояние между вторым и третьим равно 1 мм, а кольца наблюдаются в отраженном свете.

## ФР.кр-ВШЭ?.вар2-5/5(5 мин) Кольца Ньютона в кривом зеркале

Интерференционные кольца Ньютона наблюдают в отраженном свете с длиной волны $\lambda$ в оптической системе, состоящей из двух тонких симметричных линз (двояковыпуклой и двояковогнутой), которые придвинуты вплотную друг к другу, а между ними находится воздух (см. рис. *а*). Радиусы кривизны линз равны $R_1$ и $R_2$ соответственно (как на рис. *а*). Найти радиус первого светлого кольца $r_1$.

Указание: для интерференционных колец малого порядка расстояние $h$ между сферической поверхностью линзы с радиусом кривизны $R$ и плоской подложкой, на расстоянии $r$ от точки их соприкосновения, примерно равно $\frac{r^2}{2R}$.




*а* *б*

### Solution

Разность хода между лучами, отраженными от нижней сферической поверхности $l$-й линзы и верхней сферической поверхности 2 -й линзы равна $\Delta = 2\,(h_1 - h_2) + \lambda/2$, где $h_1 = \frac{r^2}{2R_1}, h_2 = \frac{r^2}{2R_2}$, а полволны добавлено из-за отражения одного из лучей от более оптически плотной среды.

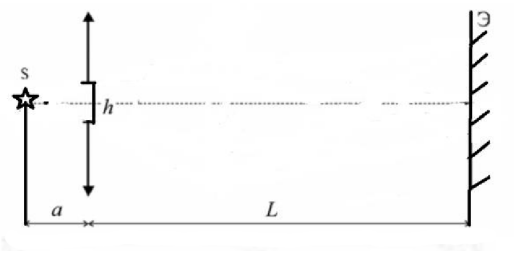
Отсюда: $\Delta = r^2\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \lambda/2$. По условию интерференции для первого максимума получаем: $\Delta = r_1^2\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{\lambda}{2} = \lambda$. Итого: $r_1 = \sqrt{\frac{\lambda}{2}\frac{R_1R_2}{R_2-R_1}}$.

Ответ: $r_1 = \sqrt{\frac{\lambda}{2}\frac{R_1R_2}{R_2-R_1}}$

Задача самая типичная, но важно прийти к верному ответу (включая знаки в нем) за 5 минут, поэтому она стоит 5-ти минут.

## ФР.кр-ВШЭ?.вар1-тренир-5/5. Билинза Бийе (!!??)

Интерференция света наблюдается по схеме с билинзой Бийе, которая сделана из двух половинок тонкой собирающей линзы с фокусным расстоянием $f$, раздвинутых на расстояние $h$ (см. рис.). На расстоянии $a$ от нее ($a = 3f/2$) помещен источник света с длиной волны $\lambda$. Экран для наблюдения интерференционных полос установлен с противоположной стороны билинзы на расстоянии $L$ от нее. Найти ширину интерференционной полосы на экране.




**Solution**

Изображение источника в половинке линзы будет на расстоянии $b = \frac{af}{a-f}$ от линзы. Соответственно, расстояние от изображения источника до экрана: $L - b = L - \frac{af}{a-f}$.

Увеличение (по модулю): $\Gamma = \frac{f}{a-f}$.

Расстояние от изображения до оси: $\frac{h}{2}\Gamma + \frac{h}{2} = \frac{h}{2}(\Gamma + 1) = \frac{h}{2}\left(\frac{a}{a-f}\right)$. Расстояние между изображениями источников: $h\left(\frac{a}{a-f}\right)$.

Ширина интерференционной полосы на экране: $\Delta x = \frac{\lambda\left\{L - \frac{af}{a-f}\right\}}{h\left(\frac{a}{a-f}\right)} = \frac{\lambda}{h}\{L(1 - f/a) - f\}$.

Для $f = 3f/2 : \Delta x = \frac{\lambda(L-3f)}{3h}$.

Ответ: $\Delta x = \frac{\lambda(L-3f)}{3h}$

## O2-3.25

Найти фокусное расстояние $f$ плосковыпуклой линзы, примененной для получения колец Ньютона, если радиус третьего светлого кольца равен 1,1 мм, $n = 1,6, \lambda = 5890$ Å. Кольца наблюдаются в отраженном свете.

## O2-3.26

При наблюдении колец Ньютона в отраженном синем свете ($\lambda_c = 4500$ Å) с помощью плосковыпуклой линзы, положенной на плоскую пластинку, радиус третьего светлого кольца оказался равным 1,06 мм. После замены синего светофильтра на красный был измерен 136 радиус пятого светлого кольца, оказавшийся равным 1,77 мм. Найти радиус кривизны $R$ линзы и длину волны $\lambda_{\text{кр}}$ красного света.

## O2-3.27

Плоскопараллельная стеклянная пластинка лежит на одной из поверхностей двояковыпуклой линзы. При наблюдении колец Ньютона в отраженном свете натриевой горелки ($\lambda = 5890$ Å) найдено, что радиус темного кольца порядка $m = 20$ (центральному темному кольцу соответствует $m = 0$) равен $r_1 = 2$ мм. Когда пластинка была положена на другую поверхность линзы, радиус темного кольца того же порядка сделался равным $r_2 = 4$ мм. Определить фокусное расстояние линзы, если показатель преломления стекла, из которого она изготовлена, $n = 1,5$.

## O2-3.28

Найти радиус $r$ цснтрального темного пятна колец Ньютона, ссли между линзой и пластинкой налит бснзол ($n = 1,5$). Радиус кривизны линзы $R = 1$ м. Показатсли претомления линзы и пластинки одинаковы. Наблюдение ведется в отраженном натриевом свете ($\lambda = 5890$ Å).

## O2-3.29

Кольца Ньютона получаются с помощью плосковыпуклой линзы с радиусом кривизны $R_1$, положенной на вогнутую сферическую поверхность с радиусом кривизны $R_2 > R_1$. Кольца наблюдаются в отраженном свете. Определить радиус $r_m m$-го темного кольца, если длина световой волны равна $\lambda$.

## O2-3.30

Кольца Ньютона получаются между двумя плосковыпуклыми линзами, прижатыми друг к другу своими выпуклыми поверхностями. Найти радиус $r_m m$-го темного кольца, если длина световой волны равна $\lambda$, а радиусы кривизны выпуклых поверхностей линз равны $R_1$ и $R_2$. Наблюдение ведется в отраженном свсте.

### O2-3.31

Тонкая симметричная двояковыпуклая линза сложена с
тонкой симметричной двояковогнутой линзой так, что в некоторой точке они соприкасаются. Показатель преломления обеих линз $n = 1,6$. Наблюдается интерференционная картина в отраженном свете на длине волны $\lambda = 0,6$ мкм. Определить фокусное расстояние $f$ системы линз, если радиус 5 -го светлого кольца $r = 2$m.

### O2-3.32

Интерференционная картина (кольца Ньютона) наблюдается в проходяшем свете (рис. 285). Показатель преломления линзы и пластинки равен $n = 1,5$. Найти отношение интенсивностей $\mathscr{I}_{\max}/_{\min}$ света в максимуме и минимуме интерференционной картины. Можно ли увидеть картину глазом, если контрастная чувствительность глаза равна $0,05$?

## 12.0.8 Problems about quasi-monochromatic light, temporal coherence

### O2-4.1 Newton's rings from two waves

Describe how the sharpness of the Newton rings will change when their numbers are increased from the central (first) ring, as well as when the Newton rings are removed from the plate behind it. The rings are observed in reflected light of the $D$-line Na, which It is not monochromatic, but represents two close spectral lines with $\lambda_1 = 5890$ Å и $\lambda_2 = 5896$ Å.

#### Solution

(a basic problem, I'll make a solution much shorter!)
The two wavelengths correspond to two systems of Newton's rings with slightly different sizes. If the lens is in contact with the surface of the plate, then in the center of the picture the light (dark) rings of one system practically coincide with the light (dark) rings of the other system. Therefore, near the center, the rings are visible almost as sharply as in monochromatic light. However, at some distance from the center, the light ring of one system may coincide in position with the dark ring of another system. Newton's rings will not be visible at the appropriate location, but they will be indistinctly visible in the vicinity of this location.
Let's determine the number of the $N$th light ring for the wavelength $\lambda_2$, which coincides in position with $(N+1)$*The*-m ring for the wavelength of $\lambda_1$. The first dark ring (more precisely, the central dark spot) for the wavelength of $\lambda_1$ corresponds to the stroke difference of $\lambda_1/2$, the second dark ring corresponds to the stroke difference of $\lambda_1 + \lambda_1/2$, etc., finally $(N+1)$*The*th dark ring is the move difference of $N\lambda_1 + \lambda_1/2$.
The same optical path of $N\lambda_1 + \lambda_1/2$ should obviously be equal to $N\lambda_2$, since there should be an overlap of the $N$-th light ring for the wavelength of $\lambda_2$ by $(N+1)$-th dark ring for the wavelength of $\lambda_1$. So,

$$N\lambda_1 + \frac{\lambda_1}{2} = N\lambda_2, \qquad \Rightarrow \qquad N = \frac{\lambda_1}{2\left(\lambda_2 - \lambda_1\right)} = \frac{5890}{12} \approx 490.$$

It follows that the rings will disappear in the vicinity of the four hundred and ninetieth ring. It is easy to see that they will be sharp again in the vicinity of the $2 \cdot 490 = 980$ ring.
When the lens is removed from the plate, the rings tighten to the center (see the solution to the problem 3.34*). (!!)
If the lens is moved to $490\lambda_1$, then 490 rings will pass through the field of view, and the rings will disappear in the center of the picture. If you move the lens to $2 \cdot 490\lambda_1 = 980\lambda_1$, the rings in the center will be sharp again; if you move to $3 \cdot 490\lambda_1 = 1470\lambda_1$, they will disappear again, etc.

### O2-4.2 Newton's rings from two waves

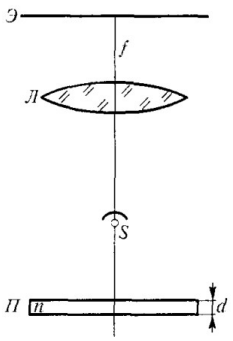
Newton's rings of $D$-lines Na are considered (there are two of them with similar wavelengths). Find the difference between these wavelengths if it is known that the sharpness of the interference pattern observed in a two-beam interferometer is minimal at the 490th, 1470th, etc., and maximal at the 1st, 980th, etc. bands. Average wavelength of $D$ lines $\lambda = 5893$ Å.

#### Solution 4.2.

$\frac{\Delta\lambda}{\lambda} = \frac{1}{980}$; $\quad \Delta\lambda = 6,02$ Å.
(см. задачу 4.1, допишу потом больше, всё ровно это же)

## О2-4.3 Кольца в особой установке

Интерференционные полосы равного наклона в фокальной плоскости линзы $L$ получаются при отражении от плоскопараллельной пластинки П, освещаемой монохроматическим источником света $S$.




Прямой свет источника на линзу не попадает. Длина световой волны $\lambda = 6000$ Å, толщина пластинки $d = 1,6$ мм; показатель преломления $n = 1,5$; фокусное расстояние линзы $f = 40$ cm. Найти радиус $r$ первого видимого на экране Э темного интерференционного кольца, если центр колец темный. Какова максимально допустимая ширина линии $\Delta\lambda$, освещающей пластинку, чтобы при указанных параметрах схемы можно было наблюдать интерференционные кольца?

### Solution 4.

3. $\Delta\lambda = 0,75$ Å.
(??? почему такая особая установка? можно было бы как вседа сделать)
Разность хода между лучами, отразившимися от передней и задней поверхностей пластинки, равна $\Delta = 2dn\cos\psi + \lambda/2$. (это типичный результат такой модели). Так как центр колец темный, то эта величина должна содержать нечетное число полуволн.
Первому темному кольцу соответствует такое приращение угла преломления $\psi$, что разность хода уменьшается на $\lambda$. Это дает $\Delta_{(2)} - \Delta_{(1)} = 2dn(1 - \cos\psi) = \lambda$, или $4dn\sin^2(\psi/2) \equiv \lambda$.
Для малых углов $dn\psi^2 = \lambda$. Малые углы падения и преломления связаны соотношением $\varphi = n\psi$. Таким образом, $\varphi^2 = n\lambda/d$.
Для радиуса первого темного кольца получаем $r = f\varphi = f\sqrt{n\lambda/d} = 9,5$ мм. (?? не уверен, почему так??)
Величина $\Delta\lambda$ находится обычным способом по порядку интерференции, который равен $2dn/\lambda$. Это дает $\Delta\lambda = \lambda^2/(2dn) = 0,75$ Å (?? не уверен, почему так??)

## О2-4.4 Кольца в особой установке

Сколько темных колец $N$ можно наблюдать в условиях предыдущей задачи, если диаметр линзы $D = 8$ cm, а источник $S$ помещен посередине между линзой и пластинкой на расстоянии $f$ от линзы?

### Solution 4.

4: $N = 2$.
Решение. Максимальный угол падения $\varphi_{\max} = \frac{D}{6f}$. (?????)
Поэтому из результатов решения предыдущей задачи получаем $N = \frac{dD^2}{36nf^2\lambda} = 2$.

## О2-4.5 Кольца в особой установке

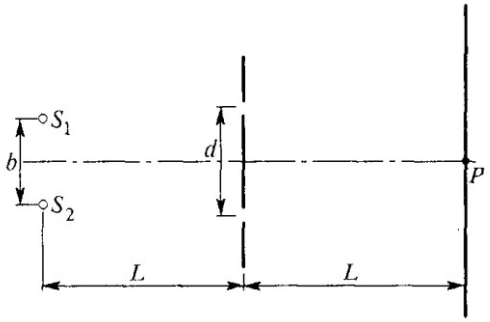
Какова должна быть минимальная толщина пластинки, чтобы в условиях предыдущей задачи можно было получить по крайней мере одно темное кольцо?

### Solution 4.

5. $d_{\min} = \frac{36nf^2\lambda}{D^2} = 0,81$mM.

## O2-4.6 Две излучающие точки

Свет от двух одинаковых некогерентных точечных источников $S_1$ и $S_2$ ($\lambda = 500$ nm, $\Delta\lambda = 50$ nm) падает на непрозрачный экран с двумя отверстиями, расстояние между которыми $d = 1$ cm.



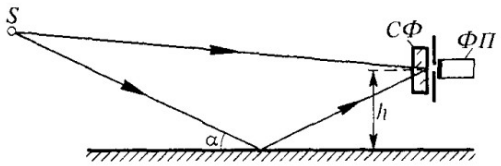

Интерференция света, прошедшего через отверстия, наблюдается вблизи точки $P$, лежащей на оси системы. Источники и точка наблюдения находятся на одинаковом расстоянии $L = 2$ м от экрана. При симметричном удалении источников от оси (т. е. при увеличении расстояния $b$ между источниками) интерференционная картина в окрестности точки $P$ периодически возникает и исчезает. Оценить число периодов восстановления интерференционной картины при увеличении расстояния $b$ от нуля до $d$ ($0 \leqslant b \leqslant d$).

### Solution 4.

6. Если бы свет был монохроматическим, то в конце процесса в точке $P$ порядок оказался бы равным $m = 50$. На самом деле в связи с немонохроматичностью света $m_{\max} = \frac{\lambda}{\Delta\lambda} = 10 < 50$. Наблюдатель увидит всего $N = 20$ полос.

## O2-4.7 Distant source and reflection

The light from a distant point source $S$ falls on the photodetector directly and is reflected from the horizontal plane:



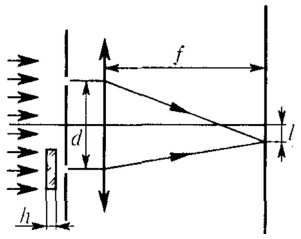

When the source is moved vertically, the photodetector $FI$ registers a change in the intensity of the light incident on it. Estimate the maximum angle $\alpha$ of the source elevation above the horizon at which changes in the photocurrent are still noticeable if an SF light filter with a bandwidth of $\Delta v = 3 \cdot 10^{11}$ Hz is installed in front of the photodetector. The photodetector inlet is located at a height of $h = 1$ cm above the reflecting plane.

### Solution 4.

7. $\alpha \leqslant \frac{c}{2h\Delta\nu} = 0,05$.

## O2-4.8 Closing the slit with a dispersion material

In a Rayleigh interferometer, a plane wave experiences diffraction at two slits. The diffraction pattern is observed in the focal plane of a lens with a focal length of $f = 100$ cm.




One of the slits is closed with a plane-parallel plate of dispersing agent with a thickness of $h = 0.01$ mm with the law of dispersion of $n(\lambda) = A - B\lambda$, where $A$ and $B$ are some constants. In this case, the white (achromatic) stripe is shifted by a distance of $l = 4$ mm. Determine the constant $A$ if it is known that the distance between the slits is $d = 1$ cm

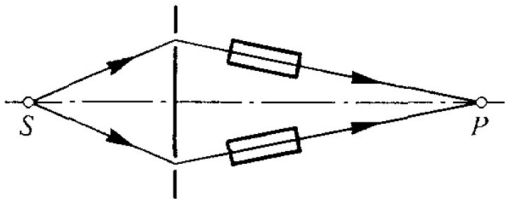
### Solution

$$\delta\Delta = nh$$

(??)
$A = 1 + \frac{dl}{fh} = 5.$

## O2-4.9 Passing through areas with different pressures

The interference of light from two small holes in an opaque screen is observed at the point $P$.




Two identical cuvettes filled with air at the same initial pressure are placed behind the holes in the path of the beams. When the pressure in one of the cuvettes changes, the change in light intensity at point $P$ has an oscillating character. Determine the pressure difference $\Delta P$ of the gas in the cuvettes at which the oscillation amplitude becomes zero if the 1st minimum intensity occurs at a pressure difference $\Delta P_1 = 10^{-3}$ mm Hg.

The radiation spectrum of the point source $S$ is uniform in the band $\Delta\omega$ and has a relative width of $\Delta\omega/\omega = 10^{-5}$.

### Solution 4.

наверное, нужно связать давление с показателем преломления, а его - с разностью хода. (???)
$\Delta p = \Delta p_1 \frac{2\omega}{\Delta\omega} = 200$ мм рт. ст.

## O2-4.10 Stripes on the air wedge

Interference fringes of equal thickness are observed on an air wedge between two glass plates with an angle at the vertex $\alpha = 1'$. The stripes are obtained in the light of a green mercury line with a wavelength of $\lambda = 5461$Åand a width of $\Delta\lambda = 0.1$Å. Identify:

1) the distance $\Delta x$ between two adjacent lanes;

2) the maximum number of $N$ stripes that could be seen on a wedge if its dimensions were not limited;

3) the distance $x$ of the last observed band from the tip of the wedge and the thickness $h$ of the wedge at this point;

4) the maximum allowable angular divergence of the rays $\delta\varphi$, at which it is possible to observe all bands.

### Solution 4.

10. 1) $\Delta x = \frac{\lambda}{2\alpha} = 0,94$ mm.

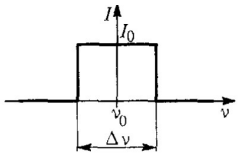
2) Решение. Допустим сначала, что линия ртути - двойная с двумя длинами волн $\lambda_1 = \lambda$ и $\lambda_2 = \lambda + \Delta\lambda$. Пусть на отрезке $x$ от вершины клина укладывается $N$ интерференционных полос с длиной волны $\lambda_1$ и $N - 1/2$ полос с длиной волны $\lambda_2$, т. е. $N\lambda_1 = (N - 1/2)\lambda_2$. Тогда на конце этого отрезка интерференционные максимумы от длины волны $\lambda_1$ наложатся на интерференционные минимумы от длины волны $\lambda_2$ - и интерференционные полосы пропадут. Число $N$ и будет искомым числом полос. Оно равно $N = \frac{\lambda_2/2}{\lambda_2-\lambda_1}$, или, пренебрегая квадратами $\Delta\lambda, N = \frac{\lambda/2}{\Delta\lambda}$. Допустим теперь, что интервал между $\lambda_1$ и $\lambda_2$ непрерывно и равномерно заполнен длинами волн. Тогда всю спектральную линию можно считать состоящей из двух линий шириной $\Delta\lambda/2$ каждая с расстоянием между ними $\Delta\lambda/2$. К этим двум линиям применимы предыдущие рассуждения. Поэтому число полос $N$ найдется из предыдущего результата заменой $\Delta\lambda \to \Delta\lambda/2$, что дает $N = \lambda/\Delta\lambda$. Таким образом, считая линию ртути сплошной, находим $N \approx \lambda/\Delta\lambda \approx 54600$.

3) $x = N\Delta x \approx 51,3$ m; $\quad h = \frac{\lambda^2}{2\Delta\lambda} \approx 14,9$ cm.

4) $\delta\varphi \approx \sqrt{(\Delta\lambda/\lambda)\alpha} \approx 0,25'$.

## O2-4.11 Видность некогерентного шаблонного источника

Определить видность $V$ интерференционной картины от двух точечных источников, спектр излучения которых одинаков и имеет вид:




Как зависит видность $V$ от ширины спектра $\Delta\nu$?

### Solution 4.

11.
(это стандартная задача теории)
$V(\Delta) = \left|\frac{\sin\left(\Delta\frac{\pi\Delta\nu}{c}\right)}{\Delta\frac{\pi\Delta\nu}{c}}\right|$, где $\Delta$ - разность хода.

## O2-4.12 Разрешение двух пучков

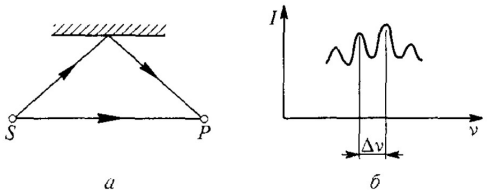
Два пучка белого света, полученные от одного точечного источника, сходятся на входной щели оптического спектрального прибора. Разность хода $\Delta = 300$ м. Оценить разрешающую способность $R$ спектрального прибора, который может обнаружить интерференцию этих пучков.

### Solution 4.

12. $R \geqslant \frac{\lambda}{\Delta\lambda} = m = \frac{\Delta}{\lambda} \approx 5 \cdot 10^8$.

## O2-4.13 Horizontal mirror

Two beams of white light from the same source arrive at the observation point $P$ (Fig. a) with a travel difference of $\Delta$.




*а* *б*

A high-resolution spectroscope is used to study the energy distribution in the spectrum of vibrations occurring at the point $P$ when both beams are superimposed. It turned out that alternating maxima and minima of spectral intensity $I(v)$ are observed, and the frequency interval between neighboring maxima is $\Delta v = 10$ MHz (Fig. b). Determine the difference in the course of $\Delta$.

### Solution 4.

13.
(я так понял, мы пишем просто формулу для интенсивности как функции разности хода, дальше очевиден ответ из неё. просто свет от двух источников, шаблонная задача)
$\Delta = \frac{c}{\Delta v} = 30$ m.

## O2-4.14

Монохроматическое излучение проходит через интерферометр Маха-Цандера, в одном из плеч которого расположена кювета с газом длиной $l$. В кювете создается избыточное давление. При этом показатель преломления газа в нем изменяется по закону $n(t) = 1 + \alpha t$. Определить спектр колебаний тока фотоприемника, расположенного в области нулевой полосы интерферометра.

### Solution 4.

14.
(??? я не понял, что за интерферометр такой???)
$\mathcal{I} \propto 2a^2\left[1 + \cos\left(\frac{\omega}{c} l\alpha t\right)\right]$. Таким образом, спектр тока фотоприемника содержит постоянную составляющую и равную ей по амплитуде переменную составляющую на частоте $\Omega = \frac{\omega}{c} l\alpha$.

## O2-4.15 Три излучающие точки

Три синфазных излучателя $l, 2, 3$ расположены вдоль прямой. Расстояние между излучателями $I$ и 2 равно $\lambda/2$, а между излучателями 2 и 3 в полтора раза больше. Амплитуды излучателей 1 и 2 одинаковы. Какова должна быть амплитуда излучателя 3, чтобы в диаграмме направленности системы существовали минимумы нулевой интенсивности? Найти направления на эти минимумы. Рещение дать аналитически и с помощью векторной диаграммы.

### Solution 4.

15. Амплитуда излучателя 3 должна быть в $\sqrt{2}$ раз больше амплитуд излучателей $I$ и 2 . Минимумы нулевой интенсивности направлены под углами $\theta = \pm 60^\circ$ к линии источников $I - 2 - 3$.

## O2-4.16 Двигающееся зеркало в интерферометре Майкельсона

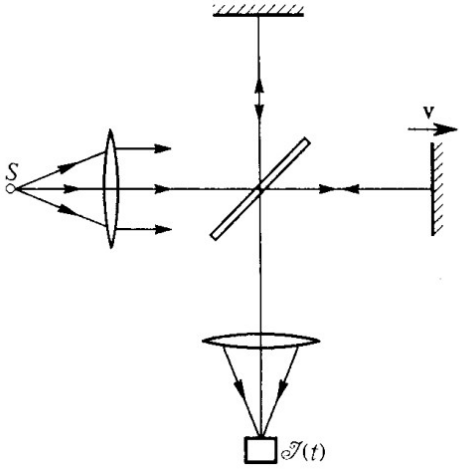
В одно из плеч интерферометра Майкельсона вместо отражающего зеркала помещена непоглощающая пластина с полупрозрачной передней и зеркальной задней стенкой (рис. 295). Толщина пластины $d =$ 2 mm, показатель преломления $n = 5$, спектр падающего излучения простирается от 0 до 110 ГГц. При перемещении зеркала во втором плече детектор регистрирует ряд пиков интенсивности излучения. Каково расстояние между пиками в единицах длины перемещения зеркала?

### Solution 4.

16. Длина когерентности $l_{\text{ког}} \approx \frac{c}{\Delta v} \approx 0,3$с. Поэтому будет виден лишь нулевой порядок: вначале при отражении от передней грани при $\Delta x = 0$, затем от задней грани при $x = 2dn = 2$ см. Затем будут следующие порядки интерференции убывающей интенсивности.

## O2-4.17 Двигающееся зеркало в интерферометре Майкельсона

В фурье-спектрометре, служащем для исследования спектрального излучения, одно из зеркал интерферометра Майкельсона перемещается со скоростью $v = 0,1$ см/с




Какова зависимость тока фотоприемника от времени $\mathcal{I}(t)$, если излучение содержит две спектральные линии $\lambda_1 = 500$ нм и $\lambda_2 = \lambda_1 + \delta\lambda$, где $\delta\lambda = 0,02$ нм, с отношением интенсивностей $I_2/I_1 = 0,5$?

Оценить минимальное время измерения, необходимое для разрешения этих линий. Нарисовать график зависимости тока от времени.

### Solution 4.

17. Ток фотоприемника $\mathcal{I} \propto 1,5I_1 + I(t)\cos\left(\frac{4\pi}{\lambda}vt\right)$,

где $I^2(t) = I_1^2 + I_2^2 + 2I_1I_2\cos\Delta\Omega t, \Delta\Omega = \frac{4\pi\delta\lambda}{\lambda^2}v$. Частота несущей $\Omega = \frac{4\pi}{\lambda}v$; период биений $\tau = \frac{2\pi}{\Delta\Omega} = 6,25$с; $T = \frac{2\pi}{\Omega} = 2,5\cdot10^{-4}$с. Время измерений должно быть $\sim$ 6с. График зависимости тока фотоприемника $\mathcal{I}(t)$ изображен на рис. 431.

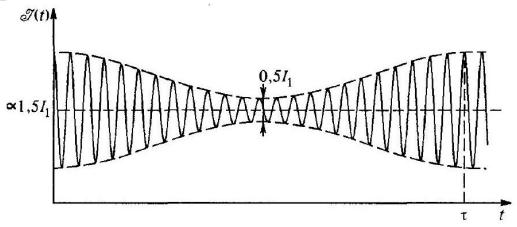
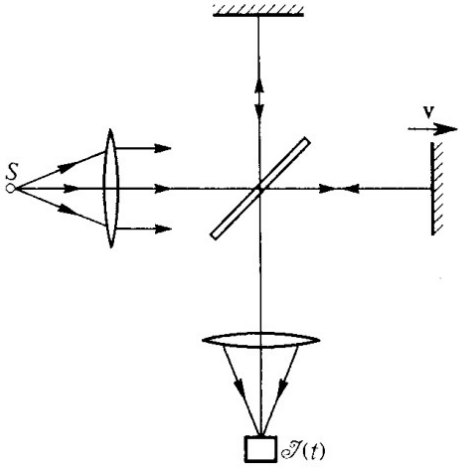



Рис. 431

(??? don't know, why it is so???)

## O2-4.18 Двигающееся зеркало в интерферометре Майкельсона

Для исследования спектрального состава излучения источника $S$ одно из зеркал интерферометра Майкельсона перемещается со скоростью $v = 0,01$ cm/c.




Какова зависимость тока фотоприемника от времени $\mathscr{Y}(t)$, если источник излучает на длине волны $\lambda = 500$ нм, причем ширина спектральной линии $\Delta\lambda = 0,01$ нм? Спектральная интенсивность внутри

спектральной линии $I(\omega) = I_0 = \text{const}$. Оценить минимальное время, необходимое для изучения спектрального состава излучения. Нарисовать график зависимости тока от времени.

### Solution 4.

18. $\mathcal{I}(t) \propto I_0 \Delta\omega \left[1 + \frac{\sin(2\pi ft)}{2\pi ft} \sin(2\pi vt)\right]$,

где $v = \frac{2v}{\lambda} = 400$ Гц; $f = \frac{\Delta\lambda}{\lambda^2} v = \frac{1}{250}$ Гц. Время измерения $\tau = \frac{1}{2f} = 125$с. График зависимости тока от времени изображен на рис. 432.

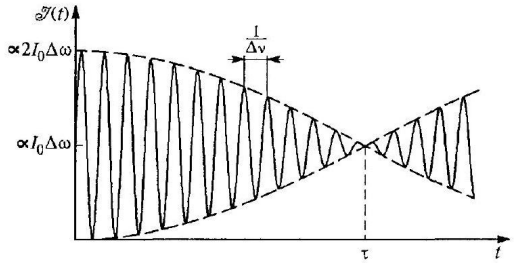
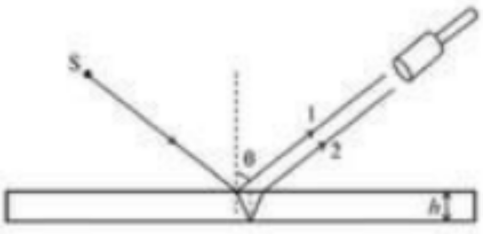



Рис. 432

### ФР.кр-ВШЭ?.вар1-5/5(5 мин) Снова полосы равного наклона

Интерференционные полосы равного наклона наблюдают с помощью телескопической трубы:




Для этого служит тонкая плоскопараллельная стеклянная пластинка с показателем преломления $n$. Длина волны падающего (монохроматического) света равна $\lambda$. Угол падения $\theta$ может изменяться от 0 до $\pi/2$. Найти толщину пластины $h$, если разность между максимальным и минимальным порядками интерференции равна $\Delta m$.

Указание: порядком интерференции называется разность хода между лучами 1 и 2 (см. рис.), деленная на длину волны падающего света $\lambda$.

### Solution

Эта типичная модель, про которую известно, что разность хода между лучами 1 и 2 :

$$\Delta(\theta) = 2h\sqrt{n^2 - \sin^2\theta} + \lambda/2.$$

Соответственно: $\Delta(0) = 2hn + \lambda/2, \Delta\left(\frac{\pi}{2}\right) = 2h\sqrt{n^2-1} + \lambda/2 \Rightarrow 2h\left(n - \sqrt{n^2-1}\right) = \lambda\Delta m.$

Ответ: $h = \frac{\lambda\Delta m}{2\left(n-\sqrt{n^2-1}\right)}$.

## 12.0.9 Problems about spatially extended light sources

### О2-5.1 Опыт Юнга с протяженным источником

Найти видность $V$ интерференционной картины в опыте Юнга при использовании протяженного источника света. Размер источника света $b$, расстояние от источника до экрана со щелями $L$, расстояние между щелями $d$. Средняя длина волны $\lambda = 500$ нм. Сравнить результат с задачей 4.11.

**Solution 5.**

1. $V(b) = \frac{\mathcal{I}_{\max} - \mathcal{I}_{\min}}{\mathcal{I}_{\max} + \mathcal{I}_{\min}} = \left| \frac{\sin\left(\frac{\pi b d}{\lambda R_0}\right)}{\frac{\pi b d}{\lambda R_0}} \right|$.

Решение. На рис. 433 изображена схема интерферометра Юнга.

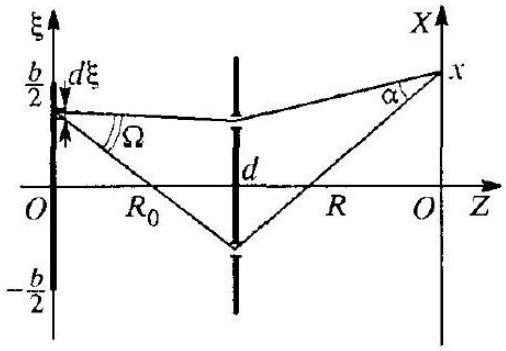



Рис. 433

Протяженный источник, размер которого равен $b$, освещает экран со щелями, находящийся на расстоянии $R_0$ от источника, расстояние между щелями равно b. Плоскость наблюдения расположена на расстоянии $R$ от интерферометра. Проведем ось $\xi$ вдоль по протяженному источнику, а ось $OX$ по экрану. Обозначим интенсивность от единицы длины протяженного источника через $I_0$. Тогда элемент $d\xi$ источника дает интенсивность $I_0 d\xi$. Пусть этот элемент (точечный источник света) размещен на расстоянии $\xi$ от центра источника, тогда интенсивность на экране в точке $x$, очевидно, будет равна

$$d_\xi = 2I_0 d\xi (1 + \cos k\Delta)$$

где $k = 2\pi/\lambda$, а разность хода $\Delta \approx \Omega\xi + \alpha x$, где в свою очередь угол $\Omega$, называемый апертурным углом или апертурой интерференции, равен $\Omega \approx d/R_0$, а угол схождения лучей $\alpha \approx d/R$.

Все элементы протяженного источника светят некогерентно и каждый из них дает свою интерференционную картину. Осталось сложить эти картины от каждого из элементов

$$\mathcal{I}(x) = \int_{-b/2}^{b/2} d\mathcal{I}_\xi = 2I_0 \int_{-b/2}^{b/2} d\xi \left\{ 1 + \cos\left[ \frac{2\pi}{\lambda} \left( \frac{d}{R_0}\xi + \frac{d}{R}x \right) \right] \right\}$$

Поскольку период интерференционной картины $\Lambda \approx \frac{\lambda}{a} = \frac{\lambda R}{d}$, то введя это обозначение, получим

$$\mathcal{I}(x) = 2I_0 \int_{-b/2}^{b/2} d\xi \left\{ 1 + \cos\left[ \frac{2\pi}{\Lambda} \left( x + \frac{R}{R_0}\xi \right) \right] \right\} \qquad q := \frac{2\pi}{\Lambda}\frac{R}{R_0}.$$

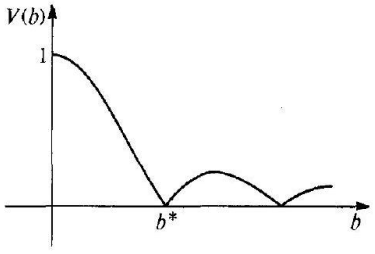
$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$,

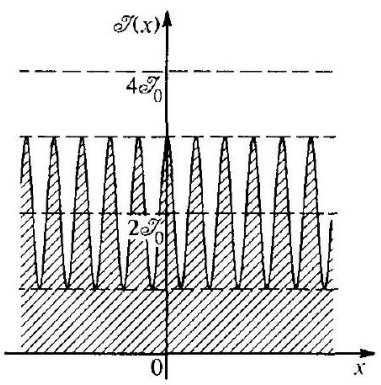
$$\mathcal{I}(x) = 2I_0 b + 2I_0 \cos\frac{2\pi}{\Lambda}x \int_{-b/2}^{b/2} \cos q\xi d\xi = 2\mathcal{I}_0 \left( 1 + \frac{\sin qb/2}{qb/2} \cos\frac{2\pi}{\Lambda}x \right).$$

$$V(b) = \frac{\mathcal{I}_{\max} - \mathcal{I}_{\min}}{\mathcal{I}_{\max} + \mathcal{I}_{\min}} = \left| \frac{\sin(qb/2)}{qb/2} \right|; \qquad \frac{qb}{2} = \frac{\pi b d}{\lambda R_0} = \frac{\pi\Omega}{\lambda b}.$$

$$V(b) = \left| \frac{\sin\frac{\pi\Omega}{\lambda/b}}{\frac{\pi\Omega}{\lambda/b}} \right|.$$

Если изменять размер $b$ источника, оставляя неизменной базу интерферометра, то при некотором значении $b^*$ видность картины превратится в ноль, а экран будет просто равномерно засвечен с интенсивностью $2\mathcal{I}0$. Это предельное значение $b^* = \frac{\lambda}{\Omega} = \frac{\lambda R_0}{d}$. Обычно полагают, что при $b \geqslant b^* V = 0$.



And если изменять only базу интерферометра $d$, то легко получить следующее неравенство. Пусть $\psi = b/R_0$ - угловой размер источника, каким мы его видим из места расположения интерферометра. Тогда осцилляции интенсивности на экране будут наблюдаться при условии, что

$$d \leqslant \frac{\lambda R_0}{b} = \frac{\lambda}{\psi}$$

Эту предельную величину $\lambda/\psi$ называют размером области когерентного освещения или размером пятна пространственной когерентности. Величину $\lambda/(2\psi) = \rho_{\text{ког}}$ часто называют радиусом пространственной когерентности.

## О2-5.2 Интерференция от Солнца (????)

На экран с двумя узкими параллельными щелями падают лучи непосредственно от Солнца. При каком расстоянии $D$ между щелями могут наблюдаться интерференционные полосы за экраном? Угловой диаметр Солнца $\alpha \approx 0,01$ рад.

### Solution 5.

The setup is exactly the same as for basic Yung's experiment, but with spartially big source, so we use the formula $d \leq \lambda/\psi$, so the answer is
$D < \frac{\lambda}{\alpha} \approx 0,05$ mm.

## О2-5.3 Интерференция от Солнца (????)

Изображение Солнца получено при помощи линзы с фокусным расстоянием $f = 50$ мм на отверстии экрана (размер отверстия равен величине изображения). За экраном помещены две узкие параллельные щели на расстоянии $D = 1$ мм друг от друга. При каком расстоянии $l$ между экраном и щелями могут наблюдаться интерференционные полосы?
Угловой диаметр Солнца $\alpha \approx 0,01$ рад.

### Solution 5.

3. $l > \frac{fD\alpha}{\lambda} \approx 100$ cm.

## me. Interference from street light

(I'll write same question as before but for a flashlight)

## О2-5.4

Какому условию должны удовлетворять размеры источника света, чтобы могли наблюдаться интерференционные полосы в установке с 1) зеркалами Френеля, 2) зеркалом Ллойда?
Различные точки источника излучают некогерентно.

### Solution 5.

4. 1) Пусть источником света служат две одинаковые некогерентные светящиеся точки $S_1$ и $S_2$ (или светящиеся линии, параллельные прямой, вдоль которой пересекаются плоскости зеркал Френеля), отстоящие друг от друга на расстояние $l$. Их мнимые изображения $S_1'$ и $S_1'', S_2'$ и $S_2''$ расположены так, как показано на рис. 436. Интерференционная картина на экране $AB$ от пары мнимых источников $S_1'$ и $S_1''$ будет смещена относительно интерференционной картины от другой пары источников $S_2'$ и $S_2''$ на расстояние $O_1O_2$, равное $l$. Если $l = \lambda/(2\varphi)$, где $\varphi$ - угловое расстояние между источниками, то максимумы одной интерференционной картины наложатся на минимумы другой, и интерференционные

полосы пропадут. Если источником служит светящаяся полоска

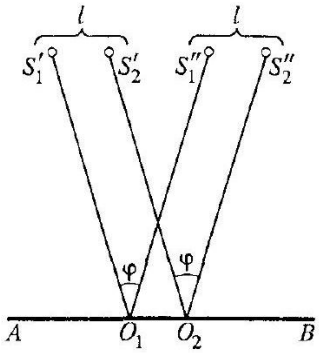



Рис. 436 шириной $l$, параллельная линии пересечения зеркал, то ее можно мысленно разбить на пары светящихся линий, расстояние между которыми равно $l/2$. В этом случае можно применить рассуждение, приведенное выше, заменив $l$ на $l/2$. Полосы пропадут, когда $l = \lambda/\varphi$. Если источник имеет неправильную форму, то для получения резких интерференционных полос его поперечные размеры в направлении, параллельном линии, соединяющей соответственные изображения в зеркалах Френеля, должны быть малы по сравнению с $\lambda/\varphi$.
2) Размер источника в направлении, перпендикулярном к плоскости зеркала, должен быть мал по сравнению с расстоянием от источника до зеркала.

## O2-5.5

Интерференционная картина наблюдается с помощью бипризмы Френеля (преломляющий угол $\alpha = 20'$, показатель преломления $n = 1,5$). Экран и источник света ($\lambda = 600$ nm, $\Delta\lambda = 20$) на ходятся на одинаковом расстоянии от бипризмы. Оценить число интерференционных полос, которые будут видны на экране. На каком расстоянии от центра интерференционной картины интерференционные полосы размываются? Каков допустимый размер источника, при котором можно наблюдать все интерференционные полосы?

### Solution 5.

5. Число наблюдаемых полос $N \approx 60$; полосы размываются на расстоянии от центра $x = 0,6$ см; допустимый размер источника $\approx 2 \cdot 10^{-2}$ cm.

## O2-5.6

Перед линзой Л (рис. 297) установлена плоскопараллельная стеклянная пластинка П, перпендикулярная к главной оптической оси и освещаемая монохроматическим светом от протяженного источника. Описать интерференционную картину в фокальной плоскости линзы. Как изменится эта картина при наклоне пластинки на угол $\alpha = 10^\circ$ (по отношению к исходному положению)? Фокусное расстояние линзы $f = 30$ см.

### Solution 5.

6. Кольца равного наклона с центром $F$. При наклоне пластинки кольца становятся эллиптическими. Центр картины в точке схождения параллельных лучей, падающих на пластинку нормально. При наклоне пластинки на угол $\alpha$ центр картины смещается на $x = f \operatorname{tg}\alpha \approx f\alpha \approx 5,3$ см.

## O2-5.7 Билинза

Билинза Бийе изготовлена из двух половинок тонкой собирающей линзы с фокусным расстоянием $f = 10$ см. На расстоянии $x = 3f/2$ от нее помещен источник света в виде щели, освещаемой широкоугольным пучком света с длиной волны $\lambda = 5790$ Å. Экран для наблюдения интерференционных полос установлен с противоположной стороны билинзы на расстоянии $L = 330$ см от нее. При какой минимальной ширине щели $b$ интерференционные полосы на экране пропадут? Считать, что различные точки щели излучают линзы $a = 0,5$ мм.

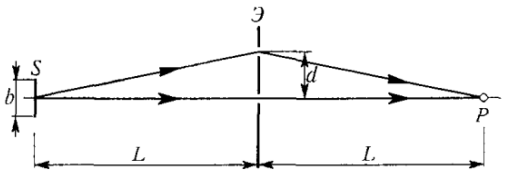
### Solution 5.

7. Первое исчезновение полос будет при ширине щели

$$b = \frac{\lambda}{a}\frac{x-f}{x}\frac{(L-f)x - Lf}{x} = \frac{L-3f}{6a}\lambda = 0,58 \text{ mm}.$$

## O2-5.8 Определение размера источника по интерференции

Свет от протяженного монохроматического источника $S$ падает на непрозрачный экран Э, в котором имеются два маленьких отверстия. Интерференция света, прошедшего через отверстия, наблюдается в точке $P$.




Источник света $S$ и точка $P$ находятся на одинаковом расстоянии $L$ от экрана. При увеличении расстояния $d$ между отверстиями изменение интенсивности в точке $P$ имеет осциллирующий характер. Определить линейный размер $b$ источника света, если 1 -й минимум интенсивности в точке $P$ наблюдается при $d = d_1 = 1$ см, а амплитуда осцилляций становится равной нулю при $d = d_2 = 20$ см (условие $d \ll L$ выполняется всегда).

### Solution 5.

8. answer: $b = \frac{2d_1^2}{d_2} = 0,1$ cm.

$$I(x) = 2I_0 \left\| 1 + \frac{\sin\frac{qb}{2}}{\frac{qb}{2}} \cos\left(\frac{2\pi}{\Lambda}x\right) \right\|$$

Решение Первый интерференционный минимум будет наблюдаться, когда разность хода $\Delta$ между интерферирующими лучами будет равна $\lambda/2$. Из геометрических соображений (теоремы Пифагора) следует, что разность хода здесь:

$$\Delta \approx \frac{d^2}{L}$$

приравниваем и получаем

$$\frac{\lambda}{2} \approx \frac{d^2}{L}$$

Амплитуда осцилляций обратится в ноль в том случае, когда видность интерференционной картины будет равна нулю. Это произойдёт в том случае, когда расстояние между щелями $d_2$ окажется равной области когерентного Бсвещения

$$d_2 = \frac{\lambda}{\psi} = \frac{\lambda L}{b}$$

Итого размер источника:

$$b = \frac{2d_1^2}{d_2} = 0.1 \text{ cm}$$

## O2-5.9 Полосы равной толщины

Наблюдаются полосы равной толщины в воздушном клине между двумя плоскопараллельными стеклянными пластинками, образующими между собой очень малый угол. Клин освещается рассеянным светом. Наблюдение ведется невооруженным глазом с расстояния ясного зрения $L = 25$ см в направлении, перпендикулярном к поверхности клина, причем глаз может смещаться перпендикулярно к ребру клина. Оценить максимальное число интерференционных полос $N$, которое может видеть глаз в монохроматическом свете при таком способе наблюдения, если диаметр зрачка глаза $d = 5$ мм. Оценить степень монохроматичности света, необходимую для того, чтобы такое максимальное число полос могло наблюдаться.

### Solution 5.

9. $N = m_{\max} = \frac{4L^2}{d^2} \approx 10^4$; $\quad \frac{\Delta\lambda}{\lambda} \approx \frac{1}{m_{\max}} \approx 10^{-4}$.

## O2-5.10

Полосы равной толщины, получающиеся в тонком стеклянном клине с показателем преломления $n = 1,5$ при освещении рассеянным монохроматическим светом с длиной волны $\lambda = 5000$ Å, проецируются линзой на экран. Перед линзой помещена квадратная диафрагма со стороной $d = 1$ см и отстоящая от клина на расстояние $L = 50$ см. Какой максимальный порядок интерференции $N$ может $L = 50$ см. Какой максимальный порядок интерференции $N$ может рующей системы приблизительно перпендикулярна к поверхности клина.

### Solution 5.

10. $N \approx \left(\frac{nL}{d}\right)^2 \approx 5600$.

## O2-5.11

Рассчитать, какую ширину $b$ может иметь источник света в интерферометре Майкельсона, если зеркала интерферометра расположены на неодинаковых расстояниях от делительной пластинки и разность этих расстояний равна $L = 2$ см. Фокусное расстояние коллиматора равна $f = 25$ см. Немонохроматичностью источника пренебречь, длина волны света $\lambda = 0,5$ мкм. Наблюдаются линии равной толщины.

### Solution 5.

11. $b < f\sqrt{\frac{\lambda}{L}} = 1,25\text{m}$.

## O2-5.12

Источник света $S$ расположен на расстоянии $L = 1$ м от тонкой слюдяной пластинки толщиной $h = 0,1$ мм с показателем преломления $n = 1,4$ (рис. 299). На таком же расстоянии от пластинки расположен небольшой экран Э, ориентированный перпендикулярно отраженным лучам, на котором наблюдаются интерференционные полосы. Угол $\varphi = 60°$. Найти порядок $m$ интерференционной полосы в центре экрана и ширину $\Delta l$ интерференционных полос. Оценить допустимый размер $b$ и допустимую монохроматичность $\Delta\lambda$ источника. Используется зеленый свет с длиной волны $\lambda = 560$ nm.

### Solution 5.

12. $m = \frac{\Delta}{\lambda} \approx 360$, где $\Delta = 2h\sqrt{n^2 - \sin^2\varphi}$; угол схождения лучей $\Omega$ в силу симметрии примерно равен апертурному углу интерференции: $\Omega = \frac{h}{L}\operatorname{tg}\psi\cos^2\varphi \approx 0,2 \cdot 10^{-4}$. Ширина интерференционных полос $\Lambda \approx \frac{\lambda}{\Omega} = 2,8$ см. Предельный размер источника $b \approx \Lambda \approx 2,8$ cm. Допустимая немонохроматичность $\Delta\lambda \approx \frac{\lambda}{m} \approx 1,6$Н.

## O2-5.13

С помощью зрительной трубы, установленной на бесконечность, наблюдают интерференционные полосы в тонкой плоскопараллельной стеклянной пластинке толщиной $h = 0,2$ мм с показателем преломления $n = 1,41$; при этом угол наблюдения $\varphi$ может изменяться от 0 до 90° (рис. 300). Найти максимальный и минимальный порядок интерференционных полос. Оценить допустимую немонохроматичность $\Delta\lambda$ источника, при которой будут достаточно четко наблюдаться все интерференционные полосы. Каков допустимый размер источника света в этом интерференционном эксперименте? Используется зеленый свет с длиной волны $\lambda = 560$ нм.

### Solution 5.

13. $m_{\max} \approx 1000$; $m_{\min} \approx 720; \Delta\lambda \leqq 0,5$ нм. Источник света может иметь любые размеры.

## O2-5.14

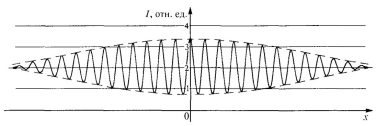
В двулучевой интерференционной схеме с равными интенсивностями интерферирующих лучей используется источник белого света, размер которого $b = 0,025$ cm. Интерференционная картина, наблюдаемая через светофильтр, изображена на рис. 301. Оценить полосу пропускания фильтра $\Delta\lambda$ и апертуру интерференции $\Omega$. Средняя длина волны равна $\lambda = 500$ нм.

### Solution 5.

14. $\Delta\lambda \approx \frac{\lambda}{m} \approx 100; \Omega = \frac{0,5\lambda}{b} = 10^{-3}$ рад.

## O2-5.15 Немонохроматичность

Интерференционная картина, полученная при интерференции двух пучков одинаковой интенсивности при апертуре интерференции $\Omega = 10^{-3}$ рад, изображена ниже.




Оценить немонохроматичность источника $\Delta\lambda$ и его линейный размер $b$. Средняя длина волны равна $\lambda = 500$ нм.

### Solution 5.

15*. $b = \frac{\lambda}{2\Omega} = 2,5 \cdot 10^{-2}$c; $m \approx 10$; $\Delta\lambda \approx \frac{\lambda}{m} = 50$ nm.

Решение. Из графнка, изображенного на рис. 300 , легко определяется видность интерференционной картины в центре: $V = \frac{3,3-0,7}{3,3+0,7} \approx 0,65$. В задаче 5.1 было получено аналитическое выражение для видности интерференционной картины протяженного источника:

$$V(b) = \left| \frac{\sin\left(\frac{\pi bd}{\lambda L}\right)}{\frac{\pi bd}{\lambda L}} \right| = \left| \frac{\sin\left(\frac{\pi\Omega}{\lambda/b}\right)}{\frac{\pi\Omega}{\lambda/b}} \right|,$$

где $\Omega == d/L$ - апертура интерференции, т. е. угол, под которым из точки, где расположен источник, видна база интерферометра. Если рассматривать видность, как функцию размера источника, то при $b = \lambda/\Omega V = 0$. Таким образом, это значение $b$ можно рассматривать как предельно допустимое: $b_{\max} = \lambda/\Omega$.

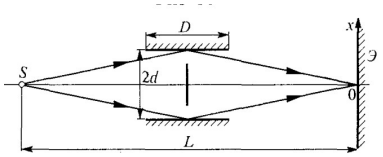
Однако в задаче требуется по виду интерференционной картины определить реальный размер $b$. Для этого необходимо решить следующее уравнение:

$$V = \left| \frac{\sin x}{x} \right| = 0,65 \approx \frac{2}{3} \approx \frac{2}{\pi}$$

Корень этого уравнения, очевидно, равен $x = \frac{\pi}{2}$, откуда следует $x = \frac{\pi\Omega}{\lambda/b} = \frac{\pi}{2}$ и $b = \frac{\lambda}{2\Omega} = 2,5 \cdot 10^{-2}$ см. Немонохроматичность источника оценивается так: $\Delta\lambda \approx \frac{\lambda}{m} \approx 50$ нм, где $m \approx 10$ мы определяем из графика - это максимально различимый порядок интерферснционной картины.

## O2-5.16 Два зеркала

В интерференционной схеме



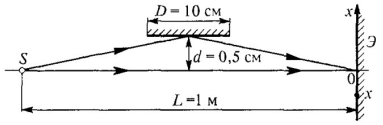

используется квазимонохроматический источник света $S\left(\lambda = 5 \cdot 10^{-5}\ \text{cm}\right)$. Отражающие зеркала расположены симметрично относительно источника $S$ и экрана Э, на котором наблюдается интерференция. Найти: 1) ширину интерференционной полосы $\Lambda$ на экране Э, 2) область локализации полос на экране, 3) максимальный и минимальный порядок интерференции и число наблюдаемых полос, 4) степень монохроматичности, при которой число наблюдаемых полос максимально, 5) допустимый размер источника $b$. Параметры схемы: $L = 1$ м, $2d = 2,5$ см, $D = 10$ см.

### Solution 5.

16. 1) $\Lambda \approx 10^{-3}$ cm; 2) $|x| \leqslant 0,25$ cm (область локализации полос); 3) $m_{\max} \approx 250, m_{\min} = 0, N \approx 500$; 4) $\Delta\lambda \approx 20$ Å; 5) $b \leq 10^{-3}$ cm.

## O2-5.17 Одно зеркало

В интерференционной схеме



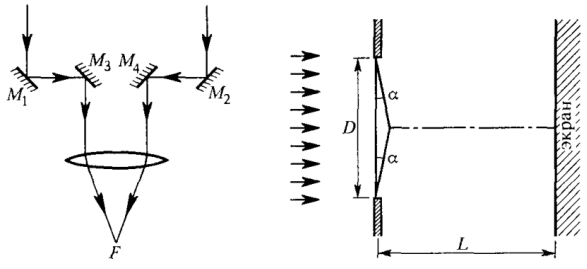

используется квазимонохроматический протяженный источник света $S$ ($\lambda = 5,0 \cdot 10^{-5}$ cm, $\Delta\lambda = 28$ Å). Полагая, что спектральная интенсивность излучения постоянна в интервале $\Delta\lambda$, найти: 1) ширину интерференционных полос $\Lambda$ на экране , 2) количество наблюдаемых 145 полос, 3) область локализации полос на экране, 4) максимальный и минимальный порядок наблюдаемых полос, 5) допустимый размер источника $b$. Параметры схемы указаны на рис. 304. Отражающес зеркало расположено симметрично относительно источника $S$ и точки О экрана .

### Solution 5.

17. 1) $\Lambda \approx 5 \cdot 10^{-3}$ cm; 2) $N \approx 100$; 3) $-0,1$ cm $< x < 0,4$ cm (область локализации) ; 4) $m_{\max} \approx 180, m_{\min} \approx 80$; 5) $b \leq 5 \cdot 10^{-3}$ cm.

## O2-5.18 Измерение угловых диаметров звезд

На рис. ниже изображена схема установки Майкельсона, предназначенной для измерения угловых диаметров звезд. Зеркала $M_1, M_2, M_3, M_4$ направляют в объектив телескопа два пучка света, интерферирующие друг с другом в фокальной плоскости объектива. При измерении углового диаметра гигантской красной звезды Бетельгейзе Майкельсон нашел, что интерференционные полосы исчезли, когда расстояние между внешними зеркалами $M_1$ и $M_2$ равнялось 306,5 см. Считая, что эффективная длина волны света от Бетельгейзе равна 5750 Å, вычислить угловой диаметр этой звезды $\psi$.



**Solution 5.**

18. $\psi \approx \frac{\lambda}{d} \approx 1,9 \cdot 10^{-7}$ рад. Более точный расчет учитывает «круглость» звезды: $\psi \approx 1,22\frac{\lambda}{d} = 2,3 \cdot 10^{-7}$ рад $= 0,047''$.

(см. Клименюка... (???))

## O2-5.19

Параллельный пучок света от удаленного источника с длиной волны $\lambda = 500$ нм падает на бипризму с преломляющим углом $\alpha = 10^{-2}$ рад и шириной $D = 2$ см, выполненную из стекла с показателем преломления $n = 1,5$ (рис. 305). 1) На каком расстоянии $L$ от бипризмы следует расположить экран, чтобы на нем можно было наблюдать максимально возможное число интерференционных полос? 2) Оценить допустимую немонохроматичность $\Delta\lambda$ света, необходимую для наблюдения всех полос. 3) Оценить допустимый угловой размер $\psi$ источника в этом интерференционном опыте.

**Solution 5.**

19. 1) $L = \frac{D}{4a(n-1)} = 1$ m; $\Lambda = 5 \cdot 10^{-3}$ cm.
2) $\Delta\lambda = \frac{\lambda}{m_{\max}} = 5$ нм, где $m_{\max} = 100$.
3) $v \leqslant \frac{2\lambda}{D} = 5 \cdot 10^{-5} = 0,17'$.

## O2-5.20

Из тонкой линзы диаметром $D = 2,5$ cm с фокусным расстоянием $f = 50$ см вырезана центральная полоска шириной $a = 0,5$ cm, после чего обе половины линзы сдвинуты до соприкосновения 146 (билинза). Источник света $S$ с длиной волны $\lambda = 500$ нм располагается на оси системы в фокальной плоскости линзы (рис. 306). 1) На каком расстоянии $L$ от билинзы следует расположить экран, чтобы на нем можно было наблюдать максимально возможное число интерференционных полос? Определить ширину $\Lambda$ интреференционных полос и их число. 2) Оценить допустимую немонохроматичность $\Delta\lambda$ источника света в этот интерференционном эксперименте, необходимую аля наблюдения всех полос. 3) Оценить допустимый размер $b$ источника света.

**Solution 5.**

20. 1) $L = \frac{D-a}{2a} = 1$ m, где $\alpha = \frac{a}{f} = 10^{-2}; \Lambda = \frac{\lambda}{a} = 5 \cdot 10^{-3}$ cm; $N = 200$.
2) $m_{\max} = 100$; $\Delta\lambda = \frac{\lambda}{m_{\max}} = 5$ nm;
3) $b \leqslant \frac{2\lambda f}{D-a} = 2,5 \cdot 10^{-3}$ cm.

## O2-5.21

Бипризма освещается монохроматическим светом с длиной волны $\lambda = 500$ нм от удаленного протяженного источника с угловым размером $\psi = 10^{-3}$ рад. Преломляющий угол бипризмы $\alpha = 5 \cdot 10^{-3}$ рад, показатель преломления $n = 1,5$. Определить видность интерференционных полос, наблюдаемых на экране, в зависимости от расстояния $L$ между экраном и бипризмой (рис. 307). При каких значениях $L$ интерференционные полосы размываются? Размер бипризмы считать достаточно большим. Источник можно считать равномерно светящейся полоской, параллельной ребру бипризмы.

**Solution 5.**

21. $V(L) = \left|\frac{\sin(0,1\pi L)}{0,1\pi L}\right|$; полосы размываются при $L = 10$ m[ см ], где $m = 1, 2, 3, \ldots$

## O2-5.22

В интерференционной схеме с двумя параллельными зеркалами (рис. 308) используется протяженный источник монохроматического света в виде однородно светящейся полоски шириной $b = 5 \cdot 10^{-3}$ cm. Длина волны излучения $\lambda = 600$ нм, расстояние между зеркалами $d = 0,2$ см. Какова видность $V$ интерференционных полос на экране, расположенном на расстоянии $z_0 = 0,5$ м от источника? В каком направлении и на какое расстояние следует переместить экран, чтобы: 1) видность увеличилась в 2 раза;

2) интерференционная картина исчезла (размылась)? Размеры зеркал считать достаточно большими. Прямые лучи от источника на экран не попадают.

### Solution 5.

22. $V_0 \approx 0,41$; при $V = V_1 = 0,827z \approx 1$ m, $\Delta z \approx 0,5$ m; интерференционная картина размывается при $z \approx 33$ cm, $\Delta z \approx 17$ cm.

## 12.0.10 Problems about Fresnel diffraction

### O2-6.1 Closing some Fresnel zones

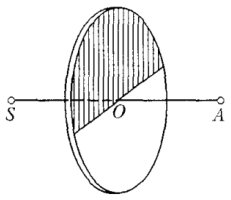
Какова интенсивность света $\mathcal{I}$ в центре дифракционной картины от круглого экрана, если он закрывает первую зону Френеля? Интенсивность света в отсутствие экрана равна $\mathcal{I}_0$.

### Solution 6.

1.
(??? forgot the method, one images a diagram...)
$\mathcal{I} \approx \mathcal{I}; E \approx 0$, в точке $A$ минимум освещенности.

### O2-6.2 Closing some Fresnel zones

Непрозрачный экран, имеющий форму полудиска, помещен между точечным источником $S$ и точкой наблюдения $A$ таким образом, что точка $O$ располагается на одной прямой с точками $S$ и $A$ (рис. 309). Экран закрывает небольшое нечетное число полузон Френеля. Какова освещенность в точке $A$?




### Solution 6.

2*
Если $E_1, E_2, E_3$ - поля, создаваемые в точке $A$ последовательными полными зонами Френеля, то искомое полс $E$ в точке $A$ представится рядом:

$$E = \frac{1}{2}(\underbrace{E_1 + E_2 + \ldots + E_{N-1}}_{\text{четное}}) + \frac{1}{2}E_N + (E_{N+1} + E_{N+2} + \ldots).$$

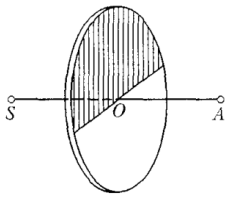
При небольшом $N$ первая скобка близка к нулю, а последняя - к $E_{N+1}/2$, так что

$$E \approx \frac{1}{2}(E_N + E_{N+1}) \approx 0$$

В точке $A$ будет минимум освещенности.

### O2-6.3 Closing some Fresnel zones

Между точечным источником $S$ и точкой наблюдения $A$ помещен диск, центр которого расположен на одной прямой с точками $S$ и $A$. Одна половина диска прозрачна, другая непрозрачна. Диск закрывает первые три зоны Френеля. Толщина прозрачной части диска $l = N\frac{\lambda}{2(n-1)}$, где $n$ - показатель преломления прозрачной части диска, $N$ - целое число. Какова освеценность в точке $A$ при четном и нечетном $N$?

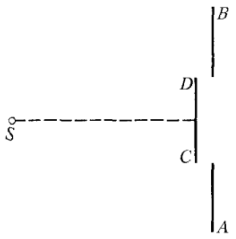



**Solution 6.**

3.
При четном $N$ в точке $A$ будет минимум освещенности ($E = 0$), при нечетном $N$ - максимум, приблизительно такой же, как при одной открытой центральной зоне.

## O2-6.4

Вдали от точечного источника $S$ электромагнитной волны поставлен бесконечный идеально отражающий экран $AB$. Пользуясь векторной диаграммой, найти, как изменится интенсивность отраженной волны в точке $S$, если из экрана вырезать диск $CD$ с центром в основании перпендикуляра, опущенного из $S$ на плоскость экрана, и смсстить этот диск по направлению к источнику на одну двенадцатую длины волны? Площадь диска составляет одну треть от площади первой зоны Френеля, Как изменится результат, если смещение произвести в противоположную сторону на ту же величину?




**Solution 6.**

4.
Колебание, вызываемое всеми зонами Френеля, изображается вектором $\overrightarrow{OC}$ колебание от одной трети первой зоны - вектором $\overrightarrow{OA}$.

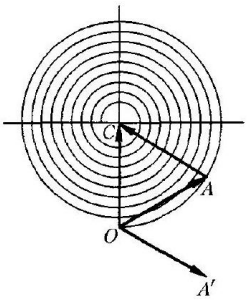
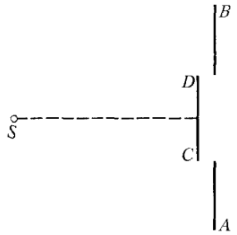



Рис. 437
Вектор $\overrightarrow{AC}$ представляет колебание, вызываемое волнами, отразившимися от внешней части экрана, расположенной за отверстием $CD$. Эти три вектора образуют равносторонний треугольник, если пренебречь уменьшением радиуса витка спирали на одном обороте. При смещении центрального круга к источнику на $\lambda/12$ фаза отраженной им волны увеличится на $2 \cdot 2\pi/12 = \pi/3$, и колебание изобразится вектором, равным и противоположно направленным вектору $\overrightarrow{AC}$. Интенсивность всей отраженной волны в точке $S$ обратится в ноль. При смещении круга $CD$ в противоположную сторону фаза колебания $\overrightarrow{OA}$ уменьшится на $\pi/3$ и вектор $\overrightarrow{OA}$ повернется в положение $\overrightarrow{OC}$. Результирующее колебание найдется сложением векторов $\overrightarrow{AC}$ и $\overrightarrow{OC}$. Таким путем найдем, что амплитуда колебаний в точке $S$ увеличится в $\sqrt{3}$ раз, а интенсивность - в три раза.
Answer: Амплитуда колебаний в точке $S$ увеличится в $\sqrt{3}$ раз, а интенсивность - в три раза.

## O2-6.5

В установке изображенной на рисунке




площадь диска составляет половину площади центральной зоны Френеля. На какое минимальное расстояние $h$ следует сместить диск в направлении от источника, чтобы интенсивность отраженной волны в точке $S$ осталась неизменной?

### Solution 6.

5.
(???)
$h = \frac{3}{8}\lambda$.

## O2-6.6

Если круглое отверстие (например, ирисовая диафрагма) увеличивается таким образом, что его радиус от радиуса одной зоны возрастает до радиуса двух зон, то освещенность в точке $P$ (точка $P$ - изображение источника света, даваемое диафрагмой) падает почти до нуля. Как согласовать Этот факт с увеличением всего светового потока через диафрагму в два раза?

### Solution 6.

6. Энергия перераспределяется, причем в одних точках плоскости изображения плотность светового потока возрастает, а в других убывает. Весь поток черсз плоскость изображения возрастает в 2 раза.

## O2-6.7

Точечный источник монохроматического света помещен на расстоянии a от круглой диафрагмы, а экран - с противоположной стороны на расстоянии $b$ от нее. При каких радиусах диафрагмы $r$ центр дифракционных колец, наблюдаемых на экране, будет темным и при каких - светлым, если перпендикуляр, опущенный из источника на плоскость диафрагмы, проходит через ее центр?

### Solution 6.

7*
Решение. Освещенность в центре дифракционной картины можно найти, разбивая волновую поверхность $ACB$ (рис. 438) на зоны Френеля.

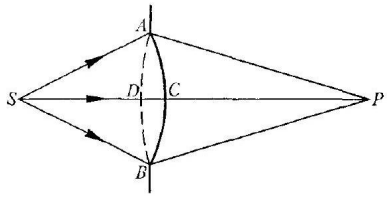



Рис. 438
Если в ней уложится четное число зон Френеля, то в $P$ получится минимум освещенности; если нечетное - максимум. Построим сферу радиусом $PA$ с центром в $P$. Число зон Френеля на волновой поверхности $ACB$, очевидно, равно длине $CD$, деленной на $\lambda/2$. Отсюда легко получить результат, приведенный в ответе.

Answer: $r = \sqrt{\frac{m\lambda}{1/a+1/b}}$. Центр колец темный, если $m$ - четное число, и светлый, если $m$ - нечетное число.

## O2-6.8

Между точечным монохроматическим источником света и точкой наблюдения перпендикулярно соединяющей их линии помещен экран, состоящий из секторов двух кругов (рис. 311). Радиус одного из них равен радиусу 1 -й зоны Френеля, другого - радиусу 2-й зоны Френеля. Определить интенсивность света в точке наблюдения, если в отсутствие экрана она равна $\mathscr{I}_0$. Рассмотреть экраны, изображенные на рис. 311 а и 311.

### Solution 6.

8. а) $\mathcal{I} = 0$; б) $\mathcal{I} = \frac{\mathcal{I}_0}{4}$.

## O2-6.9

Симметрично между источником ($\lambda = 4900$ Å) и точкой наблюдения расположен непрозрачный экран с круглым отверстием ($r = 0,35$ mm). Расстояние от источника до экрана $a = 1$ м. Во сколько раз изменится интенсивность в точке наблюдения, если источник сместить вдоль оси к отверстию на расстояние $\Delta a = 0,8$ m?

### Solution 6.

9 $\frac{\mathcal{I}'}{\mathcal{I}} = \frac{25}{9} \cdot \left(\mathcal{I} = 2\mathcal{I}_0; \mathcal{I}' = \frac{50}{9}\mathscr{O}_0\right.$, где $\mathcal{I}_0$ - интенсивность в точке наблюдения в отсутствие экрана.)

## O2-6.10

Между источником $S\left(\lambda = 4,9 \cdot 10^{-5}\text{ cm}\right)$ и точкой наблюдения $P$ расположен непрозрачный экран с круглым отверстием на оси $SP$. Расстояние от источника до экрана $a = 3$ m, между экраном и точкой наблюдения $b = 1$ м. Во сколько раз изменится интенсивность сигнала, если точку наблюдения удалить вдоль оси на расстояние $b' = 3$ м от экрана?

### Solution 6.

10. $\frac{\mathcal{I}'}{\mathcal{I}} = \frac{4}{9}$. ($\mathscr{I} = 3\mathcal{I}_0; \mathcal{I}' = \frac{4}{3}\mathcal{I}_0$, где $\mathcal{I}_0$ - интенсивность в точке наблюдения в отсутствие экрана.)

## O2-6.11

В непрозрачной пластинке имеется отверстие диаметром $d = 1$ мм. Оно освещается монохроматическим светом с длиной волны $\lambda = 500$ нм от удаленного точечного источника. Найти расстояние $L_{\max}$ от отверстия, на котором будет наблюдаться наибольшая освещенность.

### Solution 6.

11. $L = \frac{d^2}{4(2m+1)\lambda}, m = 0, 1, 2, \ldots$; при $m = 0 L_{\max} = 0,5$ m.

## O2-6.12

Два точечных некогерентных источника монохроматического света $S$ и $S'$ освещают экран (рис, 312). Для точки $C$ выполняется условие $SC = S'C = 1$ м. Во сколько раз изменится освещенность в точке $C$, если на пути лучей в точках $A$ и $A'$ расположить непрозрачные экраны с круглым отверстием, диаметр которого равен 0,6 мм? Центры отверстий совпадают с $SC$ и $S'C$; $AC = 9$ cm, $A'C = 20$ cm. Длина волны света равна 5600 Å. Как изменится ответ, если источники будут когерентными?

### Solution 6.

12. Возрастет в 2 раза. Если источники когерентны, не изменится.

## O2-6.13

Параллельный пучок монохроматического света с длино волны $\lambda = 6000$ Å нормально падает на непрозрачный экран с круглым отверстием диаметром $D = 1,2$ мм. На расстоянии $b_1 = 18$ cm за экраном на оси отверстия наблюдается пятно. На какое минимальное расстояние $\Delta b$ нужно сместиться от этой точки вдоль оси отверстия, удаляясь от него, чтобы в центре дифракционной картины вновь наблюдалось темное пятно?

### Solution 6.

13. $\Delta b = \frac{8b_1^2\lambda}{D^2 - 8b_1\lambda} = 18$ cm.

## O2-6.14

Вдали от точечного источника $S$ стоит бесконечный идеально отражающий экран. Из экрана удален диск диаметром $d_1 = 2r_1\sqrt{2/3}$, гдс $r_1$ - радиус 1-й зоны Френеля, и поставлен другой диаметром $d_2 = d_1/\sqrt{2}$. Найти интенсивность $\mathfrak{O}$ отраженной волны в точке $S$, если диск диаметром $d_2$ стоит в плоскости экрана.

### Solution 6.

14. $\mathcal{I} = 0$.

## O2-6.15

Яркий источник можно сфотографировать, поместив между ним и фотопластинкой гладкий непрозрачный шар. Обълснить явление. Диаметр шара $D = 40$m, расстояние от источника до шара $a = 12$ m, расстояние от шара до изображения $b = 18$ m, размер источника $y = 7$ мм. Определить размер изображения $y'$. Будет ли изображение испорчено, если поверхность шара испешрена множеством неправильных царапин, глубина которых $h$ порядка 0,1 мм? Можно ли шар заменить диском?

### Solution 6.

15. $y' = \frac{b}{a}y = 10,5$ мм. Опыт был поставлен Полем с параметрами, указанными в задаче. Чтобы опыт удался, необходимо, чтобы глубина неровностей $h$ удовлетворяла условию $h < \frac{\lambda}{D}\frac{ab}{a+b} = 180\lambda \approx 0,1$ мм. Шар можно

заменить диском ири условии $y < \frac{2a}{D}\sqrt{\frac{ab\lambda}{a+b}} \approx 1$ м. Соответствующий оиы был поставлен Ангрером.

## O2-6.16

Диск из стекла с показателем преломления $n$ (для длины волны $\lambda$) закрывает полторы зоны Френеля для точки наблюдения $P$. При какой толщине $h$ диска освещенность в точке $P$ будет наибольшей?

### Solution 6.

16. $h = \frac{2m+5/4}{2(n-1)}\lambda$, где $m = 0, 1, 2, \ldots$

## O2-6.17

В параллельном пучке радиоизлучения, длина волны которого $\lambda = 3$ cm, поставлен диск из диэлектрика с показателем преломления $n = 1,5$. Диск перпендикулярен направлению пучка. Диаметр диска $D = 20$ cm. При какой толщине диска $h$ и на каком расстоянии $b$ от диска вдоль его оси будет наблюдаться максимальная интенсивность излучения? Отражением излучения от диска пренебречь.

### Solution 6.

17. $h = 3 + 6k$[см], где $k = 0, 1, 2, \ldots$;

$$b_m = \frac{D^2}{4m\lambda} = 33,3; 11,1; 6,65[\text{см}]; \ldots, \text{где } m = 1, 3, 5, 7, \ldots$$

## O2-6.18

Непрозрачный диск, радиус которого $R = 0,55$ cm, освещается плоской монохроматической волной $\left(\lambda = 5 \cdot 10^{-5} \text{ cm}\right)$. Оказалось, что если в центре диска продслать маленькое отверстие, то интенсивность света в точке наблюдения, находящейся за диском на оси отверстия на расстоянии $L = 1,5$ м учетверится. Определить минимальный радиус отверстия.

### Solution 6.

18. $r_{\min} = \sqrt{m\lambda L} \approx 1,15\text{m}$, где $m = 1\frac{2}{3}$.

## O2-6.19

Непрозрачный диск диаметром $D = 1$ cm, освещается плоской нормально падающей волной $\left(\lambda = 5 \cdot 10^{-5} \text{ cm}\right)$. При каком минимальном диаметре отверстия, проделанного в центре диска, интенсивность света в точке, находящейся за диском на оси симметрии на расстоянии $L = 1,5$ м, равна нулю?

### Solution 6.

19. $d_{\min} = 2,25\text{m}$.

## O2-6.20

В плоскопараллельной стеклянной пластинке с показателем преломления $n$, на которую нормально падает плоская волна, вырезано круглое отверстие размером в одну зону Френеля для некоторой точки $P$, лежащей на оси системы. Определить, при какой толщине пластинки $h$ интенсивность $\mathscr{I}$ колебаний в точке $P$ будет максимальной. Найти $\mathcal{I}_{\max}$, если в отсутствие пластинки интенсивность равна $\mathcal{I}_0$. Длина волны падающего света равна $\lambda$.

### Solution 6.

20. $h = \frac{(2m+1)\lambda}{2(n-1)}, m = 0, 1, 2, \ldots; \quad g_{\max} = 9g_0$.

## O2-6.21

Интенсивность света в некоторой точке на оси за отверстием в непрозрачном экране, на который нормально падает параллельный пучок монохроматического света, равна $\mathscr{I}_0$, если в отверстии укладывается одна зона Френеля. С помощью векторной диаграммы найти интенсивность света в той же точке, если радиус отверстия уменьшить на $\alpha = 1/3$ первоначальной величины.

### Solution 6.

21* $\mathcal{G} \approx 0.41\mathcal{G}_0$.

Решение. Если $r$ - радиус отверстия, то разность хода между лучами, приходящими от его края и от центра, равна $r^2/(2L)$, где $L$ - расстояние от центра отвсрстия до точки наблюдения. Положим сиачала $r = r_1$, а затем $r = r_1(1-\alpha)$, где $r_1$ - радиус цеитральной зоны Френеля. Тогда соответствующие разности фаз будут $\pi$ и $\delta = \pi(1-\alpha)^2$.

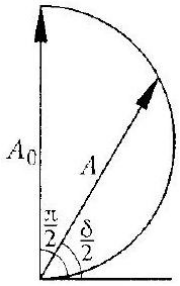



Рис. 439

Как видно из векторной диаграммы (рис. 439), амплитуды колебаний $A_0$ и $A$ в рассматриваемых двух случаях связаны соотношением $A = A_0 \sin(\delta/2)$, а интенсивности - соотнопенисм $g = \mathcal{I}_0 \sin^2(\delta/2)$. При $\alpha = 1/3$

$$\mathcal{I} = \mathcal{I}_0 \sin^2(2\pi/9) \approx \mathscr{O}_0 \sin^2 40^\circ \approx 0,41\mathcal{I}_0$$

## O2-6.22

Параллельный пучок монохроматического света ($\lambda = 5000$ Å) интенсивностью $\mathscr{I}_0$ падает на непрозрачный экран с круглым отверстием диаметром 2 мм.

1) Найти расстояния $b_1, b_2, \ldots, b_m$ от экрана до точек $P_1, P_2, \ldots, P_m$ на оси отверстия, для которых в пределах отверстия укладывается $1, 2, \ldots, m$ зон Френеля.

2) Построить приближенно график зависимости интенсивности света на оси отверстия от расстояния от точки наблюдения до экрана. 3) Насколько надо сместиться из точки $P_1$, удаляясь от экрана, чтобы интенсивность света в новой точке наблюдения стала в два раза меньше, чем в точке $P_1$?

### Solution 6.

22. 1) $b_m = \frac{r^2}{m\lambda}$.

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $b_m \cdot$ cm | 200 | 100 | 66 | 50 | 40 | 33 | 29 | 25 | 22 | 20 |

2) (См. рис. 440)

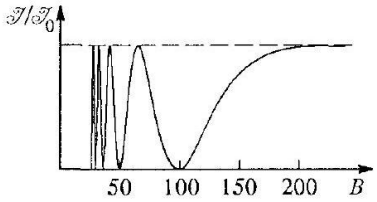



Рис. 440

3) $\Delta b = 200$ cm.

## O2-6.23

Плоская монохроматическая световая волна с интенсивностью $I_0$ падаст нормально на непрозрачный экран в виде полуплоскости с вырезом на краю, имеющем форму полукруга (рис. 313). Найти интенсивность света в точке $P$, для которой граница выреза совпадает с границей 1-й зоны Френеля.

### Solution 6.

23. $\mathcal{I}_P = \frac{9}{4}\mathcal{I}_0$.

## O2-6.24

На белой стене наблюдается тень от прямолинейного края $AB$ непрозрачного экрана, освещаемого параллельными монохроматическими лучами, падающими на экран перпсндикулярно Рис. 313 Рис. 314 ($\lambda = 5000$ Å). Плоскости стены и экрана параллельны, расстояние между ними $b = 4$ м. На краю экрана выточено углубление, имеющее форму полукруга радиусом $r = 1$ мм (рис. 314). Как изменится

интенсивность света в точке стены, являющейся геометрической тенью центра соответствующего круга по сравнению с интенсивностью в той же точке, когда углубления не было?

### Solution 6.

24. $g = g_0\left(\frac{1}{4} + 2\sin^2\frac{\delta}{2}\right)$, где $\mathcal{I}_0$ - иитенсивность света в отсутствие экрана, $\delta = \pi r^2/(b\lambda)$ - разность фаз между лучами, пришедшими в точку наблюдения от краев и центра углубления. Подставляя численные значения, получим $\delta = \pi/2, \mathcal{I} = 5]_0/4$. Интенсивность увеличится в 5 раз.

## O2-6.25

Точечный источник света с двумя монохроматическими линиями $\lambda_1 = 660$ nm и $\lambda_2 = 440$ одинаковой интенсивности расположен на расстоянии $L = 1$ м от экрана. Перед Экраном на расстоянии $a = 0,2$ м расположен прозрачный диск диаметром $D = 0,92$ мм, вносящий фазовую задержку в л для обеих компонент, причем источник света, центры диска и экрана лежат на общей оси. Как отличаются интенсивности света в центре экрана при наличии и в отсутствие диска?

### Solution 6.

25. $\mathscr{V} = \mathcal{I}_1 + \mathcal{I}_2 = 1\mathscr{C}_0$, т. е. интенсивность возрастет в 5 раз ($\mathscr{S}_1$ и $\mathcal{I}_2$ - интенсивности соответствующих спектральных компонснт в точке наблюдения, $\mathscr{O}_0$ - интенсивности этих же компонент в отсутствие диска).

## O2-6.26

Точечный источник света с двумя монохроматичсскими линиями $\lambda_1 = 6600$ Å и $\lambda_2 = 4400$ Å одинаковой интенсивности расположен на расстоянии $l = 1$ м от экрана. Перед экраном на расстоянии $a = 0,2$ м располагается непрозрачный лист с отверстием диаметром $D = 0,92$ так, что источник света, центры диска и экрана расположены на общей оси системы. Как отличаются освещенности в центре экрана при наличии и в отсутствие листа?

### Solution 6.

26. Освещенность при наличии листа в два раза больше.

## O2-6.27

Между точечным источником и приемником излучения установлен непрозрачный экран с круглым отверстием, размер которого соответствует внешнему краю зоны Френеля. Всю систему заполняют водой, показатель преломления которой $n = 1,33$. Определить интенсивность излучения $\mathcal{I}$, которую зарегистрирует приемник, если известно, что в отсутствие экрана интенсивность была равна $_0$.

### Solution 6.

27. $= 3$.

## O2-6.28

Параллельный пучок монохроматического света нормально падает из воздуха на плоскую поверхность диэлектрика. Определить максимальную величину напряженности электрического поля в диэлектрике $E_{\max}$. Оценить расстояние $l$ от поверхности диэлектрика до точки, в которой поле максимально. Диаметр светового пучка $D = 0,1$ cm, длина световой волны в воздухе $\lambda = 0,5$ мкм, плотность потока мощности в падающем пучке $S = 1$ кВт/см $^2$, показатель преломления диэлектрика $n = 2$.

### Solution 6.

28. Электрическое поле в диэлектрике максимально на оси пучка на таком расстоянии от поверхности, из которого диаметр пучка воспринимается как диаметр первой зоны Френеля $b = \frac{D^2 n}{4\lambda} = 100$ cm.

$$E_{\max} = \frac{4}{3}\sqrt{\frac{8\pi S}{c}} = 1200 \text{ В/cm}$$

## O2-6.29

Параллельный пучок монохроматического света с длиной волны $\lambda = 0,5$ мкм нормально падает на поверхность диэлектрика с показателем преломления $n = 2$. Поверхность диэлектрика покрыта черным (неотражающим) экраном, в котором имеется отверстие диаметром $D = 0,1$ cm. На каком расстоянии $l$ от поверхности диэлектрика (в воздухе) электрическое поле волны максимально? Принимая плотность потока мощности в падающей волне равной $S = 1$ кВт/см $^2$, определить максимальную величину напряженности электрического поля в воздухе $E_{\max}$ .

### Solution 6.

29. $l = \frac{D^2}{4\lambda} = 50$ cm; $E_{\max} = \frac{5}{3}\sqrt{\frac{8\pi S}{c}} = 1500$ В/cm.

## O2-6.30

### Solution 6.

30. $r = \sqrt{m\frac{\lambda f b}{b-f}} = 0,9\sqrt{m}$ мм, где $m = 1, 3, 5, \ldots$

## O2-6.31

### Solution 6.

31. $\frac{\mathcal{I}_1}{\mathcal{I}_2} = \frac{(m\pi)^2}{2} \approx 150$, где $m = \frac{r^2 D}{\lambda} = 5,5$.

## O2-6.32

### Solution 6.

32. g = 8 g.

## O2-6.33

Точечный источник света и точка наблюдения $P$ расположены симметрично на расстоянии $2L$ на оси круглого отверстия в непрозрачном экране. Отверстие оставляет открытой одну зону Френеля для точки $P$. Во сколько раз изменится интенсивность света в точке $P$, если к отверстию без нарушения осевой симметрии приложить тонкую линзу с фокусным расстоянием $f = L$?

### Solution 6.

33. Интенсивность света увеличится в два раза.

## O2-6.34

На длиннофокусную собирающую линзу ирисовой диафрагмой падает параллельный пучок монохроматического света. На расстоянии a от линзы помещен экран, на котором наблюдаются дифракционные кольца. При каких радиусах диафрагмы центр колец будет темным и при каких светлым, если фокусное расстояние Линзы равно $f'$?

### Solution 6.

34* $r = \sqrt{\frac{m\lambda}{|1/f - 1/a|}}$ Центр колец темный, если $m$-.- четное число, и светлый, если $m$ - нечетное число.

Решение. Метод решения такой же, как и в задаче 6.7. Разница только в том, что теперь сферическая волна не расходится, как было раньше, а

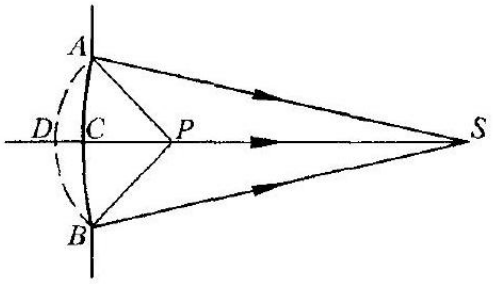



Рис. 441

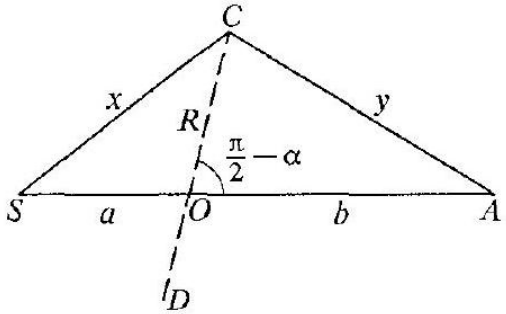



Рис. 442

сходится в точке $S$ (рис. 441). Зоны Френеля для точки $P$ строятся поэтому с вогнутой стороны сферической волновой поверхности. Число зон, укладывающихся в отверстии диафрагмы, будет равно длине $CD$, деленной на $\lambda/2$.

## O2-6.35

Зонная пластинка дает изображение источника, удаленного от нее на 3 м, на расстоянии 2 м от своей поверхности. Где получится изображение источника, если его отодвинуть в бесконечность?

### Solution 6.

35. На расстоянии $1,2$ m.

## O2-6.36

Определить фокуснос расстояние $f$ зонной пластинки для света с длиной волны 5000 Å, если радиус 5 - го кольца Этой пластинки равен 1,5 мм; определить радиус $r_1$1-го кольца этой пластинки. Что произойдет, если пространство между зонной пластинкой и экраном заполнено средой с показателем преломления $n(n > 1)$?

### Solution 6.

36. $f = 90$ см; $r_1 = 0,672$ мм. Изображения, т. е. максимумы, расположенные на оси пластинки, отодвинутся от последней.

## O2-6.37(2 min) Intensity in basic setup

Какова интенсивность света $\mathcal{I}$ в фокусе зонной пластинки, если закрыты все зоны, кроме 1-й? Интенсивность света без пластинки равна $\mathcal{I}_0$.

### Solution 6.

37.
(it is obvious from the theory)
$\mathcal{I} \approx 4\mathcal{I}_0$.

## O2-6.41

Освещенный предмет расположен на оси зонной пластинки на расстоянии $a$ от нее. Самое дальнее от зонной пластинки изображение предмета получается на расстоянии $b$ от нее. На каких расстояниях $b_k$ от пластинки получаются остальные изображения предмета?

### Solution 6.

41. Фокусы пластинки различных порядков: $f_k = \frac{ab}{(a+b)(2k+1)}$, где $k = 0, 1, 2, \ldots$ Положения всех изображений определяются формулой $\frac{1}{a} + \frac{1}{b_k} = \frac{1}{f_k}$.

## O2-6.42 Images behind zone plate

С помощью зонной пластинки надо сфотографировать освещенный предмет с угловым размером $2\alpha = 0,1$ рад. Оценить число зон пластинки, при котором будет достигнута наибольшая четкость в изображении всех частей предмста.

### Solution 6.

42. $N \sim \frac{1}{\alpha^2} \sim 400$.
Решение. Допустим сначала, что источник света $S$ точечный, а зонная пластинка $CD$ наклонена к его оптической оси под углом $\pi/2 - \alpha$ (рис. 442). Из рисунка видно $x^2 = a^2 + R^2 + 2aR\sin\alpha$. Извлекая квадратный корень и пренебрегая всеми степенями радиуса $R$, начиная с третьих, получим

$$x = a + R\sin\alpha + \frac{R^2\cos^2\alpha}{2a}.$$

Аналогично,

$$y = b - R\sin\alpha + \frac{R^2\cos^2\alpha}{2b}$$

Отсюда для разности хода между лучами $SCA$ и $SOA$ находим

$$\Delta = (x+y) - (a+b) = \frac{R\cos^2 a}{2}\left(\frac{1}{a} + \frac{1}{b}\right)$$

или $\Delta = \Delta_0 - \delta\Delta$, где $\Delta_0$ - значение $\Delta$ при отсутствии наклона зонной пластинки, а $-\delta\Delta$ - приращение величины $\Delta$, обусловленное наклоном:

$$\delta\Delta = \Delta_0\left(1 - \cos^2\alpha\right) = \Delta_0\sin^2\alpha \approx \Delta_0\alpha^2.$$

Если $\delta\Delta \ll \lambda$, то наклон пластинки не скажется существенно на работе зон, расположенных в пределах круга радиусом $R$. Если же $\delta\Delta \approx \lambda/2$, то все зоны, расположенные выше этого круга, становятся бесполезными и даже вредными. Из этого условия находится предельное значение разности хода $\Delta_0 \approx \lambda/\left(2\alpha^2\right)$. Соответствующее число зон Френеля будет $N \approx \Delta_0/(\lambda/2) \approx 1/\alpha^2$.
Допустим теперь, что фотографируемый предмет не точечный, причем его центр расположен на оси зонной пластинки. Для периферийных точек предмета, не лежащих на оси пластинки, последняя действует как наклонная под углом $\alpha$. Поэтому предельное число зон Френеля, при котором должно получиться наиболее отчетливое изображение, будет $N \sim 1/\alpha^2 \sim 400$.

## O2-6.43

Линза с фокусным расстоянием $f = 50$ cm и диаметром $D = 5$ см освещается параллельным монохроматическим пучком света с длиной волны $\lambda = 630$ нм. Найти, во сколько раз интснсивность волны I в фокусе линзы превышает интенсивность волны J 0, падающей на линзу. Оценить размер $b$ пятна в фокальной плоскости.

### Solution 6.

43. $\frac{\mathcal{I}}{\mathcal{I}_0} = \left(\frac{\pi D^2}{4f\lambda}\right)^2 \approx 3,9 \cdot 10^7; \quad b \approx \frac{2\lambda f}{D} = 10^{-3}$ cm.

## O2-6.44

Зонная пластинка с радиусом 1-й зоны Френеля $r_1 = 0,5$ мм помещена перед отверстием в экране диаметром $D = 1$ см. Пластинка освещается параллельным монохроматическим пучком света с длиной волны $\lambda = 500$ нм и интенсивностью $_0$. Определить интенсивность $\mathcal{I}$ волны в фокусе пластинки. Оценить размер $b$ пятна в фокальной плоскости.

### Solution 6.

44. Фокусное расстояние пластинки $f = \frac{r_1^2}{\lambda} = 50$ cm. Полное число зон Френеля $m = \frac{D^2}{4\lambda f} = 100 \cdot \frac{\mathcal{I}^{\mathscr{F}}}{\mathcal{I}_0} = m^2 = 10^4; b = \frac{2\lambda f}{D} = \frac{2r_1^2}{D} = 5 \cdot 10^{-3}$ cm.

## O2-6.45

Требуется изготовить отражатсльную зонную пластинку на вогнутом сферическом зеркале кольцевыми зонами Френеля. Найти радиус $m$-й зоны $r_m$, если источник света и точка наблюдения расположены на оси зеркала на расстояниях $a$ и $b$ соответствснно от его вершины, причем $a \leqslant R \leqslant b, r_m \ll a$, где $R$ - радиус кривизны поверхности зеркала.

### Solution 6.

45* $r_m = \sqrt{\frac{m\lambda}{\left|\frac{1}{a} + \frac{1}{b} - \frac{2}{R}\right|}}$

Решение. Как видно из рис. 443, $SM^2 = (a-x)^2 + r_m^2$. Извлекая квадратный корень и пренебрегая квадратом малого отрезка $x$, находим $SM \approx a + \frac{r_m^2}{2a} - x$. Аналогично, $MS' \approx b + \frac{r_m^2}{2b} - x$. Далее, $x \approx \frac{r_m^2}{2R}$. Радиус $m$-й зоны определится из условия $|(SM + MS') - (SA + AS')| = m\lambda/2$. Подставляя сюда вычисленные значения, получим

$$r_m^2 = \frac{m\lambda}{\left|\frac{1}{a} + \frac{1}{b} - \frac{2}{R}\right|}.$$

Результат можно получить проще, если заметить, что в принятом приближении зоны Френеля можно строить не на поверхности зеркала, а на поверхности сходящейся волны, отразившейся от зеркала. Радиус кривизны этой

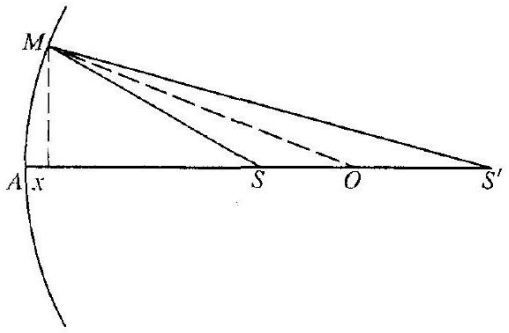



Рис. 443

поверхности $a'$ найдется по формуле зеркала $\frac{1}{a} + \frac{1}{a'} - \frac{2}{R} = 0$. Для определения $r_m$ теперь достаточно заменить в ответе к задаче 6.7 величину $a$ на $a'$.

## O2-6.46

В предыдущей задаче $R = 100$ cm, $a = 80$ cm, $\lambda = 5000$ Å, радиус 4-й зоны Френсля $r_4 = 0,2$ cm. Где будут расположсны изображения $S'$ источника нулевого и $\pm 1$-го порядков?

### Solution 6.

46. $b_0 \approx 1,33$ cm,

$$b_{+1} = 80 \text{ cm}$$
$$b_{-1} = 400 \text{ cm}.$$

Решение. Разность хода $(SMS')$ - (SAS') (рис. 443), возникающая при отражении лучей от соседних зон Френеля, равна нулю для изображения нулевого порядка и $\pm\lambda/2$ для изображений $\pm$1-го порядков. При отражении от центральной и $m$ й зон она в $m$ раз больше. Из этого условия получаем

$$\frac{r_m^2}{2}\left(\frac{1}{a}+\frac{1}{b}-\frac{2}{R}\right) = m\frac{\lambda}{2}$$

Полагая здесь $|m| = 4$, находим ответ.

## O2-6.51

На пути сферической монохроматической волны, сходящеейся в точке $O$, на расстоянии $z = R_0$ от точки $O$ устанавливается непрозрачный экран с кольцевым вырезом (рис. 318). При освещении этого экрана плоской волной с той же длиной волны экран оставлял бы открытыми 2-ю, 3-ю и 4-ю зоны Френеля для точки $O$. Найти интенсивность света в точке $O$, принимая интенсивность в плоскости экрана равной $\mathscr{O}_0$. Определить, при каких значениях $z$ будут наблюдаться локальные минимумы интенсивности, если точку наблюдения смещать от точки $O$ в область $z \leqslant R_0$. 155

### Solution 6.

51. $\varnothing = (3\pi A_0)^2 = 9\pi^2 \%_0$. Локальные минимумы наблюдаются ири $z_n = \frac{3R_0}{2n+3}$ ( $n = 1 z_1 = \frac{3}{5}R_0$; при $n = 2 z_2 = \frac{3}{7}R_0$).

## Other Solutions 6. Дифракция Френеля. Зонные пластинки

### Solution 6.

38. $g \approx g_0$.

### Solution 6.

39. $f_{\max} = \frac{r_1}{\lambda} = 8$ m; $\quad h = \frac{2m+1}{2(n-1)}\lambda, m = 0, 1, 2, \ldots; \quad \mathcal{I}_{\max} = 36\mathcal{I}_0$.

### Solution 6.

40. Основной фокус есть точка, для которой зоны, начерченные на пластинке, совпадают с зонами Френеля. Если $r$ - радиус первой зоны, начерченной на пластинке, то основной фокус определяется выражением $f_0 = r^2/\lambda$. Следующис фокусы получаются, когда в первой зоне, начерченной на пластинке, укладывается $3, 5, \ldots, 2k+1$ зон Френеля, т.е. когда $r^2/f_k = (2k+1)\lambda$. Следовательно, $f_k = \pm f_0/(2k+1)$, где $k = 0, 1, 2, \ldots$ Знаку «плюс» соответствуют действительные, а «минусу» - мнимые фокусы.

### Solution 6.

47. $r_m = \left[\frac{m\lambda R}{b_1 - b_2}\left(R + \sqrt{R^2 + (b_1 - b_2}\right)^2\right)\right]^{1/2} \approx 0,7$ cm, $(m = 5)$.

### Solution 6.

48. В задаче выполнено условие геометрической оптики: на ширине щели $l$, а тем более на расстоянии $d$ умещается число зон Френеля $m \gg 1$.

Ширина интерференииионных полос $\Lambda = \frac{\lambda}{\alpha} = \frac{\lambda f}{d} \approx 1,25 \cdot 10^{-3}$ см. Полосы исчезнут при $L \approx 2f - \frac{bd}{\lambda} = 30$ cm.

### Solution 6.

49. Длина волны должна меняться в пределах $\lambda \approx 3,7 - 3,0$ мм; при этом частота перестраивается ог 89,1 до 108,9 кГц.

### Solution 6.

50. $\lceil = (3\pi A_0)^2 = 9\pi^2 \mathscr{G}_0$. Фаза колебаний в точке наблюдения $\varphi(z) = \frac{k\rho^2}{2}\left(\frac{1}{z} - \frac{1}{R_0}\right)$, где $\rho$ - координата в плоскости экрана, отсчитываемая от центра отверстия. Локальный минимум наблюдается при $z = 3R_0$ (других нет) ; $m = 3\frac{z-R_0}{z} = 2$ - число зон Френеля, укладываемых на отверстии для локального минимума.

## 12.0.11 Problems about Fraunhofer diffraction

### O2-7.1 Radiation pattern

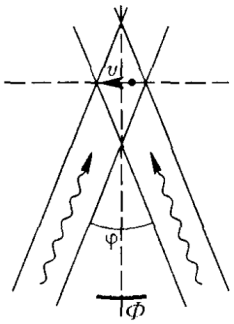
A parabolic mirror with a diameter of $D = 1$ m is used as an antenna for waves of length $\lambda = 3$ cm. Estimate the smallest distance $L_{\min}$ at which the receiver should be placed to remove the radiation pattern.

### Solution 7.

1.
(???)
$L_{\min} > \frac{D^2}{4\lambda} \approx 10$ m.

### O2-7.2 Measuring speed by lasers

Для измерения скорости частиц жидкости бесконтактным методом используется лазерный анемометр. Два когерентных лазерных пучка с длиной волны излучения $\lambda = 0,63$ мкм и углом сходимости $\varphi = 2°$ пересекаются в некоторой области жидкости, в которой небольшие взвешенные частички движутся со скоростью $v$ (рис. 319).




Определить скорость этих частиц, если известно, что при регистрации отраженного от них света частота колебаний тока фотоприемника Ф равна $v = 5,54$ кГц.

### Solution 7.

2. $v = \frac{\lambda\nu}{\varphi} = 10$ cm/c.

## O2-7.3(2min) Distinguishing car headlights

Estimate from what distance $L$ it is possible to see the light from two car headlights separately.

### Solution 7.

3.
One needs to use the basic formula of spatial coherence (??? right???)
$L \approx \frac{ld}{\lambda} \approx 10$ км, где $l \approx 1$ m - расстояние между фарами, $d \approx 5$ мм - диаметр зрачка глаза.

## O2-7.4

Пыль, взвешенная в воздухе, делает видимым узкий лазерный луч. Луч виден особенно хорошо, если смотреть почти навстречу ему в пределах угла приблизительно 10. Объяснить явление и оценить размер $b$ пылинок, если длина волны света 6300 Å.

### Solution 7.

4. $b \approx \frac{\lambda}{\theta} \approx 4$МКМ.

## O2-7.5 Зоркость птиц

О зоркости хищных птиц ходят легенды. Оценить, на основе дифракционных соображений, может ли орел, летающий над землей на высоте 1km, разглядеть мышонка размером в 2 см.

### Solution 7.

5. Не может. Диаметр зрачка глаза орла не превышает нескольких миллиметров. Если предположить, что он равен даже 10 мм, то минимальный угол, под которым орел может видеть раздельно две точки предмета, окажется примерно в 3 раза болыне угловых размеров мышонка.

## O2-7.6

Изображение точечного источника проецируется на экран с помощью тонкой линзы с малым апертурным числом двумя способами, реализуемыми при условии, что, расстояние от источника до экрана в обоих случаях остается постоянным и равным $L = 4$ м. При этом фокусное расстояние линзы $f = 0,75$ м. Как относятся освещенности в центре дифракционного изображения в этих двух случаях?

### Solution 7.

6. Освененность не измснится.

## O2-7.7

Изображение точечного источника проецируется на экран с помощью тонкой линзы с малым апертурным числом двумя разными способами. В первом расстояние от источника до линзы равно ее удвоенному фокусному расстоянию. Во втором это расстояние составляет 5/4 фокусного расстояния линзы. Во сколько раз изменится освещенность в центре дифракционного изображения?

### Solution 7.

7. $\frac{E_2}{E_1} = \frac{\Phi_2/\Phi_1}{S_2/S_1} = \frac{256}{625} \approx 0,41$, где $\Phi$ - поток энергии, а $S$ - площадь пятна.

## O2-7.8

В интерферометре Майкельсона источником света служит круглая диафрагма $S$ диаметром $d = 0,05$ mm, которая освещается параллельным пучком монохроматического света с длиной волны $\lambda = 0,6$ мкм. Длины плеч интерферометра $AB = 30$, $AC = 10$ cm (рис. 320). Интерференционная картина в виде концентрических колец наблюдается на экране Э, помещенном в фокальной плоскости линзы. Оценить число $m$ интерференционных колец, наблюдаемых в пределах главного дифракционного максимума источника.

### Solution 7.

8. $m \approx (1,22)^2 \frac{l\lambda}{d^2} \approx 70$, где $l = 20$ см - разность длин плеч интерферометра Майкельсона.

## O2-7.9

С искусственного спутника Земли, обращающегося по круговой орбите на расстоянии $h = 250$ км, проводится фотографирование земной поверхности. Разрешающая способность фотопленки $N = 500$ линий/мм. Какими параметрами должен обладать объектив фотоаппарата (диаметр $D$, фокусное расстояние $f$), чтобы при фотографировании разрешались детали с линейными размерами $l \approx 1$ м?

### Solution 7.

9. Диаметр объектива $D \geqslant 1,22\frac{\lambda h}{l} = 16,8$ cm; фокусное расстояние $f \geqslant \frac{h}{Nl} = 50$ cm. Таким образом, $\frac{D}{f} = \frac{1}{3}$.

## O2-7.10

С самолета, летящего на высоте $H = 5$ км, производится аэрофотосъемка местности. Какими следует выбрать фокусное расстояние $f$ и диаметр объектива $D$ фотоаппарата, чтобы сфотографировать объекты размером $l \approx 2,5$ см на фотопленку с разрешающей способностью $n = 500$ штрих/мм? На какое время вать затвор фотоаппарата (экспозиция), чтобы движение самолета со скоростью $V = 360/$ час не приводило к размытию изображения?

### Solution 7.

10. $f \geqslant 40$ cm; $D \geqslant 10$ cm; $\tau \approx 2,5 \cdot 10^{-3}$с.

## O2-7.11

На щель шириной $b$ положена стеклянная призма с показателем преломления $n$ и преломляющим углом $\alpha$ (рис. 321).

На грань $AB$ призмы нормально падает плоская монохроматическая волна. Найти направления на нулевой максимум и минимум в дифракционной картине Фраунгофера.

### Solution 7.

11. Нулевой максимум $\sin\theta = n\sin\alpha$, минимумы $b(\sin\theta - n\sin\alpha) = m\lambda$, где $m = \pm1, \pm2, \pm3, \ldots$

## O2-7.12

Плоская световая волна падает нормально на абсолютно черный экран, размеры которого велики по сравнению с длиной волны. Часть энергии поглощается черным экраном, а часть рассеивается из-за дифракции. Показать,что количество поглощенной энергии равно количеству рассеянной.

### Solution 7.

12* Решение. Рассмотрим щель той же ширины, что и черный экран. Если на щель и экран падает одна и та же плоская волна, то количество энергии, поглощенной экраном, будет равно количеству энергии, падающей на щель. Согласно принципу Бабине (см. задачу 7.14*), интенсивности света во всех направлениях, за исключением направления падающей волны, в обоих случаях одинаковы. Следовательно, одинаковы и энергии, рассеянные экраном и щелью. Но в случае щели вся энергия рассеивается. Значит, энергия, поглощенная экраном, равна энергии, рассеянной им.

## O2-7.13 Location of the Moon by a lazer

Оценить минимально необходимую для локации Луны энергию ∘ световой вспышки рубинового лазера ($\lambda = 0,7$ мкм), если отражение луча осуществлялось 14 призмами, установленными на Луноходе-1. Каждая из призм отражает луч на угол 180°. Отражение от призмы рассматривать, как отражение от плоского зеркала диаметром $d = 6$ cm. Посылка и прием луча осуществлялись телескопом Симеизской обсерватории. Диаметр зеркала телескопа $D = 2,6$ м. При приеме мог быть обнаружен сигнал, состоящий из двух фотонов.

### Solution 7.

13. $\mathscr{E} = \frac{4\pi\hbar\lambda^3 L^4}{Nd^4D^4} \approx 8$ Дж, где $\hbar$ - постоянная Планка, $N = 14$ (призм), $L \approx 380$ тыс. км - расстояние от Земли до Луны.

## O2-7.14

Доказать, что при дифракции Фраунгофера интенсивности дифрагированного света от дополнительных экранов совпадают во всех направлениях, за исключением направления падающей волны (принцип Бабине). Дополнительными называются два экрана, у которых непрозрачные места одного экрана по форме и положению совпадают с отверстиями другого.

### Solution 7.

14* Решение. Согласно принципу Гюйгенса-Френеля напряженность $E_1$ поля волны, дифрагировавшей на первом экране, представляется некоторым интегралом по отверстиям этого экрана. Напряженность $E_2$ поля волны, возникшей при дифракции на дополнительном экране, представляется таким же интегралом по отверстиям этого экрана. Сумма $E_1 + E_2$ представляется интегралом по всей бесконечной плоскости, в которой расположены рассматриваемые экраны. Но таким же интегралом представилось бы поле за экраном, если бы этого экрана совсем не было, т.е. волна распространялась свободно. Следовательно, $E_1 + E_2 = E$, где $E$ - напряженность поля волны в отсутствие экрана, т. е. поля падающей волны. Но в падающей волне свет распространяется только в одном направлении. Для всех других направлений $E = 0$, а следовательно, $E_1 + E_2 = 0$, откуда $E_1^2 = E_2^2$, или $\mathcal{I}_1 = \mathcal{I}_2$.

## O2-7.15

Точечный источник света находится на некотором расстоянии $a$ от щели шириной $D$. За щелью на расстоянии $b$ от нее помещен экран, плоскость которого параллельна плоскости щели. Прямая, соединяющая источник света с серединой щели, перпендикулярна к плоскости экрана. Найти приближенное выражение для расстояния $x$ между центральным максимумом и первым дифракционным минимумом на экране, считая, что углы дифракции малы. Найти условие применимости полученного приближенного выражения.

### Solution 7.

15* $x = \frac{b\lambda}{D}$ при условии $\frac{1}{2b^2}\left(\frac{b^2\lambda^2}{D^2} + \frac{D^2}{4}\right) \ll 1$.

Решение. При малых углах дифракции искомое расстояние $x$ найдется из требования, чтобы разность хода $r_2 - r_1$ крайних интерферирующих

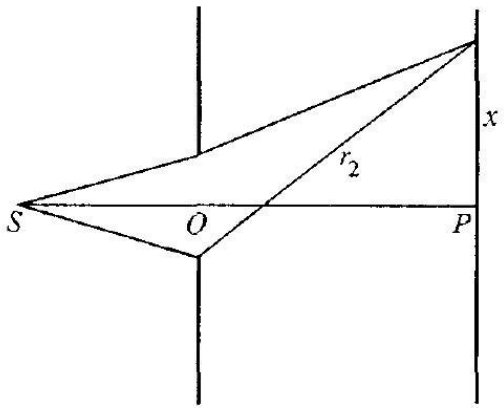



Рис. 444 лучей (рис. 444) была равна длине волны $\lambda$. Применяя теорему Пифагора и извлекая квадратные корни с помощью формулы бинома Ньютона, для этой разности нетрудно получить

$$r_2 - r_1 = \frac{xD}{b} - \frac{xD}{2b^3}\left(x^2 + \frac{D^2}{4}\right) + \dots$$

Ограничиваясь первым членом, найдем

$$x = \frac{b\lambda}{D}.$$

Ошибка при вычислении $r_2 - r_1$ должна быть мала по сравнению с $\lambda$. Отсюда получаем условие применимости предыдущей формулы

$$\frac{1}{2b^2}\left(\frac{b^2\lambda^2}{D^2} + \frac{D^2}{4}\right) \ll 1$$

## О2-7.16

Камера-обскура длиной $L = 10$ см с малым отверстием предназначена для фотографирования удаленных предметов. Оценить диаметр отверстия $D$ камеры, при котором она имеет наибольшую разрешающую способность. Длина волны $\lambda = 5000$ Å.

### Solution 7.

16 $D = \sqrt{2,44\lambda L} \approx 0,35$ мм. Вообще Рэлей доказал, что наивыгоднейший диаметр отверстия в камере обскуре подсчитывается по несколько иной

## О2-7.17

Пучок фтористо-водородного лазера, работающего в одномодовом режиме на длине волны $\lambda = 3$ мкм, формируется зеркалами диаметром $D = 3$ м. На каком максимальном расстоянии $L$ может находиться мишень, чтобы плотность потока энергии на ней была практически равна плотности потока на зеркале?

## О2-7.18

Оценить, во сколько раз отличаются напряженности электрического поля монохроматической волны $\lambda = 1$ мкм в фокусе сферического зеркала (диаметр $D = 10$ cm, радиус кривизны $R = 1$ m) и на его входе.

## О2-7.19

Плоская волна проходит через стеклянную пластинку с показателем преломления $n = 3/2$, падая на ее поверхность нормально (рис. 322). Толщина пластинки испытывает скачкообразное изменение на величину $b = 2\lambda/3$ вдоль прямой, проходящей через точку $C$ перпендикулярно плоскости рисунка. Найти интенсивность света в точке $O$, лежащей в плоскости, проходящей через точку $C$ перпендикулярно к плоскости рисунка, если интенсивность света в Этой точке в случае плоскопараллельной пластинки (т. е. при $b = 0$) равна $\mathscr{I}_0$.

### O2-7.20

Плоская волна проходит через стеклянную пластинку с показателем преломления $n$, падая на ее поверхность нормально (рис. 322). Толщина пластинки испытывает скачкообразное изменение на величину $b$ вдоль некоторой пря- мой, проходящей через точку C перпендикулярно к плоскости рисунка. При каких значениях $b$ интенсивность света в точке $O$, лежащщей в плоскости, проходящей через точку перпёндикулярно к плоскости рисунка, будет вдвое меньше интенсивности света в той же точке $O$ в случае отсутствия уступа на пластинке? Длина волны падающего света равна $\lambda$. 158

### O2-7.21

Дифракционные полосы от двух одинаковых параллельных щелей наблюдаются в фокальной плоскости линзы $L$ (рис. 323). $S_1$ и $S_2$ - бесконечно удаленные линейные источники монохроматического света, параллельные щелям. При каком угловом расстоянии между $S_1$ и $S_2$ дифракционные полосы исчезнут, если расстояние между центрами щелей равно $D$ и велико по сравнению с шириной щели и длиной световой волны $\lambda$?

### O2-7.22

В установке, описанной в предыдущей задаче источники $S_1$ и $S_2$ помещены в фокальной плоскости коллиматорной линзы с фокусным расстоянием $f$. При каком расстоянии $x$ между $S_1$ и $S_2$ дифракционные полосы исчезнут?

### O2-7.23

На рис. 324 изображена схема интерференционного опыта Юнга, в котором используется явление дифракции света на двух щелях. В качестве источника света в схеме применен лазер, работающий на длине волны $\lambda = 6328$ Å. Пучок света на выходе лазера имеет плоский волновой фронт. Диаметр пучка $d = 2$ мм. При каком расстоянии между щелями $D$ возможно наблюдение интерференционной картины на экране, если расстояние от источника до двойной щели $L = 4$ м?

### O2-7.24

Квадратное отверстие со стороной $L_0 = 0,2$ cm освещается параллельным пучком солнечных лучей, падающих нормально к плоскости отверстия. Найти форму и размер $L \times L$ изображения отверстия на экране, удаленном на 50 м от него, если плоскость экрана параллельна плоскости отверстия. Границей освещенности на экране считать положение первого дифракционного минимума наиболее сильно отклоняемых лучей (видимый спектр 7000-4000 A).

### O2-7.25

Диафрагма линзы имеет форму квадрата с длиной стороны $D$. Точечный монохроматический источник света помещается на главной оптической оси линзы. Найти распределение интенсивности света, получающееся в результате дифракции в плоскости изображения источника.

### O2-7.26

Дифракция Фраунгофера плоской волны на щели наблюдается в фокальной плоскости линзы. Во сколько раз изменится интенсивность в фокусе линзы, если щель накрыть плоскопараллельной пластинкой, амплитудный коэффицисит пропускания которой имест вид $\tau(x) = \sin(\pi x/a)$? Ось $x$ направлена перпендикулярно щели, точки $x = 0$ и $x = a$ - координаты краев щели.

### O2-7.27

Найти картину дифракции Фраунгофера плоской монохроматической неполяризованной волны на длинной прямой щели шириной $b$, вдоль оси которой симметрично расположена длинная пластинка из поляроида шириной $b/2$.

### O2-7.28

На щель шириной a нормально падает плоская волна с длиной волны $\lambda$. Lель закрыта двумя стеклянными пластинками шириной $a/2$ и толщиной $h$ с показателями преломления $n_1$ и $n_2$ и коэффициентами пропускания (по интенсивности) $\tau_1$ и $\tau_2$. Найти распределение интенсивности в дифракционной картине Фраунгофера. При каком условии в центре картины получается темная полоса?

### O2-7.29

Рассчитать и проанализировать дифракционную картину при нормальном падении света на пилообразную решетку (рис. 325), сделанную из стекла с показателем преломления $n$. Число зубьев решетки равно $N, a \gg h$. Длина волны падающего света равна $\lambda$.

### O2-7.34

Оптическая система (труба или микроскоп) дает в качестве изображения светящейся точки систему дифракционных колец. Согласно Рэлею, минимальное расстояние между двумя близкими точками, которые еще изображаются раздельно, определяется тем, что центральный светлый кружок от первой светящейся точки должен приходиться на первое темное кольцо дифракционной картины, даваемой второй светящейся точкой. Ориентировочно можно принять, что глаз способен различить две близкие точки, если максимумы освещенности в местах их геометрических изображений превосходят интенсивность посредине между ними не менее чем на 15%. Пиняв это, проверить, действительно ли при выполнении критерия Рэлея получатся раздельные изображения двух самосветящихся точек.

Указание. Для простоты расчета принять, что диафрагма квадратная. В случае круглой диафрагмы результаты мало отличаются от тех, которые получаются для квадратной диафрагмы. (См. решение задачи 7.25*)

### O2-7.35

Решить предыдущую задачу в предположении, что изображаемые точки не самосветящиеся, а освешаются одним и тем же источником света. Например, можно взять два круглых отверстия в экране, размеры которых малы по сравнению с расстоянием между ними. Рассмотреть качественно три случая: 1) отверстия освещаются пучком лучей, параллельных главной оптической оси; 2) отверстия освещаются параллельными лучами, но наклоненными к главной оптической оси; 3) отверстия освещаются дифрфузным светом.

### O2-7.36

При аэрофотосъемке местности используется объектив с фокусным расстоянием $f = 10$ cm и диаметром $D = 5$ cm. Съемка производится на фотопленку, имеющую разрешающую способность $R = 100\ \text{mm}^{-1}$. Определить, какие детали местности могут быть разрешены на фотографиях, если съемка производилась с высоты $h = 10$КМ.

### O2-7.37

Объектив телескопа имеет фокусное расстояние $f_1 = 3$ м и диаметр $D = 15$ cm. Определить фокусное расстояние $f_2$ окуляра, при котором полностью используется разрешающая способность объектива, если диаметр зрачка глаза $d = 3$ мм. Предполагая, что в системе телескоп-глаз отсутствуют аберрации, оценить, на каком расстоянии $L$ с помощью такого телескопа можно читать книгу размером букв $b \approx 2$ mm.

### O2-7.38

Современные фотопленки способны разрешать до $z = 10^4$ линий/см. Какую светосилу (т. е. отношение квадратов диаметра $D$ и фокусного расстояния $f$) должен иметь объектив фотоаппарата, чтобы полностью использовать разрешающую способность плснки?

### O2-7.39

Каково должно быть увсличение зрительной трубы для того, чтобы полностью использовать разрешающую способность ее объектива?

### O2-7.40

1) Подсчитать разрешающую способность зрительной трубы с диаметром объектива 5 см. 2) При каком увеличении будет использована полная разрешающая способность этой трубы? Диаметр зрачка глаза $d = 5$ мм.

### O2-7.41

Самый большой в мире телескоп был сооружен в России и установлен в астрономической обсерватории на северных отрогах Кавказского хребта, вблизи станции Зеленчукская. Диаметр зеркала Этого телескопа $D = 6$ м. Найти разрешаемое им угловое расстояние $\delta\theta$ для длины волны $\lambda = 5500$ Å.

### O2-7.42

В чем выгода применения телескопов для рассматривания звезд, если они не дают увеличения по сравнению с невооруженным глазом?

### O2-7.43 Update of a telescope

В принципе можно построить телескоп сколь угодно высокой разрешающей способности без объектива, заменив объектив круглым отверстием. Какова должна быть длина $L$ такого телескопа, чтобы он имел ту же разрешающую способность, что и обычный телескоп с диаметром объектива $D = 1$ м? Чему будет равна светосила $S$ такого телескопа?

### O2-7.50

Луч лазера фокусируется идеальной оптической системой с отношением $f/D = 1$. Оценить мощность $N$ лазера, при которой в электрическом поле в фокусе системы электроны смогут приобретать энергию порядка энергии покоя $m_0c^2$. Как зависит $N$ от длины волны $\lambda$? Какое магнитное поле $B$ будет при этом в фокусе?

### O2-7.51

Ракета удаляется от Земли и перестает быть видимой на фоне неба в телескоп с объективом диаметром $D_1 = 80$m, когда она находится на расстоянии $L_1 = 2 \cdot 10^4$ км от Земли. На каком расстоянии $L_2$ от Земли удается заметить эту ракету в телескоп с объективом диаметром $D_2 = 200$ мм при той же контрастной чувствительности глаза?

### O2-7.52

: Излучение лазера непрерывного действия на длине волны $\lambda = 0,63$ мкм мощностью $N = 10$ мВт направляется на спутник с помощью телескопа, объектив которого имеет диаметр $D = 30$ см. Свет, отраженный спутником, улавливается другим таким же телескопом и фокусируется на фотоприемник с пороговой чувствительностью $N_{\text{пор}} = 10^{-14}$В. Оценить максимальное расстояние $L_{\max}$ до спутника, на котором отраженный сигнал еще может быть обнаружен. Поверхность спутника равномерно рассеивает падающий свет

### O2-7.53

Оценить расстояние $L$, с которого можно увидеть невооруженным гтазом свет лазера, генерирующего в непрерывном режиме мошность $N = 10$ Вт на частоте $v = 6 \cdot 10^{14}$ Гц, если для формированищ луча используется параболическое зеркало диаметром $D = 50$ cm. Глаз видит источник в зеленой части спектра, если в зрачок (диаметр зрачка $d = 5$ мм) попадает $n = 60$ квантов в секунду.

### O2-7.54

В фокальной плоскости объектива телескопа помещена фотопластинка. Освещенность изображения звезды на фотопластинке в $\alpha = 10$ раз меньше освещенности дневного неба. Во сколько раз надо увеличить диаметр объектива, чтобы освещенность изображения звезды на фотопластинке стала в $\beta = 10$ раз больше освещенности изображения неба?

### O2-7.55

При наблюдении в телескоп с нормальным увеличением освещенность изображения звезды на сетчатке глаза в $\alpha = 10$ раз меньше освещенности дневного неба, рассматриваемого в тот же телескоп. Во сколько раз надо увеличить диаметр объектива для того, чтобы освещенность изображения звезды на сетчатке стала в $\beta = 10$ раз больше освещенности изображения неба, если вместе с объективом телескопа заменен и окуляр таким образом, что увеличение телескопа осталось нормальным?

### O2-7.56

Каково должно быть увеличение микроскопа, чтобы полностью использовать разрешающую способность его объектива?

### O2-7.57

Проволочная сетка, ячейки которой имеют форму квадратов, проецируется собирающей линзой на экран. В задней фокальной плоскости линзы помещена узкая щель. Как будет меняться картина на экране при вращении щели вокруг главной оптической оси линзы? Как изменится картина при увеличении ширины щели (опыты Аббе)?

### O2-7.58

С помощью объектива микроскопа получена микрофотография малого объекта (например, растительных клеток или бактерий) с линейным увеличением $N$. Тот же объектив был использован для проецирования полученной микрофотографии на удаленный экран. Каково должно быть минимальное значение увеличения $N$, чтобы полностью использовать разрешающую способность микроскопа? Диаметр апертурной диафрагмы объектива равен $D$, диаметр зрачка глаза $d$. Изображение на экране рассматривается с места нахождения объектива.

### O2-7.59

Определить минимальное разрешаемое расстояние $\partial$ микроскопа при наилучших условиях освещения для 1) безыммерсионного объектива с числовой апертурой $a = 0,9,2)$ того же объектива, но с масляной иммерсией $(n = 1,6)$. Длина волны при визуальных наблюдениях $\lambda = 5500$ Å.

### O2-7.63

С помощью объектива телескопа с диаметром $D$ и фокусным расстоянием $f$ производится фотографирование удаленных объектов на мелкозернистой пластинке, помещенной в фокальной плоскости объектива. Полученное изображение рассматривается в микроскоп с числовой апертурой $n \sin \alpha$ и увеличением $N$. Каким условиям должны удовлетворять числовая апертура и увеличение микроскопа, чтобы полностью использовать разрешающую способность объектива телескопа?

### O2-7.64

Угловая апертура электронного микроскопа $\Omega = 10^{-4}$, а оптического $-\Omega_{\text{опт}} \approx 1$. Оценить напряжение $U$, ускоряющее электроны, при котором разрешающая способность этих приборов будет одинакова.

## 12.0.12 Problems about spectral devices 1

### МФТИ.маг-вступ-2023.v1.4!(2 мин) Разрешение света из трубки с толщиной

Дифракционная решетка шириной $L = 5$с, имеющая $n = 200$ штрих/мм освещается нормально падающим параллельным пучком света. Найти наименьшую разность длин волн, которую может разрешить данная решетка вблизи длины волны $\lambda = 550$ нм.

#### Solution

### O2-8.1 Разрешающая сила для двух длин волн

Какую разрешающую силу должен иметь спектральный аппарат для разрешения дублета $D$-линии натрия ($\lambda_1 = 5890$ Å, $\lambda_2 = 5896$ Å)?

#### Solution 8.

1. $R \approx 1000$.

### O2-8.2

Найти величину наименьшего основания призмы $b$, изготовленной из стекла, дисперсия которого вблизи $D$-линии натрия $dn/d\lambda = 956$ cm$^{-1}$, чтобы призма смогла разрешить желтый дублет натрия ($\lambda_1 = 5890$; Å, $\lambda_2 = 5896$ Å).

#### Solution 8.

2. $b \approx 1$ cm.

## O2-8.3

Спектрограф имеет стеклянную призму с основанием $a = 10$ cm и преломляющим углом $A = 60°$, устанавливаемую при работе на угол наименьшего отклонения вблизи длины волны $\lambda = 5000$ Å. Показатель преломления стекла призмы $n = 1,73$; фокусное расстояние объектива коллиматора $f = 25$ cm. Какова должна быть ширина коллиматорной щели $b$, чтобы можно было практически нолностью использовать теоретическую разрешающую способность призмы?

### Solution 8.

3. $b \ll 2,5 \cdot 10^{-3}$m.

Решение. Пусть $\delta\lambda$ - минимальная разность длин волн двух спектральных линий, разрешаемая призмой при бесконечно узкой коллиматор-

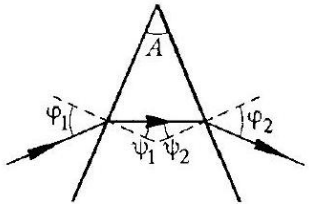



Рис. 448 ной щели. Согласно формуле для теоретической разрешающей способности $\lambda/\delta\lambda = a|dn/d\lambda|$, разность показателей преломления для этих спектральных линий $\delta n = \lambda/a$. Благодаря различию в показателях преломления первоначально параллельный пучок лучей по выходе из призмы сделается расходящимся. Рассчитаем угловое расхождение вышедшего пучка в предюоложении, что на призму падал параллельный пучок. Имеем (рис. 448) $\sin\varphi_1 = n\sin\psi_1$. Отсюда при постоянном $\varphi_1$
$\delta n\sin\psi_1 + n\cos\psi_1\delta\psi_1 = 0$.
Так как $\psi_1 + \psi_2 = A = \text{const}$ и, следовательно, $\delta\psi_1 + \delta\psi_2 = 0$, то $\delta n\sin\psi_1 = n\cos\psi_1\delta\psi_2$.

Далее, из $\sin\varphi_2 = n\sin\psi_2$ находим искомое угловое расхождение вышедшего пучка:

$$\partial\varphi_2 = \frac{\sin\psi_2}{\cos\varphi_2}\partial n + \frac{n\cos\psi_2}{\cos\varphi_2}\partial\psi_2 = \left(\frac{\sin\psi_2}{\cos\varphi_2} + \frac{n\cos\psi_2}{\cos\varphi_2}\frac{\sin\psi_1}{n\cos\psi_1}\right)\delta n,$$

а при установке на угол наименьшего отклонения ($\varphi_1 = \varphi_2 = \varphi$; $\psi_1 = \psi_2 = \psi$) :

$$\delta\varphi_2 = 2\frac{\sin\psi}{\cos\varphi}\delta n = 2\frac{\sin\psi}{\cos\varphi}\frac{\lambda}{a}.$$

Для полного использования теоретической разрешающей способности призмы необходимо, чтобы угловая ширина коллиматорной щели $\alpha = b/f$ была мала по сравнению с $\delta\varphi_2$. Это дает

$$b \ll 2\frac{\sin\psi}{\cos\varphi}\frac{f\lambda}{a}.$$

Для $A/2 = \psi = 30°, n = 1,73$ получаем

$$b \ll \frac{2}{\sqrt{4-n^2}}\frac{f\lambda}{a} = \frac{2f\lambda}{a} = 2,5\cdot 10^{-3}\text{ mm}.$$

## O2-8.4

Найти угловое распределение интенсивности света при фраунгоферовой дифракции на решетке из $N$ щелей и с периодом $d$ при условии, что световые лучи падают на решетку нормально, а ширина щели равна $b$.

### Solution 8.

4. $\mathcal{I} = C\left[\frac{\sin\left(\frac{N\pi d}{\lambda}\sin\theta\right)}{\sin\left(\frac{\pi d}{\lambda}\sin\theta\right)}\right]^2\left[\frac{\sin\left(\frac{\pi b}{\lambda}\sin\theta\right)}{\frac{\pi b}{\lambda}\sin\theta}\right]^2$,

где $C$ - постоянная, а $\theta$ - угол между нормалью к решетке и заданным направлением, к которому относится интенсивность $\mathscr{A}$.

Решение. Так как требуется рассчитать распределение интенсивности в дифракционной картине Фраунгофера на решетке, то можно считать,

что экран, на котором наблюдается эта картина, отодвинут достаточно далеко. Лучи, приходящие из различных щелей решетки в любую точку экрана, можно считать параллельными. Выберем такую точку на экране, чтобы приходящие в нее лучи образовывали уголя $\theta$ с нормалью к решетке. Разность фаз между соседиими интерферирующими лучами равна $\delta = (2\pi d\sin\theta)/\lambda$. На векторной диаграмме колебание, приходящее в рассматриваемую точку от какой-либо щели решетки, можно изобразить

вектором. Эти векторы образуют ломаную линию с равными звеньями и равными углами (рис. 449). Результирующее колебание изобразится геометрической суммой всех эих векторов. Сложение векторов на пиоскости формально тождественно со сложением комплексных чисел.

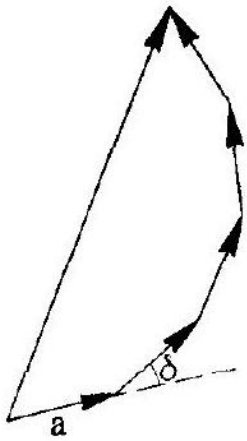



Рис. 449 Если первый вектор изобразить комплексным числом $a$, то последующие изобразятся комплексными числами $ae^{i\delta}$, $ae^{2i\delta}$ и т.д. Их сумма равна

$$A = a\left[1 + e^{i\delta} + e^{2i\delta} + \ldots + e^{i(N-1)\delta}\right] = a\frac{1 - e^{iN\delta}}{1 - e^{i\delta}}$$

Квадрат модуля комплексного числа $A$ дает относительную интенсивность света в рассматриваемой точке экрана:

$$g = |A|^2 = AA^* = |a|^2\frac{1 - e^{iN\delta}}{1 - e^{i\delta}}\frac{1 - e^{-iN\delta}}{1 - e^{-i\delta}} = \\ = |a|^2\frac{2 - \left(e^{iN\delta} + e^{-iN\delta}\right)}{2 - \left(e^{i\delta} + e^{-i\delta}\right)} = |a|^2\frac{1 - \cos(N\delta)}{1 - \cos\delta}$$

Так как $|a|^2$ определяет интенсивность дифрагированного света на одной щели, то но извсстной формуле

$$|a|^2 = C\left[\frac{\sin\left(\frac{\pi b}{\lambda}\sin\theta\right)}{\frac{\pi b}{\lambda}\sin\theta}\right]^2$$

Подставляя это значение в предыдущее выражение, легко получить ответ. Постоянная пропорииональна интснсивности падающего света и квадрату ширины щели $b$. Ни от каких других параметров решетки она не зависит.

## O2-8.5

Дополнительными, или взаимно дополнительными, дифракционными решетками называются такие щелевые решетки, у которых непрозрачные места одной решетки совпадают с прозрачными в другой. Убедиться непосрсдственным расчстом, что при освещснии таких решеток одним и тем же световым пучком положения и интенсивности главных дифракционных максимумов одного и того же порядка совпадают, за исключением главных максимумов нулевых порядков. Если же число штрихов решетки очень велико, то совпадают также интенсивности дифрагированного света для всех направлений, за исключенисм направления падающего света.

## Solution 8.

5* Решение. Представим ответ предыдущей задачи в виде

$$g(\theta) = Cb^2\left(\frac{\sin\alpha}{\alpha}\right)^2\left(\frac{\sin N\beta}{\sin\beta}\right)^2,$$

где $\alpha = \pi b\sin\theta/\lambda, \beta = \pi d\sin\theta/\lambda$, а постоянная $C$ от параметров решетки не зависит. Подставляя значения $\alpha$, получим

$$\mathcal{I}(\theta) = C\lambda^2\frac{\sin^2\alpha}{\pi^2\sin^2\theta}\left(\frac{\sin N\beta}{\sin\beta}\right)^2.$$

Для дополнительной решетки

$$g'(\theta) = C\lambda^2 \frac{\sin^2 a'}{\sin^2 \theta} \left( \frac{\sin N\beta}{\sin \beta} \right)^2$$

где $a' = [\pi(d-b)\sin\theta]/\lambda$, так что $a + a' = (\pi d \sin\theta)/\lambda = \beta$. Для главных максимумов $d\sin\theta = m\lambda, \alpha + \alpha' = m\lambda$. Поэтому $\sin^2\alpha = \sin^2\alpha' u$, следовательно, $\mathcal{I}'_{\text{г}} = \mathcal{I}_{\text{гл}}$ . При $m = 0$ рассуждение неприменимо, так как в этом случае в знаменателе $\sin\theta = 0$. Допустим теперь, что $\theta$ не есть уюол на главный максимум. При больших значениях $N$ главные максимумы очень резкие. Практически весь свет концентрируется в главных максимумах, заиимающих очень узкие интервалы углов $\Delta\theta$, стремящисся к нулю при $N \to \infty$. Практнческое значенис имеют только уюлы, удовлетворяющие условию $d\sin\theta = m\lambda$, а также углы, отлчающнеся от них ничтожно мало, а в этих случаях применимо доказательство, приведенное выше.

## O2-8.6

Параллельный пучок монохроматического света падает на дифракционную решетку с заданной полной шириной ее штрихованной поверхности. При каком значении отношения $b/d$ ширины щели $b$ к периоду решетки $d$ интенсивность главных дифракционных максимумов 2-го поnялка максимальна?

### Solution 8.

6. $\frac{b}{d} = \frac{1}{4}$ или $\frac{b}{d} = \frac{3}{4}$ (см. предыдущую задачу).

## O2-8.7

Показать, что для обычной амплитудной щелевой дифракционной решетки справедливо неравенство $\mathscr{V}_{\text{диф}} \leqslant \mathscr{V}_{\text{пад}}/4$, где $\mathscr{V}_{\text{пад}}$ - полный поток энергии, падающий на решетку, $_{\text{диф}}$ - поток энергии дифрагированного света, приходящийся на все дифракционные максимумы, кроме максимума нулевого порядка. Максимальное значение дифрагированного потока энергии достигается в случае, когда ширина щелей решетки равна половине ее периода.

### Solution 8.

7* Решение. Введем обозначения $q = b/d, q(d-b)/d = 1-q$. Суммарный ноток энергии прошедшего света, распределяюиийся по всем максимумам, пропорционален $q$, интенсивность нулевого максимума пропорциональна $q^2$, а потому

$$\frac{\mathcal{I}_{\text{прош}}}{q} = \frac{\mathcal{I}'_{\text{прош}}}{q'}, \quad \frac{\mathcal{I}_0}{q^2} = \frac{\mathcal{I}'_0}{q^2}.$$

По теореме Бабине для дополнительных решеток $\mathcal{G}_{\text{диф}} = \mathcal{G}'_{\text{диф.}}$ Кроме того, $\mathcal{I}_{\text{прош}} = \mathcal{I}_0 + \mathscr{F}_{\text{диф.}}$ . Исключив из написанных соотношений $\mathcal{I}'_{\text{диф и}}\,\mathcal{I}'_0$, найдем $\mathscr{I}_{\text{диф}} = (1-q)\mathscr{V}_{\text{прон}}$ Учитывая также, что $\mathcal{I}_{\text{прон}} = q\mathscr{O}_{\text{пад}}$ , получим $_{\text{диф}} = q(1-q)\mathscr{O}_{\text{пад}}$ . Максимум для $\mathscr{O}_{\text{диф}}$ получается при $q = 1/2$. При этом $\mathcal{I}_{\text{диф}} = \mathcal{I}_{\text{пад}}/4$.

## O2-8.8

Найти угловое распределение дифракционных минимумов при дифракции на решетке, период которой равен $d$, а ширина щели равна $b$.

### Solution 8.

8. $Nd\sin\theta = n\lambda$, но $d\sin\theta \neq k\lambda$, $b\sin\theta = m\lambda$, где $n, m, k$ - цепые чнсла. Условие минимума - выполнение хотя бы одного из этих равенств.

## O2-8.9

Найти условие появления главного дифракционного максимума при наклонном падении лучей на решетку (угол падения $\theta_0$). 166 Какой вид принимает это условие, если $d \gg \lambda$, а порядок спектра $m \ll d/\lambda$?

### Solution 8.

9. $d(\sin\theta - \sin\theta_0) = n\lambda$. Если $d \gg n\lambda$, то условие максимума принимает вид $d\cos\theta_0(\theta - \theta_0) \approx n\lambda$, т. е. постоянная репетки как бы уменьшилась по сравнению со случаем нормального падения и стала равной $d\cos\theta_0$ вместо $d$. При этом углы $\theta - \theta_0$, определяющие направления на максимумы, отсчитываются от направления падающего света (или отраженного в отражающей решетке).

## O2-8.10

Пучок рентгеновских лучей падает на решетку с периодом 1 мкм под углом $89°30'$. Угол дифракции для спектра второго порядка равен 89°. Найти $\lambda$.

### Solution 8.

10. $= 0,573$ Å.

## O2-8.11

Найти условие равенства нулю интенсивности $m$-го максимума для дифракционной решетки с периодом $d$ и шириной щели $b$.

### Solution 8.

11. $m = \frac{nd}{b}$, где $n = 1, 2, 3, \ldots$

## O2-8.12

Описать характер спектров дифракционной решетки, ссли ее постоянная $d$ равна 1) удвоенной, 2) утроенной, 3) учетверенной ширине щели $b$.

### Solution 8.

12. Исчезнут: 1) спектры порядков $2, 4, 6, \ldots$
2) спектры порядков $3, 6, 9, \ldots$
3) спектры порядков $4, 8, 12, \ldots$

## O2-8.13

Какой максимальный порядок спектра может наблюдаться при дифракции света с длиной волны $\lambda$ на решетке с периодом $d$?

### Solution 8.

13. Максимальный порядок равен наибольшему из целых чисел, не превосходящих $d/\lambda$.

## O2-8.14

Определить длину волны спектральной линии, изображение которой, даваемое дифракционной решеткой в спектре 3-го порядка, совпадает с изображением линии $\lambda = 4861$ Å в спектре 4-го порядка.

### Solution 8.

14. $\lambda = 6481$ Å.

## O2-8.15

Чем опрсделяется максимальная длина волны, которая может получиться в спектре дифракционной решетки? Определить постоянную, которую должна иметь решетка, способная давать инфракрасный спектр с длинноволновой границей 100 мкм.

### Solution 8.

15. При нормальном падении $\lambda_{\max} = d$. Период ренетки должен быть не менее 0,01 см, т. е. решетка должна иметь не более 10 штрих/мм. При скользящем падении на пропускающую решетку $\lambda_{\max} = 2d$.

## O2-8.16

При каком условии можно наблюдать зеркальное отражение от шероховатой поверхности при малых и больших углах падения?

### Solution 8.

16* Если угол падения близок к л/2, зеркальное отражение наблюдается всегда. При мальх углах падения зеркальное отражение может наблюдаться, если шероховатость поверхности $h \ll \lambda$.

Р с и с иие. Пусть падающис лучи образуют угол 0 с нормалью к плоскости, изображенной на рис. 450 птриховой линисй. Рассмотрим интерференцию вторичных волн, исходящих от поверхности тела под углом $\theta$ к нормали. Разность хода каких-либо двух лучей, идуиих в рассматриваемом направ:ении, определяется выражением

$$\Delta = AD - BC = a\left(\sin\theta' - \sin\theta\right) + h\left(\cos\theta' + \cos\theta\right).$$

Здесь а может принимать какие угодно значсния. Поэтому если $\theta' \neq \theta$, то разность хода $\Delta$ может иринимать также какие угодно значения, и притом для неправильной поверхности тела эти значения будут встречаться одинаково часто. Это показывает, что правильное отражсние, если таковое возможно, может происходить лишь под углом $\theta' = \theta$. В таком случае

$$\Delta = 2h\cos\theta.$$

Отсюда видно, что, каков бы ни был

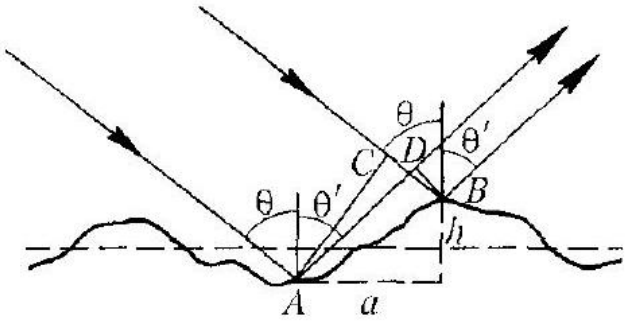



Рис. 450 размер $h$, можно подобрать достаточно большой угол $\theta$, для которого $\Delta \ll \lambda$. При этом условии отражение будет правильное. При нормальном падении $\Delta = 2h$ и правильное отражение возможно лииь ири соблюдении условия $h \ll \lambda$.

## O2-8.17

Могут ли перекрываться спектры 1 -го и 2 -го порядков дифракционной решетки при освещении ее видимым светом (7000 – 4000 Å)?

### Solution 8.

17. Не могут.

## O2-8.18

На одномерную фазовую дифракционную решетку (рис. 330) нормально падает плоская монохроматическая волна. На участках и $b$ оптические пути равны соответственно $l_1 = n_1 z_1$ и $l_2 = n_2 z_2$ ($n_1$ и $n_2$ - показатели преломления, $z_1$ и $z_2$ - соответствующие толщины). Как меняются интенсивности дифракционных максимумов и соответствующие им углы дифракции при $n_1 \to n_2$ и $z_1 \to z_2$?

### Solution 8.

18. Углы дифракиии не меняются. Интенсивности стремятся к нулю.

## O2-8.19

Прозрачная периодическая структура, профиль которой изображен на рис. 331, освещается сверху плоской монохроматической волной, падающей нормально на верхнюю границу. Ширина уступов и впадин структуры одинакова. При заданном показателе преломления $n$ подобрать глубину $h$ таким образом, чтобы главные фраунгоферовы дифракционные максимумы 1 -го порядка имели наибольшую интенсивность. Какова при этом интенсивность нулевого главного максимума?

### Solution 8.

19. $h = \frac{2m-1}{2(n-1)}$, где $m = 1, 2, 3, \ldots$ Интенсивность нулевого главного максимума равна нулю.

## O2-8.20

На одномерную дифракционную решетку со щелями (рис. 332) падает плоская монохроматическая волна. Щели $b$ решетки совсем прозрачные, а участки $a$ имеют коэффициент пропу- скания $\alpha$. Как меняются интенсивности дифракционных максимумов и соответствующие им углы дифракции, когда $\alpha \to 1$?

### Solution 8.

20. Угыы дифракции не меняются. Иитенсивности стремятся к нулю.

## O2-8.21

На дифракционную решетку нормально падает параллельный пучок монохроматического света с длиной волны $\lambda$. Решетка имеет $N$ щелей шириной $b$ и периодом $d$. Найти угловое распределение интенсивности света за решеткой, если закрыть центральную ее часть непрозрачным экраном так, что по краям решетки открытыми остаются по $n$ щелей. Нарисуйте (качественно) график зависимости интенсивности $\mathcal{I}$ от $\sin\theta$.

### Solution 8.

21. $\mathcal{I}(\sin\theta) =$

$$= g_1 \frac{1-\cos n\delta}{1-\cos\delta} 2[1+\cos(N-n)\delta] = 4\delta_1 \frac{\sin^2(n\delta/2)}{\sin^2(\delta/2)} \cos^2\left(\frac{N-n}{2}\delta\right).$$

где $\delta = \frac{2\pi}{\lambda} d\sin\theta, g_1 = g_0\left(\frac{\sin u}{u}\right)^2, u = \frac{\pi b \sin\theta}{\lambda}$. График изображен на рис. 451 .

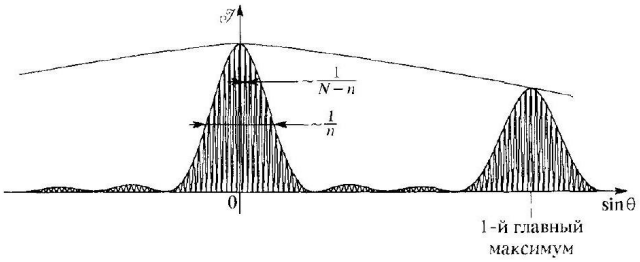



Рис. 451

## O2-8.22

На дифракционную решетку с шириной щелей $b$, равной половине периода $d(d = 2b)$, нормально падает плоская монохроматическая волна. Найти долю энергии падающей волны, рассеянную в нулевой и в каждый из 2 -х и 1 -х порядков $(d = 10\lambda)$.

### Solution 8.

22. $\frac{\mathcal{I}_0}{\mathcal{I}_{\text{пад}}} = \frac{(Nb)^2}{(Nd)^2} = \frac{1}{4}$; $\quad \frac{\mathcal{I}_{\pm 1}}{\mathcal{I}_{\text{пад}}} = \frac{4N^2b^2}{\pi^2N^2d^2} \approx 0,1$; $\quad \frac{\mathcal{I}_{\pm 2}}{\mathcal{I}_{\text{пад}}} = 0$.

## O2-8.23

Дифракционная решетка шириной $a = 3$ см с числом штрихов $N = 10^4$ освещается параллельным пучком света от натриевой лампы $J$. Пучок формируется с помощью щели $S$ шириной $\delta = 0,05$ мм, помещенной в фокусе линзы $L$ с фокусным расстоянием $f = 10$ cm (рис. 333). Лалее следует стандартное фраунгоферово расположение. В каком порядке спектра на экране Э может быть разрешен желтый дублет натрия (588, 996 и 589,593 нм)?

### Solution 8.

23. $m > \frac{\delta}{f}\frac{a}{N\delta\lambda} = 2,5$; $\quad m \geqslant 3$.

## O2-8.24

Дифракционная решетка шириной a освещается параллельным пучком света от натриевой лампы, при этом средняя длина волны, а $\partial\lambda$ - расстояние между линиями дублета. Далее на расстоянии $l$ помещена линза $\pi$ диаметром $D$, а в ее фокальной плоскости - экран наблюдения (рис. 334). Какому условию должно удовлетворять расстояние $l$, чтобы желтый дублет натрия (588, 996 и 589, 593 nm) разрешался на экране?

### Solution 8.

24. $l < \frac{Da\delta\lambda}{2\lambda^2}$. Кроме того, $l_{\min} \gg D$ (условие малости углов).

## O2-8.25

Спектр некоторого вещества в видимой области содержит ряд спектральных линий в диапазоне от 400 нм до 600 нм с минимальной разницей длин волн $\delta\lambda = 0,5$ Å Спектр изучается с помощью достаточно большой дифракционной решетки с периодом $d = 0,01$ мм. С помощью линзы спектр проецируется на экран, расположенный в ее фокальной плоскости, и рассматривается затем невооруженным 168 глазом с расстояния наилучшего зрения ($L = 25$ cm). Определить минимальные значения диаметра линзы $D$ и ее фокусного расстояния $f$, при которых наблюдатель может разглядеть все линии спектра. Диаметр зрачка глаза принять равным $d_{зр} = 0,5$ cm.

### Solution 8.

25. $D_{\min} = Nd = 5$ cm; $\quad f_{\min} = 250$ cm.

## МФТИ-ГКЭ.2021.4 Разглядеть Плутон (!!!!!!!???????)

4. Человек невооруженным глазом способен разглядеть на ночном небе подобную Солнцу звезду, находящуюся на расстоянии не более $L_{зв} = 15$ пк. Плутон - покрытая льдом карликовая планета Солнечной системы - рассеивает равномерно во все стороны $\alpha = 60\%$ попадающего на неё солнечного света. Оцените максимальное расстояние $(L_{\text{П}})_{\max}$ на котором человек может разглядеть Плутон в телескоп. Оцените диаметр телескопа $D_{\text{Т}}$ , в который можно с Земли гарантировано увидеть глазом реальный Плутон, если максимальное расстояние от него до Солнца $L_{\text{П}} = 50$ а.е. Диаметр Плутона $D_{\text{П}} = 2400$ км. Расстояние от Земли до Солнца $L_C = 1$ а.е. $= 1,5 \cdot 10^8$k. Угловой размер (диаметр) Солнца приближенно равен $\gamma = 0,01$ рад. Диаметр зрачка глаза принять равным $D_{\text{Г}} = 3$ мм. Примечание: 1 парсек (пк) - расстояние, с которого радиус орбиты Земли виден под углом в одну угловую секунду.

### Solution

(???)

Пусть яркость Солнца равна $B_0$. Яркость видимого дифракционного пятна от звезды

$$B_{3\text{ в}} = B_0 \left(1,22\frac{\lambda}{D_\text{r}}\right)^{-2} \left(\frac{D_\text{C}}{L_{3\text{ в}}}\right)^2 = B_0 \left(1,22\frac{\lambda}{D_\text{r}}\right)^{-2} \left(\gamma\frac{L_\text{C}}{L_{3\text{ в}}}\right)^2 \approx 3\cdot 10^{-10} B_0$$

Для того, чтобы планету можно было разглядеть в телескоп, яркость ее поверхности должна превысить установленный порог видимости. Тогда запишем

$$(B_\text{П})_\text{min} = B_0 \left(\frac{D_\text{C}/2}{(L_\text{П})_\text{max}}\right)^2 = \alpha B_0 \left(\frac{\gamma L_\text{C}/2}{(L_\text{П})_\text{max}}\right)^2 = B_{3\text{ в}}$$

откуда

$$(L_\text{П})_\text{max} = \frac{\gamma L_\text{C}}{2}\sqrt{\frac{\alpha B_0}{B_{3\text{ в}}}} \approx 220 \text{ а.е.}$$

Яркость поверхности реального Плутона

$$B_\text{П} = (B_\text{П})_\text{min} \left(\frac{(L_\text{П})_\text{max}}{(L_\text{П}}\right)^2$$

Яркость изображения планеты в телескопе должна быть не менее $(B_\text{П})_\text{min}$

$$B_\text{П} \left(1,22\frac{\lambda}{D_\text{T}}\right)^{-2} \left(\frac{D_\text{П}}{L_\text{П}}\right)^{-2} = (B_\text{П})_\text{min} = B_\text{П} \left(\frac{L_\text{П}}{(L_\text{П})_\text{max}}\right)^{-2}$$

откуда

$$D_\text{T} = 1,22\lambda \frac{L_\text{П}}{D_\text{П}} \frac{L_\text{П}}{(L_\text{П})_\text{max}} = 45 \text{ cm}$$

## O2-8.26

Над центром изношенной граммофонной пластинки помещен точечный источник света на высоте $h_1 = 1$ см. Глаз наблюдателя, расположенный на расстоянии $a = 110$ cm от оси пластинки и на высоте $h_2 = 10$ cm, видит, помимо геометрического изображения источника, систему дифракционных полос на поверхности пластинки. Определить расстояние $\Delta x$ между ними, если расстояние между бороздками $d = 0,5$ мм. Длина волны $\lambda = 5500$ Å.

### Solution 8.

26* $\Delta x = 1$ см.
Решение. Условие максимума $m$-го порядка (рис. 452)

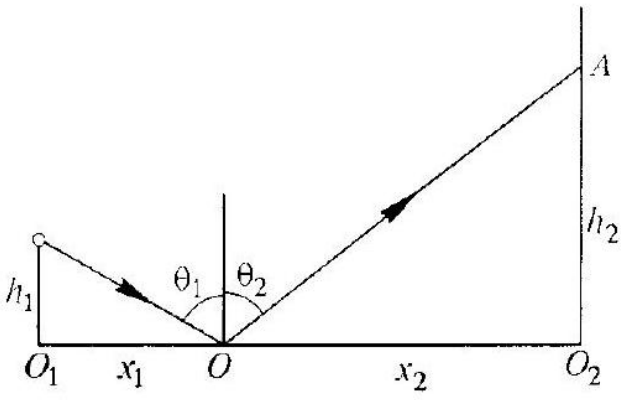



Рис. 452

$$d\,(\sin\theta_1 - \sin\theta_2) = m\lambda$$

а максимума $(m+1)$-го порядка

$$d\left(\sin\theta_1' - \sin\theta_2'\right) = (m+1))$$

При переходе от одного максимума к другому угиы $\theta_1$ и $\theta_2$ получают приращения $\Delta\theta_1$ и $\Delta\theta_2$, связанные соотношением

$$d\left(\cos\theta_1\Delta\theta_1 - \cos\theta_2\Delta\theta_2\right) = \lambda$$

Кроме того, $x_1 = h_1\,\mathrm{tg}\,\theta_1, x_2 = h_2\,\mathrm{tg}\,\theta_2$,
причем $x_1 + x_2 = a = \mathrm{const}$. Из этого условия получается второе соотношение

$$\frac{h_1}{\cos^2\theta_1}\Delta\theta_1 + \frac{h_2}{\cos^2\theta_2}\Delta\theta_2 = 0$$

Из этих двух соотношений находим $\Delta\theta_1$ и $\Delta\theta_2$, а после этого расстояние между максимумами:

$$\Delta x = \Delta x_1 = \Delta x_2 = \frac{h_1}{\cos^2\theta_1}\Delta\theta_1 = -\frac{h_2}{\cos^2\theta_2}\Delta\theta_2$$

Вычисления можно упростить, заметив, что углы $\theta_1$ и $\theta_2$ мало отличаются от угла падения $\theta_0$, соответствуюнего правильному отражению света от пластинки. Заменив эти углы на $\theta_0$, находим

$$\Delta x = \frac{h_1h_2}{h_1+h_2}\frac{\lambda}{d\cos^2\theta_0}$$

При этом $\cos\theta_0 \approx h_1/x_1 \approx h_2/x_2$, или $\cos\theta_0 = (h_1+h_2)/(x_1+x_2) = (h_1+h_2)/a = 1/10$. Подстановка числовых значений дает ответ.

**(???? условия задач опущены, потом допишу их!!!)**

### Solution 8.

27. 1) 8,1 угл. с/ Å; 2) $0,0197$m/ Å; 3) $50,7$ Å /mm.

### Solution 8.

28. Около 0,12 мм.

### Solution 8.

29. $D = \frac{n}{d\cos\theta} = \frac{n}{d\sqrt{1-(n\lambda/d)^2}} = 0,63\cdot 10^4$pa/cm $= 13$ угл. с/ Å.

### Solution 8.

30. 600 штрих/мм .

### Solution 8.

31. $\Delta\theta = \frac{\lambda}{Nd\cos\theta}$.

### Solution 8.

32. Около 1000 .

### Solution 8.

33. 12000 и 48000 .

### Solution 8.

34. Нет.

### Solution 8.

35. Разрешающая способность не изменится. Дисперсионная область уменьшится вдвое.

### Solution 8.

36. $\Delta x = \frac{f\lambda}{l} = 3,3\text{mK}$.

### Solution 8.

37. $b \ll \frac{f\lambda}{Nd} \approx 0,001$ cm.

## O2-8.38

Свет от газоразрядной трубки, диаметр которой $D = 1$ cm, непосредственно падает на дифракционную решетку, расположенную на расстоянии $L = 100$ см. Оценить, какой будет в этих условиях максимальная разрешающая способность $R_{\max} = \lambda/\delta\lambda$.

### Solution 8.

38* $R_{\max} \approx 100$.
Р еш ение. Радиус пространственной когерентности в плоскости решетки

$$r_{\text{Kor}} \approx \frac{\lambda L}{D}.$$

Эффективно работающее число щелей $N_{\text{эфф}} \approx r_{\text{ког}}/d$, где $d$ - период решетки. Разрешающая способность решетки

$$R = \frac{\lambda}{\delta\lambda} = mN_{\text{Эф}\phi}, \quad m_{\max} \approx \frac{d}{\lambda}$$

Отсюда максимальная разрешающая способность

$$R_{\max} \approx \frac{d}{\lambda}\frac{r_{\text{КОТ}}}{d} \approx \frac{L}{D} = 100$$

## МФТИ.маг-вступ-2023.4v1-dem (2 мин) Разрешение света из трубки с толщиной

Свет от газоразрядной трубки, диаметр которой $D = 0,1$ см, падает на дифракционную решётку. Оценить, на каком минимальном расстоянии $L_{\min}$ от трубки нужно расположить решётку, чтобы можно было разрешить две спектральные линии с разностью длин волн между ними $\delta\lambda = 5$ нм при средней длине волны излучения $\lambda = 500$ нм.

### Solution

(!!!! нужно знать формулы!!!! я их не знаю, так что нужно тренировать еще 1ю часть!!!)
Радиус пространственной когерентности:

$$\rho_{\text{ког}} = \frac{\lambda}{\psi} = \frac{\lambda}{D/L} = \frac{\lambda L}{D}$$

($\psi$ - угловой размер трубки)

$$\frac{\lambda}{\delta\lambda} \leq m_{\max} N = \frac{d}{\lambda}\frac{\rho_{\text{ког}}}{d} = \frac{\rho_{\text{ког}}}{\lambda} = \frac{L}{D}; \delta\lambda_{\min} = \frac{\lambda D}{L} = 4\text{нм}$$

### O2-8.39

Свет от газоразрядной трубки, диаметр которой $D = 0,1$ cm, непосрсдственно падает на дифракционную решетку. Оценить, на каком минимальном расстоянии $L_{\min}$ от трубки нужно расположить решетку, чтобы при этом можно было разрешить две спектральные линии с расстоянием между ними $\partial\lambda = 5$ нм при $\lambda = 500$ нм.

### Solution 8.

39. $L_{\min} \approx D\frac{\lambda}{\delta\lambda} = 10$ cm.

### O2-8.40

Свет от удаленного источника, угловой размер которого составляет $\psi = 10^{-3}$ рад, непосредственно падает на дифракционную решетку. Оценить, какую максимальную разрешающую способность $R_{\max}$ можно получить в таких условиях.

### Solution 8.

40. $R_{\max} \approx \frac{1}{\psi} = 10^3$.

## 12.0.13 Problems about spectral devices 2

### O2-8.41

Удаленный протяженный источник испускает две узкие спектральные линии $\lambda_1 = 500$ нм и $\lambda_2 = 500,2$ нм равной интенсивности. Свет от источника непосредственно падает на дифракционную решетку. Оценить угловой размер $\psi$ источника, при котором можно разрешить эти две линии.

### Solution 8.

41. $\psi \lesssim \frac{\lambda_2 - \lambda_1}{\lambda_1} = 4 \cdot 10^{-4}$ рад.

### O2-8.42

Наблюдается дифракция параллельного пучка монохроматического излучения с частотой $v = 10^{15}$ Гц, падающего нормально на дифракционную решетку с числом штрихов $N = 1,5 \cdot 10^4$. Во сколько раз изменится угловая расходимость в 1-м порядке, если падающее на решетку излучение промодулировать так, чтобы были сформированы короткие импульсы длительностью $\tau = 10^{-12}$ с?

### Solution 8.

42. $\frac{\delta\theta_2}{\delta\theta_1} = \frac{\frac{1}{v\tau}\operatorname{tg}\theta}{\frac{1}{N}\operatorname{tg}\theta} = \frac{N}{\nu\tau} \approx 15$.

## O2-8.43

:" Дифракционная решетка с числом штрихов $N = 10^5$ имеет заводской дефект: ее период на разных участках неодинаков и изменяется в пределах $0,1\%$. Можно ли с помощью такой решетки обнаружить простой эффект Зеемана в магнитном поле с индукцией $B = 10^4 Гс$ на длине волны $\lambda = 600$ нм? Найти минимальное расстояние между линиями, которое может разрешить такая решетка.

### Solution 8.

43* Нельзя. $\delta\lambda_{деф} \approx 3$ Å.
Решение. Зеемановское расщепление спектральной линии

$$\Delta\omega = \frac{eB}{2m_e c}, \quad \text{откуда} \quad \Delta\lambda = \frac{\lambda^2 eB}{4\pi m_e c^2} = 0,16 \text{ Å} .$$

Если решетка не имеет дефектов, то минимально разрешаемое $\delta\lambda = \lambda/(mN) = 0,06$ Å $< \Delta\lambda$, т. е. зеемановское расщепление разрешается бездефектной решеткой с теми же параметрами.
Для решетки с дефектом максимум 1 -го порядка размыт между углами $\theta_1$ и $\theta_2$, где $d_1 \sin\theta_1 = \lambda, d_2 \sin\theta_2 = \lambda$. Или $\theta_1 \approx \lambda/d_1; \theta_2 \approx \lambda/d_2$, откуда угловая полуширина максимума 1 -го порядка

$$\frac{\Delta\theta}{2} = \frac{d_2 - d_1}{2d_1 d_2}\lambda \approx \frac{\Delta d}{2d^2}\lambda,$$

где $\Delta d/d = 10^{-2}$ - относительное изменение периода решетки. При малых углах $\theta$ угловая дисперсия в первом порядке

$$\frac{d\theta}{d\lambda} \approx \frac{m}{d} = \frac{1}{d},$$

## O2-8.46

Импульсное излучение с длительностью импульсов 1 пс и длиной волны $\lambda = 0,53$ мкм падает на дифракционную решетку разрешающей способностью $R = 3400$. Оценить отношение длительности импульсов за решеткой к длительности падающих импульсов.

## O2-8.47

Излучение неодимового лазера с длиной волны $\lambda = 1,06$ мкм представляет собой последовательность ультракоротких импульсов, следующих с интервалом $\tau = 1$ нс. Излучение падает нормально на решетку с числом штрихов $n = 1500$ штрих/см. Каков минимальный размер $L$ решетки, с помощью которой можно разрешить структуру спектра излучения во 2-м порядке дифракции?

## O2-8.48

Ультрафиолетовое излучение состоит из двух близких спектральных линий ($\lambda \approx 300$ nm, $\Delta\lambda = 0,1$). Излучение падает под углом $\alpha = 86^\circ$ к нормали на отражательную дифракционную решетку с периодом $d = 4$ мкм. Оценить диаметр падающего пучка лучей, при котором возможно разрешение спектральных линий во 2-м дифракционном порядке.

## O2-8.50 Two diffraction gratings

Two short time-matched pulses of light with average wavelengths of $\lambda_1 = 0.55$ microns and $\lambda_2 = 0.56$ microns normally falls on a reflective diffraction grating $\left(d = 2 \cdot 10^{-4}\text{cm}\right)$. The light diffracted in the first order hits a second similar grating parallel to the first, on which it diffracts again in a direction parallel to the original one. How long will the pulses of different wavelengths shift relative to each other if the distance between the gratings is $l = 50$cm?

## O2-8.51

Найти период $d$ отражательной дифракционной решетки $\sigma$, если входная щель $A$ и выходная щель $B$ спектрального прибора (рис. 336) расположены на окружности радиусом $R = 1$ m, центр которой совпадает с центром решетки $M$. На щель $A$ последовательно направляют пучок CO-лазера (длина волны излучения $\lambda_1 = 5$ мкм) 171 и пучок $CO_2$-лазера ровавшего пучка CO-лазера возвращается во входную щель, а 1 -й порядок дифрагировавшего пучка от $CO_2$-лазера падает на выходную щель. Расстояние от выходной щели до нормали $M$ щ решетке а равно $0,5$ m.

### O2-8.52

Импульсное излучение лазерас длительностью импульсов $\tau = 10^{-12}$ с проходит через спектрометр с дифракционной решеткой с максимальной оптической разностью хода $\Delta = 10$ cm. Найти изменение ширины полосы излучения $\Delta\omega_2/\Delta\omega_1$.

### O2-8.53

Лазер испускает световые импульсы с центральной длиной волны 0,6 мкм с длительностью 1 пс и скважностью (отношение периода повторения импульсов к длительности каждого из них) $10^3$. Это излучение пропускается через монохроматор с разрешающей способностью $5 \cdot 10^4$. Оценить скважность импульсов $Q$ по выходе из монохроматора.

### O2-8.54

Электрон движется в вакууме со скоростью $v$ вблизи поверхности дифракционной решетки с периодом $d$. Скорость электрона параллельна поверхности решетки и перпендикулярна ее штрихам. Определить длины волн, которые могут излучаться под углом $\theta$ к нормали решетки из-за взаимодействия электрона с решеткой (Эффект Смита-Парселла).

### O2-8.55 Temperature of diffraction grating

One of the conditions for the stability of the diffraction pattern obtained using a diffraction grating is the constancy of temperature. Estimate the maximum allowable temperature change $\Delta T$ of the grating, at which its resolution is still almost fully utilized if the spectra are photographed in the 1st order. The temperature coefficient of linear expansion of the lattice material is $\alpha = 10^{-5}\mathrm{K}^{-1}$, the total number of strokes is $N = 10^5$.

#### Solution

(интересная задача, но не готов еще к ней. вообще не знаю, мб это вообще не связано???)

### O2-8.56

Для выделения одной моды из большого числа мод, генерируемых газовым лазером, предлагается использовать модифицированный резонатор Фабри-Перо, одно из зеркал которого заменено наклонной отражательной дифракционной решеткой ДР (рис. 337). Внутри резонатора располагается наряду с активной средой $AC$ телескопическая система $TC$, предназначенная для расширения светового пучка, падающего на решетку. Найти минимальный размер $D$, до которого нужно расширить пучок, чтобы было можно выделить одну моду. Длина резонатора равна $L = 15$ cm, длина волны генерации $\lambda = 1,2$ мкм, решетка имеет $N = 1600$ штрих/мм. Используется дифракция в 1-м порядке.

### O2-8.57

Для сужения спектра излучения лазера на красителе, имеющего широкий спектр в окрестности длины волны $\lambda = 600$ нм, одно из зеркал резонатора заменено дифракционной решеткой, нормаль к которой составляет угол $\varphi = 60^\circ$ к оси резонатора. Определить ширину спектра $\delta\lambda$ излучения такого лазера, если диаметр кюветы с красителем $D = 0,5$ см.

### O2-8.60

Имея в виду решение предыдущей задачи, получить выражение для разрешающей способности (одномерной) дифракционной решетки в рентгеновской области спектра.

### O2-8.61

Рентгеновское излучение с длиной волны $\lambda = 2,8$ Å дифрагирует на кристалле каменной соли, испытывая отражение от ряда кристаллических плоскостей под углом $\varphi = 30^\circ$ к нормали. Толщина кристалла $L = 0,56$ мм. Определить угловую расходимость дифракционного максимума.

### O2-8.62

На дифракционную решетку с числом штрихов $n = 1000$ штрих/см падает плоская монохроматическая волна ($\lambda_0 = 1$ мкм), частота которой начинает медленно изменяться со временем по закону $\omega = \omega_0(1 + at)$. За решеткой расположена линза с фокусным расстоянием $f = 100$ см. Оценить время, в течение которого положение 6-го максимума в фокальной плоскости достигнет положения 5-го

в начальный момент времени. Известно, что максимум интенсивности перемещается вдоль фокальной плоскости со скоростью $v = 1\text{K/c}$.

### O2-8.66

. Стеклянная призма с основанием $b = 10$ cm изготовлена из тяжелого флинта, дисперсия которого в окрестности $\lambda = 6000$ Å равна $dn/d\lambda = 1000$ cm$^{-1}$. Какую максимальную разрешающую способность может иметь дифракционная решетка, ширина заштрихованной части которой равна длине основания этой призмы? Сравнить разрешающую способность такой решетки с разрешающей способностью призмы.

### O2-8.67

Ширина заштрихованной части дифракционной решетки равна длине основания призмы из каменной соли. Разрешающая способность решетки в 1 - м порядке равна разрешающей способности призмы для длины волны $\lambda = 5150$ Å. Определить период решетки $d$, если показатель преломления каменной соли для длины волны $\lambda_1 = 4861$ Å равен $n_1 = 1,5537$, а для длины волны $\lambda_2 = 5461$ Å $n_2 = 1,5477$.

### O2-8.68

Параллельный пучок света падает на призму с угловой дисперсией $d\varphi/d\lambda = 10^3$ cm$^{-1}$. Свет, прошедший через призму, нормально падает на дифракционную решетку с периодом $d$. Размер решетки превышает поперечное сечение пучка лучей, выходящих из призмы. При каком значении $d$ разрешающая способность системы в спектре 2-го порядка будет в два раза больше разрешающей способности одной призмы? Считать дифракционные углы малыми.

### O2-8.73

При нормальном падении на плоскопараллельную пластинку из непоглощающего материала параллельного пучка монохроматического излучения, длина волны которого перестраивается непрерывно, коэффициент пропускания пластинки «осциллирует» так, что его соседние максимумы приходятся на длины волн в 720 мкм, 840 мкм и т. д. Найти показатель преломления материала пластинки, если ее толщина $d = 1,2$ mm.

### O2-8.74

Определить условие максимума, угловое расстояние между максимумами, угловую дисперсию и дисперсионную область эталона Фабри-Перо (ввести угол падения $\varphi$ и расстояние между зеркалами $L$).

### O2-8.75 Дисперсионная область Фабри-Перо (?????)

Определить дисперсионную область эталона Фабри-Перо при $h = 1$ cm, $\lambda = 5000$ Å. Угол падения $\varphi$ считать малым.

#### Solution

(не знаю. эталон - это просто интерферометр, странно, зачем так его называют?)

### O2-8.76 Порядок спектра Фабри-Перо

Чему равен порядок спектра при работе с эталоном Фабри-Перо в зеленой части спектра ($\lambda = 5500$ Å), если расстояние между пластинками равно 1 cm? Угол падения очень мал.

#### Solution

(не знаю.)

### O2-8.77 Разрешающая способность Фабри-Перо

Разрешающую способность интерферометра Фабри-Перо можно определить, пользуясь следующим критерием. Для разрешения двух спектральных линий $\lambda$ и $\lambda'$ необходимо, чтобы в интерференционной картине, даваемой интерферометром, эти линии были разведены на расстояние не меньше полуширины линии. Пользуясь Этим критерием, найти выражение для разрешающей способности интерферометра Фабри-Перо.

### Solution

(в теории точно должна быть решена, но я не особо помню это... переучивать нужно.)

## O2-8.81

Определить время установления колебаний и добротность в оптическом резонаторе, используемом в лазерах (длина волны излучения $\lambda = 0,63$ мкм) и состоящем из двух плоскопараллельных зеркал, расположенных на расстоянии $L = 100$ cm друг от друга и имеющих коэффициент отражения по энергии $\rho_1 = 100\%$ и $\rho_2 = 80\%$. Явлениями дифракции на краях зеркал пренебречь.

## O2-8.82

В интерферометре Фабри-Перо среда между зеркалами обладает дисперсией. При изменении длины волны света $\lambda$ на величину $\Delta\lambda$ наблюдаемые интерференционные кольца сдвигаются так, что каждое кольцо перемещается на место соседнего кольца. Предполагая, что база интерферометра $L \gg \lambda$, вычислить дисперсию показателя преломления среды $dn/d\lambda$.

## O2-8.83

: Определить разрешающую способность спектрометра инфракрасного диапазона, работающего по следующему принципу. Излучение исследуемого ИК-источника в диапазоне $\lambda_{\text{ик}} \approx 3$ мкм смешивается в нелинейном кристалле с излучением стабильного аргонового лазера. При этом возникает излучение на суммарной частоте, лежащей в оптическом диапазоне. Последнее анализируется с помощью интерферометра Фабри-Перо, зеркала которого отстоят друг от друга на расстояние $L = 1$ cm отражения по интенсивности $\rho = 0,9$.

## O2-8.85

Интерферометр Фабри-Перо образован двумя зеркалами с коэффициентом отражения по энергии $\rho = 0,9$, разделенными кольцом из инвара толщиной $d = 100$ мм. Интерференционная картина фиксируется на фотопластинке. Оценить допустимое изменение температуры в лаборатории за время экспонирования, если коэффициент линейного расширения инвара $\alpha = 9 \cdot 10^{-7}\text{K}^{-1}$. Длина волны $\lambda = 5000$ Å.

## O2-8.86

Интерферометр Фабри-Перо образован двумя зеркалами с коэффициентом отражения по энергии $\rho = 0,95$, разнесенными на расстояние $L = 10$ мм. Интерференционная картина фиксируется на фотопластинке. Оценить величину допустимых изменений атмосферного давления в лаборатории за время экспозиции. Показатель преломления воздуха связан с атмосферным давлением $P$ (в паскалях) соотношением $n = 1 + 28 \cdot 10^{-10}P$. Длина волны света $\lambda = 5000$ Å.

## O2-8.87

На интерферометр Фабри-Перо, состоящий из двух одинаковых зеркал диаметром $D = 1$ см, падает свет с длиной волны $\lambda = 500$ нм. Интерференционная картина наблюдается с помощью зрительной трубы, установленной на бесконечность, и имеет вид концентрических колец. Первое кольцо имеет угловой диаметр $\varphi = 10^{-2}$ рад. Оценить максимальную разрешающую способность спектрального прибора в этих условиях.

## O2-8.88

На интерферометр Фабри-Перо, состоящий из двух одинаковых зеркал, падает пучок света с длиной волны $\lambda \approx 0,5$ мкм. Интерференционная картина наблюдается в фокальной плоскости линзы диаметром $D = 2,5$ cm с фокусным расстоянием $f = 10$ cm и имеет вид концентрических колец. Первое кольцо имеет диаметр $d = 1$ см. Оценить максимальную разрешающую способность спектрального прибора в этих условиях.

## O2-8.91

На резонатор Фабри-Перо с расстоянием между зеркалами (базой) $L = 0,5$ cm и разрешающей способностью $R = 10^6$ падает ультракороткий световой импульс длительностью $\tau = 10^{-11}$ с и м длиной волны $\lambda = 500$ нм. Определить зависимость от времени сигнала, который зарегистрирует фотоприемник, установленный за резонатором Фабри-Перо.

### O2-8.92

: Плоский пучок монохроматического света шириной $D$ падает нормально на плоский резонатор Фабри-Перо. Оценить добротность резонатора $Q$, при которой диаметр пучка на выходе резонатора возрастает приблизительно в 3 раза. Длина волны падающего света отвечает максимуму пропускания резонатора.

### O2-8.93

Спектр лазерного импульса анализируется с помощью интерферометра Фабри-Перо, образованного зеркалами с коэффициентами отражения по энергии $\rho \sim 1$, находящимися на расстоянии $L$ друг от друга. Оценить длительность импульса после прохождения через интерферометр.

### O2-8.94

Найти максимальную и минимальную амплитуды колебаний знутри резонатора Фабри-Перо, настроенного в резонанс нормально падающей монохроматической волной амплитуды $A_0$. Коэффициент отражения зеркал по интенсивности $\rho = 0,95$. Поглощением света пренебречь.

### O2-8.95

Оценить разрешающую способность интерферометра Фабри-Перо, если отношение максимальной и минимальной амплитуд колебаний поля внутри интерферометра, освещаемого плоской нормально падающей волной ($\lambda = 600$ нм), равно $\alpha = A_{\max}/A_{\min} = 90$. Расстояние между зеркалами $L = 4$ cm.

### O2-8.96

На испаряющуюся прозрачную пленку нормально падает излучение лазера с длиной волны $\lambda = 630$ нм. Контроль толщины пленки осуществляется путем измерения интенсивности прошедшего излучения, которая периодически изменяется в процессе испарения, так что $\mathcal{I}_{\min}/\mathscr{I}_{\max} = 0,84$. Пренебрегая поглощением света в пленке, определите, при какой минимальной толщине интенсивность прошедшего света достигает максимального значения.

### O2-8.97

На подложке осаждается прозрачная пленка, контроль толщины которой производится путем измерения доли (по энергии) отраженного назад излучения лазера с длиной волны $\lambda = 0,63$ мкм. По мере роста пленки эта доля осциллирует, принимая минимальное значение, равное $2 \cdot 10^{-2}$. Какова толщина пленки в эти моменты? Излучение падает нормально к поверхности пленки. Коэффициент отражения на границе пленка-подложка (по амплитуде) $s = -1/7$. Рассмотреть двулучевое приближение.

### O2-8.98

Импульс видимого света длительностью $\tau$ падает на интерферометр Фабри-Перо параллельно его оси и затем фокусируется на чувствительную площадку фотоприемника. Расстояние между зеркалами интерферометра $L = 15$ cm, коэффициент отражения 178 зеркал по энергии $\rho = 0,99$. Оценить, при каком значении величи н ы рактер. Оценить частоту осцилляций $v$, характерное время $T$ затухания и число $N$ колебаний фототока за время $T$. Инерционность фотоприемника считать достаточно малой.

## 12.0.14 Problems about Fourier optics and holography

### O2-9.1

Три плоские монохроматические волны с амплитудами 1, $a$ и $a(a \ll 1)$ падают на плоскость $z = 0$ под углами $0, \alpha$ и $-\alpha$ (рис. 342) так, что в точке $x = 0$ колебания оказываются синфазными. При смещении плоскости наблюдения в область $z > 0$ происходят периодические изменения контраста интерференционной картины. Объяснить явление. Найти положения плоскости наблюдения, в которых контраст картины максимальный и минимальный. Чему он равен?

#### Solution

(потом добавлю картинку, утеряно условие пока что)

## O2-9.2 Спектр за синусоидальной решеткой

Найти спектр плоских волн $F(u)$ за синусоидальной решеткой, освещаемой нормально падающей плоской волной. Амплитудный коэффициент пропускания решетки $\tau(x) = 1 + a\cos\Omega x (a < 1)$

### Solution

(не знаю, не готов к такому)

## O2-9.3

Найти спектр плоских волн за щелью шириной a, освещаемой нормально падающей плоской волной. Решить ту же задачу, если шель перекрыть решеткой с периодом $d$ и размером прозрачных участков $b (a = Nd$, где $N$ - число штрихов решетки).

### Solution

## O2-9.4

Два плоских когерентных монохроматических пучка света с длиной волны $\lambda = 500$ нм и с амплитудами $A_0$ и $2A_0$ падают под углами $\alpha = \pm 0,05$ рад на синусоидальную решетку с амплитудным коэффициентом пропускания $\tau(x) = (1 + \cos\Omega x)/2$. В точке $x = 0$

эти волны создают противофазные колебания (рис. 343). Период решетки $d = 10^{-3}$ cm. Определить пространственный спектр волны за решеткой.

### Solution

## O2-9.5

. В кювету, имеющую форму прямоугольного параллелепипеда, налит толуол, в котором возбуждаются ультразвуковые волны с помощью колебаний пластинки пьезокварца. Пластинка кварца установлена параллельно боковым стенкам кюветы. Ультразвуковые волны, возбуждаемые пластинкой, отражаются от одной из боковых стенок кюветы. В результате в жидкости образуется стоячая ультразвуковая волна. Чему равен пространственный период изменения тразвуковой волны?

## O2-9.6

На рис. 344 изображена схема установки для наблюдения дифракции света на ультразвуке. Стоячие ультразвуковые волны образуются в кювете $K$. Пластинка кварца $P$ установлена параллельно стенке $AC$, так что волны, излучаемые ею, распространяются в направлении, параллельном $AB$. Дифракционные максимумы и минимумы наблюдаются в трубу $T$, установленную на бесконечность. Показать, что угол дифракции $\theta$ для максимума $m$-го порядка определяется из условия $\Delta\sin\theta = m\lambda$.

Ука з ан и е. Принять во внимание, что частота ультразвуковых колебаний весьма мала по сравнению с частотой световых колебаний.

## O2-9.7

Дифракция света на ультразвуковой волне в толуоле наблюдается на установке, описанной в предыдущей задаче. В качестве источника света использована зеленая линия ртути ($\lambda = 5461$ Å). Вместо трубы $T$ за кюветой поставлена собирающая линза с фокусным расстоянием $f = 30$ cm. Дифракционные полосы получаются в фокальной плоскости линзы и рассматриваются в микроскоп, снабженный шкалой. Определить скорость звука $v$ в толуоле, если расстояние между двумя соседними максимумами $\Delta x = 0,546$ mm, а частота ультразвука $v = 4000$ кГц.

## O2-9.8

Можно ли по характеру дифракционной картины Фраунгофера на плоской ультразвуковой волне решить, происходит ли дифракция на бегущей или стоячей ультразвуковой волне? Поглощением ультразвука пренебречь.

## O2-9.9

Если жидкость, в которой установилась стоячая ультразвуковая волна, рассматривать в микроскоп, то благодаря неоднородности жидкости будут видны светлые и темные полосы. Чему равно расстояние между двумя соседними светлыми или темными полосами?

## O2-9.10

Горизонтальный луч He-Ne лазера с $\lambda = 0,63$ мкм падает нормально на тонкую стеклянную кювету с водой. В воде возбуждена стоячая ультразвуковая волна. Направление распространения ультразвука перпендикулярно направлению падающего луча. В результате дифракции света на ультразвуке первые дифрагированные волны отклонились на угол $\varphi = 5,7°$. Дифрагированный свет анализируется интерферометром Фабри-Перо толщиной $L = 1$ cm. Определить спектральный состав дифрагированного света и оценить коэффициент отражения зеркал интерферометра, при котором можно наблюдать исследуемую структуру спектра. Скорость звука в воде $v = 1,5$km/с.

## O2-9.11

Два плоских монохроматических когерентных пучка света с длиной волны $\lambda = 600$ нм и равными амплитудами $A_0$ падают под углом $\alpha = \pm 0,06$ рад на синусоидальную решетку с амплитудными коэффициентом пропускания $\tau(x) = (1 + \sin \Omega x)/2$ ( рис. 343). В точке $x = 0$ эти волны создают синфазные колебания. Период решетки $d = 10^{-3}$ см. Определить пространственный спектр волн за решеткой.

## O2-9.12

Плоский монохроматический пучок света интенсивностью $\mathscr{I}_0$ и длины волны $\lambda$ дифрагирует на двух последовательно расположенных синусоидальных решетках с амплитудным коэффициентом

## O2-9.14

Предлагается следующая схема спектРис. 346 рометра - устройства для исследования спектрального состава излучения источника света, содержащего две спектральные компоненты. Сколлимированный пучок света дифрагирует на двух последовательно расположенных одинаковых синусоидальных решетках, одна из которых перемещается с постоянной скоростью. (рис. 347$a$). В фокальной плоскости Рис. 347 линзы $J_2$ исследуется зависимость $\mathscr{V}(t)$ интенсивности первого спектрального максимума от времени. Оказалось, что эта зависимость имеет вид, изображенный на рис. 347б. Определить относительное расстояние $\Delta\lambda/\lambda$ между двумя спектральными компонентами $\lambda$ и $\lambda + \Delta\lambda$ в излучении источника.

## O2-9.15

При наблюдении фазовых (прозрачных) структур методом темного поля в общей фокальной плоскости линз $J_1$ и $J_2$ (рис. 348) на оптической оси устанавливается проволока П. Фазовая решетка создается в жидкости стоячей ультразвуковой волной частоты $v = 20$М ц. Найти расстояние $\Delta l$ между интерференционными полосами на экране Э, а также максимально допустимое удаление $L_{\max}$ экрана от линзы $J_2$, при котором еще возможно наблюдение 180 интерференционной картины. Диаметр линзы $J_2 D = 4$ cm, скорость звука в жидкости $v = 1,5$km/с. Рещетка освещается нормально падающей плоской волной ($\lambda = 0,5$ мкм).

## O2-9.16

При наблюдении фазовых (прозрачных) структур методом темного поля в общей фокальной плоскости линз $J_1$ и $J_2$( рис. 348) Рис. 348 на оптической оси устанавливается проволока П. Оценить ее допустимый диаметр ($d_{\max}$ и $d_{\min}$) для наблюдения на экране Э интерференционной картины от фазовой синусоидальной решетки с периодом $\Lambda = 2$ мм, освещаемой нормально падающей плоской волной длины $\lambda = 0,5$m. Диаметр линзы $J_2$ равен $D = 2$ cm, фокусное расстояние $f = 20$ cm.

## O2-9.17

Один из методов наблюдения фазовых (прозрачных) обьектов состоит в следующем: в общей фокальной плоскости линз $J_1$ и $J_2$ на оптической оси устанавливается прозрачная пластинка Рис. 349 П, вносящая фазовуго задержку в $\pi/2$ (рис. 349). Найти распределение интенсивности $\mathcal{E}(x)$ в плоскости изображения (в задней фокальной плоскости линзы $J_2$), если предмет - фазовая синусоидальная решетка с амплитудным коэффициентом пропускания $\tau(x) = \exp(im \cos \Omega x), m \ll 1$ - расположен в передней фокальной плоскости линзы $J_1$. Как изменится картина интенсивности, если использовать пластинку с задержкой в $3\pi/2$? Как изменится контраст, если пластинка обладает коэффициентом поглощения $k_{\text{п}}$?

## O2-9.18

Фазовая синусоидальная решетка с глубиной модуляции $m \ll 1$ установлена в передней фокальной плоскости линзы $J_1$ и освещается параллельным пучком монохроматического линейно поляризованного света с длиной волны $\lambda$. Определить толщину и ориентацию маленькой кристаллической пластинки, установленной в 183 фурье-плоскости (рис. 350), при которой контраст изображения на экране (видность)

максимален. Какова при этом максимальная и Рис. 350 минимальная видности при вращении поляроида вокруг оптической оси (поляроид расположен между линзой $J_2$ и экраном)?

## O2-9.19

Для наблюдения фазового (прозрачного) объекта в общей фокальной плоскости линз $J_1$ и $J_2$ устанавливают на оптической оси небольшой идеальный поляроид, который вносит для проходящего света фазовую задержку в $2\pi$ (рис. 351). Определить видность Рис. 351 интерференционной картины в задней фокальной плоскости линзы $J_2$, если расположенную в передней фокальной плоскости линзы $J_1$ фазовую решетку с амплитудным коэффициентом пропускания $\tau(x) = \exp(i \cdot 0,2 \cos \Omega x)$ освещать плоским пучком монохроматического света, поляризованного по кругу.

## O2-9.25

При нормальном освещении амплитудной синусоидальной решетки монохроматическим параллельным пучком света в плоскостях саморепродукции $P$ воспроизводятся ее изображения Рис. 356 (рис. 356). Если за дифракционной решеткой поместить однородную плоскопараллельную пластинку толщиной $h = 30$ мм с показателем преломления $n$, то все плоскости изображения сместятся по оси $z$ на $\Delta z = 10$ мм. Определить $n$, если отношение длины волны $\lambda$ к периоду решетки $d$ достаточно мало $(\lambda/d \ll 1)$.

## O2-9.26

Дифракционная решетка размером $L$ и периодом $d$ освещается нормально падающей плоской полной (длина волны $\lambda \ll d$). За ней воспроизводятся ее изображения в плоскостях саморепродукции. Оценить число изображений, наблюдаемых за решеткой, в которых еще различима ее структура.

## O2-9.28

Две последовательно расположенные вплотную друг к другу решетки с функциями пропускания $\tau_1(x) = (1 + \cos \Omega x)/2$ и $\tau_2(x) = \exp(\ \mathrm{im} \cos \Omega x), m \ll 1$ освещаются плоской нормально падающей монохроматической волной. Как изменится отношение интенсивностей волн, дифрагировавших в $\pm 1$-е порядки дифракции, если сдвинуть первую решетку вдоль оси $x$ на четверть периода? Какова разность фаз колебаний поля в $\pm 1 - x$ порядках дифракции? $\lambda = 5000$ Å) дифрагирует на амплитудной синусоидальной решетке, а затем проходит через интерферометр Фабри-Перо с расстоянием между зеркалами $L = 2$ см. При каком максимальном периоде решетки $d$ контраст интерференционной картины, возникающей в плоскости наблюдения, максимален?

## O2-9.30

Оптическая система состоит из двух одинаковых линз с общей фокальной плоскостью; $f$ - фокусное расстояние линз. Амплитудная синусоидальная решетка находится в передней фокальной плоскости 1-й линзы, а изображение рассматривается в задней фокальной плоскости 2-й линзы. Как зависит характер изображения от соотношения между частотой решетки $\Omega$ и диаметром линзы $D$?

## O2-9.33

Найти амплитудный коэффициент пропускания $\tau(x)$ голограммы точечного источника света, если в качестве опорной волны используется нормально падающая на плоскость голограммы плоская волна. Расстояние от источника до голограммы равно L. Считать, что прозрачность голограммы пропорциональна интенсивности света при записи. Найти положение действительного и мнимого изображений при восстановлении изображения нормально падающей плоской волной. Как изменится положение восстановленных изображений, если при записи использовать наклонный опорный пучок с углом наклона $\theta$? Оценить минимальный размер $a_{\min}$ голограммы, при котором полностью используется разрешающая способность фотоэмульсии, равная $n$ [линий/мм]. Найти размер $b$ восстановленного изображения.

## O2-9.34

Голограмма записана на пластинке радиусом $r = 5$ см. Она освещается монохроматическим светом длины волны $\lambda = 0,5$ мкм, а изображение получается на расстоянии $L = 1$ м. Найти допустимую немонохроматичность света $\Delta\lambda$, при которой еще полностью используется теоретическая разрешающая способность голограммы.

### O2-9.35

Получена голограмма небольшого предмета, расположенного на расстоянии $L = 50$ см от нее. Каким должен быть размер $D$ фотопластинки, чтобы записать на голограмме детали размером $b \approx 0,01$ мм? Какая немонохроматичность света $\Delta\lambda$ допустима при записи голограммы? Длина волны света $\lambda = 0,5$ мкм.

### O2-9.36

При записи голограммы предмета, находящегося на расстоянии $L = 1$, используется излучение He-Ne лазера ($\lambda \approx 6300$ Å). Восстанавливается изображение с помощью протяженного квазимонохроматического источника с угловым размером $\alpha = 10^{-4}$ рад. Каков минимальный размер деталей в восстановленном изображении? Какова при этом требуемая монохроматичность?

### O2-9.37

Излучение He-Ne лазера ($\lambda \approx 6300$ Å) используется для записи голограммы. Расстояние от предмета до голограммы $L = 1$ м. Какого минимального размера детали $d$ можно восстановить с помощью немонохроматического источника с шириной полосы $\Delta\lambda = 9$ Å? Каков необходимый для этого размер голограммы $D$?

### O2-9.38 Максимальное число колец на голограмме

Точечный объект движется параллельно фотопластинке, на которую записывается его голограмма. Оценить скорость $v$ объекта, при которой голограмма будет иметь максимальное число колец. Объект и фотопластинка освещаются плоской волной с длиной волны $\lambda = 0,5$ мкм, нормальной к плоскости фотопластинки. Размер фотопластинки $D = 0,1$ м, расстояние между объектом и фотопластинкой $L = 1$ m, время экспозиции $\tau = 0,01$с. Найти также разрешающую способность голограммы $\Delta l$ в направлении движения объекта.

## 12.0.15 Problems about dispersion

### O2-10.1 Групповая скорость двух волн (!?!??!)

Рассматривая импульс, представляющий собой суперпозицию двух гармонических волн $\sin(\omega t - kx)$ и $\sin(\omega' t - k' x)$, найти групповую скорость $u$. Считать, что $\omega \approx \omega', k \approx k'$.

### O2-10.2(2 мин (?)) Формула групповой скорости

Выразить групповую скорость $u = d\omega/dk$ через фазовую скорость света $v$ и $dv/d\lambda$, а также через $v$ и $dn/d\lambda$.

### O2-10.3 График для групповой скорости (????)

Изобразим кривой зависимость фазовой скорости волны $v$ от длины волны $\lambda$ (рис. 362). Показать, что отрезок $OA$ на оси $v$, отсекаемый касательной к этой кривой в точке $\lambda_0$, равен групповой скорости для длины волны $\lambda = \lambda_0$. (Построение П. С. Эренфеста).

### O2-10.4

Плоское волновое возмущение распространяется в среде с линейным законом дисперсии: $v = a + b\lambda$, где $v$ - фазовая скорость, а и $b$ - постоянные. Показать, что каково бы ни было возмущение, форма его, непрерывно изменяясь, будет периодически восстанавливаться по истечении времени $\tau = d\lambda/dv = 1/b$. Показать, что отношение пути $s$, пройденного возмущением за промежуток времени $\tau$, продолжительности этого промежутка равно групповой скорости.

Ук а за н и е. Любое плоское возмущение в любой момент времени может быть получено суперпозицией синусоид. Каждая из синусоид перемещается со своей фазовой скоростью в одном и том же направлении. Вследствие этого непрерывно деформируется форма возмущения. Утверждение, сформулированное в задаче, будет доказано, если показать, что существует такое время $\tau$, по истечении которого точно восстанавливается первоначальное взаимное расположение синусоид. Достаточно провести рассуждение для трех синусоид; обобщение на большее число синусоид тривиально. 190

### O2-10.5

Вычислить групповую скорость для различных законов дисперсии ($v$ - фазовая скорость): 1) $v = a (a = \text{const})$ - недиспергирующая среда, например звуковые волны в воздухе; 2) $v = a\sqrt{\lambda}$ - волны на поверхности воды, вызываемые силой тяжести (гравитационные волны); 3) $v = a/\sqrt{\lambda}$ - капиллярные волны; 4) $v = a/\lambda$ - поперечные колебания стержня; 5) $v = \sqrt{c^2 + b^2\lambda^2}$ - электромагнитные волны в ионосфере ($c$ - скорость света в вакууме, $\lambda$ - длина волны в среде, см. задачу 10.16) 6) $v = c\omega/\sqrt{\omega^2\varepsilon\mu - c^2\alpha^2}$ - электромагнитные волны в прямолинейном волноводе, заполненном диспергирующей средой с диэлектрической проницаемостью $\varepsilon = \varepsilon(\omega)$ и магнитной проницаемостью $\mu = \mu(\omega)$($c$ - скорость света в вакууме, $\alpha$ - постоянная, зависящая от размеров и формы поперечного сечения волновода).

### O2-10.6

При каком законе дисперсии немагнитной среды $\varepsilon = \varepsilon(\omega)$, заполняющей прямолинейный волновод или бесконечное пространство, связь между фазовой и групповой скоростями электромагнитных волн принимает вид $vu = c^2$? Ук а з а и с. См. прсдыдушую задачу.

### O2-10.7

Показать, что в условиях предыдущей задачи, а также в том случае, когда внутри волновода вакуум, фазовая скорость электромагнитных волн в волноводе превосходит скорость света в вак у уе.

### O2-10.8

Найти групповую скорость и ренгеновского излучения в среде, если предельный угол полного внутреннего отражения при падении этих волн на среду из воздуха равен $\alpha$. Показатель преломления ренгеновских волн определяется выражением $n^2 = 1 - \omega_0^2/\omega^2$, где $\omega_0$ - постоянная.

### O2-10.9

Майкельсон измерил скорость света в сероуглероде по методу вращающегося зеркала. Показатель преломления сероуглерода для средней длины волны видимого спектра равен $n = 1,64$, а величина $1 + \frac{\lambda}{n}\frac{dn}{d\lambda} = 0,93$. Определить, какое следует ожидать значение для отношения скорости света в вакууме к измеренной этим методом скорости света в сероуглероде.

### O2-10.10

Получить формулу для диэлектрической проницаемости $\varepsilon(\omega)$ ионизованного газа в монохроматическом электрическом поле $E = E_0 \cos \omega t$. Столкновениями электронов и ионов пренебречь.

### O2-10.11

Может ли показатель преломления быть меньше единицы?

### O2-10.12

Диэлектрическая проницаемость плазмы $\varepsilon(\omega)$ (см. задачу 10.10) отрицательна, если $\omega < \omega_0$. В этом случае показатель преломления $n = \sqrt{\varepsilon}$ - чисто мнимая величина. Выяснить физический смысл чисто мнимого показателя преломления.

### O2-10.13

Радиоволна распространяется вверх. Волны каких частот могут проходить через ионосферу? Какие волны будут полностью отражаться?

### O2-10.14

Радиосигнал определенной частоты $v = \omega/2\pi$ посылается вверх и отражается на определенной высоте. Определить концентрацию электронов в точке отражения.

## O2-10.15

Концентрация электронов на Солнце на расстоянии $r = 0,06R$ от границы фотосферы ($R = 6,95 \cdot 10^{10}$ cm$-$ радиус Солнца) примерно равна $N = 2 \cdot 10^8$ cm$^{-3}$. Могут ли радиоволны из Этой области Солнца достигать Земли, если длина волны (в вакууме) равна 1) $\lambda = 1$ m, 2) $\lambda = 10$ m, 3)$\lambda = 50$ m?

## O2-10.16

Получить выражение для фазовой скорости радиоволны в ионосфере в зависимости от длины волны $\lambda$ в ионосфере. (См. задаuу 10.10)

## O2-10.17

Лазер на $CO_2$ со средней длиной волны $\lambda = 10,6$ мкм излучает две близкие частоты $v_1$ и $v_2$. Излучение такого лазера смешивают в нелинейном кристалле излучением лазера на Nd ($\lambda_3 = 1,06$m). Анализ излучения на комбинационных частотах ($v_1 + v_3$ и $v_2 + v_3$) показал, что соответствующие им длины волн отличаются на $\delta\lambda = 0,5$ нм. Определить разность длин волн $\Delta\lambda = \lambda_1 - \lambda_2$ излучения лазера.

## O2-10.18

. Показатель преломления ионосферы для радиоволн с частотой $v = 10$М ц равен $n = 0,90$. Найти концентрацию $N$ электронов в ионосфере, а также фазовую $v$ и групповую и скорости для этих радиоволн.

## O2-10.19 Толщины слоев плазмы, зная отношение коэффициентов пропускания через них

При изучении прохождения плоской электромагнитной волны частоты $v = 8$ МГц через плоские однородные слои плазмы с концентрацией свободных электронов $N = 10^6$ cm$^{-3}$ найдено, что энергетические коэффициенты пропускания волн отличаются в 10 раз для слоев плазмы, толщины которых отличаются в 2 раза. Пренебрегая интенсивностью волн, отраженных от задней границы каждого слоя, найти их толщины $d_1$ и $d_2$.

### Solution

Плазма - сразу же найти показатель преломления.

$$n^2 = 1 - \frac{\nu^2}{m\omega^2} = 1 - \frac{Ne^2}{\pi m\nu^2} = -0.258 \Rightarrow n = \pm i\kappa = \pm i \cdot 0.508$$

то есть происходит затухание, тогда волна

$$E \propto \exp(ikx) = \exp\left(i\frac{2\pi}{\lambda}nx\right) = \exp\left(\pm\frac{2\pi}{\lambda}\kappa x\right)$$

В итоге

$$I \propto \exp\left(-\frac{4\pi}{\lambda}\kappa x\right) = \exp\left(-\frac{4\pi\nu}{c}\kappa x\right)$$

и пользуемся условием $d_2 = 2d_1$ и про интенсивности

$$10 = \left(\frac{E_2}{E_1}\right)^2 = \exp\left(-\frac{4\pi\nu}{c}\kappa\left(d_2 - d_1\right)\right) = \exp\left(-\frac{4\pi\nu}{c}\kappa d_1\right) \Rightarrow d_1 = \frac{\ln 10}{4\pi\nu\kappa} \approx 13.5\text{M}$$

Задача составлена специально замороченно, можно было бы написать проще условие.

## O2-10.20

Для оценки интегральных и средних характеристик межзвездной плазмы можно использовать экспериментальный факт, установленный сразу же после открытия пульсаров. Оказалось, что из-за дисперсии плазмы импульсы радиоизлучсния пульсаров на более низких частотах всегда запаздывают по отношению к импульсам более высоких частот. Рассмотрите следующий идеализированный пример. Два монохроматических сигнала с длинами волн $\lambda_1 = 3$ cm и $\lambda_2 = 5$ cm распространяются в плазме. Определить полное число $n$ свободных электронов на пути сигналов (т. е. их число в цилиндре площадью 1 cm$^2$ и высотой, равной расстоянию источник-приемник), если испущенные одновременно сигналы запаздывают друг относительно друга на время $\Delta t = 10^{-5}$с. Концентрация электронов хотя и не постоянна вдоль пути сигналов, но показатель преломления везде весьма близок к единице. Определить также среднюю концентрацию $N$ свободных электронов на пути сигналов, если их относительное запаздывание $\Delta t/t_0 = 10^{-15}$ ($t_0$ - время распространения от источника до приемника). 192

## O2-10.21 Расстояние до пульсара

Импульсное излучение пульсара CP1919 + 21 на частоте $\nu_1 = 80$ МГц достигает Земли на $\Delta t = 7$ с позже, чем соответствующий импульс на частоте $\nu_2 = 2000$ МГц. Оценить расстояние $L$ до пульсара, если принять среднюю концентрацию электронов в межзвездном пространстве равной $N \approx 0,05$ см$^{-3}$.

### Solution

Сразу считаем показатель преломления по концентрации
Плазменная частота:

$$\omega_p = \sqrt{\frac{4\pi N e^2}{m}} \approx 0.4 \cdot 10^4 \text{Гц}$$

Таким образом частота излучения в обоих случаях много больше плазменной, что означает, что показатель преломления будет действителен. Но так как он будет различен для разных частот, то и фазовая и групповая скорости каждого импульса будут отличаться.

Произведение фазовой на групповую скорость для плазмы дает квадрат скорости света: $u \cdot v = c^2$, и также $v = \frac{c}{\sqrt{1-\omega_p^2/\omega^2}}$, поэтому

$$u = c\sqrt{1 - \frac{\omega_p^2}{\omega^2}} \approx c\left(1 - \frac{\omega_p^2}{2\omega^2}\right)$$

Окончательно разность длительностей будет: (?? проверю потом математику)

$$\Delta t = L\left(\frac{1}{u_1} - \frac{1}{u_2}\right) = \frac{L}{c}\frac{\omega_p^2}{2}\left(\frac{1}{\omega_1^2} - \frac{1}{\omega_2^2}\right)$$

Отсюда выражаем $L \approx 6.6 \cdot 10^{20}$ cm, что примерно 700 световых лет

Расширение Вселенной не учитываем??? других эффектов не будет????.

## O2-10.22 Погрешность измерения скорости ракеты

Измерение скорости ракеты при вертикальном взлете проводится импульсным радиолокатором, расположенным в точке старта. На экране локатора по оси времени фиксируются моменты посылки двух последовательных радиоимпульсов и их приема после отражения от ракеты. Поскольку скорость распространения радиоволн в ионосфере точно неизвестна, возникает погрешность в определении скорости ракеты. Найти относительную погрешность в определении скорости $(\Delta v/v)$ ракеты, принимая максимальную концентрацию электронов в ионосфере равной $N = 10^6$ см$^{-3}$, а рабочую частоту радиолокатора $v = 400$МГц.

### Solution

## O2-10.23 Частота для измерений орбиты

С целью проверки теории относительности предполагается с помощью радиоволн точно измерить параметры орбиты спутника Земли. Однако из-за преломления радиоволн в ионосфере, где средняя концентрация электронов $N = 10^5$ см$^{-3}$, возникают ошибки измерений. Оценить минимальную частоту $v_{\min}$, на которой следует проводить такие наблюдения.

### Solution

## O2-10.24

Определить число свободных электронов на атом Ag, если пленка серебра прозрачна для ультрафиолета, начиная с энергии $\mathscr{E} = 5$ эВ. Для серебра относительная атомная масса равна $A = 108$, плотность $\rho = 10,5$г/см$^3$.

### Solution

## O2-10.25

Параллельный пучок рентгеновского излучения длины волны $\lambda = 0,1$ нм падает на тонкую двояковыпуклую линзу из бериллия (плотность бериллия $\rho = 1,85$ г/см $^3$, порядковый номер $Z = 4$, относительная атомная масса $A = 9$) с поверхностями одинаковых радиусов кривизны $R = 40$ см. Диаметр линзы считать равным $D = 9$ см. Найти угол расхождения пучка после линзы.

**Solution**

## O2-10.26

На экран из $^{9}_{4}$Be (плотность $\rho = 1,85\text{г/cm}^3$) падает плоская волна $\lambda = 100$ Å (от рентгеновского лазера). В экране имеется круглое отверстие диаметром $d = 2,45 \cdot 10^{-2}$ cm. Расстояние до точки $P$ наблюдения $L = 1$ м. Определить толщину экрана $h$, когда интенсивность в точке $P$ максимальна (поглощением и отражением пренебречь). Определить величину максимальной интенсивности. (Для рентгеновского излучения электроны Be можно считать свободными).

**Solution**

## O2-10.27

На круглую пластинку из $^{9}_{4}$Be (плотность $\rho = 1,85\text{г/cm}^3$) падает плоская монохроматическая волна $\lambda = 100$ Å (от рентгеновского лазера). Расстояние до точки $P$ наблюдения $L = 1$ м. Определить диаметр и толщину пластинки, при которых интенсивность в точке $P$ максимальна. Определить величину максимальной интенсивности. Поглощением в пластинке пренебречь. (Для рентгеновского излучения электроны Be можно считать свободными).

## O2-10.32

Для одновременной передачи множества сигналов используют геостационарный спутник Земли в качестве ретранслятора. Оценить, сколько телефонных каналов с шириной полосы $\Delta v = 3$k ц можно одновременно передавать по такой линии связи на средней частоте 10 ГГц. Средняя концентрация свободных электронов на пути сигналов $N = 10^5$ cm$^{-3}$. Влияние ионов можно не учитывать.

## O2-10.33

«Рентгеновод» представляет собой полый капилляр, стенки которого выполнены из твердого вещества ($SiO_2$). Захват 104 рентгеновского кванта в «рентгеновод» происходит за счет полного «внутреннего» отражения. Оценить, какими должны быть внутренний диаметр капилляра и радиус изгиба рентгеновода», чтобы он был пригоден для транспортировки рентгеновских лучей с энергией $f = 10\text{K} \ni \text{B}$.

## O2-10.34

. Найти показатель преломления $n$ газа и его градиент по высоте на поверхности Венеры, атмосфера которой состоит из углекислого газа $CO_2$ с поляризуемостью молекул $\alpha = 2,7 \cdot 10^{-23}$ cm$^3$. Давление на Венере $p_0 = 100$ атм, температура $t = 500°$C. Найти радиус кривизны $r$ светового луча, пущенного горизонтально. К каким особенностям атмосферной оптики планеты приводит найденное значсние? Ускорение свободного падсния на Венере $g_{\text{B}} = 0,84g_3$. Указ ан и е. Радиус $r$ кривизны горизонтального луча определяется соотношением $\frac{1}{r} = \frac{1}{n}\frac{dn}{dh}$.

## O2-10.35

Во сколько раз должна увеличиться плотность атмосферы Земли для того, чтобы в ней, как на Венере, возникла круговая рефракция, при которой луч света обойдет Землю? Показатсль преломления воздуха при атмосферном давлении равен $n_0 = 1,0003$ (см. предыдущую задачу).

## O2-10.36

Световой луч распространяется в поле тяжести параллельно поверхности Земли. Пренебрегая движением воздуха, определить величину отклонения луча на пути 1 км. Считать давление $P = 1$ атм, температуру воздуха $T = 300$K, коэффицисит преломления воздуха при этих условиях $n = 1 + 3 \cdot 10^{-4}$.

## O2-10.37

На плосковыпуклую линзу с радиусом кривизны $R = 100$ см падает плоская монохроматическая волна, частота которой возрастает со временем по закону $\omega = \omega_0(1 + at)$, $\frac{\omega - \omega_0}{\omega_0} \ll 1$, $\lambda_0 = 1$ мкм. Определить константу $a$, если фокус перемещается со скоростью $v = 3\text{Km/c}$. Показатель преломления линзы $n = 1,5$. Дисперсия линзы $dn/d\lambda = -10^3$ cm$^{-1}$.

## O2-10.38

Линейно поляризованный импульс квазимонохроматического излучения проходит через электрооптическую ячейку Поккельса длиной $l = 10$ cm. Показатель преломления ячейки увеличивают по закону $n(t) = n_0 + \alpha t$. Как изменится длительность импульса и его средняя частота после прохождения через ячейку, если $\alpha = 3 \cdot 10^8 \text{с}^{-1}$?

## O2-10.39

Показатель преломления некоторой прозрачной среды вблизи частоты $\bar{\omega}$ изменяется по закону $n(\omega) = n_0 - A/(\omega - \omega_0)$, где $n_0 = 1,5, \omega_0 = 4 \cdot 10^{14}\text{с}^{-1}, A = \text{const}$, $\omega < \omega_0$. Через слой такого вещества толщиной $l = 3$ см проходит короткий световой импульс, срелняя частота которого равна $\bar{\omega}$, а спектральная ширина $\Delta\omega \ll |\bar{\omega} - \omega_0|$. Известно, что $|\bar{\omega} - \omega_0| \approx 10^{12}\text{с}^{-1}$, и $|n(\bar{\omega}) - n_0| = 0,01$. Оценить время прохождения импульса через слой и сравнить это время со временем прохождения такого же расстояния в вакууме.

## O2-10.40

Определить время прохождения импульса света через слой вещества толщиной $l = 1$ cm, показатель преломления которого вблизи средней частоты $\bar{\omega}$ импульса описывается как $n(\omega) = n_0 - A\,(\omega - \omega_0)\,,$ где $n_0 = 1,5,$ $A$ - константа и $\omega_0 = 4 \cdot 10^{14}\text{с}^{-1}$ - резонансная частота атомов вещества. Рассмотреть случай $\bar{\omega} < \omega_0, |\bar{\omega} - \omega_0| \approx 10^{12}\text{с}^{-1}, |n(\bar{\omega}) - n_0| \approx 0,1$. Считать спектральную ширину импульса $\Delta\omega \ll |\bar{\omega} - \omega_0|$. Поглощением света пренсбречь.

## O2-10.41

. Для того чтобы короткий импульс-сигнал, описываемый функцией $f(t)$, передать через диспергирующую среду без искажений, предлагается на входе в среду сформировать плоское волновое возмущение, периодически (с периодом $T$) повторяя сигнал $f(t)$. На каком минимальном расстоянии от входной плоскости повторяется неискаженная форма сигнала? Закон дисперсии среды в полосе частот, охватываемых сигналом, имеет вид $k(\omega) = B\omega^2$

## O2-10.42

Для того чтобы короткий импульс-сигнал, описывасмый функцией $f(t)$, передать через диспергирующую среду (толщина среды $L$) без искажений, предлагается на входе в среду сформировать плоское волновое возмущение, периодически повторяя сигнал $f(t)$. Закон дисперсии среды в полосе частот сигнала имеет вид $k(\omega) = B\omega^4$. Какова необходимая минимальная частота повторения, при которой на выходе из среды повторяется неискаженная форма сигнала?

## O2-10.43

Плазма заполняет полупространство $x > 0$, причем концентрация электронов растет вглубь по закону $n(x) = \mu x$; $\mu = \text{const}$. Перпендикулярно границе $x = 0$ падает электромагнитный волновой пакет со средней частотой $\omega$, уходит в плазму, отражается от зоны критической плотности и через некоторое время $\tau$ регистрируется при $x = 0$. Определить время $\tau$.

### Solution

(????)

## O2-10.47(2 мин(??)) Движущийся источник в среде

Найти $\Delta\lambda/\lambda$, если источник движется со скоростью $v$ в среде с показателем преломления $n$.

### Solution

(????)

## О2-10.48 Детектирование рассеяния на воде (!?!!!)

Излучение рубинового лазера рассеивается на звуковых колебаниях в воде. При рассеянии света происходит доплеровское смещение частоты. Оценить число штрихов $N$ дифракционной решетки, с помощью которой в 1-м дифракционном порядке можно обнаружить смещение частоты в свете, рассеянном под прямым углом. Скорость звука в воде $v = 1400$ м/с, показатель преломления $n = 1,3$. Считать, что в воде есть звуковые волны всевозможных направлений.

### Solution

(????)

## О2-10.49

Какое число штрихов $N$ должна иметь дифракционная решетка, чтобы с ее помощью в спектре испускания Солнца (во время полного солнечного затмения) можно было обнаружить гравитационное смешение спектральных линий во 2-м порядке? Данные, необходимые для числовых расчетов, можно получить, зная угловой диаметр Солнца (при наблюдении с Земли) $\alpha_{\text{С}} \approx 0,01$ рад и скорость движения Земли по орбите $v_{\text{З}} = 30$km/с.

### Solution

(????)

## О2-10.50

Спектральные линии, излучаемые нагретым газом, оказываются уширенными вследствие того, что атомы газа движутся с разными скоростями относительно наблюдателя (эффект Доплера). Считая распределение скоростей атомов газа максвелловским, оценить размеры $L$ дифракционной решетки с периодом $d = 1$ мкм, которую надо использовать для изучения формы спектральных линий, излучаемых неоном $^{20}$Ne при температуре $T = 1000$K.

### Solution

(????)

## МФТИ-ГКЭ.2020.1А Детектирование спутника

Нобелевская премия по физике за 2019 г. присуждена за открытие несветящихся спутников (экзопланет) у звёзд. Экзопланеты могут быть обнаружены с помощью доплеровского смещения спектра излучения звезды за счёт движения вокруг общего центра масс. Оценить, могут ли экзопланетные физики обнаружить наличие у Солнца спутника (планеты Юпитер), если они используют спектрометр с разрешающей способностью $R = 4 \cdot 10^7$?

Период обращения Юпитера $T_{\text{Ю}} = 11,9$ года, масса $M_{\text{Ю}} = 1,8 \cdot 10^{30}$ г. Масса Солнца $M_C = 2 \cdot 10^{33}$ г. Считать, что наблюдение ведётся в плоскости движения системы Солнце-Юпитер.

### Solution

Мы сравниваем минимальную скорость, которую можем детектировать и скорость вращения источника (Солнца) в такой системе. Первую находим из определения разрешающей способности, вторую из законов Кеплера.

Минимальная скорость, которую способен зарегистрировать по доплеровскому сдвигу спектрометр с такой разрешающей способностью:

$$V_{\min} = c(\Delta f/f) = c(\Delta\lambda/\lambda) = c/R = 7,5 \text{ м/с}$$

(??? додумаю эти формулы, нужно повторить опр. разреш. спос. допл. приборов!)

Оценим скорость Солнца, совершающего круговое движение вокруг центра масс системы Солнце–Юпитер в приближении $M_{\text{Ю}} \ll M_{\text{С}}$ . Используя третий закон Кеплера, находим орбитальную скорость Юпитера

$$V_{\text{Ю}} = 2\pi r_{\text{Ю}} / T_{\text{Ю}} = 2\pi \left[ G M_{\text{С}} (T)^2 / 4\pi^2 \right]^{1/3} / T = 13\text{км/с}$$

(???? проверю, по идее обычная механика это и выдаст)

Следовательно, скорость движения Солнца вокруг центра масс системы Солнце–Юпитер:

$$V_{\text{С}} = V_{\text{Ю}} \ (M_{\text{Ю}} / M_{\text{С}}) = 12\text{км/с}$$

(???? проверю, по идее обычная механика это и выдаст)

Регулярное изменение скорости Солнца относительно экзопланеты имеет вид: $V(t) = V_{\text{С}} \sin(2\pi t l/T)$, т.е. амплитуда колебаний скорости составляет $V_{\text{С}} = 12$ m/c $> V_{\min}$, что превышает предел чувствительности спектрометра. Таким образом, экзопланетные коллеги нобелевских лауреатов 2019 г. Майора и Кело вполне могут обнаружить Солнечную планетную систему.

## МФТИ-ГКЭ.2020.1Б Параметры спутника по излучению

Экзопланетный физик, наблюдая Солнечную систему, заметил возникающее с периодом $T = 11,9$ лет кратковременное (длительностью около земных суток) уменьшение светимости Солнца на $\delta = 1\%$. Также он обнаружил происходящие с тем же периодом колебания доплеровского сдвига линии атомарного водорода в спектре Солнца с амплитудой $|\Delta\lambda/\lambda| \approx 4 \cdot 10^{-8}$, обусловленные движением Солнца вокруг центра масс системы. Оценить из этих данных массу и радиус этой планеты, если масса Солнца $M_{\text{С}} = 2.\ 10^{33}$ г, а его радиус $R_{\text{С}} = 7 \cdot 10^{10}$ cm. Считать, что наблюдение ведётся в плоскости движения системы Солнце-планета.

### Solution

(??? нужно додумать светимость!!)

Изменение светимости происходит из-за того, что «холодная» планета перекрывает для удаленного наблюдателя некоторую долю излучения от Солнца. Поскольку светимость пропорциональна площади поверхности Солнца, то изменение светимости

$$\Delta L/L = \delta \cong R_{\text{пл}}^2 / R_{\text{С}}^2$$

откуда $R_{\text{пл}} = R_{\text{С}}\sqrt{\Delta L/L} = 0.1 R_{\text{С}}$.

Вклад в светимость различных участков Солнца дается интегралом по сферическим кольцам на его поверхности

$$L = 2\pi R_{\text{С}}^2 \int\limits_0^{\pi/2} \sin\theta d\theta I_0 \cos\theta = \pi R_{\text{С}}^2 I_0 \sin^2\theta \Big|_0^{\pi/2} = \pi R_{\text{С}}^2 I_0.$$

$I_0$ яркость излучения перпендикулярно поверхности, при написании интеграла использован закон Ламберта. Если планета расположена в центре диска Солнца прямо напротив наблюдателя, то нижний предел в интеграле следует заменить на $R_{\text{пл}} / R_{\text{С}} \ll 1$. Откуда следует формула $\Delta L/L = \delta \cong R_{\text{пл}}^2 / R_{\text{С}}^2$.

Поскольку Солнце и планета вращаются вокруг общего центра масс, то доплеровский сдвиг будет максимален, когда Солнце движется к наблюдателю и минимален, когда Солнце движется от наблюдателя. Следовательно, линейная скорость вращения Солнца находится из соотношения $V_{\text{С}} = c\Delta\lambda/\lambda = 1200$ см/с. Будем, считать, что масса планеты много меньше массы Солнца и центр масс системы практически совпадает с центром Солнца. Тогда радиус орбиты планеты находится из третьего закона Кеплера

$$R_{\text{пл}} = G M_{\text{С}} T^2 / 4\pi^2$$

откуда $R_{\text{пл}} = 7,8 \quad 10^{13}\text{с} = 7,8 \quad 10^8$ км $= 5,2$ а.е., а линейная скорость движения планеты по орбите $V_{\text{пл}} = 2\pi R_{\text{пл}} / T = 13\text{k/с}$. Из равенства импульсов планеты и Солнца, получаем $m_{\text{пл}} = M_{\text{С}} V_{\text{С}} / V_{\text{пл}} = 1,8510^{30}$ г. Полученные данные соответствуют Юпитеру.

## МФТИ-ГКЭ.2020.1Б Параметры спутника по излучению

В 2016 г. у ближайшей к нам звезды красного карлика Проксима Центавра ($M_{\Pi} = 0,12 M_{\text{С}}, R_{\Pi} = 0,14 R_{\text{С}}, T_{\Pi} = 3050$ K) было обнаружено периодическое изменение доплеровского смещения линий излучения атомарного водорода, которое было интерпретировано как следствие движения её центра масс по круговой орбите со скоростью $v_{\Pi} = 1.5$ m/c и с периодом $\tau_{\Pi} = 11,2$ сут. Считая, что это движение обусловлено существованием у звезды спутника (планеты земного типа Проксима Центавра b), оценить его радиус орбиты (в астрономических единицах, а.е.), массу и температуру поверхности.

### Solution

(??? нужно додумать светимость!!)

Для системы планета-звезда по третьему закону Кеплера $L_b^3 = GM_\Pi \tau_\Pi^2/4\pi^2$, откуда $L_b = \left(GM_\Pi \tau_\Pi^2/4\pi^2\right)^{1/3} = 7.2$ млн. км $= 4,8 \cdot 10^{-2}$ а.е. (1 а.е. $= 150$ млн. км - расстояние от Земли до Солнца и предполагается, что масса звезды намного больше массы планеты).

Радиусы орбит планеты и звезды связаны соотношением $L_\Pi M_\Pi = L_b M_b$, откуда $M_b = \frac{L_\Pi M_\Pi}{L_b} = \frac{\tau_\Pi v_\Pi M_\Pi}{2\pi L_b} = 7.7 \cdot 10^{27}\Gamma = 1.28 M_3$.

Приравнивая величины падающей в единицу времени на поверхность планеты энергии и энергии ей излучаемой, получаем

$$4\pi R_\Pi^2 \sigma T_\Pi^4 \frac{\pi R_b^2}{4\pi L_b^2} = 4\pi R_b^2 \sigma T_b^4$$

откуда

$$T_b = T_\Pi \left(\frac{R_\Pi}{2L_b}\right)^{1/2} = 3050 \left(\frac{0,14 \cdot 7,0 \cdot 10^5}{2 \cdot 4,8 \cdot 10^{-2} \cdot 150 \cdot 10^6}\right)^{1/2} = 251 \text{ K}$$

## O2-10.51 Детектирование системы двух звезд

Двойная звезда состоит из двух близких по массе звезд, вращающихся относительно общего центра с периодом $\tau = 10$ сут и отстоящих друг от друга на расстояние $L = 2 \cdot 10^7$km. Определить, какое число штрихов $N$ дифракционной решетки необходимо для того, чтобы при наблюдении видимого спектра водорода в излучении этих звезд можно было во 2-м порядке заметить вращение системы. Можно ли, в принципе, таким способом заметить относительное вращение таких звезд, если период $\tau = 10$ лет?

Температура поверхности звезд $T = 6000$K.

### Solution

(аналогично пред. задаче решится скорее всего)

## O2-10.55

Оценить температуру водородной лампы, используемой в качестве источника света в интерферометре Майкельсона, если при смещении одного из зеркал число наблюдаемых интерференционных максимумов составляет $N = 7 \cdot 10^4$. Первоначально зеркала интерферометра были расположены на одинаковых расстояниях от делительной пластинки. Влиянием протяженности источника пренебречь.

### Solution

(????)

# 12.0.16 Problems about polarization effects 1

## O2-11.1

Найти наименьшую толщину $d$ пластинки кварца, вырезанной параллельно оптической оси, чтобы падающий плоско поляризованный свет выходил поляризованным по кругу ($n_\text{e} = 1,5533$, $n_0 = 1,5442, \lambda = 5 \cdot 10^{-5}$ cm)

## O2-11.2

При какой толшине пластинка из исландского шпата является пластинкой в четверть волны для света с длиной волны $\lambda_1 = 5880$ Å и может поворачивать плоскость поляризации на 90° для света с длиной волны $\lambda_2 = 5740$ Å? Разность показателей преломления для обыкновенного и необыкновенного лучей принять равной 0,2 для обеих длин волн. Считать, что обыкновенный и необыкновенный лучи идут по одному направлению.

## O2-11.3

Параллельный пучок света падает нормально на пластинку исландского шпата, вырезанную параллельно оптической оси. Определить разность хода $\Delta$ обыкновенного и необыкновенного лучей, прошедших через пластинку. Толщина пластинки равна 0,03 мм; $n_0 = 1,658, n_e = 1,486$

## O2-11.4

Какова должна быть наименьшая толщина $d$ пластинки слюды, чтобы она могла служить в качестве пластинки в 1/4 волны для света натриевого источника, если для этого света показатели преломления волн, идущих перпендикулярно к пластинке, соответственно равны $n_1 = 1,5941, n_2 = 1,5887$?

## O2-11.5

Почему, если через плоскопараллельную двоякопреломляющую кристаллическую пластинку смотреть на удаленный предмет, видно одно изображение, а не два, как в случае близко расположенных предметов?

## O2-11.6

Наблюдатель смотрит на близкий предмет через плоскопараллельную двоякопреломляющую пластинку из исландского шпата и видит два прямых увеличенных изображения предмета, когда между пластинкой и предметом помещена собирающая линза на расстоянии 4 см от предмета. После того как к линзе вплотную приложили собирающее очковое стекло с оптической силой в 5 дптр, стало видно только одно изображение предмета. Определить фокусное расстояние $f$ линзы.

## O2-11.7

Ветровое стекло и фары автомашин снабжаются пластинками из поляроида. Как должны быть расположены эти пластинки, чтобы шофер мог видеть дорогу, освещенную светом его фар, и не страдал бы от света фар встречных машин?

## O2-11.8

Узкий пучок исполяризованного света падает нормально на пластинку исландского шпата и затем нормально на вторую такую же пластинку, главная плоскость которой образует с главной плоскостью первой пластинки угол 30°. Затем свет падает на экран. Описать полученную картину и найти относительную интенсивность наблюдаемых на экране пятен.

При меч ание. Пластинки вырезаны так, что оптическая ось составляет угол $\gamma$ с плоскостью пластинки. При этом $0 \leqslant \gamma < 90°$.

## O2-11.9

В интерференционном опыте Юнга между щелью $S$ и щелями $S_1$ и $S_2$ (рис. 367) введен поляроид $P$, главные оси которого параллельны или перпендикулярны к щелям $S_1$ и $S_2$. Как изменится интерференционная картина на экране, если щели $S_1$ и $S_2$ прикрыть пластинками в полволны, ориентированными взаимно пергендикулярно друг к другу (параллельно и перпендикулярно к щелям)? Что произойдет, если поляроид $P$ повернуть на 90°? Какая картина будет наблюдаться, если убрать поляроид? Рассмотреть ту же задачу, если вмсто пластинки в полволны используется пластинка в четверть волны. Щели $S_1$ и $S_2$ предполагаются узкими (порядка длины волны), а расстояние между ними большим по сравнению с их шириной.

## O2-11.10

: Частично линейно поляризованный свет рассматривается через николь. При повороте николя на 60° от положения, соответствующего максимальной яркости, яркость пучка уменьшается в два раза. Найти степень поляризации пучка $\Delta = \frac{\max - \mathfrak{d}_{\min}}{\mathfrak{f}_{\max} + \lceil_{\min}}$ и отношение интенсивностей естественного и линейно поляризованного света ($\mathscr{V}_{\max}$ и $\mathscr{I}_{\min}$ - максимальная и минимальная интенсивности света, проходящего через николь). 199

## O2-11.11

Определить, во сколько раз изменится интенсивность частично поляризованного света, рассматриваемого через николь, при повороте николя на 60° по отношению к положению, соотвстствующему максимальной интенсивности. Степень поляризации света $\Delta = \frac{\mathfrak{f}_{\max} - \mathscr{V}_{\min}}{\mathfrak{d}_{\max} + \mathfrak{m}_{\min}} = 0,5$.

## O2-11.12

Один поляроид пропускает 30% света, если на него падает естественный свет. После прохождения света через два таких поляроида интенсивность падает до 9%.Найти угол $\varphi$ между осями поляроидов.

## O2-11.13

Некогерентная смесь линейно поляризованного света и света, поляризованного по кругу, рассматривается через поляроид. Найдено положение поляроида, соответствующее максимальной интенсивности прошедшего свста. При повороте поляроида из этого положения на угол $a = 30°$ интенсивность света уменьшалась на $p = 20\%$ Найти отношение интенсивности света $I_{\text{к}}$, поляризованного по кругу, к интенсивности линейно поляризованного света $\pi_{\pi}$.

## O2-11.14

Некогерентная смесь линейно поляризованного света и света, поляризованного по кругу, рассматривается через николь. Найдено положение николя, при котором интенсивность проходящего свста максимальна. При поворотс николя от этого положения на некоторы угол вокруг оси пучка интенсивность проходящего света уменьшается в $m = 2$ раза по сравнению с максимальной и во столько же раз увеличивается по сравнению с минимальной. Найти отношение интенсивности $\mathscr{V}_{\text{k}}$ света, поляризованного по кругу, к интенсивности света $I_{\text{л}}$ линейно поляризованного.

## O2-11.15

Как изменится результат предыдущей задачи, если линейно поляризованный свет и свет, поляризованный по кругу, будут когерентны?

## O2-11.16

Смесь света, поляризованного по кругу, и естественного рассматривается через кристаллическую пластинку в чстверть волны и николь. При вращении николя вокруг оси светового пучка найдено, что максимальная интенсивность света, прошедшего чсрез систему в $m = 3$ раза, превосходит минимальную интенсивность. Найти отношение интенсивности света к, поляризованного по кругу, к интенсивности естественного света $_{\text{e}}$.

## O2-11.17

Параллельный пучок монохроматического света с длиной волны $\lambda$, поляризованный по правому кругу, падает нормально на пластинку в полволны. Найти состояние поляризации света, прошедшего через эту пластинку.

## O2-11.18

Параллельный пучок монохроматического света с длиной волны $\lambda$ падает нормально на поляроид, а затем на пластинку в полволны. Главная плоскость поляроида (в которой лежит электрический вектор пропускаемой им волны) составляет угол $\alpha$ с осью этой пластинки. Найти состояние поляризации прошедшего света на выходе из пластинки в полволны.

## O2-11.19

Параллельный пучок монохроматического света проходит через два николя, главные плоскости которых повернуты друг относительно друга на угол $\alpha = 20°$. Между николями ставится пластинка одноосного кристалла, вырезанная параллельно оптической оси и вносящая разность хода $\lambda/2$ между обыкновенным и необыкновенным лучами. Какой угол $\beta$ должна составлять оптическая ось пластинки с главным направлением первого николя, чтобы свет через эту систему не прошел?

## O2-11.20

Для сравнения яркостей двух повсрхностей, освещаемых неполяризованным светом, одну из них рассматривают непосрсдственно, а другую через два николя. Каково отношение этих яркостей, если освещенность обеих поверхностей кажется одинаковой при угле между николями $\alpha = 60°$? Считать, что потери света в каждом николе на отражение и поглощенис составляют $p = 10\%$ от падающего света.

### O2-11.21

Имеется горизонтальный параллельный пучок эллиптически поляризованного света. Обнаружено, что при прохождении пучка чсрез пластинку в $\lambda/4$ при определенной ее ориентации свет оказывается линейно поляризованным под углом $\alpha_1 = 23^\circ$ к вертикали. Если пластинку повернуть на угол $90^\circ$, то весь свет снова оказывается линейно поляризованным под углом $\alpha_2 = 83^\circ$ вертикали. Найти отношение $a/b$ полуосей эллипса поляризации и угол $\varphi$ наклона большой оси.

### O2-11.22

Некогерентная смесь неполяризованного, линейно поляризованного света и света с круговой поляризацией анализируется при помощи быстро вращающегося поляризатора и фотоприемника, ток которого зависит от интенсивности света. Оказалось, что глубина модуляции фототока равна $m_1 = 0,1$. После установки на пути лучей пластинки $\lambda/4$ было выяснено, что свет по-прежнему представляет собой некогерентную смесь неполяризованного, линейно поляризованого света и света, поляризованного по кругу, но теперь глубина модуляции фототока составила $m_2 = 0,2$. Отределить степень поляризации света 1.

### O2-11.26

Плоская волна монохроматического света длины $\lambda$, поляризованного по кругу, создает в точке $\bar{P}$ интенсивность $I_0$. На пути волны ставят большую пластинку из идеального поляроида, как показано на рис. 369. Показатель преломления вещества поляроида $n$. Найти толщину $d$ пластинки, при которой интснсивность свста в точке $P$ будет максимальной. Чему равна $\mathcal{I}_{\max}$?

### O2-11.27

Плоская волна монохроматического света, поляризованного по кругу, создает в точке $P$ интенсивность $\mathscr{I}_0$. На пути волны ставят две большие пластинки в $\lambda/4$, как показано на рис. 370. Главные направления пластинок ориентированы взаимно перпендикулярно. Найти интенсивность $\mathcal{Y}$ в точке $P$.

### O2-11.32

Круглое отверстие в непрозрачном экране содержит для точки наблюдения $P$ одну зону Френеля. Отверстие закрыто поляроидом так, что направления колебаний в первой и второй половинах зон взаимно перпендикулярны. Отверстие освещается светом, поляризованньм по кругу. Определить интенсивность света $\int$ в точке $P$, если в отсутствие экрана она равна $_0$. Как будет поляризован свет в точке наблюдения? Считать, что в поляроидах нет поглощения света разрешенной поляризации.

### O2-11.33

Определить интенсивность света $\mathscr{V}$ в точке $P$ экрана, на который падает монохроматический свет интенсивностью $\vartheta_0$, если на пути поставить диск из оптически активного вещества, закрывающий полторы зоны Френеля и поворачивающий плоскость поляризации на $90^\circ$. Отражением и поглощением света пренебречь.

### O2-11.34

Параллельный пучок неполяризованного монохроматического света падает на пластинку в $\lambda/4$. Интенсивность света в некоторой точке $P$ за пластинкой равна $\mathscr{V}_0$. Из пластинки вырезают диск, закрывающий одну зону Френеля для точки $P$. Диск повернули вокруг луча на угол $90^\circ$ и поставили на место. Какой стала интенсивность $\circ$ в точке $P$?

### O2-11.35

Параллельный пучок неполяризованного монохроматического света падает на пластинку в $\lambda/2$. Интенсивность света в некоторой точке наблюдения $P$ за пластинкой равна $\mathscr{Q}_0$. Из пластинки вырезают диск, закрывающий полторы зоны Френеля для точки $P$. Диск повернули вокруг луча на угол $\pi/2$ и поставили на место. Какой стала интенсивность $\mathscr{I}$ в точке $P$?

### O2-11.36

Из кристаллической пластинки $\lambda/2$ вырезаны диски диаметром в одну и две зоны Френеля для точки $P$. Диски вносят в пучок света вплотную друг к другу, так что у них совпадают 1) разноименные главные направления, 2) одноименные главные направления. При этом для света, поляризованного по одному из главных направлений, ни амплитуда, ни фаза колебаний не изменились. 203 Во сколько раз изменится интенсивность света той жс поляризации в случаях 1) и 2), если малый диск повернуть на $90^\circ$?

### O2-11.37

Плоская монохроматическая волна, поляризованная по кругу, падает на диск, вырезанный из пластинки $\lambda/2$. Для точки наблюдения на оси диск закрывает три первые зоны Френеля. Толщина диска подобрана так, что он вносит дополнительный оптический путь в $3\lambda/2$ для необыкновенного луча. Во сколько раз изменится интенсивность в точке наблюдения, если диск убрать? Поглощением и отражением света пренебречь.

## 12.0.17 Problems about polarization effects 2

### O2-11.38

Зонная пластинка сделана из поляроида. Во всех четных зонах поляроид ориентирован вертикально, во всех нечетных - горизонтально. Какова будет интенсивность света в основном фокусе пластинки, если она освещается неполяризованным светом?

### O2-11.39

В непрозрачном экране, на который нормально падает плоская линейно поляризованная волна интенсивностью e, вырезано круглое отверстие размером в одну зону Френеля для некоторой точки наблюдения, лежащей на оси системы. В отверстие вставлены пластинки $\lambda/4$ в форме полудисков, одноименные оси которых ориентированы взаимно перпендикулярно. Направление колебаний в падающей волне составляет 45° с главными направлениями пластинок в обоих случаях. Какова интенсивность колебаний в точке наблюдения??

### O2-11.40

В непрозрачном экране, на который нормально падает зано круглое отверстие размером в две зоны Френеля для некоторой точки наблюдения, лежашей на оси системы. Первая зона перекрыта пластинкой $\lambda/4$ в форме диска, а вторая - пластинкой $\lambda/4$ в форме кольца. Одноименные оси пластинок ориентированы взаимно перпендикулярно. Главные направления пластинок в обоих случаях составляют угол 45° с направлением колебаний в падающей волне. Какова интенсивность колебаний в точке наблюдения?

### O2-11.42

Параллелыный пучок линейно поляризованного света длиной волны $\lambda = 500$Н с вектором $\mathbf{E}_0$, лежащим в плоскости рис. 374, нормально падает на три шели, закрытые одинаковыми двоякопреломляющими пластинками в $\lambda/2$. Причем первые две рядом стоящие щели закрыты одинаково ориентированными пластинками так, что вектор E падающей волны совпадает с одним из главных направлений этих пластинок. А третья пластинка повернута относительно первых двух на 90°. Интерференционная картина рассматривается на экране с помощью линзы с фокусным расстоянием $f = 2$ м. Считая размер щелей пренебрежимо малым по сравнению с расстоянием $d$ между ними ($d = 0,5$ cm), найти распределение интенсивности в фокальной плоскости линзы. Определить также видность интерференционной картины и ее период на экране .

### O2-11.43

Параллельный пучок поляризованного по кругу монохроматического света падает на решетку с периодом $d$ и шириной щелей $b = d/2$. Каждая щель перекрыта двумя полосками поляроида одинаковой ширины $b/2$ с взаимно перпендикулярными разрешенными направлениями. Какова поляризация света в нулевом и боковых дифракционных максимумах ($\pm 1 - , \pm 2 - , \ldots$)?

### O2-11.44

Как изменится разрешающая способность дифракционной решетки, если одну ее половину прикрыть поляроидом, ориентированным параллельно штрихам решетки, а другую - поляроидом, ориентированным перпендикулярно к штрихам? Будет ли зависеть разрешающая сила решетки от поляризации падающего света?

### O2-11.45

В предыдущей задаче перед и за решеткой дополнительно ставятся два поляроида, главные направления которых параллельны друг другу и образуют угол 45° с направлением штрихов решетки. Как изменится разрешающая способность такой решетки по сравнению с ничем не прикрытой решеткой?

## O2-11.46

Параллельный пучок циркулярно поляризованного света с длиной волны $\lambda$ падает нормально на дифракционную решетку. Как изменится расстояние между главными максимумами и интенсивность нулевого максимума, если щели решетки перекрыть пластинками в $\lambda/4$, причем так, что одноименные главные направления на четных и нечетных щелях взаимно перпендикулярны? Какова поляризация света в нулевом, четных и нечетных дифракционных максимумах?

## O2-11.48

Наблюдается дифракция Фраунгофера монохроматического циркулярно поляризованного излучения на плоской решетке. Как изменятся расстояния между дифракционными максимумами и интенсивность нулевого максимума, если щели закрыть пластинками в $\lambda/2$ так, что главные оси пластинок в соседних щелях будут взаимно перпендикулярны?

## O2-11.49

Плоская монохроматическая линейно поляризованная световая волна падает нормально на периодическую структуру, состоящую из чередующихся полосок поляроида, главные плоскости которых взаимно перпендикулярны. Отношение ширины полосок разных типов равно трем. Направление колебания электрического вектора волны составляет угол $\alpha$ с главной плоскостью полоски ⌋ меньшей шириной. Определить направления на главные максимумы первого порядка и их интенсивности, если интенсивность волны в нулевом максимуме равна $\mathscr{I}_0$.

## O2-11.50

: Линейно поляризованное лазерное излучение с длиной волны $\lambda = 0,63$ мкм падает нормально на тонкую магнитно-оптическую прозрачную пленку с чередующейся одномерной доменной структурой. Свет, проходящий через соседние домены, испытывает поворот плоскости поляризации в противоположные стороны на угол $\pi/2$. Ширина доменов одинакова и равна $d$. Дифракционная картина наблюдается в фокальной плоскости линзы. Определить интенсивность света в дифракционных максимумах нулевого, 1-го и 2-го порядков, если известно, что при освещении тем же излучением обычной амплитудной решетки, состоящей из чередующихся прозрачных и непрозрачных полосок шириной $d$, интенсивность в максимуме 1 -го порядка оказывается равной

## O2-11.51

Световые волны с одинаковыми частотой и линейной поляризацией распространяются в плечах двулучевого интерферометра. На выходе из него фотоприемник регистрирует интенсивность света $I$ из плеч она колеблется между $\mathscr{I}_{\max}$ и $\mathscr{I}_{\text{мин}}$ так, что видность картины $V = (\mathscr{I}_{\max} - \mathscr{I}_{\text{мин}}) / (\mathscr{I}_{\max} + \mathscr{I}_{\text{мин}}) = 1$. Какой будет ее наименьшая величина, если в одно из плеч интерферометра поместить прозрачную кристаллическую пластинку $\lambda/4$? Считать, что поток проходит через пластинку однократно.

## O2-11.59

Между двумя скрещенными поляроидами расположена анизотропная кристаллическая пластинка с заданными $\Delta$ и $d$, вырезанная параллельно оптической оси. Разрешенные направления поляроидов составляют угол $45°$ с оптической осью. Система освещается параллельным пучком амплитудно-модулированного света. При какой частоте модуляции система пропустит боковые гармоники, отфильтровав несущее колебание?

## O2-11.60

Показатель преломления кристаллического кварца для длины волны $\lambda = 589$ равен $n_0 = 1,544$ для обыкновенного луча и $n_e = 1,553$ для необыкновенного луча. На пластинку из кварца, вырезанную параллельно оптической оси, нормально падает линейно поляризованный свет указанной длины волны, занимающий спектральный интервал $\Delta\lambda = 40$ нм. Найти толщину пластинки $d$ и направление поляризации падающего света, если свет после пластинки оказался неполяризованным.

## O2-11.61

Расположив пластинку, вырезанную из исландского шпата, параллельно его оптической оси, между скрещенными николями, можно осуществить монохроматор, позволяющий, например, задержать одну из линий дублета натрия и пропустить другую. Найти, какой должна быть при этом минимальная толщина $d_{\min}$ пластинки и как ее нужно ориентировать. Показатели преломления исландского шпата для линии $\lambda_1 = 589,0$ нм равны $n_{\text{el}} = 1,48654$ и $n_{\text{ol}} = 1,65846$, для линии $\lambda_2 = 589,6$ nm $- n_{\text{e2}} = 1,48652$ и $n_{02} = 1,65843$.

## O2-11.62

Кварцевая пластинка $\lambda/2$ используется как анализатор степени поляризации лазерных импульсов. Оценить минимальную длительность лазерных импульсов, для которых еще можно пользоваться таким анализатором, если длина волны света $\lambda = 0,63$ мкм, а коэффициенты преломления для обыкновенного и необыкновенного

лучей в кварце $n_0 = 1,5442$ и $n_c = 1,5533$ соответственно. Дисперсией показателей преломления пренебречь.

## O2-11.63

Ограниченный импульс длительностью зованного излучения нормально падает на двулучепреломляющую пластинку толщиной $d$. Пластинка вырезана параллельно оптической оси. Плоскость поляризации света составляет угол 30° относительно одного из главных направлений. Найти значения $d$, при которых за пластинкой будут наблюдаться два раздельных импульса. Как поляризованы эти импульсы и каковы максимальные значения их амплитуд, если максимальное значение амплитуды исходного импульса равно $E_0$. Разность $\Delta n = n_0 - n_e$ считать известной и не зависящей от частоты света.

## O2-11.67

Два николя $N_1$ и $N_2$ повернуты один относительно другого на угол $\alpha$; между ними помещен николь $N_3$ (рис. 381). На систему падает параллельный пучок неполяризованного света. Предполагая, что необыкновенный луч проходит через каждый николь без потерь, найти ориентацию николя $N_3$ относительно николя $N_1$, при которой интенсивность проходящего света максимальна. Определить 209 интенсивность проходящего света в этих положениях, если интенсивность падающего света равна $I_0$

## O2-11.68

На пластинку кварца, вырезанную параллельно оптической оси, нормально падает белый свет, поляризованный по кругу. За пластинкой поставлен поляроид, главное направление которого составляет угол 45° с осью пластинки. Прошедший свет попадает на щель спектрографа. Сколько темных полос $k$ получится в спектре, если толщина кварцевой пластинки $d' = 2$ mm, $n_e = 1,55, n_o = 1,54$. Падающий свет занимает интервал длин волн от $\lambda_1 = 4000$ Å до $\lambda_2 = 5000$ Å, в котором $n_e - n_0 = \text{const}$.

## O2-11.69

Между скрещенными николями помещена пластинка кварца, вырезанная параллельно оптической оси. Оптическая ось пластинки составляет угол 45° с главными направлениями николей. Рассчитать минимальную толщину пластинки, при которой одна линия водорода $\lambda_1 = 6563$ Å будет сильно ослаблена, а другая $\lambda_2 = 4102$ Å будет обладать максимальной интенсивностью. Величина анизотропии кварца $\Delta n = 0,009$.

## O2-11.72

На кварцевую пластинку толщиной 3 мм, вырезанную параллельно оптической оси, нормально падает пучок белого линейно поляризованного света, плоскость поляризации которого составляет угол 45° с осью пластинки. Выходящий из пластинки свет сначала вновь проходит через николь, скрещенный с первичным поляризатором светового пучка, а затем падает на щель спектроскопа. Сколько темных полос будет наблюдаться в спектре между длинами волн $\lambda_D = 5890$ Å и $\lambda_F = 4860$ Å, если обыкновенный ($n_0$) и необыкновенный ($n_e$) показатели преломления кварца для этих длин волн имеют следующие значения: для $\lambda_D - n_0 = 1,5442, n_e = 1,5533$, а для $\lambda_F - n_0 = 1,5497, n_e = 1,5589$.

## O2-11.73

Между двумя скрещенными николями помещена кристаллическая пластинка толщиной $d = 0,045$ мм с показателями преломления $n_e = 1,55, n_o = 1,54$. Пластинка вырезана параллельно оптической оси кристалла и ориентирована так, что главное направление первого николя составляет угол $\alpha = 30°$с ее оптической осью. 210 На систему нормально падает естественный свет с длиной волны $\lambda = 6000$ Å и интенсивностью $\mathscr{I}$ прошедшего через описанную систему.

## O2-11.74

Два когерентных пучка квазимонохроматического неполяризованного света равной интенсивности дают на экране интерференционные полосы. Какой толщины кристаллическую пластинку надо ввести на пути одного из этих пучков, чтобы интерференционные полосы исчезли и притом так, чтобы их нельзя было восстановить никакой стеклянной пластинкой, вводимой в другой пучок? Как изменится

картина, если за кристаллической пластинкой поставить поляроид? При каком положении поляроида интерференционных полос не будет?

## O2-11.78

Между скрещенными николями помещена кристаллическая пластинка толшиной $d_1 = 0,02$ m с величиной анизотропии $\Delta n_1 = 0,05$. На нее в параллельном положении положена другая пластинка толщиной $d_2 = 0,02$m $\Delta n_2 = 0,025$. В какой цвет окрашено поле зрения? В какой цвет будет оно окрашено, если верхнюю пластинку и верхний николь повернуть на $90^\circ$ от первоначального положения?

При меч ание. Параллельным положением называется такая ориентировка пластинок, при которой направление колебаний волны, распространяющейся с большей (меньшей) скоростью в одной пластинке, совпадает с направлением колебаний волны, распространяющейся с большей (меньшей) скоростью во второй пластинке.

## O2-11.79

Поверхности кристаллической пластинки толщиной $d = 1$ мм, обладающей слабой анизотропией, покрыты отражающими покрытиями ⌋ коэффициентом отражения по энергии $\rho = 0,9$. Пластинка вырезана параллельно оптической оси и освещается параллельным пучком неполяризованного света $\left(\lambda = 5 \cdot 10^{-5} \text{ cm}\right)$. Оценить минимальную величину анизотропии $\Delta n$, при которой свет на выходе из пластинки окажется практически полностью поляризованным $(n_0 = 1,62)$.

## O2-11.80

Поверхности плоскопараллельной кварцевой пластинки толщиной $d = 0,5$ мм, вырезанной параллельно оптической оси, покрыты высокоотражающими покрытиями с коэффициентом отражения по энергии $\rho = 0,9$. Пластинка освещается параллельным пучком света с круговой поляризацией (длина волны $\lambda = 6 \cdot 10^{-5}$ cm). 211 Какова поляризация и интенсивность света, прошедшего пластинку, если интенсивность падающего света $\mathscr{I}_0$? Главные показатели преломления $n_0 = 1,5442, n_\text{e} = 1,5534$.

## O2-11.82

Клин из двоякопреломляющего вещества помещен на пути монохроматического света, поляризованного по кругу. Оптическая ось параллельна ребру клина. Описать наблюдаемую через николь картину, когда клин неподвижен и когда он поворачивается вокруг направления распространения света.

## O2-11.83

Две одинаковые плоские густые решетки из параллельных тонких идеально проводящих проволок расположены в параллельных плоскостях на расстоянии $l = 1,5$ cm одна от другой так, что образующие их проволоки взаимно перпендикулярны. Плоскость поляризации волны $\lambda = 8,5$ cm, падающей на систему решеток под углом $\alpha = 45^\circ$, наклонена на угол $\theta = 45^\circ$ к плоскости падения. Определить характер поляризации отраженной волны для случая, когда плоскость падения параллельна направлению проволок 1 -й решетки.

## O2-11.84

Параллельный пучок монохроматического излучения (длина волны в вакууме $\lambda = 496$ мкм), поляризованного по кругу, падает нормально на решетку, изготовленную в виде натянутых проволочек с расстоянием между ними $d \ll \lambda$. При таких условиях решетка полностью пропускает излучение, поляризованное так, что электрический вектор направлен перпендикулярно проволочкам, и отражает излучение с поляризацией, повернутой на $90^\circ$. Найти вращающий момент $M$ и силу $F$, действующие на решетку, если интенсивность потока в пучке $\mathscr{I} = 10\text{В/cm}^2$, а облучаемая поверхность решетки $S = 10 \text{ cm}^2$.

## O2-11.85

Две одинаковые плоские густые решетки из параллельных тонких идеально проводящих проволок расположены в параллельных плоскостях на расстоянии $l = 2$ см одна от другой так, что образующие их проволочки взаимно перпендикулярны. Волна $\lambda = 11,3$ см с правой эллиптической поляризацией падает на систему решеток под углом $\alpha = 45^\circ$. Большая ось эллипса лежит в плоскости падения, отношение осей эллипса равно 1,73. Определить характер отраженной волны для случая, когда плоскость падения параллельна направлению проволок первой решетки. Описать изменение ее характера при вращении системы решеток (их взаимное расположение и плоскости сохраняются).

### O2-11.87

Пространство между зеркалами интерферометра ФабриПеро с разрешающей способностью $R = \lambda/\delta\lambda = 10^8$ заполнено химически чистым нитробензолом. При наложении однородного поперечного электрического поля нитробензол становится слабоанизотропной средой, причем оптическая ось совпадает с направлением поля (эффект Керра). Интерферометр освещается монохроматическим пучком неполяризованного света ($\lambda = 600$ нм). Оценить минимальную величину электрического поля, при которой на выходе интерферометра будет наблюдаться почти полностью линейно поляризованный свет. Постоянную Керра для нитробензола принять равной $B = 2 \cdot 10^{-5}$ ед. СГСЭ.

П р и меч а н и е: Постоянной Керра называют константу в выражении $n_e - n_0 = \lambda B E^2$.

### O2-11.88

Нелинейный интерферометр Фабри-Перо представляет собой тонкую пластинку из вещества, показатель преломления которого пропорционален квадрату напряженности электрического поля: $n = n_0 + n_2 E^2$. Пластинка покрыта высокоотражающими покрытиями с коэффициентом отражения по энергии $\rho = 99\%$. Определить уровень плотности мощности $S$ лазерного излучения с $\lambda = 1,051$ мкм, пропускаемого таким интерферометром, если $n_0 = 3,5, n_2 = 10^{-9}$ ед. СГСЭ, а толщина пластинки $d = 12$ мкм.

### O2-11.89

Явлене самофокусировки объясняется зависимостью показателя преломления от интенсивности света $\left(n = n_0 + n_2 E_0^2\right.$, где $E_0$ - амплитуда напряженности электрического поля в световой волне). Одним из самых больших значений $n_2$ обладает сероуглерод $\left(n_2 = 2 \cdot 10^{-11}\right.$ ед. СГСЭ). Мощный пучок лазерного излучения с параболической зависимостью интенсивности от расстояния от центра пучка $\left(\mathcal{I} = \mathscr{I}_0\left(1 - r^2/r_0^2\right)\right.$ при $r < r_0$ и $\mathscr{I} = 0$ при $r > r_0$) проходит сквозь слой сероуглерода толщиной $L = 5$ см. Найти, на каком расстоянии от кюветы с сероуглеродом сфокусируется лазерный пучок $\mathcal{I}_0 = 5 \cdot 10^8 \text{Вт/см}^2, r_0 = 5\text{мМ}.1$

### O2-11.90 Мощность для фокусировки

Гауссов пучок неодимового лазера ($\lambda = 1$ мкм) с радиальным распределением поля по сечению: $E = E_0 \exp\left[-r^2/\left(R^2\right)\right]$ ($R$=3 мм) и с плоским волновым фронтом падает на плоскопараллельную пластинку толщиной $d = 1$ cm, сделанную из нелинейного вещества, показатель преломления которого зависит от интенсивности: $n = n_0 + n_2 E^2$ ($n_2 = 10^{-11}$ ед. СГСЭ). Оценить, при какой мощности лазера возможно уменьшить диаметр пучка (фокусировка) после прохождения пластинки.

### O2-11.91

У большинства веществ показатель преломления уменьшается с ростом температуры (вследствие теплового расширения). Пучок лазерного излучения видимого диапазона спектра проходит сквозь слой слабопоглощающей жидкости толщиной $L = 1$ cm, так что в жидкости устанавливается распределение температуры $t = t_0 + t_1\left(1 - r^2/r_0^2\right)$ при $r < r_0$ и $t = t_0$ при $r > r_0$, где $r$ - расстояние от центра пучка, $t_1 = 2°\text{C}, r_0 = 2$ мм. Найти фокусное расстояние и знак наведенной в жидкости линзы, если $dn/dt = -4 \cdot 10^{-4}\text{K}^{-1}$.

# 13 Questions about checking knowledge

## 13.0.1 Questions about understanding typical optics

### Visibility

What is a visibility for the following intensity
Answer:
it is always one here, since $I_{\min} = 0$.

### Почему происходит потеря полволны при отражении от более плотной среды? (?????)

Answer:
(отвечу потом. В школе об этом говорят, а в вузе так и не до понимания этого)

#### How a Lens creates an Image?

Answer:
(I'll write my answer later, for now see https://www.youtube.com/watch?v=SS2AbZVdk2A&ab_channel=HuygensOptics)

### 13.0.2 Questions about understanding details in optics

#### How big is a visible photon?

Answer:
(I'll write my answer later, for now see https://www.youtube.com/watch?v=SDtAh9IwG-I&t=1105s)

### 13.0.3 Questions about understanding quantum optics

# 14 Preparation to some parts of optics

## 14.1 Background for optics

### 14.1.1 Problems for mastering mathematics for optics

(?? I'll write later)

### 14.1.2 Problems for mastering field theory for optics

(?? I'll write later)

### 14.1.3 Problems about some generally obvious proofs

If one really likes these proofs, one can open a book by a real geometer, like academic Alexandrov, Labochevsky, and find there hundreds of pages with these proofs. They tell us nearly nothing about optics, so there is no sense of doing it.

#### O2-1.2 Basic geometry

Доказать геометрически, что если луч света, исходящий из точки $A$, попадает в точку $B$ после отражения от плоского зеркала, то длина пути этого луча меньше, чем длина любого другого пути, проходящего от $A$ к зеркалу, а затем к $B$.

#### O2-1.3 Вывести элементарный принцип Ферма

Показать, что если луч света, исходящий из точки $A$, попадает в точку $B$ после преломления на плоской границе раздела двух сред, то оптическая длина этого луча меньше оптической длины любого пути, соединяющего $A$ и $B$ (принцип Ферма).
(может, это много где и решено?)

## 14.2 Easy optics for preparation

### 14.2.1 Problems about easy geometrical optics

(School is here)

### 14.2.2 Problems about proofs in geometrical optics

#### O2-1.6

Вывести формулу сферического зеркала и формулу тонкой линзы из принципа таутохронизма.

# 15 Problems about specific setups, phenomena

## 15.1 Problems about specific well-known phenomena

### 15.1.1 Problems about Pressure of Light

#### O2-2.31 Давление на плоское зеркало (??)

Исходя из представления, что свет состоит из фотонов, каждый из которых обладает импульсом $h\nu/c$, определить давление $P$ световой волны на плоское зеркало, предполагая, что коэффициент отражения зеркала равен $r$, а угол падения равен $\varphi$. Определить также тангенциальную силу $T$, действующую на единицу поверхности зеркала со стороны падающего излучения.

(возможно, в термаче уже это решали, пока ничего не помню.)

#### O2-2.32 Давление на матовую поверхность (??)

Исходя из представления, что свет состоит из фотонов, каждый из которых обладает импульсом $h\nu/c$, определить давление $P$ световой волны на матовую поверхность, предполагая, что коэффициент отражения зеркала равен $r$, а угол падения равен $\varphi$. Определить также тангенциальную силу $T$, действующую на единицу поверхности зеркала со стороны падающего излучения. Решить в предположении, что поверхность, на которую падает световая волна, идеально матовая (удовлетворяет закону Ламберта).

(доработаю условие, удалю скорее всего r, t)

#### O2-2.33 Сила на площадку

На плоскую поверхность раздела вакуум-стекло (показатель преломления $n = 1,5$) из вакуума падает перпендикулярно к поверхности световой пучок интенсивностью $\mathcal{I} = 10\text{Вт}/\text{см}^2$. Найти величину и направление силы $F$, действующей на единицу площади поверхности раздела сред.

#### O2-2.34

Линейно поляризованная световая волна с направлением электрического вектора в плоскости падения и с интенсивностью $\mathscr{I} = 1\text{Вт}/\text{см}^2$ падает из вакуума под углом Брюстера на круглую плоскопараллельную стеклянную пластинку толщиной $d = 3\text{m}$ и диаметром $D = 10$ cm. Показатель преломления стекла $n = 1,5$. Найти момент сил $M$, действующий на пластинку. В какую сторону будет разворачиваться пластинка?

(??? она закреплена в центре???? я хз.)

#### O2-2.35

Показать, что давление излучения при нормальном падении света на идеальное зеркало равно $2u$, а на полностью поглощающую поверхность - $u$, где $u$ - плотность энергии падающего излучения.

#### O2-2.36 Давление света на идеальное зеркало (???)

Каково давление света на идеальное зеркало, если излучение изотропно?

#### O2-2.40

Определить силу светового давления $F_1$ солнечного излучения на поверхность земного шара, считая ее абсолютно черной. Найти отношение этой силы к силе гравитационного притяжения Солнца $F_2$. Средняя плотность Земли $\rho = 5,5$ г/см $^3$.

#### O2-2.41

Принимаs во внимание, что давление света на идеально отражающес зеркало при нормальном падении равно удвоенной плотности энергии падающей волны, найти закон изменения плотности свстовой энсргии при отражснии от медлснно движущсгося идеального зеркала.

### O2-2.42

Существуют лазеры, мощность излучения которых в импульсе составляет сотни мегаватт. Допустим, что мощность лазера равна $W = 500$МВ, а площадь поперечного сечения светового пучка $S = 1\ \text{см}^2$. Пусть луч сфокусирован идеальной линзой с фокусным расстоянием $f = 5$ см. Оценить напряженность электрического поля $E$ и давление света $P$ в фокусе такой линзы. Длина волны $\lambda = 6943$ Å

### O2-2.43

Лазер на рубине излучает в импульсе длительностью $\tau = 0,5$ мс энергию $\mathcal{E} = 1$ Дж в виде почти параллельного пучка с сечением $S = 1\ \text{см}^2$. Рабочая длина волны лазера $\lambda = 6943$ Å. Onределить следующие величины: 1) давление несфокусированного пучка света на площадку, перпендикулярную к пучку; 2) давление света на площадку, перпендикулярную к пучку, при максимально возможной концентрации светового пучка (при фокусировке в область с площадью поперечного сечения порядка $\lambda^2$); 3)напряженность электрического поля $E$ в области максимально возможной концентрации светового пучка.

Пр и мечан и е. Считать излучение лазера во время импульса равномерным.

## 15.1.2 Simple Problems about Nonlinear Optics

## 15.1.3 Problems about Radiation and Black Body

### O3.1-1. Излучение solutions

**1.**

1: Решение. Из полного числа фотонов $dn_\omega$ с частотами вблизи $\omega$, находящимися в $1\ \text{см}^3$ вакуумированной полости в пределах телесного угла $d\Omega = 2\pi \sin\theta d\theta$, к стенке будут лететь

$$dn_{\omega,\theta} = dn_\omega \frac{d\Omega}{4\pi} = \frac{1}{2} dn_\omega \sin\theta d\theta$$

фотонов. Нормальная компонента плотности их потока $dj_{\omega,\theta}\left[\text{ фотон}/(\text{см}^{\ 2} \cdot \text{с})\right]$ получается умножением $dn_{\omega,\theta}$ на нормальную компоненту скорости

$$dj_{\omega,\theta} = dn_{\omega,\theta} \cdot c \cos\theta$$

Если считать, что фотоны отражаются от стенки зеркально (это не сказывается на

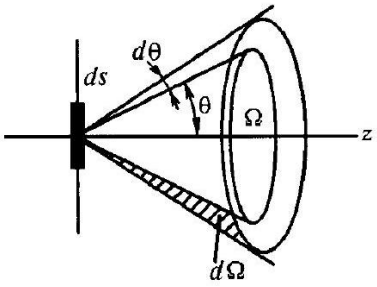



Рис. 199 общности результата), то каждый из них передает стенке импульс $2p_n = 2\frac{\hbar\omega}{c}\cos\theta$. Таким образом, фотоны создают давление

$$dP_{\omega,\theta} = 2p_n dj_{\omega,\theta} = d\rho_\omega \cos^2\theta \sin\theta d\theta,$$

где $d\rho_\omega = \hbar\omega dn_\omega$ - плотность энергии фотонов с частотами вблизи $\omega$. Давление, создаваемое всеми фотонами частотой $\omega$,

$$dP_\omega = \int_{(\theta)} dP_{\omega,\theta} = d\rho_\omega \int_0^{\pi/2} \cos^2\theta \sin\theta d\theta = \frac{1}{3} d\rho_\omega$$

Полное давление

$$P = \int_{(\omega)} dP_\omega = \frac{1}{3}\int_{(\omega)} d\rho_\omega = \frac{1}{3}\rho.$$

Малое отверстие в полости ведет себя как абсолютно черное тело, поскольку излучение, попавшее через это отверстие внутрь полости, в резуль тате многократного отражения от стенок практически полностью поглощается, так что коэффициент поглощения отверстия можно считать равным единице.

Излучение, покидающее полость через малое отверстие, по интенсивности и спектральному составу идентично излучению абсолютно черного

тела с температурой $T$. Вычислим, какую энергию в интервале частот $d\omega$ испускает в единицу времени полость через отверстие площадью $dS$ при плотности излучения $d\rho_\omega$ в полости. Энергия этого излучения, выходящего из отверстия под углом $\theta$ к нормали к площадке $dS$ в телесном угле $d\Omega = 2\pi \sin\theta d\theta$ (рис. 199), очевидно, равна

$$d\rho_\omega c dS \cos\theta \frac{d\Omega}{4\pi} = \frac{1}{2} d\rho_\omega c dS \cos\theta \sin\theta d\theta$$

Суммарное количество энергии в интервале частот $d\omega$, исходящее из отверстия в единицу времени, найдем интегрированием по полупространству

$$\frac{1}{2} d\rho_\omega c dS \int_0^{\pi/2} \cos\theta \sin\theta d\theta = \frac{1}{4} d\rho_\omega c dS$$

Интегрирование по всем частотам дает полное количество энергии, исходящее из отверстия в единицу времени: $\frac{\rho c}{4} dS$. С другой стороны, отверстие имеет ту же излучательную способность, что и абсолютно черная поверхность той же площади: $j dS$, где $j$ [эрг /$\text{см}^2 \cdot \text{с}$] - плотность потока энергии излучения. Таким образом, мы получаем

$$j(T) = \frac{\rho(T) c}{4} \qquad (*)$$

Иногда (в фотометрии, например) интегральную (по всем частотам) плотность потока энергии, излучаемой с единицы поверхности протяженного источника излучения в телесный угол $d\Omega$ под углом $\theta$ к нормали к поверхности, записывают так:

$$dj_\theta = B_\theta \cos\theta d\Omega$$

Коэффициент пропорциональности $B_\theta$ называют в этом случае яркостью протяженного источника в направлении угла $\theta$. Если $B_\theta$ не зависит от угла $\theta$, т. е. $B_\theta = B$, то такой источник называют ламбертовыл. Абсолютно черное тело - ламбертов источник излучения. Вычислим суммарную плотность потока энергии излучения $j$, посылаемую единицей поверхности абсолютно черного тела в телесный угол $2\pi$:

$$j = \int_0^{\pi/2} B \cos\theta \cdot 2\pi \sin\theta d\theta = \pi B \qquad (**)$$

Кроме того, эта же величина в соответствии с законом Стефана-Больцмана равна $\sigma T^4$, где $\sigma = 5,67 \cdot 10^{-5}$ эрг/(ссм $^2 \cdot$ К$^4$). Приравнивая (**) к полученному ранее выракению (*), получаем, что яркость

$$B = \frac{\rho c}{4\pi} = \frac{\sigma T^4}{\pi}$$

Плотность потока энергии, излучаемой на всех частотах с единицы площади абсолютно черного тела, имеющего температуру $T$, в телесный угол $2\pi$, называется энергетической светимостью (интегральной излучательной способностью)

$$R(T) = \frac{\rho(T) c}{4} = \sigma T^4$$

В соответствии с вышеприведенными рассуждениями излучательную способность абсолютно черного тела на частоте $\omega$ под углом $\theta$ к нормали, можно записать так

$$R_{\omega,\theta} = \frac{c}{4\pi} \rho_\omega \cos\theta$$

Это выражение называется законом Ламберта.

## 1.

2*: $\rho(T) = aT^4; s(T) = \frac{4a}{3} T^3$, где $a = \frac{4\sigma}{c}$.

Решение. Пусть равновесное тепловое излучение содержится в объеме $V$. Тогда его внутренняя энергия $U$ и энтропия $S$ соответственно равны

$$U(V,T) = \rho(T)V; S(V,T) = s(T)V \qquad (*)$$

Воспользуемся термодинамическими тождествами

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial P}{\partial T}\right)_V; T\left(\frac{\partial S}{\partial T}\right)_V = C_V = \left(\frac{\partial U}{\partial T}\right)_V$$

Подставляя в первое из них уравнение состояния излучения $P = \rho(T)/3$ и соотношение (*), находим, что

$$s(T) = \frac{1}{3}\frac{d\rho}{dT}. \qquad ()$$

Подставляя полученный результат и первое из соотношений (*) во второе гождество, находим уравнение для $\rho(T)$

$$\frac{d^2\rho}{dT^2} = \frac{3}{T}\frac{d\rho}{dT}$$

откуда $\rho(T) = aT^4$, а также

$$s(T) = \frac{1}{3}\frac{d\rho}{dT} = \frac{4}{3}aT^3.$$

Полученная постоянная $a$ в этих выражениях может быть выражена через постоянную Стефана-Больцмана $\sigma = 5,67 \cdot 10^{-5}$ эрг/ $\left(\text{с} \cdot \text{см}^2 \cdot \text{К}^4\right)$. Интегральная плотность потока теплового излучения $j$ [эрг/(см.с)] зависит от температуры по закону Стефана-Больцмана $j = \alpha T^4$. Кроме того, $j = c\rho/4$ (см. задачу 1.1), откуда постоянная $a = 4\sigma/c = 7,56 \cdot 10^{-15}$ эрг/$\left(\text{см}^3 \cdot \text{К}^4\right)$.

Заметим, что соотношение $j = c\rho/4$ легко получить, если рассматривать излучение как идеальный газ фотонов, движущихся со скоростью $c$. Как известно из молекулярно-кинетической теории газов, плотность потока частиц, т. е. число частиц $z$, движущихся со средней тепловой скоростью $\bar{v}$ и пересекающих за 1 с единицу площади поверхности стенки сосуда, равно $z = n\overline{v}//4$. В этой формуле $n$ - концентрация молекул газа, в нашем случае - это число фотонов в единице объема.

**1.**

3. $S_2 - S_1 = \frac{16\sigma}{3c}T^3\,(V_2 - V_1)$.

**1.**

4. $A = \frac{4\sigma}{3c}; c_V = \frac{16\sigma}{3c}T^3; P = \frac{4\sigma}{3c}T^4; c_V = 8{c_V}^{\mu,2}; \Phi = 0$.

**1.**

5. $A = \frac{16\sigma}{3c}T_1^3\,(T_1 - T_2)\,(V_2 - V_1)$.

**1.**

6. $P = \frac{32\sigma}{3c}T^4 = 1,26 \cdot 10^{-3}$ дин /с$^2$ $= 1,26 \cdot 10^{-4}\Pi \approx 10^{-6}$ мм рт. ст.

**1.**

7. $C_P = \infty; VT^3 = \text{const}$ или $PV^{4/3} = \text{const}$.

**1.**

8. $T = \left(\frac{3c}{4\sigma}P\right)^{1/4} \approx 1,4 \cdot 10^5$ К.

**1.**

9. $n = \frac{4\sigma}{3ck_{\text{E}}}T^3 \approx 1,8 \cdot 10^7 \text{ см}^{-3}$.

**1.**

10. $\frac{C_V^{\text{цд}}}{C_V^{\text{из}}} = \frac{1}{8}$.

**1.**

11: $n = \frac{16\sigma}{3ck_{\text{F}}}T^3 = 7,3 \cdot 10^{10} \text{ см}^{-3}, T_{\text{r}} = \frac{T}{4} = 250 \text{ К}$.

Решение. Температуры в обеих частях сосуда разные. Запишем условие механического равновесия поршня

$$P_1 = nk_{\mathbf{5}}T_1 = \frac{4\sigma}{3c}T_2^4 = P_2 \qquad (*)$$

где индекс 1 относится к газу, а 2 - к излучению. Естественно предположить, что $T_1 \ll T_2$, так что давлением излучения при $T_1$ можно пренебречь. По условию задачи при малых изменениях температуры в обеих частях со суда на одну и ту же величину направление смещения поршня не зависит от знака этой величины. Это означает, что в зависимости объема одной из частей сосуда от температуры имеется экстремум (минимум, если система устойчива), в окрестности которого $\partial V/\partial T = 0$. Так как $N = nV = \text{const}$, то в линейном приближении по отклонениям от равновесия $n\frac{\partial V}{\partial T} + V\frac{\partial n}{\partial T} = 0$, и, следовательно, $\partial n/\partial T = 0$. Тогда из соотношения (*) следует

$$nk_{\mathbf{6}}dT_1 = \frac{16}{3c}\sigma T_2^3 dT_2$$

По условию задачи $dT_1 = dT_2$, откуда

$$n = \frac{16\sigma T_2^3}{3ck_{\text{B}}} = 7,3 \cdot 10^{10} \text{ см}^{-3}$$

Начальная температура газа $T_1 = T_2/4 = 250$ К. Поэтому отношение давлений излучения $\frac{P_1^{\text{изл}}}{P_2^{\text{изл}}} = \left(\frac{T_1}{T_2}\right)^4 = \frac{1}{256} \ll 1$, что оправдывает сделанное предположение о малости температуры газа в сравнении с температурой излучения.

**1.**

12*

$$T^4(r) = T_0^4 \frac{R^2 h}{2\left(h^2 + r^2\right)^{3/2}}$$

Решение. Поток энергии, излученный элементом поверхности $dS_1$, и падающий на элемент другой поверхности $dS_2$, определяется соотношением

$$d^2\Phi = \frac{B dS_1 ds_2 \cos\theta_1 \cos\theta_2}{R^2}$$

где в соответствии с рис. 200$\mathbf{n}_1$ и $\mathbf{n}_2$ - единичные векторы нормалей к соответствующим площадкам, $R$ - расстояние между площадками, $B$ - коэффициент пропорциональности, называемый поверхностной яркостью (см. задачу 1.1), вообще говоря, зависящий от угла $\theta_1$ (а также азимутального угла $\varphi_1$, не изображенного на рисунке). Для абсолютно черного тела $B = \text{const}$ (закон Ламберта) и $B = \frac{\sigma T^4}{\pi}$. Если ввести телесный угол $d\Omega_2 = \frac{dS_2 \cos\theta_2}{R^2}$, под которым видна площадка $dS_2$ из центра первой площадки, то

$$d^2\Phi = \frac{\sigma T^4}{\pi} dS_1 \cos\theta_1 d\Omega_2.$$

На рис. 201 изображен шар - источник равновесного излучения, а под ним - за-

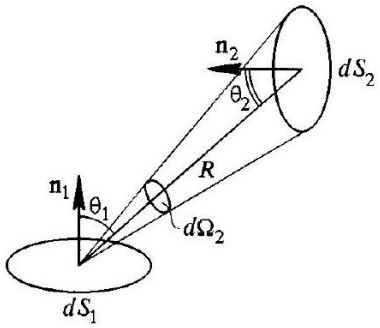



Рис. 200 черненная с обеих сторон поверхность, где и надо найти распределение температур. Из соображений симметрии ясно, что $T = T(r)$. При условии $h \gg a$ телесный угол, под которым из центра нара видно кольцо радиусом $r$ на плоскости

$$d\Omega_2 \simeq \frac{2\pi r dr \cos\theta_2}{R^2} = \frac{2\pi r dr}{h^2 r^2} \frac{h}{\sqrt{h^2 + r^2}}.$$

Кроме того, условие $h \gg a$ позволяет считать, что $\theta_1 \approx \varphi$, где $\varphi$ - угол между нормалью к $dS_1$ (т. е. радиусом) и вертикалью. Поэтому $dS_1 \cos\theta_1 \approx\approx dS_1 \cos\varphi = dS_1'$, где $dS_1'$ - проекция $dS_1$ на экваториальную плоскость шара. Суммирование (интегрирование) по всем элементам полусферы дает л $a^2$. Поэтому поток излученной шаром энергии и поглощенный кольцом

$$d\Phi = \pi a^2 \frac{\sigma T_0^4}{\pi} \frac{2\pi r dr h}{(h^2 + r^2)^{3/2}}.$$

В равновесии этот поток переизлучается кольцом по обе стороны плоскости:

$$d\Phi = 2\pi r dr \cdot 2\sigma T^4(r).$$

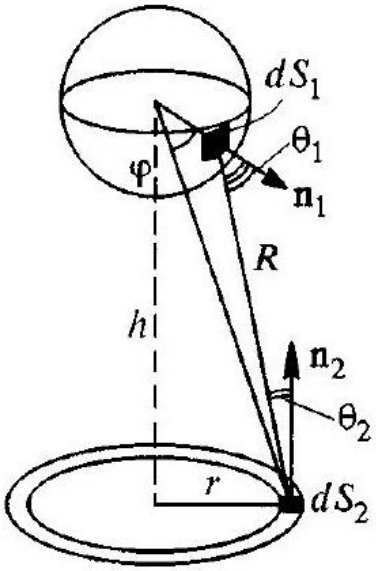



Рис. 201
Из равенства этих потоков и следует ответ

$$T^4(r) = T_0^4 \frac{a^2 h}{2\,(h^2 + r^2)^{3/2}}$$

## 1.

13. $T^4(r) = T_0^4 \frac{Rh}{2(h^2+r^2)}$, где $r$ - расстояние на плоскости до проекции оси цилиндра.

**1.**

14. $T^4(r) = T_0^4 \frac{a^2h^2}{2(h^2+r^2)^2}$.

**1.**

15. $T_2 = T_1\sqrt{\frac{r}{R}}\sqrt[4]{\cos\theta_1\cos\theta_2}$; $\quad\frac{T_2(\theta=0)}{T_2(\theta=\pi/4)} = 2^{1/8} \approx 1,09$.

**1.**

16. $T_2 = T_1\sqrt[4]{\frac{1}{2}\left(1 - \frac{l}{\sqrt{l^2+D^2/4}}\right)} \approx 1125$ K.

**1.**

17. $A \approx 1 - \frac{\rho d^2}{4D^2}$; $\frac{d}{D} \leqslant \frac{1}{16}$.

**1.**

18* $T = \frac{T_\text{C}}{2}\sqrt{\frac{D}{F}} = 1500$ K.

Решение. Вычислим величину телесного угла $\Omega$, под которым была бы видна линза из центра Солнца. Пусть $L$ - расстояние от Солнца до Земли, а $D$ - диаметр линзы. Угол $\alpha$, под которым из центра Солнца виден радиус линзы, равен $\alpha = D/2L$, отсюда

$$\Omega = 2\pi(1-\cos\alpha) = 4\pi\sin^2\frac{\alpha}{2} \approx \pi\alpha^2 = \pi\left(\frac{D}{2I}\right)^2.$$

Плотность потока энергии, излучаемой с поверхности Солнца, $j = \sigma T_\text{C}^4$. Поток энергии, излучаемой с поверхности Солнца в единицу телесного угла, равен $\frac{\sigma T_\text{C}^4 4\pi R_\text{C}^2}{4\pi}$. Мощность солнечного излучения, падающего на линзу,

$$N_\text{C} = \sigma T_\text{C}^4 R_\text{C}^2 \pi\left(\frac{D}{2L}\right)^2$$

Нагреваясь, шарик, излучаег. Когда поток энергии, полученной от Солнца, сравняется с потоком излученной энергии, установится равновесная температура шарика. Поток энергии с поверхности шарика диаметром $d$

$$N_\text{m} = \sigma T^4 4\pi\frac{d^2}{4}.$$

По условию размер изображения Солнца $\alpha_C F$ равен диаметру $d$ шарика, где $\alpha_\text{C} = \frac{2R_C^2}{L}$ - угловой размер Солнца, а $F$ - фокусное расстояние линзы. Таким образом,

$$N_\text{m} = \sigma T^4\pi d^2 = \sigma T^4\pi\alpha_\text{C}^2F^2 = \sigma T^4\pi F^2\frac{4R_\text{C}^2}{L^2}.$$

Приравнивая потоки энергии $N_\text{C}$ и $N_\text{ші}$ , получаем

$$\sigma T_\text{C}^4 R_\text{C}^2\pi\frac{D^2}{4L^2} = \sigma T^4F^2\pi\frac{4R_\text{C}^2}{L^2},$$

огкуда и находим искомую температуру шарика

$$T = \frac{T_\text{C}}{2}\sqrt{\frac{D}{F}} = 1500 \text{ K}.$$

**1.**

19. $T = T_0 \left(1 + \frac{j_C D^2}{4\sigma T_0^4 d^2}\right)^{1/4} \approx 1980$ K.

**1.**

20. $\frac{j_{кор}}{j_{\cdot g}} = \frac{n(m_p + m_e)v^3}{2\alpha_C^2 \sigma T^4} \approx 1,5 \cdot 10^{-8}$.

**1.**

21. $T_{ра} = \sqrt[4]{\frac{J_C}{\sigma}\left(\frac{L}{R_C}\right)^2} \approx 5850$ K.

**1.**

22. $D = \frac{2L}{T^2}\sqrt{\frac{j}{\sigma}} = 1,55 \cdot 10^6$с $= 15,5$ км, где $T \approx \frac{0,29}{\lambda_{max}[см]} = 1,45 \cdot 10^7$ K.

**1.**

23. $T_2 = \left(T_1^4 + T_{cT2}^4 - T_{cT1}^4\right)^{1/4} \approx 340$ K.

**1.**

24. $T = T_C\sqrt{\frac{R\alpha_C}{2r}} = 2120$ K.

**1.**

25. $T = T_C\sqrt{\frac{R_C}{2L}}$, где $L$ - расстояние от Солнца до соответствующей планеты; $T_B \approx 340$ K; $T_З \approx 290$ K; $T_M \approx 230$ K; $T \approx 130$ K.

Перегрев поверхности Венеры возникает вследствие парникового эффекта. Атмосфера Венеры толще и плотнее земной, в ее составе значительную долю составляет $CO_2$. Атмосфера пропускает солнечный свет с $\lambda \simeq \simeq 4800$Å(6000 K), но не пропускает $\lambda \simeq 10$ мкм (300 K).

**1.**

26. $T_C = 2T_З\alpha_C^{-1/2} \approx 6000$ K.

**1.**

27. $\frac{dT}{dt} = \frac{W - (1-\varepsilon)S\sigma T^4}{Mc} \simeq 0,25 \cdot 10^{-3}$ K/с $\simeq 0,96$ K/ ч, где принято: масса человека $M = 70$ кг, поверхность человеческого гела $S \sim 1$ m$^2$, удельная теплоемкость тела человека $c = 4,2$ Дж/(гК) (удельная теплоемкость воды). Можно также оценить максимальную длительность пребывания человека в открытом космосе (без принудительной вентиляции). Если прииять, что максимально возможная температура для человеческого тела $T_{max} = 315$ K (42°C), то тогда $\Delta T_{max} = 5,4$ K. Полагая найденную $\frac{dT}{dt} = \text{const}$, получим $\tau \simeq 6$ ч. Учет нелинсйности производной только увели чит $\tau$ $\left(\frac{dT}{dt} \to 0\right.$ с ростом $T$ и поэтому значение вырастет).

**1.**

28. 1) $L = \frac{R_\text{C}}{2}\frac{T_\text{C}^2}{T^2} = (0,905 \div 2,53)\cdot 10^8$КМ.
2) $\frac{R_\text{C}^2}{L^2} = 4\left(\frac{T}{T_\text{C}}\right)^4 - \frac{(1-\eta)N}{\pi r^2 \sigma T_\text{C}^4 (1-\alpha)}$.

В случае 1) ответ не зависит от отражательной способности $\alpha$; в случае 2) диапазон $\alpha$ определяется диапазоном температур и условием $(R_\text{C}/L)^2 > 0$ (правая часть положительна). Отражательная способность $\alpha$ должна лежать в преде лах $0,773 < \alpha < 0,971$.

Отсюда следует, что при $\alpha \geqq 0,971$ станция не может работать в непрерывном режиме на любых расстояниях от Солнца (и без Солнца будет перегрев); при $\alpha \leqslant 0,773$ получаем $L \geqq 0,96\cdot 10^8$ км, т. е. станция может работать на сколь угодно большом расстоянии от Солнца; при $\alpha = 0$ $L \approx (0,93 \div 2,9)\cdot 10^8$КМ.

**1.**

29. $u_\text{max} = \frac{2RT}{cA} \approx 0,1$ см /с, где $A = 56$ г/моль - атомная масса железа.

Решение. Согласно решению задачи 1.1 плотность потока энергии, излученной с поверхности пластинки под углом $\theta$ к нормали в телесный угол $d\Omega = 2\pi \sin\theta d\theta$, равна

$$dj = \frac{\rho c}{4\pi}\cos\theta d\Omega$$

Так как пластинка излучает лишь в полупространство, то она получает составляющую плотности потока импульса в направлении угла $\theta$:

$$dp = \frac{dj}{c} = \frac{\rho}{4\pi}\cos\theta d\Omega \left[\frac{\text{г}\cdot\ \text{см/с}}{\text{см}^2\cdot\text{с}}\right]$$

Понятно, что только проекция составляющей импульса $dp_z$ дает движение пластинки вдоль по оси $z$ (рис. 202)

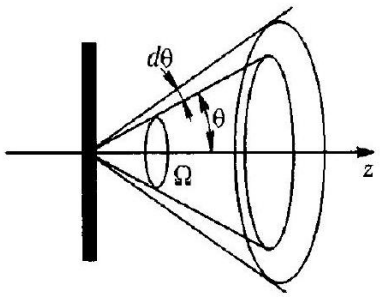



Рис. 202

$dp_z = \frac{\rho}{4\pi}\cos^2\theta d\Omega = \frac{\rho}{4\pi}\cos^2\theta\cdot 2\pi\sin\theta d\theta$.

Полная плотность потока импульса, которая в силу симметрии направлена нормально к пластинке, равна

$$p_z = \frac{\rho}{2}\int_0^{\pi/2}\cos^2\theta\sin\theta d\theta = \frac{\rho}{6} = \frac{2}{3}\frac{\sigma}{c}T^4$$

Найденная величина - это нормальная компонента импульса с единицы площади в единицу времени, т. е. нормальное давление. Полный импульс, полученный пластинкой, находится интегрированием нормального давления по площади пластины $S$ и по времени излучения $t$. В соответствии с законом сохранения энергии он равен

$$p_\text{полн} = \int p_z dSdt = \frac{2}{3c}\int \sigma T^4 dSdt = \frac{2}{3c}\mathscr{E}$$

где $\mathscr{E}$ - полная внутренняя энергия пластинки, переходящая в излучение. В соответствии с законом Дюлонга-Пти $\mathcal{E} = 3k_\mathbf{5}TN$, где $N = \frac{M}{m} = N_\text{A}\frac{M}{A}$ - число атомов железа в пластинке, $M$ - ее масса, $A$ - атомная масса. Использованное для & приближение является довольно грубым, т. к. не учитывает уменьшения теплоемкости при остывании пластины ниже температуры Дебая (для железа примерно 500 К). Тем не менее его можно использовать для оцен ки максимальной скорости, развиваемой пластинкой. Таким образом,

$$v_\text{max} = \frac{p_\text{полн}}{M} = \frac{2RT}{cA} \approx 0,1\ \text{см/с}$$

**1.**

30. $\Delta T = T_3 \frac{W}{4W} \approx 7 \cdot 10^{-3}$ K, где $T_3$ - средняя температура поверхности Земли. $W_{\max} \approx 1,5 \cdot 10^{15}\text{Bt} = 150W$.

**1.**

31. $T = \sqrt[4]{\frac{L_\text{C}}{16\pi R^2 \text{O}}} = 278$ K; $\Delta R = -2R\frac{\Delta T}{T} \sim -10^9$M.

Истинная орбита Земли эллиптическая: $R_{\max} \approx 152$ млн км, а $R_{\min} \approx \approx 148$ млн км. Полученная оценка допустимых отклонений по порядку величины совпадает с этими отклонениями. Следовательно, реальная оценка допустимого температурного перегрева Земли несколько занижена.

**1.**

32. $T = \left(\frac{\mathscr{E}_a \lambda_a N}{\sigma S}\right)^{1/4} \approx 11,1$ K, где $N = N_\text{A}\frac{m}{A}$ - число ядер в кусочке урана, $\lambda_\alpha = \frac{\ln 2}{T^a_{1/2}}$ - постоянная распада, $S = 4\pi\left(\frac{3m}{4\pi\rho}\right)^{2/3}$ - площадь поверхности кусочка. Вклад осколков деления в разогрев шарика несущественен в силу большой разницы их парциальных времен полураспада.

**1.**

33. Теплообмен падает в $N+1$ раз,

$$T_n^4 = \frac{(N+1-n)T_\text{г}^4 + nT_\text{х}^4}{N+1}, n = 1, 2, \ldots N.$$

**1.**

34* 1) $T_{A2} = 950$ K; 2) $T_{A2} = 1230$ K.

Решение. 1) В первом случае ток не изменится. При повышении напряжения вдвое мощность тока увеличится также вдвое. В стационарном состоянии излучаемая мощность возрастает в $\sqrt[4]{2}$ раз и будет равной $\simeq 950$ K. Аналитически это записывается следующим образом

$$W_\text{A} = \mathcal{I}_\text{A} V_\text{A}; \quad _\text{A} = \text{ const }; \quad \frac{W_\text{A1}}{W_\text{A2}} = \frac{\mathcal{I}_\text{A} V_\text{A1}}{\mathscr{I}_\text{A} V_\text{A2}} = \frac{\sigma T_\text{A1}^4}{\sigma T_\text{A2}^4},$$

откуда $T_\text{A2} = T_\text{A1}\left(\frac{V_\text{A2}}{V_\text{A1}}\right)^{1/4} \simeq 950$ K.

2) Во втором случае ток возрастет в $2^{3/2}$ раз, а мощность в $2^{5/2}$ раз. При этом температура анода повысится в $2^{5/8}$ раз и достигнет величины $\simeq 1230$ K. Запишем это:

$$T_\text{A2} = T_\text{A1}\left(\frac{W_\text{A2}}{W_\text{A1}}\right)^{1/4} = T_\text{A1}\left(\frac{\mathcal{I}_\text{A2} V_\text{A2}}{\mathcal{I}_\text{A1} V_\text{A1}}\right)^{1/4} = T_\text{A1}\left(\frac{V_\text{A2}}{V_\text{A1}}\right)^{5/8} \approx 1230 \text{ K}.$$

**1.**

35. $T_\text{A2} = T_{A1}\left(\frac{W_2}{W_1}\right)^{1/4} = 1083$ K,

где $W_1 = \mathcal{I}_\text{Al} V_\text{A1} + W_\text{накл1} \approx \frac{\mathscr{E}}{R_\text{H}} V_\text{Al} + W_\text{накл1} = 200$ В, $W_2 = \mathcal{I}_\text{A2} V_\text{A2} + W_\text{нак 2} = \mathcal{I}_\text{A2}\left(\mathscr{E} - \mathcal{I}_\text{A2} R_\text{H}\right) + W_\text{нак 2} = 2550$ Вт.

**1.**

36. $\frac{\rho_\text{изл}}{\rho_\text{пок}} = \frac{4\text{o}T^4}{c^3 m_\text{p} n} \sim 10^{-3}$.

**1.**

37. $\frac{\rho_{\text{изл}}}{\mathbb{E}_{\text{маг}}} = \frac{32\pi\sigma T^4}{cB^2} \approx 3,8.$

**1.**

38. $\frac{\Delta E}{E} = \frac{\Delta j}{j} \approx \frac{2}{5}\frac{\hbar\omega}{k_{\mathbf{E}}T}\frac{\Delta V}{V} \approx 0,38(38\%)$ - с учетом зависимости сопротивления нити от температуры, $\frac{\Delta E}{E} \approx \frac{1}{2}\frac{\hbar\omega}{k_{\text{B}}T}\frac{\Delta V}{V} \approx 0,48$ (48%) - без учета этой зависимости.

**1.**

39*. $\frac{\Delta B}{B} \approx \frac{\Delta\rho}{\rho} = \frac{\hbar\omega}{k_{\text{E}}T_0}\frac{\Delta T}{T_0} \approx 0,0256(2,56\%),$
$\Delta T = \frac{5}{16}\frac{n}{\sigma T^2}\sqrt{\frac{k_{\text{F}}T_1}{\pi m_{\text{N}}}} \approx 5,6$ K
Решение. При данной температуре диссоциация азота не происходит, т.к. 9,74 эВ $\gg 0,216$ эВ ($\kappa_{\mathbf{6}}T_0$). Концентрация молекул азота $n = \frac{P}{k_{\text{D}}T} = 1,93\cdot 10^{15}$ cm$^{-3}$. Длина свободного пробега молекул азота $\Lambda = \frac{1}{n\sigma_{\text{N}}} \simeq 2,8$ cm $> 2$ cm - размера колбы лампы. Таким образом, движение молекул внутри колбы - бесстолкновительное. Для оценки будем считать, что всю энергию, полученную от спирали, молекула теряет при столкновении с холодной стенкой, которая имеет температуру $T_1 \simeq 300$ K. Число ударов в единицу времени о единицу поверхности стенки $z = \frac{n\bar{v}}{4}$, где $\bar{v} = \sqrt{\frac{8RT}{\pi\mu}}$ на среднюю энергию одной молекулы $\overline{\varrho} = c_V T$ е. интенсивность потери энергии $N_{\text{пот}} = z\overline{\mathcal{I}}$), а вращательные возбуждены, т. к.

$$\frac{\hbar^2}{2I} = \frac{\hbar^2}{m_{\text{N}}d^2} = 5\cdot 10^{-4}9 \text{ В} \ll 0,216 \text{ эВ } (k_{\text{B}}T_0)$$

(здесь $m_{\text{N}}$ - масса атома азота). Теплоемкость, приходящаяся на одну молекулу, $c_V = c_{\text{пост}} + c_{\text{вр}} = \frac{5}{2}k_{\mathbf{6}}$, т. е. $\overline{\mathcal{E}} = \frac{5}{2}k_{\mathbf{6}}T_0$. Таким образом, интенсивность потерь энергии

$$N_{\text{пот}} = \frac{5}{8}n\overline{n}\overline{\pi}k_{\mathbf{B}}T_0.$$

Уравнение теплового баланса

$$\sigma\left(T_0 - \Delta T\right)^4 + N_{\text{пот}} = \sigma T_0^4.$$

Считая $\Delta T \ll T_0$, получаем

$$\Delta T \simeq \frac{N_{\text{пот}}}{4\text{o}T_0^3} = \frac{5}{16}\frac{n}{\sigma T_0^2}\sqrt{\frac{k_{\text{L}}T_1}{\pi m_{\text{N}}}} \approx 5,6 \text{ K}.$$

При $\lambda = 5000$Å отношение $\frac{\hbar\omega}{k_{\text{E}}T_0} = \frac{hc}{\lambda k_{\text{b}}T_0} = 11,5$, поэтому $e^{\hbar\omega/k_{\text{B}}T_0 \gg 1}$ и $\rho(\omega) \propto \omega^3 e^{\hbar\omega/k_{\text{E}}T_0}$. Отсюда следует, что при изменении температуры на $\Delta T$

$$\frac{\Delta\rho}{\rho} = \frac{\hbar\omega}{k_{\text{B}}T_0}\frac{\Delta T}{T_0} \approx 0,0256(2,56\%).$$

Поскольку излучательная способность (яркость лампы) однозначно определяется спектральной плотностью излучения, то искомое отношение равно $\frac{\Delta B}{B} = 0,0256$.

**1.**

40. $T_2 \approx 80$ K.
Решение. Частота, соответствующая граничной длине волны $\lambda_0 = 5$ мкм, равна $\omega_0 = 2\pi/\lambda_0 = 3.8\cdot 10^{14}\text{с}^{-1}$. Поскольку пластинка облуча-
ется потоком солнечного излучения, то необходимо вычислить частоту, на которой спектральная плотность солнечного излучения $\rho_T(\omega)$ максимальна. Если считать Солнце абсолютно черным телом, то в соответствии с законом смещения Вина эта частота

$$\omega^* \approx 2,8\frac{k_{\text{b}}T_{\text{C}}}{\hbar} \approx 23\cdot 10^{14}\text{с}^{-1}$$

откуда видно, что $\omega_0 \ll \omega^*$.
На рис. 203 показана спектральная плотность излучения Солнца $\rho_T(\omega)$. Из рисунка вид-

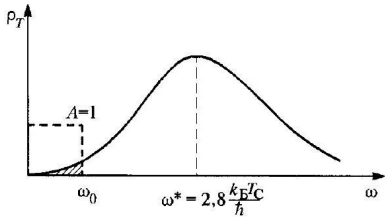



Рис. 203 но, что кварцевая пластинка поглощает лишь только малую часть излучения солнечного спектра (область, обозначенная штриховкой). Поток энергии, приходящий на кварцевую пластинку площадью $S$ от Солнца, равен

$$\Phi = jS = S\frac{c}{4}\int_0^{\omega_0} \frac{\hbar\omega^3}{\pi^2 c^3}\frac{d\omega}{e^{\hbar\omega/k_{\text{B}}T_C - 1}} \approx S\frac{c}{4}\int_0^{\omega_0} \frac{\omega^2 k_{\text{D}} T_{\text{C}} d\omega}{\pi^2 c^3} =$$
$$= S\frac{\omega_0^3 k_{\text{E}} T_{\text{C}}}{12\pi^2 c^2} \approx 4,3\cdot 10^8 \cdot S\left[\frac{\ni\ \text{pr}}{c}\right].$$

Рассчитаем теперь поток энергии, получаемый абсолютно черной пластинкой (в тех же условиях)

$$\Phi_{\text{черн}} = S\sigma T_{\text{C}}^4 \approx 734,8\cdot 10^8 \cdot S[\text{ эрг /с}],$$

т. е. поток энергии, поглощаемый абсолютно черной пластинкой, больше такового для кварцевой пластинки в $\Phi_{\text{черн}}/\Phi \approx 170$ раз.
Черная пластинка имеет равновес-

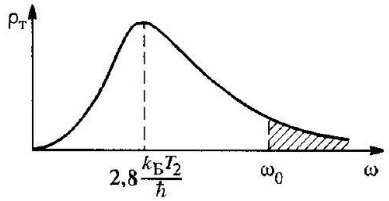



Рис. 204 ную температуру $T_1 = 300$ К. В соответствии с законом Стефана-Больцмана поток излучаемой ею энергии (с двух сторон)

$$\Phi'_{\text{черн}} = 2S\sigma T_1^4 = 4,59\cdot 10^5 \cdot 2S[\text{ эрг /с}].$$

Кварцевая пластинка, находящаяся в тех же условиях, излучает поток

$$\Phi' = 2S\left[\sigma T_2^4 - \frac{c}{4}\int_{\omega_0}^{\infty} \rho(\omega)d\omega\right] \qquad (*)$$

На рис. 204 показана спектральная плотность равновесного теплового из лучения кварцевой пластинки при температуре $T_2$. Штриховкой показана область частот, в которой пластинка не излучает. Предполагаем, что $\omega_0 \gg 2,8k_{\text{B}}T_2/\hbar$. В этом случае интегралом $\int^{\infty}\rho_T(\omega)d\omega$ в формуле (*) можно пренебречь. Условием теплового равновесия является равенство поглощаемого и излучаемого потоков энергии, т. е. $\Phi = \Phi'$ и $\Phi_{\text{черн}} = \Phi'_{\text{черн.}}$. Поэтому

$$\Phi' \approx 2S\frac{c}{4}\int_0^{\omega_0}\rho_T(\omega)d\omega \approx \frac{1}{170}\Phi'_{\text{черн}}$$

откуда $T_2^4 = \frac{T_1^4}{170}$ или $T_2 = \frac{T_1}{\sqrt[4]{170}} \approx 80$ К.
В заключение убедимся, что $\omega_0 \gg 2,8k_{\mathbf{6}}T_2/\hbar$. В самом деле $2,8k_{\mathbf{5}}T_2/\hbar \approx$ $\approx 0,3\cdot 10^{14}\text{с}^{-1} \ll 3,8\cdot 10^{14}\text{с}^{-1} = \omega_0$. Таким образом, наше приближение оправдано, и задача решена.

**1.**

41. $\Phi \approx S\frac{15}{\pi^4}\sigma T^4 x_0^3 e^{-x_0} \approx 8,1\cdot 10^{-6}$Вт, где $x_0 = \frac{\hbar\omega_0}{k_{\text{B}}T} \approx 16,2$, $S$ - площадь окошка.

**1.**

42. $x = \frac{5}{\pi^4}\left(\frac{\mathscr{E}}{k_{\text{E}}T}\right)^3 \approx 0,003$.

**1.**

43. $x = \frac{15}{\pi^4}\left(\frac{\mathscr{E}}{k_{\text{B}}T}\right)^3 e^{-\frac{8}{k_{\text{B}}T}} \approx 1,63 \cdot 10^{-7}$.

**1.**

44. $T = \frac{\pi^4}{5}\frac{k_{\text{E}}^3 T^{*4}}{\hbar^3\omega_0^3} = \frac{\pi^4}{5}\frac{T^{*4}}{\theta^3} \approx 5840$ K.

**1.**

45. $T = \frac{\pi^4}{5}\left(\frac{k_{\text{E}}T^*}{\hbar\omega_1}\right)^3 T^* \approx 370$ K.

**1.**

46. $d = \frac{1}{\varkappa}\ln\frac{I_0 - I_\lambda}{I_0} = 23$ cm.

**1.**

47. $\left\langle \Delta N_\lambda^2 \right\rangle \approx \langle N_\lambda \rangle \approx \frac{\tau\alpha_c^2 S\pi c}{2\lambda^3}\frac{\Delta\lambda}{\lambda}\exp\left(-\frac{hc}{k_{\text{B}}T\lambda}\right) = 5,9 \cdot 10^9$. Относительная флуктуация $\frac{\sqrt{\left\langle \Delta N_\lambda^2 \right\rangle}}{\langle N_\lambda \rangle} = 1,3 \cdot 10^{-5}$.

Заметим, что в задаче требуется определить среднеквадратичную флуктуацию сигнала, т. е. энергии, принимаемой приемником. Покажем, что при условии $\hbar\omega \gg k_{\mathbf{5}}T$ это будет флуктуация числа фотонов. Строгая термодинамическая формула для флуктуации энергии имеет вид

$$\left\langle \Delta\mathscr{E}^2 \right\rangle = k_{\mathbf{5}}T^2\frac{d\langle\mathscr{E}\rangle}{dT}$$

В интервале частот от $\omega$ до $\omega + d\omega$
Соответственно

$$\langle\mathcal{E}\rangle \equiv V\rho_T(\omega)d\omega = \frac{V}{\pi^2c^3}\frac{\hbar\omega^3}{\exp\left(\frac{\hbar\omega}{k_{\text{E}}T}\right) - 1}d\omega$$

$$\left\langle \Delta\mathscr{E}^2 \right\rangle = \frac{V}{\pi^2c^3}\frac{\hbar^2\omega^4\exp\left(\frac{\hbar\omega}{k_{\text{B}}T}\right)}{\left[\exp\left(\frac{\hbar\omega}{k_{\text{B}}T}\right) - 1\right]^2}d\omega$$

При $\hbar\omega/\left(k_{\mathbf{5}}T\right) \gg 1$ (область Вина)

$$\left\langle \Delta\mathfrak{E}^2 \right\rangle \approx \frac{V}{\pi^2c^3}\hbar^2\omega^4\exp\left(-\frac{\hbar\omega}{k_{\text{E}}T}\right)d\omega = \hbar\omega\langle\mathfrak{E}\rangle$$

Сравнивая с результатом для флуктуации числа частиц идеального газа $\left\langle \Delta n^2 \right\rangle = \langle n \rangle$, мы видим, что в этой области справедливо представление о фотонах как газе независимых частиц

$$\langle\mathfrak{E}\rangle = \langle n\rangle\hbar\omega$$

Напротив, в области Рэлея-Джинса справедливы волновые представления о свете и

$$\left\langle \Delta\mathscr{E}^2 \right\rangle = k_{\mathbf{5}}T\langle\mathscr{E}\rangle$$

**1.**

48. $T = \frac{2\pi\hbar c}{1{,}62k_{\text{E}}\lambda_2} = 2,96$ K; при $\lambda_3 = 30$ cm сигнал уменьшится в 1000 раз.

**1.**

49. В обоих случаях $T = \frac{2\varepsilon\lambda^2}{k_\text{Б}\ \pi^2\tau\Delta\lambda} = 4,7 \cdot 10^{17}$ К.

Решение. Энергия, уносимая излучением за время $dt$ с площадки $dS$ под углом $\theta$ к ней в телесный угол $d\Omega$

$$d\mathscr{E} = I_{\omega,\theta} d\omega d\Omega dS dt$$

где $I_{\omega,\theta} = \frac{cp_w}{4\pi}\cos\theta$ - излучательная способность АЧТ на частоте $\omega$ под углом $\theta_{\text{к нормали. Согласно распределению Планка}}$

$$\rho_\omega = \frac{\hbar}{\pi^2 c^3}\frac{\omega^3}{e^{\hbar\omega/k_\text{Б}T} - 1}$$

Если рассматривать излучение вдоль направления нормали к площадке, то $\theta \approx 0; \cos\theta \approx 1$.

Переходя к конечным приращениям, запишем

$$\Delta\mathscr{E} = \frac{c\rho_\omega}{4\pi}\Delta\omega\Delta\Omega\Delta S\Delta t, \quad \Delta\mathscr{E} = \mathscr{E}, \quad \Delta t = \tau \qquad (*)$$

Случай $a$) : В отсутствие фокусировки $\Delta S = S = \frac{\pi D^2}{4}$, $\quad \Delta\Omega = 2\pi(1 - -\cos\varphi) = 4\pi\sin^2\frac{\varphi}{2}$, где $\varphi$ - угол дифракции на выходном отверстии. Полагая, что $\sin\varphi \approx \varphi \approx \frac{\lambda}{D}$, получим $\Delta\Omega \approx \frac{\pi^2\lambda^2}{4S}$. В свою очередь,

$$\Delta\omega = \left|\Delta\left(\frac{2\pi c}{\lambda}\right)\right| = \frac{2\pi c}{\lambda^2}\Delta\lambda$$

После подстановки найденных величин в (*) получаем

$$\mathscr{E} = \frac{\pi^3\hbar c^2\Delta\lambda\tau}{\lambda^3\left[\exp\left(\frac{\hbar\omega}{k_\text{Е}T_{j\phi\phi}}\right) - 1\right]}.$$

Вычислим выражение в скобках:

$$\exp\left(\frac{\hbar\omega}{k_5 T_{3\ \text{d}\phi}}\right) - 1 = \frac{\pi^3\hbar c^2\Delta\lambda\tau}{\varrho\lambda^3} \approx 0,4 \cdot 10^{-15} \ll 1$$

Полученный результат означает, что $\frac{\hbar\omega}{k_\Sigma T_{\text{эфф}}} \ll 1$. Тогда

$$T = \frac{2\delta\lambda^2}{\pi^2 c\tau\Delta\lambda k_\text{В}} = 4,7 \cdot 10^{17}\ \text{К}$$

Заметим, что полученный ответ - чисто «классический», т. к. постоянная Планка в нем отсутствует. Это означает, что можно было воспользоваться для вычисления $\rho_\omega$ формулой Рэлея-Джинса, а не Планка.

Случай б): При максимальной фокусировке с помощью линзы с фокусным расстоянием $F$ площадь пятна в фокусе $\Delta S' = \pi\left(F\frac{\lambda}{D}\right)^2$, а телесный угол схождения лучей в фокус $\Delta\Omega' = \frac{\pi D^2}{4F^2}$. Отсюда следует $\Delta\Omega'\Delta S' = \frac{\pi^2\lambda^2}{4}$, т. е. такая же величина, что и в первом случае. Получен ный результат $\Delta\Omega\Delta S =$ const является общим свойством оптических систем: поверхностная яркость предмета и его изображения одинаковы. Можно считать, что изображение «излучает» в телесный угол $\Delta\Omega'$. Поскольку все остальные параметры при фокусировке не меняются, то ответ будет таким же, как и в случае $a$).

Фактически ответ можно получить и из качественных рассуждений. Энергия $\mathfrak{E}$ в интервале $\Delta\omega$ излучается лазером с площадки $S$ в телесный угол $\Delta\Omega \sim \lambda^2/S$ за время $\tau$. Энергия, излучаемая абсолютно черным телом с той же площадки, в тот же угол и за то же время есть $\rho_\omega\Delta\omega\Delta\Omega\tau S$, где $\rho_\omega \sim \frac{k_\text{Б}\ T_{\text{эфф}}}{c\lambda^2}$. Сравнивая, получаем $kT_{\text{эфф}} \sim \frac{\mathscr{E}}{\tau\Delta\omega}$. Линза не меняет спектрального состава излучения, а значит, не меняет эффективную температуру.

**1.**

50* $T = \frac{2\mathscr{E}}{\pi^2 k_\text{Б}} = 1,47 \cdot 10^{22}$ К.

Решение. Поскольку энергия лазерного излучения $\mathscr{E} = 1$ Д ж (что соответствует температуре $\mathscr{E}/k_\text{Б} \sim 10^{23}$ К) сосредоточена в очень узком спектральном интервале, то можно с уверенностью считать, что для абсолютно черного тела, имеющего температуру $T_{\text{эфф}}$ , $k_\text{Б}\ T_{\text{эфф}} \gg \hbar\omega$ ($\omega$ - частота излучения лазера, $\hbar\omega \simeq 2$ эВ для красного цвета). Таким образом, спектральную плотность потока энергии излучения черного тела (спектральную светимость $R_T(\omega)$) можно записать, ограничившись классической формулой Рэлея-Джинса.

$$R_T(\omega) = j_T(\omega) = \frac{c}{4}\rho_T(\omega) = \frac{\omega^2 k_\Sigma T_\text{J}}{4\pi^2 c^2} = \frac{k_\text{В} T_{\text{фф}}}{\lambda^2}$$

Спектральная яркость излучения абсолютно черного тела $B_T(\omega)$ связана со спектральной светимостью соотношением $R_T(\omega) = \pi B_T(\omega)$, являющимся следствием закона Ламберта. Откуда $B_T(\omega) = \frac{k_\text{Б} T_\text{эфф}}{\pi\lambda^2}$.

Найдем спектральную яркость лазерного излучения $B_\text{лаз}(\omega)$. Если обозначить плотность потока энергии на выходе лазера как $q$, сечение пучка на выходе $S = \pi D^2/4$, а время импульса излучения т, то, тогда очевидно, что

$$q = \frac{\mathscr{E}}{S_\tau} = \frac{4\mathcal{I}}{\pi D_\tau^2}$$

Ширина спектра импульсного излучения $\Delta v \approx \frac{1}{\tau}$, а телесный угол, в котором распространяется лазерный пучок,

$$\Delta\Omega = 2\pi(1-\cos\theta) \approx \pi\theta^2 \approx \pi\left(\frac{\lambda}{D}\right)^2$$

Таким образом, искомая спектральная яркость лазера

$$B_\text{лаз}(\omega) = \frac{q}{\Delta\omega\Delta\Omega} = \frac{q\tau}{2\pi^2\left(\frac{\lambda}{D}\right)^2} = \frac{2\mathscr{E}}{\pi^3\lambda^2}$$

По условию $B_T(\omega) = B_\text{лаз}(\omega)$, откуда

$$T_\Phi = \frac{2\mathscr{E}}{\pi^2 k} = 1,47\cdot 10^{22}\ \text{К}$$

Эта колоссальная температура полностью оправдывает применение классической формулы Рэлея-Джинса.

Как и в 1.49 можно привести качественную оценку. Поскольку здесь $\Delta v\cdot\tau \simeq 1$, то $k_\mathbf{\Sigma} T_9 \simeq \mathscr{E}$.

**1.**

51. $T_\ni = \frac{jd^2\lambda}{2\pi c k_\text{В}}\frac{v}{\Delta v} = 2,65\cdot 10^{16}$ К.

Как и в 1.49 , считая, что $\mathfrak{E}/\tau \simeq p$, получаем $k_\text{Б}\, T_\text{эфф} \simeq p/\Delta v$.

**1.**

52. $T_\text{эфф} = \frac{2W}{\pi^2 k_5 \delta v} = 1,47\cdot 10^{15}$ К.

Как и 1.49 , поскольку $\mathscr{E} = W\tau$, то $k_\text{Б}\, T_\text{эфф} \sim W/\delta\nu$.

**1.**

53* $T_\text{эфф} = aT = 3$ К.

Решение. Антенный тракт представляет собой волновод, по которому электромагнитное излучение передается к приемнику. Поглощение возникает главным образом из-за конечной проводимости стенок волновода. Пусть $a_T(\omega)$ - коэффициент поглощения (поглощательная способность), а $R_T(\omega)$ - излучательная способность материала волновода при температуре $T$ на частоте $\omega = 2\pi c/\lambda$. Тогда по закону Кирхгофа

$$R_T(\omega) = a_T(\omega)\varepsilon_T(\omega)$$

где $\varepsilon_T(\omega)$ - излучательная способность на частоте $\omega$ абсолютно черного тела, имеющего температуру $T$. По определению эффективной температуры $R_T(\omega) = R^{\text{откуда}}_{T^{\text{зачт}}_{\text{эфф}}}\ a_T(\omega)\varepsilon_T(\omega) = \varepsilon_{T'}(\omega)$. Очевидно, что $T < T$. Предположим, что $\hbar\omega \ll k_\text{Б}\, T_\text{эфф.}$ . Тогда в приближении Рэлея-Джинса и при условии, что $a_T(\omega) \equiv a$, получаем

$$a\frac{\omega^2}{\pi^2 c^3}k = \frac{\omega^2}{\pi^2 c^3}k_\text{Б}\, T,$$

откуда и следует ответ $T_\text{эфф} = aT = 3$ К.

Так как $\frac{\hbar\omega}{k_\text{Б}\, T_\text{эфф}} = 0,16 \ll 1$, использованное приближение оправданно.

**1.**

54* $j_0 \approx \frac{\pi^2\hbar c}{\lambda^3\tau} = 9,5\cdot 10^4 \rho\Gamma/\left(\text{см}^2\cdot\text{с}\right) = 9,5\text{МВТ/см}^2$, $a_{\max} = \frac{\hbar N_\text{A}}{2A_{\text{NaT}}\lambda^2} = 9,2\cdot 10^7$ см/с$^2$.

Решение. Существование силы светового давления обусловлено изменением импульса атома за счет процессов «ноглощение + спонтанное испускание». При поглощении фотона импульс атома получает приращение $\hbar k$. В то же время при спонтанном испускании фотон излучается в произвольном направлении, и поэтому при многократном повторении процессов поглощения и испускания средний

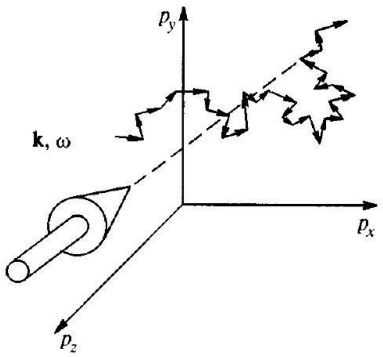



Рис. 205 импульс, уносимый фотонами, равен нулю. Поэтому в среднем каждый процесс «поглощение + спонтанное испускание» изменяет импульс атома на величину $\hbar\mathbf{k}\,(k=\omega_0/c)$. Однако, наряду со снонтанным излучением, в потоке излучения может происходить и вынужденное (индуцированное) излучение, при котором атом отдает импульс вдоль заданного направления (волнового вектора излучения). В результате совокупного действия поглощения фотонов и их последующего индуцированного и спонтанного излучения движение атома имеет характер систематического дрейфа в направлении вектора $k$ (рис. 205) (длины стрелок одинаковы).

По определению сила, действующая на атом, равна

$$\mathbf{F} = \frac{\Delta\mathbf{p}}{\Delta t} \qquad ()$$

где $\Delta\mathbf{p}$ - изменение импульса атома. Промежуток времени $\Delta t$, когда систему можно описывать классически, выбирается достаточно большим, чтобы за это время произошло много актов поглощения и испускания. Имеется два характерных масштаба времени: т и $1/(\sigma\Phi)$. Характерное время спонтанного испускания $\tau$ связано с шириной линии и вероятностью спонтанного испускания $W_\text{сп} \sim 1/\tau$. Характерное время процессов вынужденного поглощения и испускания $1/(\sigma\Phi)$ определяется сечением взаимодействия о фотона с атомом и плотностью потока фотонов $\Phi\left[\text{см}^{-2}\cdot\text{с}^{-1}\right]$. При этом $\sigma\Phi \sim W \equiv W_\text{исп}$ , где $W_\text{п}$ и $W_\text{исп}$ - вероятности вынужденного поглощения и вынужденного испускания фотона (рис. 206). Таким образом, промежуток времени $\Delta t$ должен удовлетворять условию $\Delta t \gg \text{т}, 1/(\alpha\Phi)$.

Пусть $n_1$ и $n_2$ - относительные заселенности уровней нижнего и верхнего состояний соответственно (рис. 206). Они удовлетворяют системе уравнений

$$\begin{cases} \frac{dn_1}{dt} = -\alpha\Phi n_1 \\ \frac{dn_2}{dt} = -\frac{n_2}{\tau} - \alpha\Phi n_2 \\ n_1 + n_2 = 1 \end{cases} \qquad (**)$$

В стационарном состоянии скорости переходов «вниз» и «вверх» должны быть равны, и тогда из (**) следует условие баланса $\alpha\Phi n_1 = \frac{n_2}{\tau} + \sigma\Phi n_2$ и выражения для $n_1$ и $n_2$:

$$n_1 = \frac{\sigma\Phi + \frac{1}{\tau}}{2\sigma\Phi + \frac{1}{\tau}}; \quad n_2 = \frac{\sigma\Phi}{2\sigma\Phi + \frac{1}{\tau}}$$

Видно, что с ростом Ф заселенность нижнего уровня $n_1 \to \frac{1}{2} + 0$, а верхнего $-n_2 \to \frac{1}{2} - 0$, что соответствует насыщению. В качестве предельного значения можно взять $\Phi_0 \approx \frac{1}{2\sigma r}$.

Плотность потока энергии насыщения $j_0 = \hbar\omega\Phi_0 \approx \frac{\hbar\omega}{2\sigma\tau}$. Поскольку сечение поглощения фотона - процесс резонанс-

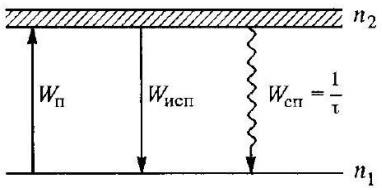



Рис. 206 ный, то согласно формуле Брейта-Вигнера $\sigma = 4\pi\hat{t}^2 = \frac{\lambda^2}{\pi}$ и $j_0 \approx \frac{\pi^2\hbar c}{\lambda^3\tau} = 9,5\cdot 10^4$ эрг $/\left(\text{см}^2\cdot\right) = 9,5$ мВт /см$^2$. Здесь сечение процесса поглощения считается как сечение упругого рассеяния. Это связано с тем, что в начальном состоянии были фотон и атом, а в конечном состоянии - атом и фотон. Как следует из вышесказанного,

$$\mathbf{F} = \frac{\Delta \mathbf{p}}{\Delta t} = \left(\frac{\Delta \mathbf{p}}{\Delta t}\right)_{\text{спонт}} + \left(\frac{\Delta \mathbf{p}}{\Delta t}\right)_{\text{инд}} = 0 + \hbar \mathbf{k} \sigma \Phi \left(n_1 - n_2\right) = \hbar \mathbf{k} \frac{\sigma \Phi}{2\sigma \Phi \mathrm{t} + 1}$$

Следовательно, при $\Phi \to \infty$, получаем

$$|\mathbf{F}_{\max}| = \frac{\hbar k}{2\mathrm{r}}, \text{ откуда } a_{\max} = \frac{F_{\max}}{M} = \frac{\hbar N_A}{2A_{\mathrm{Na}}\mathrm{r}\lambda} = 9,2 \cdot 10^7 \frac{\mathrm{cm}}{\mathrm{c}^2} \left(\sim 10^5 \text{ g}\right)$$

Выражение для силы $F$ может быть интерпретировано следующим образом: $F = \hbar k \frac{1}{2\tau}$, т. е. произведение импульса одного кванта $\hbar k$ на вероятность спонтанного излучения в единицу времени $1/(2\tau)$. Последняя в свою очередь есть произведение вероятности того, что атом находится в возбужденном состоянии $(= 1/2)$ на вероятность спонтанного излучения при условии, что атом возбужден $(= 1/\tau)$.

Отметим, что если бы спонтанное излучение отсутствовало, то было бы $F = 0$ (при $\tau \to \infty, F \to 0$). За большой $(\Delta t \gg \tau \to)$ промежуток времени

атом поглотит и излучит в том же направлении одинаковое число квантов. Спонтанное излучение нарушает этот баланс, и в результате импульс атома равен импульсу тех поглощенных квантов, которые были переизлучены спонтанно.

**1.**

55. Обозначая вероятность индуцированного перехода с $n$-го на $m$-й уровень $p_{nm}\,(p_{nm} = p_{mn})$, а спонтанного перехода $\boldsymbol{W}_{nm}$, в случае атома

$$\frac{p_{nm}}{W_{nm}} = \frac{1}{e^{\hbar\omega/k_{\mathrm{B}}T} - 1} \approx e^{-\hbar\omega/k_{\mathrm{B}}T} \approx 1,6 \cdot 10^{-17} \left(\hbar\omega \gg k_{\mathbf{5}}T\right).$$

Для электронного спина в магнитном поле

$$\frac{p_{nm}}{W_{nm}} \approx \frac{k_{\mathrm{B}}T}{2\mu_{\mathrm{F}}B} \approx 2,2 \cdot 10^3 \left(k_{\mathrm{E}}T \gg 2\mu_5 B\right).$$

**1.**

56. $\lambda < \frac{hc}{k_{\mathrm{B}}T \ln 101} \approx 10,6 \cdot 10^{-4}$ cm.

**1.**

57. $x(T) = x_0 \mathrm{th}\frac{\hbar\omega}{2k_{\mathrm{B}}T}; 1\Big)\, x(T) \approx x_0 \frac{\hbar\omega}{2k_{\mathrm{B}}T}$;
2) $x(T) \approx x_0 \left[1 - 2\exp\left(-\frac{\hbar\omega}{k_{\mathrm{B}}T}\right)\right]$.

**1.**

58. $\frac{W_{\text{сп}}}{W^0_{\text{сп}}} \simeq \frac{\pi^2 Q c^3}{V\omega^3}$.

Решение. В квантовой физике спонтанное излучение возникает из-за возмущающего влияния на молекулу нулевых колебаний электромагнитного поля. Вероятности спонтанного и вынужденного излучений выражаются через коэффициенты Эйнштейна $A$ и $B$ и плотность энергии излучения $\rho(\omega)$

$$W_{\text{вын}} = \rho(\omega)B; \quad W_{\text{сп}} = A;$$

при этом $B/A = [\hbar\omega\mathscr{D}(\omega)]^{-1}$, где $\mathscr{D}(\omega)$ - спектральная плотность мод (осцилляторов, числа колебаний) поля в единичном объеме. Таким образом,

$$\frac{W}{W_{\mathrm{cI}}} = \frac{\rho(\omega)}{\hbar\omega\mathscr{D}(\omega)} = n(\omega, T)$$

где $n(\omega, T)$ - число фотонов данной моды колебаний.

Как известно, для невырожденных уровней одной моды молекулы (атома) вероятности вынужденного излучения и поглощения равны. Рассмотрим молекулу в поле излучения со спектральной плотностью потока энергии $j(\omega) \propto c\rho(\omega)$. Здесь точное значение коэффициента пропорциональности зависит от углового распределения излучения. Так для параллельного потока $j(\omega) = c\rho(\omega)$, а для изотропного

излучения $j(\omega) = c\rho(\omega)/4$. В резонаторе этот коэффициент зависит от пространственного распределения поля, но по порядку величины он близок к единице.

Поглощаемая молекулой мощность

$$\frac{d\mathscr{E}}{dt} = \int j(\omega)\sigma(\omega)d\omega$$

где $\sigma(\omega)$ - сечение поглощения молекулы. Поскольку функция $\sigma(\omega)$ быстро спадает при удалении от резонансной частоты $\omega_0$, то можно записать

$$\frac{d\&}{dt} \simeq j\left(\omega_0\right) \int \sigma(\omega)d\left(\omega \propto c\rho\left(\omega_0\right)S\right.$$

где $S$ - площадь под кривой поглощения $\sigma(\omega)$. В обоснование вышеизложенного отметим, что в вакууме $j(\omega)$ является плавной функцией $\propto \omega^3$, а в резонаторе $j(\omega)$ занимает частотный интервал $\Delta\omega = \omega_0/Q$, который по условию задачи превосходит собственную ширину линии Г, т. е. ширину кривой поглощения $\sigma(\omega)$. Если теперь записать $d\&/dt = \hbar\omega_0 W_{\text{вын}}$ , то из сравнения получим

$$W_{\text{вын}} \propto \frac{cp\left(\omega_0\right)S}{\hbar\omega_0}, \text{ откуда } B \propto \frac{cS}{\hbar\omega_0}$$

и не зависит с точностью до коэффициента от того, где находится молекула: в свободном пространстве или в резонаторе.

Вероятность спонтанного излучения

$$W_{\text{сп}} = W_{\text{вын}} \frac{1}{n(\omega,T)} = W_{\text{вын}} \frac{\hbar\omega D(\omega)}{\rho(\omega)} = \hbar\omega\mathscr{D}(\omega)B.$$

Запишем эту вероятность для двух случаев - в резонаторе ($W_{\text{сп}}$ ) и в свободном пространстве ($W_{\text{с}}^0$)

$$W_{\text{сп}} = \hbar\omega_0\mathscr{D}\text{ре}_{\text{з}}\left(\omega_0\right)B; \quad W_{\text{сп}}^0 = \hbar\omega_0\mathscr{D}\left(\omega_0\right)B$$

Искомое отношение этих вероятностей есть

$$\frac{W_{\text{сг}}}{W_{\text{сг}}^0} = \frac{\mathscr{D}_{\text{рез}}\left(\omega_0\right)}{\mathscr{D}\left(\omega_0\right)}$$

т. е. вероятность перехода пропорциональна плотности конечных состояний. В свободном пространстве (вакууме) спектральная плотность мод $\mathcal{D}\left(\omega_0\right) \simeq \frac{\omega_0^2}{\pi^2c^3}$. А в резонаторе, настроенном на частоту перехода, в интервале частот $\Delta\omega$ имеется только одно колебание (мода). Тогда спектральная плотность числа мод (осцилляторов) поля имеет порядок $\mathscr{D}_{\text{рез}}\left(\omega_0\right) \sim \sim \frac{1}{\Delta\omega\cdot\boldsymbol{V}} \sim \frac{Q}{\omega_0\boldsymbol{V}}$, откуда следует ответ

$$\frac{W_{\text{с}\pi}}{W_{\text{с}\pi}^0} \simeq \frac{\pi^2Qc^3}{V\omega^3} \qquad ()$$

В случае 1) получаем $\frac{W_{\text{сп}}}{W_{\text{сп}}^0} \simeq 4\cdot10^2$, а в случае 2) получим $\frac{W_{\text{сп}}}{W_{\text{сп}}^0} \simeq 4\cdot10^{-8}$. Как видно, в зависимости от конкретных данных возможны различные ситуации.

Приведем еще одно решение этой задачи. Вероятность спонтанного излучения молекулы равна вероятности поглощения, т. е. вероятность вынужденного испускания под действием одного фотона. В свою очередь вероятность поглощения пропорциональна времени нахождения фотона вблизи молекулы в резонаторе $\tau_\rho$. Из соотношения неопределенностей

$$\tau_{\mathbf{p}}\Delta\omega \sim 1, \text{ откуда } \tau_{\mathbf{p}} \sim \frac{1}{\Delta\omega} = \frac{Q}{\omega} = \frac{Q\lambda}{2\pi c}$$

Минимальный объем резонатора при длине волны $\lambda$ соответствует условию $\lambda/2 \sim \sqrt[3]{V}$. Таким образом, $\tau_{\mathbf{p}} = \frac{QV^{1/3}}{\pi c}$. Если взять такой же объем $V$, но в свободном пространстве, то время нахождения вблизи молекулы т $\tau_{\text{св}} \sim V^{1/3}/c$.

Искомое отношение вероятностей спонтанного перехода для этих двух случаев

$$\frac{W_{\text{с}\pi}}{W_{\text{сП}}^0} = \frac{\tau_{\text{р}}}{\tau_{\text{сВ}}} \sim \frac{Q}{\pi}$$

Очевидно, что это тот же результат, в чем легко убедиться, подставив $V = (\lambda/2)^3, \omega = 2\pi c/\lambda$ в соотношение (*).

**1.**

59. $\frac{N_2}{N_0} = \frac{1}{2}\left[1 - \frac{\ln(1-\tau)}{2lk_n}\right] = 0,5055$, где $N_2$ - число атомов хрома, находящихся в возбужденном состоянии, $N_0$ - полное число атомов хрома.

**1.**

60. $\Delta v = 2v_0\sqrt{\frac{2RT}{\mu c^2}\ln\left(\frac{\beta_0 - 1}{1-R}\right)} = 3,5$ МГц.

Ук азание. При а $\to 1 + 0 \quad \ln\beta \simeq \beta - 1$.

**1.**

61. Сдвиг частоты принимаемого сигнала при изменении направления движения происходит за счет эффекта Доплера и составляет $\Delta\omega \simeq 2\frac{v}{c}\omega$, где $\omega$ - частота, на которой производится измерение. При этом сдвиг распределения излучения абсолютно черного тела $\Delta\rho \simeq \frac{d\rho}{d\omega}\Delta\omega$. Отсюда видно, что измерения надо проводить на тех частотах, где максимальна величина $\left|\omega\frac{d\rho}{d\omega}\right|$.

Таким образом,

$$\left|\frac{\Delta\rho}{\rho}\right|_{x=x_1} \simeq 1,55\cdot 10^{-3}; \quad \left|\frac{\Delta\rho}{\rho}\right|_{x=x_2} \simeq 2,52\cdot 10^{-3}$$

**1.**

62. $t = t_1\frac{T_1}{T_0} \sim 10^7$ лет.

**1.**

63. $P_{\text{СВ}} = \frac{4\sigma}{3c}T^4 \simeq 4,6\cdot 10^{17}$ дин $/\text{см}^2 \simeq 4,6\cdot 10^{11}$ атм; $P_{\text{газ}} = \frac{N_{\text{А}}\rho Z}{A}k_{\text{Б}}\,T = 7\cdot 10^{10}$ атм $< P_{\text{св}}$ . Реально при $kT = 10$ кэВ происходит лишь частичная ионизация. Заметим, что в водородной бомбе именно световое «сдавливание» термоядерного горючего (т. н. имплозия) играет решающую роль.

**1.**

64. $T_{\text{пл}} = T_3\left(\frac{\tau_3}{\tau_{\text{пл}}}\right)^{1/3} \approx 1350$ K, где $\tau_3$ - период обращения Земли, а $T_3$ - температура ее поверхности.

**1.**

65. $L_{\max} = RT_0^2\sqrt{\frac{\sigma}{5j}} = 1,5\cdot 10^5$ m $\approx 150$k.

**1.**

66. $T = T_{\text{C}}\sqrt{\frac{a_c}{2}}\left(\frac{1+d^2/\left(16r^2\right)}{1+r^2/R^2}\right)^{1/4} \approx 430$ K.

**1.**

67. $F = \frac{\sigma T^4 \pi a^4}{cL^2} = 6 \cdot 10^{-7}$ дин (притяжение). Если бы в полости был один шарик, то никаких сил на него не действовало бы (симметрия). Наличие второго шарика приводит к нарушению симметрии. Он «затеняет» часть падающего излучения и нарушает баланс сил. В результате возникает эффективное притяжение между шариками. Собственное излучение шариков не дает вклада в силу из-за его изотропности.

**1.**

68. $L \leq \lambda = \frac{1}{n\sigma} = 2,5 \cdot 10^{25}$ cm $\simeq 10^7$ пк (парсек), т. е. много больше размеров нашей Галактики ($\sim 10^3$ пк), где концентрация реликтовых фотонов $n = \int_0^\infty \frac{\omega^2 d\omega}{\pi^2 c^3 [\exp(\hbar\omega/k_B T)-1]} \approx 400$ cm$^{-3}$.

**1.**

69. $\omega \geqslant \pi \left(\frac{c^4 \sigma}{2a^4}\right)^{1/5}$. Отсюда предельно возможная частота в резонаторе $\omega_{\max} = 1,43 \cdot 10^{11}$с$^{-1}$; а минимальная частота колебаний в кубическом резонаторе $\omega_{\min} = c\frac{\pi}{a}\sqrt{2} = 4,16 \cdot 10^9$с$^{-1}$. Резонансные свойства пропадут, когда расстояние между пиками станет больше или равно ширине самого пика (аналог критерия Рэлея). Величина $\omega_{\max}$, где $\tau = \frac{mo}{ne^2} \sim 10^{-14}$с, $\omega_{\max}\tau \lesssim 10^{-3}$, т. е. можно считать постоянной.

**1.**

70. $\frac{\mathscr{E}}{\Delta v} = kT\frac{2L}{c} = 5,2 \cdot 10^{-21}\frac{\text{эрг}}{\Gamma\mu}$.

Замкнутый с обоих концов волновод с указанными размерами представляет собой длинный резонатор, в котором закон дисперсии $\varepsilon\frac{\omega^2}{c^2} = k^2 = k_x^2 + k_y^2 + k_z^2$, где $k_x = m\frac{\pi}{a}, k_y = n\frac{\pi}{b}, k_z = l\frac{\pi}{L}; m, n, l = 0, 1, 2, 3, \ldots$. Поскольку размеры $a$ и $b \ll L$, то в диапазоне длин волн $\lambda = 10$ см поперечные моды сильно разрежены по сравнению с продольными. Продольные моды расположены намного чаще, и вся энергия сосредоточена в основном в них. Моды продольных колебаний поля распределены также, как колебания в одномерной цепочке. Резонансные частоты продольных колебаний находятся из условия $L = m\frac{\lambda}{2}, m$ - целое число. Число мод в частотном интервале $\Delta v$ равно $\frac{2L}{c}\Delta v$. Поскольку $hv \sim 10^{-5}$ эВ $\ll kT \sim 5 \cdot 10^{-2}$ эВ, то энергия одной моды колебаний равна $kT$. Поэтому плотность электромагнитной энергии в волноводе $\frac{\mathscr{E}}{\Delta v} = kT\frac{2L}{c} = 5,2 \cdot 10^{-21}\frac{\text{эр}}{\Gamma}$.

**1.**

71. $t = \frac{M^3}{3A} = 0,84 \cdot 10^{20}$ с, где через $A$ обозначено $A = \frac{\hbar c^4}{15360\pi\gamma^2} \simeq \simeq 0,4 \cdot 10^{25} r^3$/с.

**1.72.** $H \geqslant \frac{RT}{\mu_к g}\ln\left[\frac{P_0\sigma\tau}{k}\sqrt{\frac{8R}{\pi\mu_к T}}\right] \approx 1,6 \cdot 10^5$ м $= 160$ км, где $\mu_в$ и $\mu_к$ — молярные масса воздуха и кислорода, $P_0 \approx 10^5$ Па - атмосферное давление у поверхности Земли.

**1.**

73* $n \leqslant 825$.

Решение. Образование устойчивой атомной системы с уровнями $\mathscr{E}_n$ возможно, если уширение уровней меньше расстояния между уровнями

$$\partial\mathscr{E}_n \leqslant \mathscr{E}_n - \mathscr{E}_{n-1} = \hbar\omega$$

В поле магнитотормозного излучения время жизни возбужденных атомов определяется двумя факторами - временем спонтанных переходов $\tau_{\text{сп}}$ и временем индуцированных переходов. Т. о. ширина атомных уровней $\delta\mathscr{E}_n$ состоит из двух слагаемых - естественной ширины, определяемой временем спонтанного перехода, и ширины, обусловленной временем индуцированного излучения

$$\delta\delta_n \simeq \frac{\hbar}{\tau_{\text{изл}}} = \hbar\left(\frac{1}{\tau_{\text{сп}}} + \frac{1}{\mathfrak{r}_{\text{кнд}}}\right).$$

Время индуцированных переходов с $n$-уровня под действием фотонов магнитотормозного излучения равно

$$\tau_{\text{инд}} = \frac{\tau_{\text{сI}}(n)}{N(\omega)} = \tau_{\text{с}\pi}(n)\frac{\hbar\omega\mathcal{D}(\omega)}{\rho(\omega)} = \tau_{\text{с}\pi}(n)\frac{\hbar\omega^3}{4\pi^2c^2I(\omega)}.$$

Здесь $N(\omega)$ - среднее число фотонов магнитотормозного излучения, $D(\omega) = \omega^2/\left(\pi^2c^3\right)$ - плотность состояний, $\rho(\omega) = 4I(\omega)/c$ - спектральная плотность изотропного излучения на частоте $\hbar\omega \approx 2\text{Ry}/\text{n}^3$.

В отсутствии вынужденного излучения, как это следует из условия, при $n \gg 1$ естественная ширина уровней пропорциональна $n^{-5}$, и поэтому она убывает быстрее, чем расстояние между уровнями, которое пропорциональ но $n^{-3}$. Поэтому спонтанное излучение не может привести к ограничению на величину $n$.

Однако фотоны магнитотормозного излучения существенно увеличивают ширину уровней, и предельное значение номера уровня находится из условия

$$\frac{\hbar}{\tau_{\text{инд}}} \leqslant \frac{2\text{Ry}}{n^3},$$

откуда

$$n^7 \leqslant \frac{4Ry^4 6\cdot 10^{-12}}{\pi^2c^2\hbar^3I(\omega)} = 0,26\cdot 10^{21},$$
$$n_{\max} \simeq 825.$$

Отметим, что магнитотормозное излучение не является равновесным и среднее число фотонов не описывается формулой Планка, а находится из заданной спектральной интенсивности

$$I(\omega) = \frac{c\rho(\omega)}{4} = \frac{c}{4}\hbar\omega\cdot\mathscr{D}(\omega)N(\omega)$$

откуда $N(\omega) \simeq 1,7\cdot 10^{11}$. Поскольку в космическом пространстве есть и реликтовое излучение с температурой $T = 2,7$ K, то в принципе эти фотоны также могут вызывать индуцированные переходы. Среднее число таких фотонов в интересующем нас интервале частот равно

$$\bar{N}(\omega) = \frac{1}{\exp\left(\hbar\omega/k_{\text{B}}T\right)} \simeq \frac{k_{\text{B}}T}{\hbar\omega} \simeq 4,8\cdot 10^3 \ll N(\omega),$$

Фактически эта оценка показывает, что эффективная температура магнитотормозного излучения, соответствующая данной частоте, равна

$$T_{\text{эфф}} = T\frac{N(\omega)}{\bar{N}(\omega)} \simeq 9,4\cdot 10^7 \text{ K}$$

**1.**

74. $\sqrt{\langle\delta^2\rangle} = 2\lambda_{\text{e}}\sqrt{\frac{\alpha}{\pi}\ln\frac{2}{\alpha}} = 8,5\cdot 10^{-12}$ см, где $\alpha = 1/137$ - постоянная тонкой структуры, $\lambda_{\text{e}}$ - комптоновская длина волны электрона.

Ренение. Рассмотрим нулевое колебание (моду) с частотой $\omega$: $\mathbf{E}(\omega) = \mathbf{E}_0\cos\omega t$. Уравнение движения свободнго электрона в поле этой волны имеет вид $m\frac{d^2\delta_\omega}{dt^2} = e\mathbf{E}(\omega)$, где $\delta_\omega$ - смещение электрона. Интегрируя это уравнение, получаем $\delta_\omega = -\frac{e}{m\omega^2}\mathbf{E}_0\cos\omega t$. Средний по времени квадрат смещения электрона $\overline{\delta_\omega^2} = \frac{e^2E_0^2}{m^2\omega^4}\overline{\cos^2\omega t} = \frac{e^2E_0^2}{2m^2\omega^4}$. Амплитуда колебания $E_0$ может быть найдена по принципу соответствия классического и квантового выражений для энергии моды

$$\int\frac{\overline{E^2(\omega)} + \overline{H^2(\omega)}}{8\pi}dV = \frac{1}{2}\hbar\omega.$$

Здесь $V$ - объем системы. В плоской электромагнитной волне $\overline{E^2(\omega)} = \overline{H^2(\omega)} = \frac{E_0^2}{2}$, поэтому $E_0^2 = \frac{4\pi}{V}\hbar\omega$. Следовательно $\overline{\delta_\omega^2} = \frac{2\pi}{\boldsymbol{V}}\frac{\hbar e^2}{m^2\omega^3}$. Поскольку нулевые колебания различных частот независимы, то надо просуммировать (проинтегрировать) это выражение по всем возможным колебаниям. Число мод нулевых колебаний (число осцилляторов поля) в объеме $V$, приходящихся на интервал $d\omega$, есть $dN_\omega = V\frac{\omega^2}{\pi^2c^3}d\omega$. Поэтому

$$\left\langle\delta^2\right\rangle = \int\overline{\delta_\omega^2}dN_\omega = \frac{2}{\pi}\frac{e^2}{\hbar c}\left(\frac{\hbar}{mc}\right)\int_{\text{Ry}/\hbar}^{2^{2mc}c^2/\hbar}\frac{d\omega}{\omega} = \frac{4}{\pi}\alpha\hat{\pi}_{\text{e}}^2\ln\frac{2}{\alpha}$$

Здесь $\alpha$ - постоянная тонкой структуры, $\lambda_e$ - комптоновская длина волны электрона. Таким образом,

$$\sqrt{\langle \delta^2 \rangle} = 2\hat{\pi}_e \sqrt{\frac{\alpha}{\pi} \ln \frac{2}{\alpha}} = 8,5 \cdot 10^{-12} \text{ cm}.$$

Такое смещение электрона на «орбите» может быть описано введением донолнительного, по отношению к кулоновскому потенциалу ядра, отталкивательного потенциала. Этот потенциал должен быть сферически симметричен и, как следует из полученной оценки, его эффективный радиус намного меньше боровского радиуса. Поэтому он главным образом возмущает $2s$-состояние, волновая функция которого в отличие от $2p$-состояния конечна при $r = 0$. В результате уровень $2s_{1/2}$ смещается вверх по энергии относительно уровня $2p_{1/2}$.

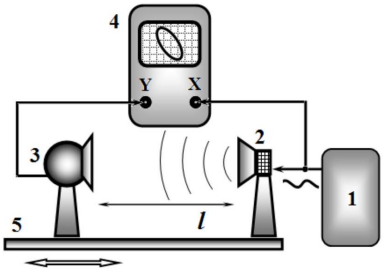
## 15.1.4 Problems about Specific Optical Devices

(I'll put them here from the previous section)

## 15.1.5 Problems of other parts of physics that uses concepts from optics

### BelPho.2016.resp-11-1 Измерение скорости звука с помощью осциллографа.

Для точного измерения скорости звука используется следующая экспериментальная установка.
Гармонический синусоидальный электрический сигнал с заданной частотой $v$ от генератора 1 подается на звуковой динамик 2 , испускающий звук той же частоты. Этот звук улавливается микрофоном 3. Динамик и микрофон размещены на подставках, микрофон может перемещаться по рельсу 5 , тем самым, изменяя расстояние $l$ между динамиком и микрофоном. Сигнал от генератора также подается на вход $X$ осциллографа. Микрофон под действием звука выдает электрический сигнал, напряжение которого пропорционально давлению звуковой волны, падающей на микрофон. Этот сигнал подается на вход Y осциллографа.



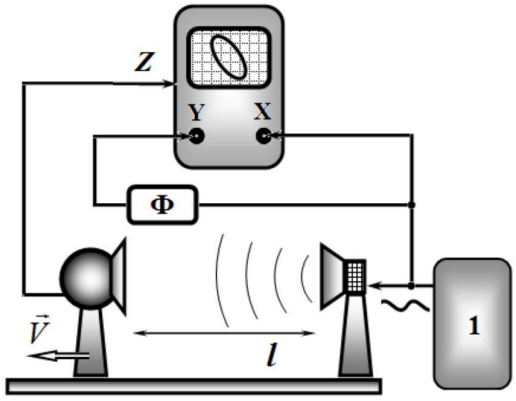

Горизонтальное смещение луча на экране осциллографа пропорционально напряжению, подаваемому на вход $X$, а вертикальное смещение пропорционально напряжению, подаваемому на вход Y. Осциллограф настроен так, что максимальные смещения луча по горизонтали и вертикали равны.

#### Solution pr. Часть 1. Изменение фигуры.

1.1 Запишите уравнение звуковой волны, распространяющейся от динамика. Амплитуда волны $A$, частота $v$, скорость звука $c$.
1.2 Запишите функции от времени, описывающие сигналы, подаваемые на входы $X$ и $Y$ осциллографа, в зависимости от расстояния $l$.
1.3 Изобразите траектории луча на экране осциллографа при различных расстояниях $l = m\frac{\lambda}{8}$ $m = 0, 1, 2 \ldots 8$, где $\lambda$ - длина звуковой волны. В каждом случае укажите направление движения луча.

#### Solution pr. Часть 2. Движение выреза.

В осциллографе также имеется вход $Z$, напряжение на котором регулирует яркость луча. Если напряжение на этом входе превышает пороговое значение $U_{\text{пор.}}$ , то луч гаснет. На вход $Z$ подается сигнал с микрофона. Пороговое значение отрегулировано так, что $U_{\text{пор.}} = 0,9U_{\max}$, где $U_{\max}$ - амплитудное значение сигнала микрофона. Синусоидальный сигнал от генератора по-прежнему подается на динамик и на вход $X$ осциллографа. Этот же сигнал проходит через фазовращатель




(на схеме обозначен Ф), который изменяет фазу сигнала на $\frac{\pi}{2}$, не изменяя его амплитуды (для этого используется обычный конденсатор).

2.1 Изобразите траекторию луча на экране при отключенном микрофоне.

При включенном микрофоне на траектории луча появляется вырез (из-за того, что вход $Z$ гасит луч).

2.2 Чему равен угловой размер образующегося выреза?

2.3 Микрофон перемещают с постоянной скоростью $\vec{V}$, удаляя его от динамика. Модуль этой скорости значительно меньше скорости звука. Определите угловую скорость движения выреза на экране осциллографа, при частоте генератора равной $v$.

2.4 При частоте генератора $v = 600Гu$ и скорости движения микрофона $V = 2,5\frac{MM}{c}$ вырез делает полный оборот за время $\tau = 220$ с. Определите по этим данным скорость звука в воздухе.

## Solution sol.Часть 1. Изменение фигуры.

1.1 Уравнение звуковой волны, распространяющейся от динамика, имеет традиционный вид

$$p = A\cos(\omega t - kx) = A\cos\left(2\pi v\left(t - \frac{x}{c}\right)\right). \tag{1}$$

1.2 Сигналы, которые подаются на входы осциллографа, описываются функциями

$$\begin{aligned} U_x &= A\cos\omega t \\ U_y &= A\cos(\omega t - kl)\cdot \end{aligned} \tag{2}$$

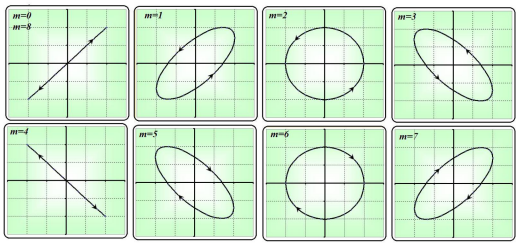
1.3 Выражение для сигнала, снимаемого с микрофона и подаваемого на вход Y. Удобно представить в виде

$$U_y = A\cos(\omega t - kl) = A\cos\left(\omega t - \frac{2\pi}{\lambda}m\frac{\lambda}{8}\right) = A\cos\left(\omega t - m\frac{\pi}{4}\right) \tag{3}$$

Тогда уравнение траектории луча в параметрическом виде имеет вид

$$\begin{aligned} x &= A\cos\omega t \\ y &= A\cos\left(\omega t - m\frac{\pi}{4}\right) \end{aligned} \tag{4}$$

Траектория, описываемая этими уравнениями в общем случае является эллипсом, который при $m = 0$ и $m = 4$ вырождается в отрезок прямой, а при $m = 2$ и $m = 6$ является окружностью. Последовательность этих траекторий при различных $m$ показана на рисунке.



**Solution sol.Часть 2. Движение выреза.**

2.1 Так как сдвиг фаз между колебаниями вдоль перпендикулярных осей равен $\frac{\pi}{2}$, то траекторией луча будет окружность.
2.2 Угловой размер образующегося выреза можно определить из условия его образования $U_Z > U_{пор.}$. Из которого следует, что угловые границы выреза удовлетворяют условию

$$\cos\varphi = 0,9 \Rightarrow \Delta\varphi = 2\arccos 0,9 = 0,90 \approx 52^\circ \tag{5}$$

2.3 Не сложно сообразить, что вырез делает полный оборот, если микрофон смещается на расстояние равное длине волны звука. На это требуется время

$$T = \frac{\lambda}{V} = \frac{1}{v}\frac{c}{V} \tag{6}$$

Следовательно, угловая скорость движения выреза равна

$$\Omega = \frac{2\pi}{T} = \frac{2\pi}{v}\frac{V}{C} \tag{7}$$

2.4 При частоте генератора $v = 600\text{Г}$ и скорости движения микрофона $V = 2,5\frac{\mathcal{M}\mathcal{M}}{c}$ вырез делает полный оборот за время $\tau = 220c$. Определите по этим данным скорость звука в воздухе.
Из формулы (6) следует, что

$$c = \nu V\tau = 600 \cdot 2,5 \cdot 10^{-3} \cdot 220 = 330\frac{\mathcal{M}}{c} \tag{8}$$

# Part V
# —— Special Optics in a Nutshell ——

## 16 On Other Properties and Methods of Optics

### 16.1 Other properties and methods for typical phenomena

#### 16.1.1 Diffraction: other properties and methods

**Other well-known derivations in a nutshell**

**Diffraction of a spherical wave by Sivukhin**

(??? maybe the advantage is that we can obtain the normalization constant??? but it is also an interesting proof)

$$E = \int \frac{K(\alpha)}{r_0 \rho} e^{i(\omega t - k r_0 - k\rho)} dF$$

$$dF = 2\pi r_0^2 \sin\theta d\theta \quad \rho^2 = r_0^2 + (r_0 + r)^2 + ...$$

$$E = \frac{2\pi}{r_0 + r} e^{i(\omega t - k r_0)} \int_r^{r_{\max c}} K(\rho) e^{-ik\rho} d\rho$$

$r_{max} = r + 2r_0$,

$$K_n \int_{r+(n-1)\lambda/2}^{r+m_2/2} e^{-ikp} d\rho = (-1)^{n+1} \frac{2K_n}{ik} e^{-ikr}$$

$E = \sum_{i=1}^{N} E_i$,

$$E_n = (-1)^{n+1} \frac{4\pi K_n}{ik\,(r_0 + r)} e^{i[\omega t - k(r_0 + r)]}$$

$$E = \frac{1}{2} E_1 = \frac{2\pi K_1}{ik\,(r_0 + r)} e^{i[\omega t - k(r_0 + r)]}$$

$$E = \frac{1}{r_0 + r} e^{i[\omega t - k(r_0 + r)]}$$

$$K_1 = \frac{ik}{2\pi} = \frac{k}{2\pi} e^{i\pi/2}$$

**Формализм дифракции Френеля по Ландау**

(??? I'll look at it again later, it is maybe not an efficient approach)

$$u \sim \exp\left[ik\sqrt{y^2 + z^2 + (D_q + y\,\mathrm{tg}\,\alpha)^2}\right]$$

$$R = \sqrt{y^2 + (z - d)^2 + (D_p - y\,\mathrm{tg}\,\alpha)^2}$$

$$u_P \sim \int_{-\infty}^{+\infty}\int_{0}^{\infty} \exp\left[ik\sqrt{\left(D_q + y\,\mathrm{tg}\,\alpha\right)^2 + y^2 + z^2} + \sqrt{\left(D_p - y\,\mathrm{tg}\,\alpha\right)^2 + y^2 + (z-d)^2}\right] dydz$$

$$\sqrt{\left(D_q + y\,\mathrm{tg}\,\alpha\right)^2 + y^2 + z^2} \approx D_q + \frac{y^2+z^2}{2D_q} + y\,\mathrm{tg}\,\alpha$$

$$\sqrt{\left(D_p - y\,\mathrm{tg}\,\alpha\right)^2 + y^2 + (z-d)^2} \approx D_p + \frac{y^2+(z-d)^2}{2D_p} - y\,\mathrm{tg}\,\alpha.$$

$$u_P \sim \int_{0}^{\infty} \exp\left[ik\left(\frac{1}{2D_q}z^2 + \frac{1}{2D_p}(z-d)^2\right)\right] dz$$

$$u_P \sim \exp\left[ik\frac{d^2}{2\left(D_p + D_q\right)}\right] \int_{0}^{\infty} \exp\left[ik\frac{\frac{1}{2}\left[\left(\frac{1}{D_p}+\frac{1}{D_q}\right)z - \frac{d}{D_p}\right]^2}{\frac{1}{D_p}+\frac{1}{D_q}}\right] dz$$

$$u_P \sim \int_{-w}^{\infty} e^{i\eta^2} d\eta$$

$$w := d\sqrt{\frac{kD_q}{2D_p\left(D_q + D_p\right)}}$$

$$I = \frac{I_0}{2}\left|\sqrt{\frac{2}{\pi}}\int_{-w}^{\infty} e^{i\eta^2} d\eta\right|^2 = \frac{I_0}{2}\left[\left(C\left(w^2\right)+\frac{1}{2}\right)^2 + \left(S\left(w^2\right)+\frac{1}{2}\right)^2\right]$$

$$C(z) = \sqrt{\frac{2}{\pi}}\int_{0}^{\sqrt{z}} \cos\eta^2 d\eta, \quad S(z) = \sqrt{\frac{2}{\pi}}\int_{0}^{\sqrt{z}} \sin\eta^2 d\eta$$

$$\int_{|w|}^{\infty} e^{i\eta^2} d\eta = -\frac{1}{2i|w|}e^{iw^2} + \frac{1}{2i}\int_{|w|}^{\infty} e^{i\eta^2}\frac{d\eta}{\eta^2}$$

$$\int_{|w|}^{\infty} e^{i\eta^2} d\eta = e^{iw^2}\left(-\frac{1}{2i|w|} + \frac{1}{4|w|^3} - \ldots\right)$$

$$I = \frac{I_0}{4\pi w^2}$$

$$\int_{-w}^{\infty} e^{i\eta^2} d\eta = \int_{-\infty}^{+\infty} e^{i\eta^2} d\eta - \int_{-\infty}^{-w} e^{i\eta^2} d\eta = (1+i)\sqrt{\frac{\pi}{2}} - \int_{w}^{\infty} e^{i\eta^2} d\eta$$

$$\int_{-w}^{\infty} e^{i\eta^2} d\eta \approx (1+i)\sqrt{\frac{\pi}{2}} + \frac{1}{2iw}e^{iw^2}$$

$$I = I_0\left[1 + \sqrt{\frac{1}{\pi}}\frac{\sin\left(w^2 - \frac{\pi}{4}\right)}{w}\right]$$

Таким образом, в освещенной области, вдали от края тени, интенсивность имеет неограниченный ряд максимумов и минимумов, так что отношение $I/I_0$ колеблется в обе стороны от единицы.

В области геометрической тени интенсивность спадает монотонно при удалении от границы тени (на самой этой границе $I/I_0 = \frac{1}{4}$). При положительных $w$ интенсивность имеет чередуюЩиеся максимумы и минимумы.

В первом, наибольшем из максимумов, $I/I_0 = 1,37$

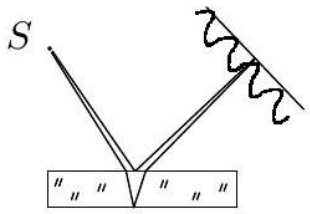
## 16.2 Other well-known but not so useful phenomena

### 16.2.1 Radio waves from labnik

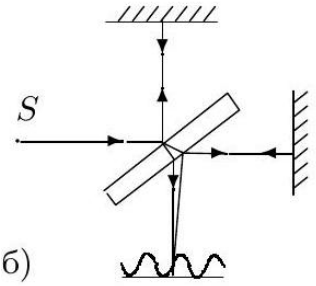
**1.**

() $\Delta\lambda \approx (4 \div 8) \cdot 10^{-5}$
c
-.
-.
()

6.1. 6.1a 6.1




а)



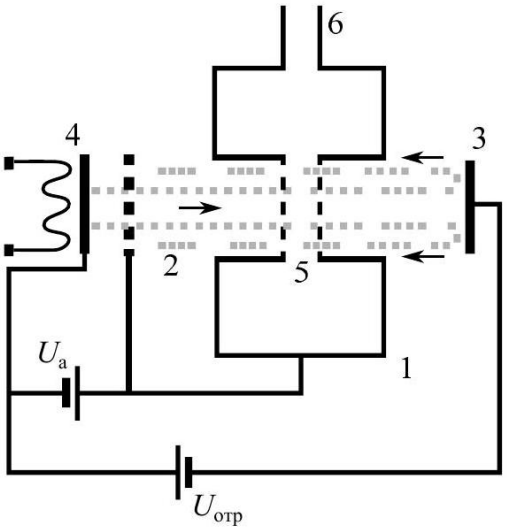

. 6.1.
$\Delta\nu$ $\tau$

$$\tau \cdot \Delta\nu \sim 1 \tag{6.1}$$

$\tau$ $10^{-10}$ .
( )
- (6.1). $\tau$
$\lambda$ $10^4 \div 10^5$
$\tau$ $10^{-5} \div 10^{-3}$
$c\tau$ $\gamma < \lambda/\rho$ $\rho$ - .
)01 $< \lambda \leqslant$ 1) 30 $\leqslant \nu <$ 300 ()
)1 $< \lambda \leqslant$ 10) 3 $\leqslant \nu <$ 30 ()
03 300 1 m 1

**2. -**

- (5 5)
6.2
1 2. 3 $U$
4 $U_{\text{a}}$



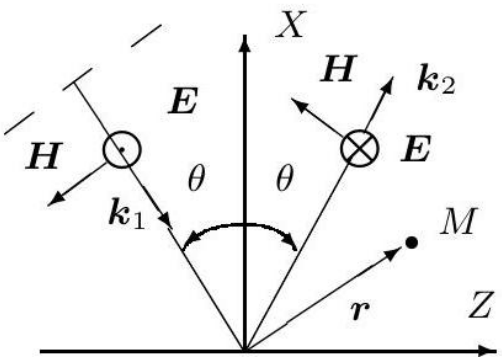

. 6.2.
5
$\approx 1\%$ $\approx 20\%$
- 6 -

**3. -**

$$\boldsymbol{E}(\boldsymbol{r}t) = \boldsymbol{E}_0 e^{i(\boldsymbol{k}\boldsymbol{r}-\omega t)} = \boldsymbol{A}(\boldsymbol{r})e^{-i\omega t} \tag{6.2}$$

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{v} \tag{6.3}$$

$$\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{E}_0 e^{i\boldsymbol{k}\boldsymbol{r}} = \boldsymbol{E}_0 e^{i(k_x x + k_y y + k_z z)} \tag{6.4}$$




$$\boldsymbol{E} = \boldsymbol{E}_0 e^{i(\boldsymbol{k}_1\boldsymbol{r}-\omega t)} \quad \boldsymbol{E} = -\boldsymbol{E}_0 e^{i(\boldsymbol{k}_2\boldsymbol{r}-\omega t)}$$

$k_1 = k_2 = k = \omega/c.$

$$\boldsymbol{E} = \boldsymbol{E}_0 \left( e^{i\boldsymbol{k}_1\boldsymbol{r}} - e^{i\boldsymbol{k}_2\boldsymbol{r}} \right) e^{-i\omega t} \tag{6.5}$$

$\boldsymbol{k}_{12}\boldsymbol{r}$ (6.5)

$$\boldsymbol{k}_{12}\boldsymbol{r} = k(\mp x\cos\theta + z\sin\theta) \tag{6.6}$$

$$\boldsymbol{E} = -2i\boldsymbol{E}_0 \sin(kx\cos\theta) e^{i(kz\sin\theta - \omega t)} \tag{6.7}$$

$$\boldsymbol{E}_{\mathrm{Re}} = 2\boldsymbol{E}_0 \sin(kx\cos\theta)\sin(kz\sin\theta - \omega t) \tag{6.8}$$

$$k_z = k\sin\theta \tag{6.9}$$

$$v_\Phi = \frac{\omega}{k_z} = \frac{c}{\sin\theta} \tag{6.10}$$

$$k_\perp = k_x = k\cos\theta \tag{6.11}$$

$$x_n = \frac{m\pi}{k\cos\theta}; \quad m = 012\ldots \tag{6.12}$$

$x_1 = \pi/k\cos\theta = a$

$$\cos\theta_m = \frac{m\pi}{ka} = \frac{m\lambda}{2a} = \frac{m\pi c}{a\omega} \tag{6.13}$$

(6.3)

$$\cos\theta_m \leqslant 1 \tag{6.14}$$

$$k_{\mathrm{Kp}} = \frac{\pi}{a} \quad \omega_{\mathrm{Kp}} = \frac{c}{a} \quad \lambda_{\mathrm{Kp}} = 2a \tag{6.15}$$

$$k_z = k\sin\theta = \frac{\omega}{c}\sqrt{1 - (\omega_{\mathrm{Kp}}/\omega)^2} \tag{6.16}$$

$$v_\Phi = \frac{\omega}{k_z} = \frac{c}{\sin\theta} = \frac{c}{\sqrt{1-\cos^2\theta}} = \frac{c}{\sqrt{1 - (\omega_{\mathrm{Kp}}/\omega)^2}} \tag{6.17}$$

$v = d\omega/dk_z$

$$\omega = \sqrt{\omega_{\mathrm{Kp}}^2 + (ck_z)^2} \tag{6.18}$$

$$v_\Gamma = c\sqrt{1 - \left(\frac{\omega_{\mathrm{Kp}}}{\omega}\right)^2} = \frac{c^2}{v_\Phi} \tag{6.19}$$

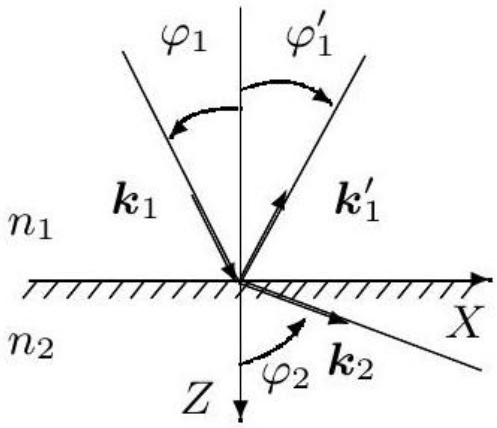
$$k_\perp^2 = k_x^2 + k_y^2 = \left(m\frac{\pi}{a}\right)^2 + \left(n\frac{\pi}{b}\right)^2 \quad mn - \tag{6.20}$$

$$k_z = \pm\sqrt{k^2 - k_\perp^2} \tag{6.21}$$

**4. -**

$$\boldsymbol{E} = \boldsymbol{a} e^{\mp \varkappa z} e^{i(k_x x + k_y y)} e^{-i\omega t} \tag{6.22}$$

$$E_{1\tau} = E_{2\tau} \quad D_{1n} = D_{2n} \quad H_{1\tau} = H_{2\tau} \quad B_{1n} = B_{2n} \tag{6.23}$$




$$\begin{gathered} \boldsymbol{E}_1 = \boldsymbol{a}_1 e^{ik_1(x\sin\varphi_1 + z\sin\varphi_1)} e^{-i\omega t} \\ \boldsymbol{E}_1' = \boldsymbol{a}_1' e^{ik_1'\left(x\sin\varphi_1' + z\sin\varphi_1'\right)} e^{-i\omega_1' t} \\ \boldsymbol{E}_2 = \boldsymbol{a}_2 e^{i(k_{2x}x + k_{2z}z)} e^{-i\omega_2 t} \end{gathered} \tag{6.24}$$

$$\boldsymbol{a}_{1\tau} e^{ik_1 x\sin\varphi_1} e^{-i\omega t} + \boldsymbol{a}_{1\tau}' e^{ik_1' x\sin\varphi_1'} e^{-i\omega_1' t} = \boldsymbol{a}_{2\tau} e^{ik_{2x}x} e^{-i\omega_2 t}$$

$$\omega_1 = \omega_1' = \omega_2 \tag{6.25}$$

$$k_1\sin\varphi_1 = k_1'\sin\varphi_1' = k_{2x} \tag{6.26}$$

$k_1 = k_1'$

$$\sin\varphi_1 = \sin\varphi_1'$$

$k_{2x} = k_2\sin\varphi_2$

$$\frac{\sin\varphi_2}{\sin\varphi_1} = \frac{k_1}{k_2} = \frac{v_2}{v_1} = \frac{n_1}{n_2} = n \tag{6.27}$$

$n = n_1/n_2$
$\sin\varphi_2 \leqslant 1$ (6.27) - $(n_1 > n_2)$ $\varphi_1$ $\varphi$

$$\sin\varphi \ = \frac{k_2}{k_1} = \frac{n_2}{n_1} = \frac{1}{n} \tag{6.28}$$

$\varphi_1$
(6.26)
$\varphi_1 > \varphi$

$$k_1\sin\varphi_1 > k_1\sin\varphi_{\pi\mathrm{p}} = k_1\frac{k_2}{k_1} = k_2$$

(6.25)

$$k_{2x} > k_2 \tag{6.29}$$

Но

$$k_2^2 = k_{2x}^2 + k_{2z}^2 \tag{6.30}$$

$k_{1y} = k_{2y} = 0$. (6.30) $k_{2x}$ (6.29)

$$k_{2z} = \pm\sqrt{k_2^2 - k_{2x}^2} = \pm i\sqrt{k_{2x}^2 - k_2^2} = \pm i\sqrt{k_1^2 \sin^2\varphi_1 - k_2^2} \tag{6.31}$$

$k_{2z}$ (6.22) $k_y = k_{2y} = 0 k_x = k_{2x} = k_1 \sin\varphi_1$ $\varkappa$

$$\varkappa = \sqrt{k_1^2 \sin^2\varphi_1 - k_2^2} \tag{6.32}$$

$\varphi_1 > \varphi$ ( ) (6.22) $\varkappa$ -
$\exp(-z/2\Lambda)$ $\Lambda = 1/2\varkappa$.

$$I \propto e^{-z/\Lambda} \tag{6.33}$$

$\Lambda$ (6.32) (6.28)

$$\Lambda = \frac{1}{2\sqrt{k_1^2 \sin^2\varphi_1 - k_2^2}} = \frac{1}{2k_2\sqrt{n^2 \sin^2\varphi_1 - 1}} = \frac{\lambda_2}{4\pi\sqrt{n^2 \sin^2\varphi_1 - 1}} \tag{6.34}$$

(6.23). (6.33) (6.34)
(6.23) -
$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$
-
(6.33).
$T \to 1$ $R \to 0$. $R \to 1$ $T \to 0$. $R + T = 1$ -
( )
$\alpha$

1. . IV. - 2005. §§12627333563 – 66.

2. . . . 2 1.
/. - 2001. III . 6§6.7; 7.

3. . - 2003. 13 16 135 137 139.

4. . 2. - 1972. §§5.75.8.

5. . - 1978. II.

## 16.2.2 Polarization, Crystal optics from labnik

**1.**

$$\operatorname{div} \boldsymbol{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0$$

$x$ $y$ $\partial E_z/\partial z = 0$ . $E_z$

$$(\operatorname{rot} \boldsymbol{H})_z = \frac{\partial H_y}{\partial x} + \frac{\partial H_x}{\partial y} = \frac{1}{c}\frac{\partial E_z}{\partial t}$$

. $\partial E_z/\partial t = 0$
$E_z$ $H_z = 0$.
$x$ - $y$ $\boldsymbol{E}$ $\boldsymbol{H}$ $E_z = H_z = 0$
$E_x$ $H_y$ - $E_y$ $H_x$

$$\begin{cases} (\operatorname{rot} \boldsymbol{E})_y = \dfrac{\partial E_x}{\partial z} = -\dfrac{1}{c}\dfrac{\partial H_y}{\partial t} \\ (\operatorname{rot} \boldsymbol{H})_x = -\dfrac{\partial H_y}{\partial z} = \dfrac{1}{c}\dfrac{\partial E_x}{\partial t} \end{cases}$$

$$\begin{cases} (\operatorname{rot} \boldsymbol{E})_x = \dfrac{\partial E_y}{\partial z} = \dfrac{1}{c}\dfrac{\partial H_x}{\partial t} \\ (\operatorname{rot} \boldsymbol{H})_y = \dfrac{\partial H_x}{\partial z} = \dfrac{1}{c}\dfrac{\partial E_y}{\partial t} \end{cases}$$

$E_x = \mp H_y = f_x(ct \pm z)$ $E_y = \pm H_x = f_y(ct \pm z)$ $f_x$ $f_y$ -

$$E_x = H_y = A_x \cos(\omega t - kz + \varphi_x) \tag{7.1}$$

$$E_y = -H_x = A_y \cos(\omega t - kz + \varphi_y) \tag{7.2}$$

$\boldsymbol{E}$ $\boldsymbol{H}$ $(E_x E_y)$ . $\boldsymbol{H}$
$\boldsymbol{E}$ -
1808 -
(7.1) (7.2) $\boldsymbol{E}$ $\boldsymbol{H}$ $(E_x E_y)$. -
$\boldsymbol{E}$
..
(. )
()

**2.**

K
(-)
- - [1] (dchroos -) -
- ( )
$(\perp)$ $(\parallel)$

$$\rho_\perp = \frac{\sin^2(\varphi - r)}{\sin^2(\varphi + r)} \quad \rho_\parallel = \frac{\operatorname{tg}^2(\varphi - r)}{\operatorname{tg}^2(\varphi + r)}$$

$\varphi$ - $r$ -
$\varphi + r = 90°$ . $\rho_\parallel$

$$\varphi_{\mathrm{B}} = \operatorname{arctg}(n) \tag{7.3}$$

$n$ - $\rho_{\parallel}$
.
$(1-\rho_{\perp})\ \big(1-\rho_{\parallel}\big)$.
90° 100%.
$1/\lambda^4$.

**3.**

.

$$I = I_0 \cos^2 \varphi \tag{7.4}$$

$\varphi$ -
(..)

$$P = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{7.5}$$

$I_{\max}$ $I_{\min}$ - $\varphi$.

$$P = \frac{I_{\max}(\varphi\delta) - I_{\min}(\varphi\delta)}{I_{\max}(\varphi\delta) + I_{\min}(\varphi\delta)}$$

$\varphi$ 0 $\pi$ $\delta$ 0 $\lambda$. $\lambda/4$ (.. $\lambda/4$).
$\varphi$

**4.**

$$U = a_x x^2 + a_y y^2 + a_z z^2$$

$a_x a_y$ $a_z$
-
- $a_y = a_z = a_{\perp}$ $a_x = a_{\parallel}$. $x$
$\boldsymbol{P}$
$\boldsymbol{D}$ $\boldsymbol{E}$

$$\boldsymbol{P} = \alpha_{\perp}\boldsymbol{E}_{\perp} + \alpha_{\parallel}\boldsymbol{E}_{\parallel} \quad \boldsymbol{D} = \varepsilon_{\perp}\boldsymbol{E}_{\perp} + \varepsilon_{\parallel}\boldsymbol{E}_{\parallel} \tag{7.6}$$

$\alpha_{\parallel}\alpha_{\perp}$ - $\varepsilon_{\parallel} = 1 + 4\pi\alpha_{\parallel}\varepsilon_{\perp} = 1 + 4\pi\alpha_{\perp}$ -

$$\operatorname{rot}\boldsymbol{H} = \frac{1}{c}\frac{\partial \boldsymbol{D}}{\partial t} \quad \operatorname{rot}\boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t}$$

$$\boldsymbol{E} = \boldsymbol{E}_0 e^{i(\omega t - \boldsymbol{kr})} \quad \boldsymbol{H} = \boldsymbol{H}_0 e^{i(\omega t - \boldsymbol{kr})} \quad \boldsymbol{D} = \boldsymbol{D}_0 e^{i(\omega t - \boldsymbol{kr})}$$

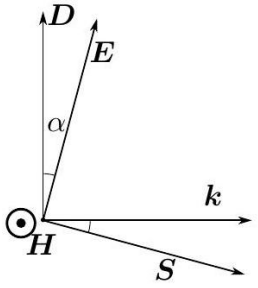
$\operatorname{rot}\boldsymbol{H} = -i\boldsymbol{k} \times \boldsymbol{H}\frac{\partial \boldsymbol{D}}{\partial t} = i\omega\boldsymbol{D}$ [2].
$\boldsymbol{DH}$ $\boldsymbol{E}$

[0]1

[02] $\frac{\partial}{\partial t} \to i\omega$ $\nabla$ $\nabla \to -i\boldsymbol{k}(\operatorname{rot}\boldsymbol{E} \equiv \nabla \times \boldsymbol{E} \to -i\boldsymbol{k} \times \boldsymbol{E}\, \operatorname{div}\boldsymbol{E} \equiv \nabla \cdot \boldsymbol{E} \to -i\boldsymbol{k} \cdot \boldsymbol{E})$.

$$\boldsymbol{D} = -\frac{c}{\omega} k \times \boldsymbol{H} \quad \boldsymbol{H} = \frac{c}{\omega} \boldsymbol{k} \times \boldsymbol{E} \tag{7.7}$$

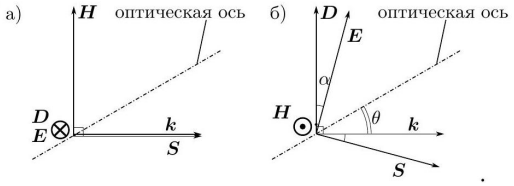
$\boldsymbol{DH}$ $\boldsymbol{k}$ . $\boldsymbol{D}$




. 7.1.
$\boldsymbol{DEkS}$ $\boldsymbol{H}$ $\boldsymbol{E}$. $\boldsymbol{E}$ $\boldsymbol{D}$ $\boldsymbol{k}$ . $\boldsymbol{H}$.
$\boldsymbol{DEk}$ $\boldsymbol{S} = \frac{4\pi}{c}\boldsymbol{E} \times \boldsymbol{H}$ 7.1.
(7.7) (7.6) $\boldsymbol{D} = \varepsilon_\perp \boldsymbol{E}_\perp + \varepsilon_\parallel \boldsymbol{E}_\parallel$ (. 7.2).

1. $\boldsymbol{D}$ $\boldsymbol{k}$ .

2. $\boldsymbol{D}$




. 7.2. () () -
-
$\boldsymbol{E}_\parallel = 0\boldsymbol{E} = \boldsymbol{E}_\perp$
$\boldsymbol{D} = \varepsilon_\perp \boldsymbol{E}_\perp$
$v = \omega/k$ $\boldsymbol{k}$

$$v_o = \frac{c}{\sqrt{\varepsilon_\perp}} = \frac{c}{n_o} \tag{7.8}$$

$n_o = \sqrt{\varepsilon_\perp}$ - [3].
$\theta$ $\boldsymbol{k}$. $D/E$ $D = |\boldsymbol{D}|$ $E_D = E\cos\alpha$ $\boldsymbol{E}$ $\boldsymbol{D}$

$$D = \varepsilon E_D \quad v_e = \frac{c}{\sqrt{\varepsilon}} = \frac{c}{n(\theta)} \qquad \varepsilon = \frac{1}{\frac{\sin^2\theta}{\varepsilon_\parallel} + \frac{\cos^2\theta}{\varepsilon_\perp}} \tag{7.9}$$

$n(\theta) = \sqrt{\varepsilon}$ - $\theta$ $\boldsymbol{S} = \frac{c}{4\pi}\boldsymbol{E} \times \boldsymbol{H}$ $\boldsymbol{k}$
(7.9). $v = \omega/k$ (7.7). 7.1 $v = \omega/k = cH/D$ $v = cE\cos\alpha/H$ $\alpha$ - $D$ $E$. $H$ $\alpha$

$$v = c\sqrt{\frac{E\cos\alpha}{D}} = c\sqrt{\frac{(\boldsymbol{E}\cdot\boldsymbol{D})}{D^2}} \tag{7.10}$$

($\boldsymbol{D} = \varepsilon\boldsymbol{E}$ (7.10) $v = c/\sqrt{\varepsilon}$.).
$\boldsymbol{D}$ $D_\parallel$ $D_\perp$ $D_\parallel = \varepsilon_\parallel E_\parallel$ $D_\perp = \varepsilon_\perp E_\perp$.

$$(\boldsymbol{ED}) = \varepsilon_\parallel E_\parallel^2 + \varepsilon_\perp E_\perp^2 = \frac{D_\parallel^2}{\varepsilon_\parallel} + \frac{D_\perp^2}{\varepsilon_\perp}$$

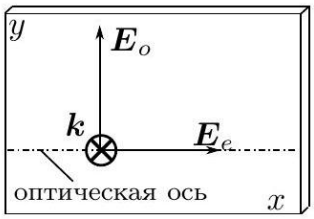
[0]3 ordinary () extraordinary ().

(. 7.2) $D_\parallel = D\sin\theta$ $D_\perp = D\cos\theta$
($\boldsymbol{ED}$) (7.10) $v_e$ $\theta$

$$v_e = c\sqrt{\frac{\sin^2\theta}{\varepsilon_\parallel} + \frac{\cos^2\theta}{\varepsilon_\perp}} \equiv \frac{c}{n(\theta)} \tag{7.11}$$

$\varepsilon \equiv n^2(\theta)$
(7.10) (7.11) (7.9).

**5.**




. 7.3
$\boldsymbol{E}$
- ($x$ 7.3). $n_o = \sqrt{\varepsilon_\perp}$ $-n_e = \sqrt{\varepsilon_\parallel}$((7.9) $\theta = 90°$).
$E_x$ $E_y$ $h$ $E_x$ $E_y$

$$\Delta\varphi = kh\,(n_e - n_o) \tag{7.12}$$

$k = 2\pi/\lambda$ -
( $E_x$ $E_y$).
$\lambda/4\lambda/2\lambda$ ($\Delta\varphi = \pi/2\pi 2\pi$).
(7.12)

**6.**

$U = ax^2$ $U = ax^2 + \beta x^3 + \gamma x^4$ $x_0 \neq 0$ $\partial^2 U/\partial x^2 = 2a + 6\beta x_0 + 12\gamma x_0^2$ $x$
$\beta \neq 0$ $x_0$
$\beta$ -
$\beta = 0$ $\gamma \neq 0$ $\varepsilon_\parallel$

## 16.3 Special professional methods in optics

### 16.3.1 Geometrical Optics by Born and Wolf

(мб почитаю, пока просто не нужно это)

**Приближение очень коротких длин волн по Борну и Вольфу**

**Общие свойства лучей**

**Принцип Ферма**

**Основные определения. Закон преломления и отражения. Принцип взаимности**

**Преломление (и отражение) на сферической поверхности**

**Фокусы сферической поверхности**

**Изображение матых предметов при преломлении на сферической поверхности**

**Увеличение. Теорема Лагранжа-Гельмгольца**

4.1.10

**Условие синусов Аббе**

(еще одно что-то совсем бесполезное, см. нпр. Кириченко)

**Центрированная оптическая система**

4.1.11

**Преломление в линзе. Общая формула линзы**

**Фокусные расстояния тонкой линзы**

**Изображение в тонкой линзе. Увеличение**

**Идеальные оптические системы**

### 16.3.2 Eikonal equation

**Eikonal main ideas and formulas (???)**

(??? don't understand yet)

### 16.3.3 Deeper Understanding of Corpuscular Optics

(как конкретно это рабоатет - про это 100 можно важных вопросов задать, напишу самые важные)

**переход в квантовую механику**

????
пока с ходу не скажу, как делается это

### 16.3.4 Specific interference

**Многолучавая интерференция**

(потом подумаю про это)

**Интерференция в случае разных интенсивностей источников**

(потом подумаю про это)

### 16.3.5 Specific dispersions

**ЛЛ8-103. Пространственная дисперсия**

**ЛЛ8-104. Естественная оптическая активность**

**ЛЛ8-105. Пространственная дисперсия в оптически неактивных средах**

**ЛЛ8-106. Пространственная дисперсия вблизи линии поглощения**

### 16.3.6 Specific reflectance

**Goos-Hänchen Effect (????)**

Это явление продольного сдвига (относительно траектории распространения волнового пучка) при полном внутреннем отражении. То есть волна не просто отражается, а еще и сдвигается.

(??? когда он используется??? мб он очень мелкий???? пока просто слышал об этом. сперва в теории напишу про это.)

### 16.3.7 Green's function method for optics (?)

(for PDF in optics it is used a lot)

### 16.3.8 $S$-matrix for optics (?)

(there are special theories with it)

### 16.3.9 On PT-symmetry in optics (!!!?!?!??!?)

(статья пухова и др. мб на самом деле там много всего??? пока пусть как заготовка будет. пока я все-таки не въезжаю, что тут важного?)

A A Zyablovsky, A P Vinogradov, A A Pukhov, A V Dorofeenko, A A Lisyansky

DOI 10.3367/UFNe.0184.201411b.1177

2. PT symmetry. Basic concepts and definitions 1064 2.1 Parity and time reversal operators; 2.2 Necessary and sufficient condition for real eigenvalues of a PTsymmetric Hamiltonian; 2.3 Phase transition in PT-symmetric systems

3. PT-symmetric optical systems 1067 3.1 PT-symmetry concept for optical systems; 3.2 Restrictions imposed by the causality principle. Kramers±Kronig relations and the possibility of PT-symmetric optical systems existing in a finite frequency range

4. Two-dimensional PT-symmetric optical systems 1068 4.1 Analogy between the two-dimensional Helmholtz equation and the one-dimensional SchroÈdinger equation; 4.2 Phase transition in two-dimensional PT-symmetric optical systems; 4.3 'Loss-induced transparency'; 4.4 Refraction asymmetry in the PT-symmetric optical system

5. One-dimensional PT-symmetric optical systems 1072 5.1 Phase transition in a one-dimensional optical system; 5.2 Lasing in PT-symmetric optical systems and the 'hidden' phase transition in a one-dimensional optical system

6. 'Quasi-diode' PT-symmetric optical systems. The reciprocity principle

## 16.3.10 Fourier Optics and Holography (??)

(уменьшил приоритет этих тем, потому что как-то вообще нигде мне это не встречалось, хотя в задачнике это и известные темы.)

**Метод Релея**

**Дифракция на периодических структурах**

**Дифракция на синусоидальной решетке**

**Эффект саморепродукции**

**Функция пропускания клина**

**Функция пропускания линзы**

**Принципы фурье-оптики**

**Фурье-плоскость**

**Теория Аббе формирования изображений**

**Принципы пространственной фильтрации**

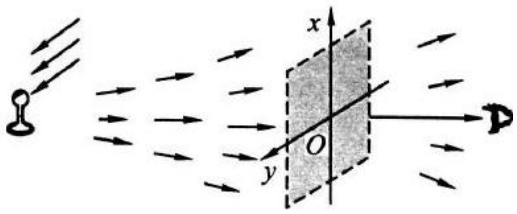
**Методы наблюдения фазовых объектов**

**Метод фазового контраста Цернике**

**Метод тмного поля**

**3. Holography**

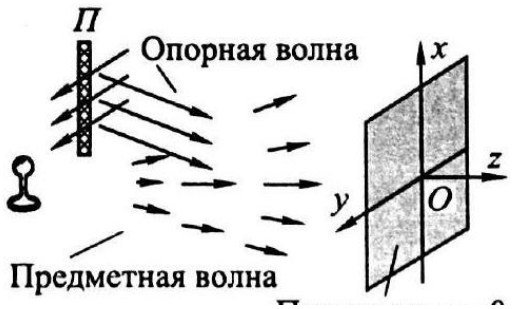
(
) pe . (.)
( 3.47)
$z = 0$




. 3.47.

$$f_{\text{П}}(xy) = a(xy)e^{i\varphi(xy)}$$

$z = 0$ ( )
. $I(xy) = |a(xy)|^2$. $\varphi(xy)$
(1948)
$f\,(xy)$ $f_{\text{o}}(xy) = a_{\text{o}}(xy)e^{i\varphi_{\text{o}}(xy)}$ (. 3.48). - 3.48 -




$z = 0$
. 3.48. П
.

$$f(xy) = f_{\text{П}}(xy) + f_{\text{о}}(xy)$$

( )

$$t(xy) \propto I(xy) = |f_{\text{П}}(xy) + f_{\text{О}}(xy)|^2$$

$$t \propto |f_{\text{П}}|^2 + |f_{\text{о}}|^2 + f_{\text{П}} f_{\text{о}}^* + f_{\text{П}}^* f_{\text{О}} = a^2 + a_{\text{о}}^2 + 2aa_{\text{о}} \cos(\varphi - \varphi_{\text{о}})$$

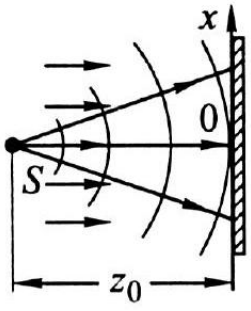
(
) $\varphi(xy)$.

**3.1.**

$S$ .

$$f_{\text{П}} = \frac{a_{\text{П}}}{r} e^{ikr}$$

$r = \sqrt{z_0^2 + x^2 + y^2}$ - $S$ $(xy)$




. 3.49.
$z$ $z = 0$ (. 3.49)

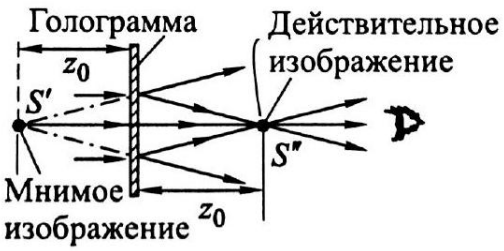
$$f_{\text{о}} = ae^{ikz}$$

.
$z = 0$ $f_{\text{о}} = a$.
.. $f \approx ae^{ikr}$.

$$f = ae^{ikr} + a.$$

$$t(xy) \sim I(xy) = \left|a + ae^{ikr}\right|^2 \tag{3.83}$$

-




-

. 3.50. ()
$z = 0$ (. 3.50).
. $f_-(xy) = 1$.

$$f_+(xy) = f_-(xy)t(xy) = \left|a + ae^{ikr}\right|^2 \tag{3.84}$$

(3.84)

$$f_+(xy) = 2a^2 + a^2e^{ikr} + a^2e^{-ikr} = 2a^2(1 + \cos(kr)) \tag{3.85}$$

$z = 0_+$ $z > 0$ $z = 0_+$ $z > 0$.
$2a^2$ $f_1 = 2a^2e^{ikz}$ - $z > 0$ $z$. $f_2 = a^2e^{ikr}$ - $z = 0$

$z_0$ $z > 0$ $f_2$ ( $S$). $(z > 0)$ $z_0$ $S'$ . $S'$
- - -
(3.85) $f_3 = a^2e^{-ikr}$. $f_2$ $f_2$ $(xy)$ $z = 0$ $f_3$ $(xy)$ $z = 0$ $S''$ $z_0$ (. 3.50). $f_3$
$S''$. - $z_0$ -
;
; $z$
(3.85)

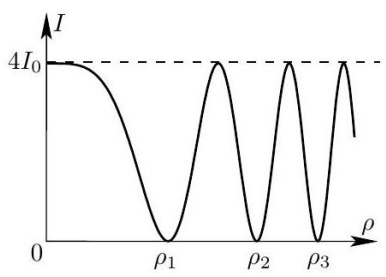
$$r \approx \sqrt{z_0^2 + x^2 + y^2} \approx z_0 + \frac{x^2 + y^2}{2z_0} = z_0 + \frac{\rho^2}{2z_0}$$

$\rho^2 = \sqrt{x^2 + y^2}$ - $(xy)$ $(xy)$

$$\varphi = kr \approx kz_0 + \frac{k}{2z_0}\rho^2 \tag{3.86}$$

$z$ $z = 0$ $\varphi = \frac{k}{2z_0}\rho^2$.
() $(x = 0y = 0)$

$$I(\rho) = 2a^2\left(1 + \cos\frac{k\rho^2}{2z_0}\right) \tag{3.87}$$



-
. 3.51. $I(\rho)$ $I(\rho)$ 3.51. () $\rho_1$ $k\frac{\rho_1^2}{2z_0} = \pi$ ($\cos k\frac{\rho_1^2}{2z_0} = -1$ $I(\rho_1) = 0$) $\rho_1 = \sqrt{\lambda z_0}$.

$$\rho_m = \sqrt{m\lambda z_0} \tag{3.88}$$

($m$ - $m$ -)
(3.85) 3.52 ($S''$ -)
$S'$



. 3.52.
$z_1 > z_0$ (3.86)

$$f_-(\rho) = e^{-i\frac{k}{2z_1}\rho^2} \tag{3.89}$$

$z = 0$

$$f_+(\rho) = f_-(\rho)t(\rho)$$

(3.89) (3.83)

$$f_+(\rho) = e^{-i\frac{k}{2z_1}\rho^2} + e^{-i\frac{k}{2}\left(\frac{1}{z_1}+\frac{1}{z_0}\right)\rho^2} + e^{-i\frac{k}{2}\left(\frac{1}{z_0}-\frac{1}{z_1}\right)\rho^2}$$

$b_2$

$$\frac{1}{b_2} = \frac{1}{z_0} + \frac{1}{z_1}$$

. $b_3$

$$\frac{1}{b_3} = \frac{1}{z_0} - \frac{1}{z_1}$$

**3.2.**

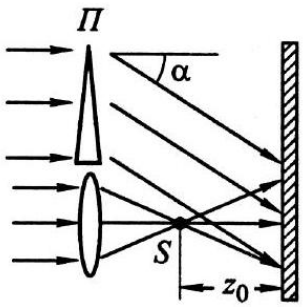
$S''$
( ) ((3.85))
$D$ (3.74). -

$$\Delta x \sim \frac{\lambda}{D} z_0 \tag{3.90}$$

$z_0$ - $D$ - $\Delta x$
$\Delta x$ $z_0$
.
10

**3.3.**

$S'$ $S''$ (. 3.50). $z$ 1962
10 () ()



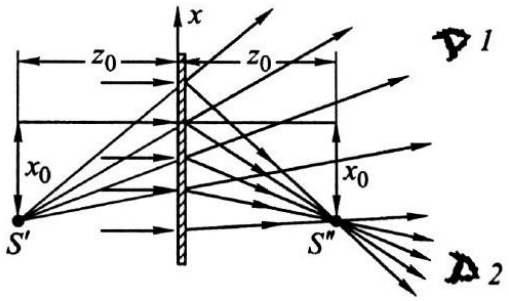

. 3.53.
3.53 $S$.

$$f_{\text{П}}(xy) = ae^{i\frac{k}{2z_0}\left(x^2+y^2\right)}$$

. $f_0 = ae^{iu_0x}$ $u_0 = k\sin\alpha$ - $xy$

$$t(xy) \sim \left|ae^{iu_0x} + ae^{i\frac{k}{2z_0}\left(x^2+y^2\right)}\right|^2$$

$$f_3(xy) \sim be^{i\frac{k}{2z_0}\left((x+z_0\sin\alpha)^2+y^2\right)} \quad f_4(xy) \sim be^{-i\frac{k}{2z_0}\left((x+z_0\sin\alpha)^2+y^2\right)}$$




. 3.54.
$b$ - $f_3(xy)$ $z = 0$ $x_0 = -z_0\sin\alpha y = 0$ $z_0$ $S'$ (. 3.54). $f_4(xy)$ $S''$
$z_0$ $z$ $x_0 = -z_0\sin\alpha$.
(1 2 3.54)
$\alpha$ (.)

1. . . . 2 1.
/. - 2001. 3 . 8.

2. . T. IV. - 2005. IV.

3. . - 2003. VIII-XI.

4. . - 1986. 7.

5. - 1973. 8.

**О голографии**

**Схемы голографии**

**Запись и восстановление плоской волны**

**Голограмма с наклонным опорным пучком**

**Разрешающая способность голограммы**

**Ограничения на допустимую ширину спектра**

**Голограмма Габора точечного источника**

**Объмные голограммы**

## 16.3.11 Talbot Effect, Other Phenomena of Diffraction (???)

**Эффект саморепродукции**

Рассмотрим экран с периодически расположенными отверстиями (например, дифракционную решетку). Будем для простоты говорить только о структуре на экране, периодичной вдоль оси $x$ и однородной вдоль оси $y$. Обобщение на случай периодичности и вдоль оси $y$ не представляет труда.

Пусть слева на экран падает плоская волна вдоль оси $z$ (рис. 8.2.2):

$$A(x,z) = A_0 e^{ikz}, z < 0$$

Вследствие периодичности структуры экрана справа от него при $z = +0$ поле может быть представлено в виде ряда Фурье:

$$A_S(x) \equiv A(x,z)|_{z=+0} = D(x)A(x,z)|_{z=-0} = \sum_{n=0}^{\infty} a_n e^{iu_n x}$$

где

$$u_n = n\Omega, \Omega = 2\pi/d,$$

$d$ - период структуры (вдоль оси $x$). Каждое слагаемое в сумме (8.2.2) порождает в области $z > 0$ волну

$$a_n e^{iu_n x + iw_n z}, w_n = \sqrt{k^2 - u_n^2}$$

так что

$$A(x,z)|_{z>0} = \sum_{n=0}^{\infty} a_n e^{iu_n x + iw_n z}.$$

Если период структуры $d \gg \lambda$, то есть $\Omega \ll k$, то для не слишком болыших номеров гармоник $n$ можно считать $u_n \ll k$ и записать:

$$w_n \approx k - u_n^2/2k.$$

Соответственно выражение для $n$-й волны принимает вид

$$a_n e^{ikz} \exp\left(iu_n x - i\frac{u_n^2}{2k} z\right)$$

Если на некотором расстоянии $z$ от экрана (периодической структуры) будет выполнено условие

$$\frac{u_n^2}{2k} z = 2\pi p$$

($p$ - целое число), или

$$n^2\frac{\Omega^2}{2k}z = 2\pi p,$$

то в соответствии с (8.2.22), (8.2.23) окажстся

$$A(x,z) = e^{ikz}\sum_{n=0}^{\infty} a_n \exp\left(iu_n x - i\frac{u_n^2}{2k}z\right) =$$

$$= e^{ikz}\sum_{n=0}^{\infty} a_n \exp\left(iu_n x\right) = e^{ikz}A_S(x).$$

Это значит, что с точностью до фазового множителя $e^{ikz}$ на таком расстоянии воспроизводится структура исходного волнового поля $A_S(x)$, поскольку относительный набег фаз всех слагаемых в сумме (8.2.20) оказывается кратным $2\pi$. Условие (8.2.24) выполняется в точках

$$z = mz_T, z_T = 2\pi\frac{2k}{\Omega^2} = \frac{2d^2}{\lambda}, m = 0, 1, 2, \ldots$$

Действительно, в этих точках

$$\frac{u_n^2}{2k}z = n^2\frac{\Omega^2}{2k}mz_T = 2\pi mn^2 = 2\pi p,$$

$p = mn^2$ - целое число. Заметим, что хотя каждое из слагаемых в сумме (8.2.20) приобретает свой набег фазы, но эти набеги для всех слагаемых кратны $2\pi$.

Таким образом, периодическая структура $A_S(x)$ периодически самовоспроизводится (повторяется) при условии

$$z = mz_T = m\frac{2d^2}{\lambda}.$$

Оценим максимальное число «копий» структуры, которое можно наблюдать при использовании решётки длиной $D$. Эффект саморепродукции возникает благодаря интерференции волн, распространяющихся от решетки под разными углами. Наименьший угол с осью $z$ образуют 0 -й и)1)-й главные дифракционные максимумы. Распространяясь от решетки

конечных размеров $D$, эти три волны перестают перекрываться на расстоянии (рис. 8.2.3):

$$L = \frac{D}{2}\operatorname{ctg}\theta$$

$\theta$ - угол дифракции, определяемый условием

$$d\sin\theta = \lambda.$$

Считая $\lambda \ll d$, заключаем, что $\theta \approx \lambda/d \ll 1$. Поэтому

$$L \approx \frac{D}{2\theta} \approx \frac{Dd}{2\lambda}.$$

На этом расстоянии число плоскостей саморепродукции составляет

$$N \sim \frac{L}{z_T} \approx \frac{Dd/2\lambda}{2d^2/\lambda} = \frac{D}{4d}$$

В действительности для установления периодичности объекта по его «копии» достаточно иметь пересечение только 0-го дифракционного максимума и хотя бы одного

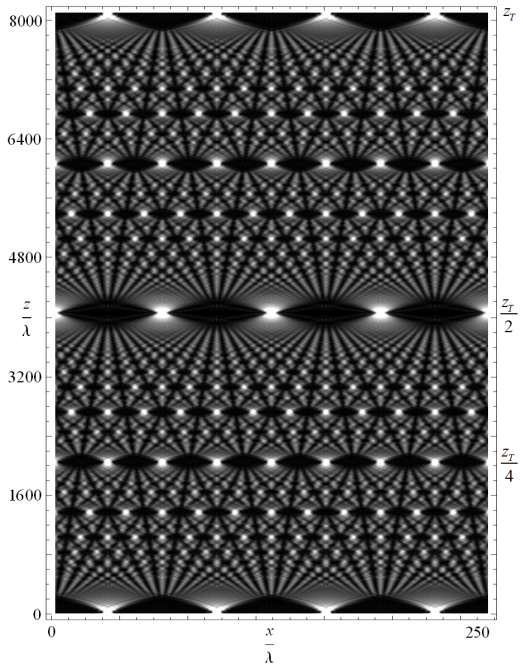
из боковых лепестков: $(+1)$-го или $(-1)$-го. Соответственно расстояние $L$ можно взять вдвое большее: $L_1 = D\operatorname{ctg}\theta$. Это даёт вдвое большее число наблюдаемых «копий» решётки:

$$N \sim \frac{L_1}{z_1} \approx \frac{D}{2d}.$$

Явление саморепродукции наблюдалось в 1836 г. английским физиком и химиком Генри Фоксом Талботом и получило название эффекта Талбота.

Рассмотренный эффект имеет место на расстояниях $z \geq z_T = 2d^2/\lambda$. По отношению к решётке в целом эта область попадает в зону дифракции Френеля. Действительно, все наблюдаемые «копии» находятся на расстоянии, не превышающем $L = D\operatorname{ctg}\theta \approx Dd/\lambda$. Условие дифракции Френеля имеет вид $D^2/\lambda > L$. Согласно предыдущему равенству это сводится к очевидному требованию $D > d$: размер решётки должен превышать её период.

В ближней зоне $(z < z_T)$ дифракционная картина оказывается весьма сложной. На рис. 8.2.4 приведено распределение интенсивности в области $0 \leq z \leq z_T$. Эту картину называют «ковёр Талбота».




**Функция пропускания линзы**

Рассмотрим на основе метода Рэлея действие линзы. Считаем линзу тонкой и ограничиваемся приближением параксиальных пучков (когда лучи образуют малые углы с главной оптической осью).

Как известно, оптические длины лучей, идущих от линзы к фокусу, одинаковы. Это значит, что плоская волна преобразуется в сферическую, сходящуюся к фокусу. Сформировавшийся сферический волновой фронт показан на рис. 8.3.1 штриховой линией. Отсюда следует, что лучи, наталкивающиеся на линзу, приобретают фазовую задержку до момента достижения сходящегося волнового Фронта:

$$\Delta\varphi = k\left(f - \sqrt{f^2 + \rho^2}\right) \approx -\frac{k\rho^2}{2f}$$

Здесь $\rho$ - расстояние от точки входа луча в линзу до оптической оси. Таким образом,

функцию пропускания линзы можно записать в виде

$$D = e^{i\Delta\varphi} = \exp\left(-\frac{ik\rho^2}{2f}\right).$$

Если линза рассеивающая, то замена $f \to -f$ даёт

$$D = \exp\left(\frac{ik\rho^2}{2f}\right)$$

В качестве применения формулы (8.3.2) рассмотрим действие собирающей линзы на плоскую волну, распространяющуюся параллельно оптической оси:

$$A = A_0 e^{ikz}.$$

Запишем волну, прошедшую через линзу:

$$A_S(x,y) = D(x,y)A_0 = A_0 \exp\left(-\frac{k\,(x^2+y^2)}{2f}\right).$$

Чтобы найти поле в области $z > 0$, нужно выполнить пространственное разложение Фурье функции (8.3.5):

$$a(u,v) = \int\limits_{x=-\infty}^{\infty}\int\limits_{y=-\infty}^{\infty} e^{-i(2x+vy)} A_S(x,y)dxdy.$$

Для вычисления воспользуемся следующим равенством:

$$\int\limits_{k_x=-\infty}^{\infty} e^{-i\left(2x+kx^2/2f\right)}dx = (1-i)\sqrt{\frac{\pi f}{k}}\exp\left(i\frac{f}{2k}u^2\right).$$

Аналогичное равенство имеет место и при интегрировании по $dy$. Соответственно получаем

$$a(u,v) = \frac{2\pi f}{ik}A_0 \exp\left[i\frac{f}{2k}\left(u^2+v^2\right)\right]$$

Чтобы восстановить поле за линзой, нужно продлить в область $z > 0$ каждую из пространственных гармоник:

$$A(x,y,z) = \int\limits_{u=-\infty}^{\infty}\int\limits_{v=-\infty}^{\infty} e^{iux+ivy+iz\sqrt{k^2-u^2-v^2}} a(u,v)\frac{dudv}{(2\pi)^2}.$$

Имея в виду параксиальное приближение: $|u| \ll k, |v| \ll k$, положим

$$\sqrt{k^2-u^2-v^2} \approx k - \frac{u^2+v^2}{2k}.$$

Таким образом, задача сводится к вычислению интеграла

$$A(x,y,z) = A_0 e^{ikz}\frac{2\pi f}{ik}\int\limits_{k_x=-\infty}^{\infty}\int\limits_{y}^{\infty}\exp\left[iux+ivy-i\frac{u^2+v^2}{2k}(z-f)\right]\frac{dudv}{(2\pi)^2}.$$

Расчёт с использованием равенства типа (8.3.7) даёт

$$A(x,y,z) = A_0 e^{ikz}\frac{f}{f-z}\exp\left(-i\frac{k\rho^2}{2(f-z)}\right), z < f$$

Заметим, что допустимость использования параксиального приближения (8.3.10) в рассматриваемой ситуации связана с быстрой сходимостью интеграла (8.3.11).

Таким образом, мы получили очевидный результат: интенсивность вдоль оптической оси меняется по закону

$$I = \frac{I_0}{(1 - z/f)^2},$$

достигая бесконечного значения непосредственно в фокусе линзы $z = f, \rho = 0$. В рассмотренной постановке задачи линза предполагалась неограниченной. Поэтому фокусное пятно стягивается практически в точку - эффекты дифракции оказываются пренебрежимо малыми.

# 17 Other Models

## 17.1 Other Well-known Models of Optics

### 17.1.1 Optical Anisotropy (!??!)

**Пьезооптический эффект**

**Эффект Поккельса**

В некоторых средах при наложении постоянного электрического поля возникает двойное лучепреломление. Изменение показателя преломления пропорционально величине поля

$$n_e - n_o = bE.$$

В этой связи данное явление называют также линейным электрооптическим эффектом.

**Эффект Керра**

В изотропных средах может наблюдаться возникновение двойного лучепреломления под действием внешнего электрического поля - эффект Керра (квадратичным электрооптическим эффектом). Этот эффект не может быть линейным по полю вследствие центральной симметрии среды.

Если до наложения поля среда была неполяризованной и изотропной, то разность показателей преломления обыкновенного и необыкновенного лучей является чтной функцией напряжнности поля $E_0$. Это значит, что в разложении по степеням $E_0$ должны присутствовать члены лишь чтных порядков, начиная с квадратичного

$$n_e - n_0 = B\lambda_0 E_0^2$$

где $\lambda_0$ - длина волны света в вакууме. Коэффициент $B$ зависит от свойств вещества и называется постоянной Керра.

**Эффект Фарадея**

**О других способах создания оптической анизотропии**

### 17.1.2 Nonlinear optics: basics (????)

(интересный раздел, нужно будет - неделю просижу на нем, разберу)

**ЛЛ8-107. Преобразование частот в нелинейных средах**

**ЛЛ8-108. Нелинейная проницаемость**

**ЛЛ8-109. Самофокусировка**

**ЛЛ8-110. Генерация второй гармоники**

**ЛЛ8-111. Сильные электромагнитные волны**

**ЛЛ8-112. Вынужденное комбинационное рассеяние**

### 17.1.3 Photonic Crystals

(see articles about it)
https://www.youtube.com/watch?v=TJcB2cJ3_k4

### 17.1.4 Scattering on Crystals

(that from solid state and ED theory)

**Основные формулы (??)**

### 17.1.5 X-ray Diffraction (?)

(потом пропишу, с твердым телом тут связка большая!)

**Рентгеновское излучение**

**Формула Брэга-Вульфа**

Введм атомные плоскости, то есть плоскости, содержащие большое число атомов (рис. 6.9.1, слева). Обозначим $d$ расстояние между парой соседних плоскостей, а $\boldsymbol{\theta}$ - угол скольжения луча.

Отражение волны от кристалла можно рассматривать как результат интерференционного усиления волн, идущих от разных атомных плоскостей. Рассмотрим две соседние параллельные плоскости (рис. 6.9.1, справа). Разность хода лучей $MON$ и $M'O'N'$ равна

$$\Delta = PO'Q = 2d\sin\theta.$$

Условие интерференционного усиления этих волн имеет вид

$$2d\sin\theta = m\lambda, \quad m = \pm1, \pm2, \ldots$$

Полученное соотношение называется формулой Брэгга-Вульфа.

**Формулы Лауэ**

Кристалл - это трхмерная структура, периодическая по трм направлениям. Обозначим периоды по осям кристалла 1, 2, 3 соответственно $d_1.d_2, d_3$. Рассмотрим цепочку атомов 1. Пусть излучение падает на не под углом $\alpha_0$, а наблюдается под углом $\alpha$ (рис. 6.9.4). Тогда разность хода лучей, дифрагировавших на соседних атомах, равна

$$\Delta = \Delta_1 - \Delta_0, \Delta_1 = d_1\cos\alpha, \Delta_0 = d_1\cos\alpha_0.$$

Отсюда получаем условие интерференционного усиления волн

$$\Delta = m_1\lambda, m = 0, \pm1, \pm2, \ldots$$

или

$$d_1 (\cos\alpha - \cos\alpha_0) = m_1\lambda.$$

Заметим, что в силу условия $|\cos\varphi| \leq 1$ число дифракционных максимумов по каждому направлению ограничено $|m_k| \leq 2d_k/\lambda$, причм более одного максимума возможно, если $\lambda < 2d_k$.

$$d_2 (\cos\gamma - \cos\gamma_0) = m_2\lambda$$
$$d_3 (\cos\gamma - \cos\gamma_0) = m_3\lambda.$$

**Эквивалентность формул Брэгга-Вульфа и Лауэ**

**Лауэграмма**

### 17.1.6 Enhancement of Reflection

(brag mirror and maybe some other things)

### 17.1.7 Enhancement of Absorption

(there are also articles about it)

### 17.1.8 Amplification of Waves

(there are also articles about it)

## 17.2 Quantum Optics: Other Well-known Models

## 17.3 Special Theoretical Methods

### 17.3.1 Path integral for optics: formalism

(see Feldbrugge for now. a big topic, advanced!)

### 17.3.2 Path integral for optics: diffraction

(here there are main about applications, how to count diffraction, etc)

### 17.3.3 Pressure of Light, Optical Tweezers

(! принципиально важные свойства укажу, все подробно про них - см. теорию)

**Давление света (????)**

(пока тут, потом мб перенесу в другое место.)

### 17.3.4 Geometrical gravitational lensing

(see a special notes about it)

### 17.3.5 Attosecond pulses

(later maybe I'll learn how it works)

# Part VI
# Optics in Continuous Media

## 18 Propagation of Light Across the Boundary of Two Media

### 18.0.1 Fresnel Formulas

(see Landsberg for example)

**Отражение и преломление света на границе двух диэлектриков**

**Отражение и преломление на границе двух диэлектриков. Формулы Френеля**

**Поляризация света при прохождении через границу двух диэлектриков. Наглядная интерпретация закона Брюстера**

### 18.0.2 Total Internal Reflection

(see Landsberg for example)

**Явление полного внутреннего отражения**

**Исследование отраженной волны**

.

**Эллиптическая поляризация**

**Исследование преломленной волны**

## 19 Interaction of Waves with Matter and Molecular Optics

Типичные оптические задачи есть задачи вида “волна падает на такой-то прибор”, сразу же при этом хочется знать теорию взаимодействия волн с веществом, которую тут и приведем

(по Ландсбергу если что многое догружу)

### 19.1 Fundamentals of Optical Wave Scattering

### 19.1.1 Rayleigh Scattering

(по идее тут его всего и обсудим)

### 19.1.2 Scattering Cross Section and Angular Distribution

дальше все скипаю, см [1].

**вкратце**

Излучение диполя такое: (почему???)

$$Q = \frac{2}{3c^3}\ddot{\mathbf{p}}^2$$

????
получаем *закон Релея* :

$$Q = \frac{(2\pi e)^4 c\beta^2 E_0^2}{3\lambda^4} \sim \frac{1}{\lambda^4}$$

но кстати знаем, что это другими словами дифракция на объектах размеров порядка длины волны. !!!!!

### 19.1.3 Absorption

(????)

### 19.1.4 Polarization of Scattered Waves

(??????)

### 19.1.5 Other Effects in Scattering

**Об эффекте Комптона**

(?? он с оптикой связан или нет? вообще по идее такое тоже может быть, чего про это не упоминается особо в оптике?)

**Об аномалиях в теории рассеяния**

(краснок что-то про это писал, пока не смотрел.)

## 19.2 Propagation of Waves Through Matter

### 19.2.1 Refractive Index

все про него.
(нужна предыдущая теория, на нее ссылки!!!)

**Основные свойства показателя преломления**

и я так-то хз пока что

**Вывод показателя преломления**

Ответим, откуда он вообще взялся?

**теория по сути**

Поставим задачу об электромагнитном поле в отсутствие источников в некоторой линейной среде, которую можно охарактеризовать константами $\varepsilon, \mu$.

Конечно, безграничность такой среды еще более проблематична, нежели безграничность пустого пространства в предыдущей главе.

Но как всегда в физических задачах, такую идеализацию следует понимать в смысле некоторого сильного неравенства.

В данном случае условие применимости следует из ответа: характерный размер образца диэлектрика или характерный масштаб его неоднородности должен многократно превосходить длину волны.

Вместо (10.12)(???) получаем систему уравнений:

$$\operatorname{rot} \mathbf{B} = \varepsilon\varepsilon_0\mu\mu_0 \frac{\partial \mathbf{E}}{\partial t}; \quad \operatorname{rot} \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

Соответственно, и (10.13),(10.14) изменяются лишь в одном коэффициенте:

$$\frac{\partial^2 \mathbf{B}}{\partial t^2} = \frac{c^2}{\varepsilon\mu}\nabla^2 \mathbf{B}; \quad \frac{\partial^2 \mathbf{E}}{\partial t^2} = \frac{c^2}{\varepsilon\mu}\nabla^2 \mathbf{E}$$

мы опять пришли к волновому уравнению, и его решением снова будет монохроматическая плоская волна (10.15) при том же законе дисперсии (10.6), но скорость распространения волн будет меньше, чем в вакууме:

$$\frac{\omega}{k} = \frac{c}{n}, \quad \text{где} \quad n = \sqrt{\varepsilon\mu}$$

Величина $n$ называется коэффициентом преломления - как мы вскоре убедимся, она как раз соответствует коэффициенту преломления в оптике.

**Оценки показателя преломления**

Природа устроила так, что во всех без исключения средах, прозрачных в оптическом диапазоне, величина $\mu$ с хорошей точностью равна единице, поэтому в оптике обычно без оговорок полагают $n = \sqrt{\varepsilon}$.

**Диэлектрическая проницаемость не везде она??**

?????? да ну???

Не следует, однако, отождествлять диэлектрическую проницаемость, которую мы вводили в электростатике, с константой $\varepsilon$ в уравнениях (11.1)-(11.3).

Это одно и то же по физическому смыслу, но не по величине.

**Количественные оценки показателя преломления**

Известно, что у чистой воды в статике $\varepsilon = 81$. Казалось бы, отсюда следует $n = 9$, но в действительности показатель преломления воды равен примерно 1,33 в красной области спектра и $1,34$ - в фиолетовой.

Не следует преувеличивать полноты информации, представленной нашим повседневным опытом.

Порядок цветов в радуге любой природы таков, что из него следует рост $n$ с частотой.

Но во-первых, как раз при $\omega = 0$ величина $\sqrt{\varepsilon}$ оказывается почти в 7 раз больше, чем в оптике, Т. е. при $\omega \sim 10^{15}\text{с}^{-1}$. А во-вторых, для электромагнитных волн существенно

более высокой частоты - жестких рентгеновскиХ лучей - коэффициент преломления любого вещества оказывается очень близок к единице.

Дисперсия электромагнитных волн - явление сложное, функция $n(\omega)$ немонотонна и включает области непрозрачности.

Но и тот простой случай, который представляют уравнения $(11.1) - (11.3)$, реализуется во многих ситуациях, представляющих практический интерес, а потому стоит изучения.

Приведем некоторые соотношения, демонстрирующие отличие волн в прозрачной среде с постоянным коэффициентом преломления, от волн b bakyyme.

Как уже было отмечено в гл.

10, восприятие цвета обусловлено частотой.

Мы знаем, что в стационарной оптической системе чистый спектральный цвет не меняется при переходе из одной среды в другую.

(Это не относится к нестационарным системам, включающим, например, движущиеся зеркала или среды с переменным во времени коэффициентом преломления, а также к нелинейным средам.) Факт сохранения частоты в линейной стационарной системе - по сути, закон сохранения энергии для световых квантов.

Не касаясь этой стороны дела, мы его примем в качестве экспериментального.

Как следствие, разная фазовая скорость волн в разных средах обусловлена зависимостью от $n$ волнового числа или, что то же, длины волны:

$$k(n) = nk_0; \quad \lambda(n) = \frac{\lambda_0}{n}$$

где $k_0 = 2\pi/\lambda_0 = \omega/c$ - характеристики волны данной частоты в вакууме.

Из соотношений (11.4) естественно следует определение относительного коэффициента преломления двух сред:

$$n_{12} = \frac{k\,(n_1)}{k\,(n_2)} = \frac{n_1}{n_2}$$

И это определение соответствует стандартному, принятому в оптике.

Получим далее соотношения между величинами электрического и магнитного полей в волне, аналогичные (10.16).

Подставим решения вида (10.15) в уравнения (11.1) и постараемся записать ответ в форме, как можно более симметричной:

$$B = \frac{n}{c}E = \sqrt{\varepsilon\varepsilon_0\mu\mu_0}E; \quad H = \sqrt{\frac{\varepsilon\varepsilon_0}{\mu\mu_0}}E.$$

Отсюда следует, что плотность энергии электромагнитной волны может быть представлена в следующем виде:

$$w = \frac{\varepsilon\varepsilon_0}{2}\left\langle E^2\right\rangle + \frac{\mu\mu_0}{2}\left\langle H^2\right\rangle = \varepsilon\varepsilon_0\left\langle E^2\right\rangle = \frac{1}{\mu\mu_0}\left\langle B^2\right\rangle$$

Представляется достаточно интересным тот факт, что электри- ческая и магнитная энергия поперечной волны остаются, как и в вакууме, равны друг другу.

Исходя из (11.6), мы можем вычислить и вектор Пойнтинга

$$\langle \mathbf{S} \rangle = \frac{k}{k}\sqrt{\frac{\varepsilon\varepsilon_0}{\mu\mu_0}}E^2$$

И, наконец, плотность импульса электромагнитной волны в прозрачной среде оказывается равной

$$\frac{d\mathbf{p}}{dV} = \frac{\mathbf{S}}{c^2} = \frac{k}{k}\sqrt{\frac{\varepsilon}{\mu}}\varepsilon_0^{3/2}\mu_0^{1/2}E^2$$

Ради общности мы всюду сохранили величину $\mu \neq 1$, хотя, как уже говорилось, для прозрачных сред магнитная проницаемость, как правило, несущественна.

## 19.2.2 Reflection and Refraction at a Plane Boundary

**мотивация**

обсудим, где эти знания будут полезны?

**типичная теория прохождения**

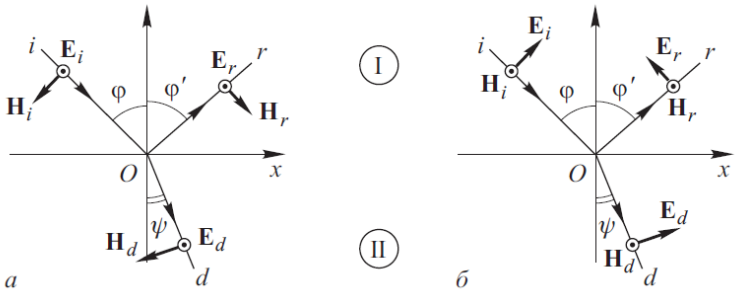
Рассмотрим задачу о прохождении электромагнитной волны через плоскую границу двух диэлектрических сред I и II (одной из них может, в частности, быть и вакуум).

Направление электрического и магнитного векторов соответствует правилу,

согласно которому $\mathbf{k}, \mathbf{E}, \mathbf{B}$ образуют правую тройку.

Мы использовали в качестве магнитного вектора H, поскольку именно для него будем писать соответствующее граничное условие.

Типичная ситуация имеет вид:




Заметим, что магнитное поле ориентировано так же, как электрическое поле на рис. 11.1 б (с точностью до знака), а на рис. 11.1 б — ортогонально плоскости рисунка.

Результаты, которые мы получим в данном параграфе, можно переносить и на случай искривленной поверхности раздела. Она лишь должна быть гладкой, а радиус кривизны ее должен многократно превосходить характерный пространственный масштаб электромагнитного поля — длину волны.

Таким образом мы опишем действие на электромагнитную волну, в частности, поверхности линзы.

Все расчеты линз, тонких и толстых, а также и сложных оптических систем базируются именно на законе преломления.

Мы будем пользоваться не понятием луча, принятым в геометрической оптике, а более корректным с точки зрения электродинамики понятием волнового фронта; «лучи» — падающий, отраженный и преломленный — изображенные на рис. 11.1, представляют в наших терминах нормали к волновому фронту, направление которых задается вектором k.

Пусть показатели преломления сред I и II равны, соответственно, $n_1, n_2$ (в вакууме - просто единице).

Мы показали на каждом из рис. 11.1 три волны: падающую (индекс i), отраженную (индекс $r$), и преломленную (индекс $d$). Это экспериментальный факт, известный каждому школьнику, но даже и не зная этого заранее, можно было его предсказать, исходя из уравнений Максвелла.

В случае падения электромагнитной волны на проводящую поверхность (см. §10.4) имеет место единственное нетривиальное граничное условие - закон сохранения тангенциальной компоненты электрического поля (9.12), а все остальные выполнялись должным образом за счет зарядов и токов, индуцированных на поверхности проводника.

Теперь такой возможности уже нет, поскольку мы имеем дело с диэлектрическими средами, а поэтому всего лишь одной волны, помимо падающей, нам просто не хватит.

В каждой из двух ситуаций на рис. 11.1 нам придется выполнить условия непрерывности $\mathbf{E}_\tau$ и $\mathbf{H}_\tau$. Третье условие - сохранение $\mathbf{B}_n$ в случае, изображенном на рис. 11.1$a$, и $\mathbf{D}_n$ в случае рис. 11.16 - будет выполнено автоматически как следствие закона преломления.

Еще один вопрос, на который целесообразно ответить заранее: правомерно ли разделение постановки задачи именно на те два случая, которые представлены на рис. 11.1$a, 6$? Не могут ли возникнуть отраженные либо преломленные волны с поляризацией, ортогональной таковой в падающей волне?

Ответ: не могут, и это прямое следствие уравнений Максвелла и граничных условий.

В силу линейности задачи, мы можем расщепить решение уравнений Максвелла на два линейно независимых, соответствующих двум различным поляризациям.

Выбирая решение с одной из поляризаций - той же, что и у падающей волны, мы оперируем с полями трех волн, что позволяет выполнить граничные условия (9.12) и (9.15).

При попытке выполнить их для другой поляризации нам опять «не хватит» переменных, т.к. в нашем распоряжении будут только две волны, без падающей, так что единственным возможным решением с такой поляризацией окажется нулевое поле.

Разумеется, эти рассуждения находятся в полном соответствии с экспериментальными данными.

Пусть все три волны записаны в виде (10.15).

Очевидно, для линейной среды зависимость $\mathbf{H}(\mathbf{r}, t)$ или $\mathbf{B}(\mathbf{r}, t)$ будет иметь точно такой же вид.

Воспользуемся для случая а граничным условием (9.12)

$$\mathbf{E}_1 = \mathbf{E}_{11} \to E_i + E_r = E_d$$

а для случая $6 -$ (9.15)

$$\mathbf{H}_\mathrm{I} = \mathbf{H}_\mathrm{u} = H_i + H_r = H_d$$

(мы учли, что в обоих случаях в соотношение входят параллельные векторы).

Поскольку дальнейшие действия для обеих поляризаций совершенно идентичны, мы ограничимся случаем рис. 11.1 а.

Пусть в какой-то момент времени граничное условие выполнено.

Однако оно сразу же нарушится, если зависимость от времени не будет одинаковой для всех трех полей. Это означает, что частота всех трех волн должна быть одинаковой (и действительно, отражение от прозрачной среды и преломление в ней «сохраняют цвет»).

Далее введем в плоскости падения вдоль границы сред координату $x$. Из поперечности волн и параллельности векторов $\mathbf{E}_i, \mathbf{E}_r, \mathbf{E}_d$ следует, что все три волновых вектора $\mathrm{k}_i, \mathrm{k}_r, \mathrm{k}_d$ лежат в одной плоскости - плоскости падения.

Вдоль оси $Ox$ произведения $\mathbf{kr}$ в формулах типа (10.15) вырождаются В $k_x \cdot x$. Таким образом, граничное условие при равных частотах сводится к следующему:

$$E_{i0} \exp(ikx \sin\varphi) + E_{r0} \exp(ikx \sin\varphi') = E_{d0} \exp(ik'x \sin\psi)$$

Мы воспользовались обозначениями углов рис. 11.1 и учли, что из равенства частот для падающей и отраженной волн следует равенство волновых чисел.

Для преломленной волны волновое число $k'$ определяется формулой (11.5).

Теперь потребуем, чтобы наше граничное условие выполнялось в любой точке оси $Ox$. Для этого необходимо, чтобы экспоненциальные множители были тождественно равны друг другу, а значит, равны должны быть их аргументы.

$$k \sin\varphi = k \sin\varphi' = k\,(n_\mathrm{II}/n_\mathrm{I}) \sin\psi$$

Мы получили аналитически из законов электродинамики хорошо известные правила вычисления углов отражения и преломления:

$$\varphi = \varphi'; \quad \frac{\sin\varphi}{\sin\psi} = n = \frac{n_{\text{II}}}{n_{\text{I}}}$$

Соотношения (11.10), в частности, подтверждают, что введенный нами коэффициент преломления $n = \sqrt{\varepsilon\mu}$ и в самом деле имеет право так называться.

Далее из (11.10) можно извлекать следствия, хорошо известные из школьного курса физики: эффект полного внутреннего отражения, отклонение световой волны призмой, фокусное расстояние тонкой линзы и т.д. Однако решение задачи о прохождении волной границы двух сред не исчерпывается этим результатом.

Те же самые граничные условия (9.12) – (9.15) позволят нам продвинуться существенно дальше - определить коэффициенты отражения и прозрачности на границе раздела двух диэлектриков.

Впервые это было сделано О.

Френелем (более чем за десять лет до рождения Максвелла).

## 19.2.3 Fresnel Formulas for Reflected and Transmitted Waves

Обсудим подробно формулы Френеля.

### Формулы Френеля вкратце и их применения

(с опытом впишу)

### Theory

Предметом нашего рассмотрения по-прежнему остаются задачи, представленные на рис.???

Для упрощения последующих выкладок положим для обеих сред $\mu = 1$; по сути, все сколько-нибудь интересные ситуации укладываются в рамки такого ограничения.

Далее, мы проведем вычисления для границы диэлектрика с вакуумом: $n_{\text{I}} = 1, n_{\text{II}} \equiv n$, а ответ представим в такой форме, что его можно будет отнести к границе сред с произвольными коэффициентами преломления.

В дальнейшем, выписывая граничные условия, мы сразу будем сокращать экспоненциальные осциллирующие множители, поскольку равенство частот и формулы (11.10) гарантируют нам такую возможность.

Теперь выполнение граничных условий потребует лишь определенного соотношения между амплитудами волн.

В этой связи заметим, что направление векторов $\mathbf{E}$ на рис. 11.1$a$ и $\mathbf{H}$ на рис. 11.1$б$ выбрано нами из соображений Как можно большей симметрии получающихся уравнений - в принципе для одной из вторичных волн - отраженной либо преломленной - ответ может включать знак «-», что, конечно, должно получиться совершенно автоматически.

Введем амплитудные коэффициенты отражения $r$ и прозрачности $t$ следующим образом:

$$E_{r0} = rE_{i0}; \quad E_{d0} = tE_{i0}$$

Мы ввели их через отношение амплитуд электрического поля в духе сделанного выше замечания о том, что при регистрации электромагнитных волн происходит измерение именно электрического поля.

Если вычислительная процедура приведет к тому, что один из них окажется отрицательным, это будет означать просто фазовый сдвиг на $\pi$. В наших

предположениях $H = n\left(\varepsilon_0/\mu_0\right)^{1/2} E$, при этом имеет смысл удерживать только фактор $n$, все прочее можно будет в наших уравнениях сократить.

для конкретной геометрии рис. 11.1$a$ (электрическое поле ортогонально плоскости падения) граничные условия (9.12), (9.15) преобразуются к виду

$$[\mathbf{E}_\tau] = 0 \Longrightarrow E_i + E_r = E_d$$

$$[\mathbf{H}_\tau] = 0 \Longrightarrow -H_i \cos\varphi + H_r \cos\varphi = -H_d \cos\psi \Longrightarrow$$

$$\Longrightarrow -E_i \cos\varphi + E_r \cos\varphi = -nE_d \cos\psi$$

Эту систему уравнений с помощью введенных нами коэффициентов можно свести к следующей:

$$1 + r = t; \quad (1 - r)\cos\varphi = t\frac{\sin\varphi}{\sin\psi}\cos\psi$$

Как и было обещано, система уравнений переписана в универсальном виде, в ней фигурируют только функции углов падения и преломления.

Решение (11.12) не представляет труда.

Мы приводим сразу же ответ:

$$r_\perp = -\frac{\sin(\varphi - \psi)}{\sin(\varphi + \psi)}; \quad t_\perp = \frac{2\sin\psi\cos\varphi}{\sin(\varphi + \psi)}$$

Здесь индекс $\perp$ означает поляризацию электрического вектора нормально к плоскости падения.

Как мы и предвидели, высокая симметрия рис.

11.1$a$ не соблюдена - отражение происходит потерей полуволны.

Надо заметить, что граничных условий в общей сложности четыре.

Но условие (9.14) в данном случае вырождено, поскольку электрический вектор не имеет составляющей, нормальной к поверхности раздела.

Обратимся к условию (9.13), которое для случая $\mu = 1$ может быть переписано для напряженности поля: $[\mathbf{H}_n] = 0$, или, что то же (см.

рис.

11.1$a$)

$$(H_i + H_r)\sin\varphi = H_d \sin\psi \Longrightarrow (E_i + E_r)\sin\varphi = E_d \sin\psi$$

Это условие тождественно удовлетворяется как следствие закона преломления и граничного условия на тангенциальные компоненты электрического поля.

Тем самым мы убедились, что ответ (11.13) вполне корректен.

Рассмотрим случай поляризации электрического вектора в плоскости падения (рис 11.1 б).

Те же самые условия (9.12),(9.15) теперь запишутся в виде

$$[\mathbf{E}_\tau] = 0 \Longrightarrow E_i \cos\varphi - E_r \cos\varphi = E_d \cos\psi$$

$$[\mathbf{H}_\tau] = 0 \Longrightarrow H_i + H_r = H_d \Longrightarrow E_i + E_r = \frac{\sin\varphi}{\sin\psi}E_d$$

В терминах коэффициентов отражения и прозрачности эта система уравнений может быть представлена следующим образом:

$$t = \frac{\sin\varphi}{\sin\psi}(1 + r); \quad \frac{\cos\varphi}{\cos\psi}(1 - r) = \frac{\sin\varphi}{\sin\psi}(1 + r)$$

Решение (11.14), как и (11.12), сводится к стандартным тригонометрическим преобразованиям, хотя и несколько более громоздким.

Ответ:

$$r_{\parallel} = \frac{\mathrm{tg}(\varphi - \psi)}{\mathrm{tg}(\varphi + \psi)}; \quad t_{\parallel} = \frac{2 \sin \psi \cos \varphi}{\sin(\varphi + \psi) \cos(\varphi - \psi)}$$

Предоставляем читателю самостоятельно убедиться в том, что граничное условие (9.14) в данном случае выполняется автоматически, а условие (9.13) вырождено,

Таким образом, задача решена полностью.

Хорошей проверкой наших ответов будет сведение их к общему предельному случаю нормального падения,

которому мы припишем коэффициенты $r_0, t_0$.

Из рис.

11.1 следует, что значения коэффициента отражения,

полученные предельным переходом из (11.13), (11.15),

должны отличаться знаком - на рис.

11.1$a$ знак «-» учитывает потерю полуволны, которая на рис.

11.16 последует автоматически,

потому что здесь сама амплитуда $E_r$ - другого знака.

С учетом этого обстоятельства искомые величины оказываются равны следующим:

$$r_0 = \frac{n-1}{n+1}; \quad t_0 = \frac{2}{n+1}$$

Формулы (11.13), (11.15), (11.16) - именно они и называются формулами Френеля (по имени французского физика О.Ж.

Френеля (1788 – 1827)) - в достаточной мере универсальны с точки зрения соотношения коэффициентов преломления,

т.е.

они равным образом могут применяться при падении волны из вакуума на границу с диэлектриком,

из диэлектрика в вакуум или на границу между двумя диэлектриками.

При этом углы $\varphi, \psi$ и коэффициенты преломления $n_1, n_{\mathrm{L}}, n$ связаны законом (11.10)

**примеры применения**

окей, пусть мы поняли это, как это применять?

и тут некоторые есть проблемы, ибо формулы большие.

нужна практика!!!

## 19.2.4 Polarization Effects

**Суть**

**Теория**

Сравнивая между собой коэффициенты отражения и прозрачности для различных поляризаций падающей волны — формулы (11.13) и (11.15), мы находим, что они отличаются друг от друга для всех углов падения, кроме нулевого — форму- ла (11.16).

Нам еще предстоит иметь дело с эффектом поляри- зации света в следующем разделе нашего курса, здесь же мы рассмотрим некоторые следствия такого рода из формул Френеля. Дело в том, что естественный свет — и равным образом любые электромагнитные волны естественного происхождения — всегда неполяризованы.

С точки зрения решений уравнений Максвелла это значит, что свет (излучение) представляет со- бой линейную смесь волн самых различных поляризаций.

Мы уже обсуждали это свойство в связи с проблемой точечного источника.

Разумеется, каждый отдельный квант, излучаемый возбужденным атомом, как-то поляризован, но если характерное время регистрации много больше времени замены одного такого кванта другим, с другой поляризацией, то мы этого не увидим. Это свойство присуще также и излучению искусственного про- исхождения, например, если его источник — тепловой природы: лампа накаливания, газоразрядная трубка (но не лазер!). Теперь представим себе, что такое неполяризованное излучение падает, как это показано на рис. 11.1, на границу двух сред с разными n.

Ввиду различия коэффициентов (11.13) и (11.15), как отраженный, так и преломленный свет уже приобретет некоторую степень поляризации.

Последнюю можно характеризовать, например, отношением средних квадратов электрического поля, соответствующего различным поляризациям.

Неполяризованный свет (или иное электромагнитное излучение) отвечает условию

$$\left\langle E_{\perp}^2 \right\rangle = \left\langle E_{\parallel}^2 \right\rangle$$

Напомним, что в данном случае усреднение проводится по промежутку времени порядка или большему характерного времени регистрации.

Для вторичных волн из (11.13) и (11.15) получаем

$$\frac{\langle E_{\perp}^2 \rangle_r}{\left\langle E_{\parallel}^2 \right\rangle_r} = \left(\frac{r_{\perp}}{r_{\parallel}}\right)^2 = \left[\frac{\cos(\varphi-\psi)}{\cos(\varphi+\psi)}\right]^2; \quad \frac{\langle E_{\perp}^2 \rangle_d}{\left\langle E_{\parallel}^2 \right\rangle_d} = \left(\frac{t_{\perp}}{t_{\parallel}}\right)^2 = \cos^2(\varphi-\psi)$$

Дадим для справки формальное определение степени поляризации электромагнитного излучения $\Delta$

$$\Delta = \frac{\left|\langle E_{\perp}^2 \rangle - \left\langle E_{\parallel}^2 \right\rangle\right|}{\langle E_{\perp}^2 \rangle + \left\langle E_{\parallel}^2 \right\rangle}$$

Особый интерес представляет возможность полной поляризации отраженного света при $\varphi + \psi = \pi/2$ - уже из (11.15) видно, что при этом $r_{\parallel} = 0$. Это означает, что в отраженном свете вектор E строго перпендикулярен плоскости падения.

С учетом (11.10) находим соответствующий угол падения

$$\varphi_{\text{B}} = \operatorname{arctg} n$$

Этот угол называется углом Брюстера, что и отражено индексом.

Заметим, что максимальная поляризация отраженного света не соответствует максимальной поляризации преломленного - это можно увидеть из второй формулы (11.17).

Еще отметим следующее обстоятельство.

Если диэлектрическая пластинка освещается белым светом, то условие (11.18) должно включать зависимость $n(\omega)$ или, что то же, $n(\lambda)$, и угол Брюстера будет разным для волн разной частоты.

Таким образом, простейший поляроид, доступный каждому, - это просто стеклянная пластинка, которую следует располагать так, чтобы свет падал на нее по возможности близко К $\varphi_{\text{B}}$. С помощью такого поляроида нетрудно, например, заметить поляризацию света, отраженного от водной поверхности.

Если же набрать пакет из таких пластинок и наблюдать источник света на просвет, то можно получить достаточно высокую степень поляризации, $\sim \cos^{4N}(\varphi-\psi)$, где $N$ - число параллельных пластинок.

(Фактор 4, вместо двойки, есть следствие того, что у каждой пластинки - две поверхности раздела сред.) Такой поляроид впервые был предложен русским физиком А. Г. Столетовым (1839 – 1896) и называется стопой Столетова.

Если $r$ и $t$ - амплитудные коэффициенты отражения и прозрачности, то $r^2$ и $t^2$ - то же для интенсивности.

В этой связи обращает на себя внимание следующее обстоятельство.

Вне зависимости от поляризации, из обеих формул (11.13) и (11.15) следует $r^2+t^2 \neq 1$. Легче всего это увидеть из формул (11.16)

Если $r$ и $t$ - амплитудные коэффициенты отражения и прозрачности, то $r^2$ и $t^2$ - то же для интенсивности.

В этой связи обращает на себя внимание следующее обстоятельство.

Вне зависимости от поляризации, из обеих формул (11.13) и (11.15) следует $r^2+t^2 \neq 1$. Легче всего это увидеть из формул (11.16) ДлЯ нормального падения, и в этом же случае легче всего прокомментировать.

Никакого парадокса здесь нет, поскольку в задаче такого рода следует подводить баланс не энергии как таковой, а потока энергиИ.

Предположим, что электромагнитная волна падает из вакуума на границу диэлектрика.

Падающий на граниЦу поток энергии равен, согласно (11.8),

$$|\langle \mathbf{S} \rangle_i| = \sqrt{\frac{\varepsilon_0}{\mu_0}} \left\langle E_i^2 \right\rangle$$

отраженный, соответственно,

$$|\langle \mathbf{S} \rangle_r| = r_0^2 \sqrt{\frac{\varepsilon_0}{\mu_0}} \left\langle E_i^2 \right\rangle$$

а уходящии с преломленнои волнои,

$$|\langle \mathbf{S} \rangle_d| = t_0^2 \sqrt{\frac{\varepsilon\varepsilon_0}{\mu_0}} \left\langle E_i^2 \right\rangle = n t_0^2 \, |\langle \mathbf{S} \rangle_i|$$

Таким образом, баланс потоков энергии в данной задаче требует следующего равенства:

$$r_0^2 + n t_0^2 = 1$$

которое и в самом деле выполнено для коэффициентов отражения и прозрачности (11.16).

### 19.2.5 Energy Flux Through the Boundary

(???)

## 19.3 Dispersion and Absorption of Light (According to Landsberg)

(write his theory here later!!!)

## 19.4 Dispersion

(подробно её обсудим.)

### 19.4.1 Phase and Group Velocity

**О законе дисперсии**

При распространении не монохроматических волн у каждой компоненты имеется своя фазовая скорость $v_f = \omega/k$, ранее она не зависела от $\omega$ и была одинакова для ращазных волн.

Однако некоторых средах существует *закон дисперсии*

$$\omega = \omega(k)$$

В этом случае $v_f = v_f(k) = v_f)\lambda)$.

такие среды называют *диспергирующими.*

Мы будем рассматривать только диэлектрические такие среды (что в не диэлектрических?)

В них рассмотрим волну из двух компонент:

$$u_1(x,t) = a\sin(k_1)\omega_1)x - \omega_1 t)$$
$$u_2(x,t) = a\sin(k_2(\omega_2)x - \omega_2 t)$$

Сама волна в итоге имеет вид::

$$u(x,t) = 2a\cos\left[\frac{k_2-k_1}{2}x - \frac{\omega_2-\omega_1}{2}t\right]\sin\left[\frac{k_2+k_1}{2}x - \frac{\omega_2+\omega_1}{2}t\right]$$

В случае, если эти волны не сильно отличаются друг от друга по $\omega$ bb $\mathbf{k}$, то:

$$u(x,t) = 2a\cos\left[\frac{\Delta k}{2}\left(x - \frac{d\omega}{dk}t\right)\right]\sin(kx - \omega t)$$

(заметим, что тут не спроста так cos b sin с разными аргументами, их нельзя менять местами.)

**биения**

(нужно тут вспомнить теорию колебаний, там есть биения.

там объяснение этого перехода, пока просто верим.

я хз, этот переход сейчас не понимаю!!!)

Таким образом.

**биения с периодом**

$$\tau = \frac{2\pi}{|\omega_2 - \omega_1|} \approx \frac{2\pi}{\Delta k|d\omega/dk|}$$

(тут прямо серьезные переходы, над которыми стоит подумать, переходим В СО с $v_{gr}$ - так это амплитудная модуляция??)

(кириченко съел этот момент, почему, а у меня нет времени)

если перейти в со с $v_{gr} = d\omega/dk$, то там вид волны покоится, таким образом вся волна распространяется с *групповой скоростью*

$$v_{gr} = d\omega/dk$$

амплитудная модуляция дается членом с групповой скоростью.

(а если не близкие волны, то что?

хотя, в свете фиолетовый и желтый может быть считаются близкими, так что всегда все близко???)

(добавления про период биений,

про суперпозицию многих волн

какие-то мелочи, см [1])

(про расплывание пакета потом можно подумать)

Изучим подробно групповую скорость.

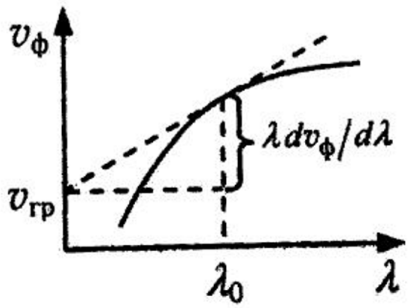
(я не знаю, зачем это и что значит этот переход, но просто так вот пока можно написать. потом подумаю.)

**формула Релея**

Первое свойство зависимости от фазовой скорости и типа волны (**k**) - называется *формула Релея:*

$$v_{gr} = d(kv_f)/dk = v_f + k\frac{dv_f}{dk} = v_f - \lambda\frac{dv_f}{d\lambda}$$

Это легко иллюстрировать на картинке. Производная - это тангенс наклона, так что при домножении ее на $v_f$ мы получаем часть групповой скорости.




(?? таким образом, зная для данной среды, как проходит через нее каждая волна, мы сразу по этому графику видим, как через нее будет проходить волна, содержащая пару близких длин волн.)

(как с пр Лежандра связано??)

Заметим, что $\omega(k)$ происходит из-за зависимости показателя преломления от длины волны.

так что имеет смысл понять, как групповая скорость зависит от $n$:

$$v_{\text{гр}} = v_\phi - \lambda\frac{d}{d\lambda}\left(\frac{c}{n}\right) = \frac{c}{n} + \frac{c\lambda}{n^2}\frac{dn}{d\lambda} = v_\phi\left(1 + \frac{\lambda}{n}\frac{dn}{d\lambda}\right)$$

(место для каких-то выводов формулы, которые мб потом пойму)

(я правильно все понимаю?) Таким образом, зная $n)\omega)$ мы сразу находим групповую скорость для пакета близких волн с известными $\omega$:

$$v_{\text{гр}} = \frac{c}{n}\left(1 + \frac{\lambda}{n}\frac{dn}{d\lambda}\right) = \frac{c}{n + \omega dn/d\omega}$$

Таким образом, при дисперсии групповая скорость меньше или фазовой за счет дополнительного слагаемого в знаменателе.

(и пара мыслей, что же мы получили, какие формулы, что дают?

пока нет времени думать)

И также тут видно, что знак $dn/d\omega$ дает нам принципиально разные случаи дисперсии.

**применения формулы Релея дя дисперсии**

получу опыт - пойму, как применять ее.

пока хз.

## 19.4.2 Classical Theory of Dispersion

Попробуем понять, какие свойства материи являются причиной дисперсии света.
простейшее объяснение дает так называемая *классическая теория дисперсии.*

**главная идея классической теории дисперсии**

**границы применимости классической теории дисперсии**

Электрон считается точечной частицей.
когда это не так??
как посчитать дисперсию, если электрон волна?
В классической теории дисперсии добавляем внешнее поле от волны, типа оно действует на электрон.
Рассматриваем электрон в атоме,
в общем его уравнение движения с учетом любых возможных сил и внешнего поля имеет вид:

$$\ddot{\mathbf{r}} + 2\Gamma\dot{\mathbf{r}} + \omega_0^2\mathbf{r} = \frac{e\mathbf{E}_0}{m}e^{-i\omega t},$$

(??)
Подставляя разложение фурье
(???)
находим уравнение движения:

$$\mathbf{r} = \frac{e/m}{\omega_0^2 - \omega^2 - 2i\Gamma\omega}\mathbf{E}$$

Заметим, что тут три частоты: $\omega_0$ - частота колебаний электрона, $\omega_p$ - просто константа для числителя, $\omega$ - частота падающей волны.
(тут забытая мной теория из электричества, пока - верю.
потом - все осмыслю нормально.
к чему тут $\mathbf{p} = \beta\mathbf{E}$?
тут эл на 100% используется, нужно его переучить, так эту тему я и не взял!!! и в теории поля кстати тоже она рассматривается)
Молекула приобретает дипольный момент $\mathbf{p} = e\mathbf{r}$,
поэтому его поляризация $\mathbf{P} = N\mathbf{p}$, а следовательно напряженность поля в среде $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P} \equiv \varepsilon\mathbf{E}$, поэтому

$$\varepsilon = 1 + \frac{\omega_p^2}{\omega_0^2 - \omega^2 - 2i\Gamma\omega}, \qquad \omega_p^2 = \frac{4\pi N e^2}{m}$$

Мы ввели *плазменную частоту* - коэффициент пропорциональности для всех проходящих $\omega$.
Заметим, что если частота волны совпадает с плазменной - то происходит резонанс, то есть $n$ станет сильно отличен от показателей преломления для других частот.
Чаще всего в этой формуле для просто ты будем пренебрегать зависимостью от $\omega$ и писать ее для $\omega = 0$, это не сильно повлияет на результат.
(потому что частота мала по величине?) В этом случае у среды нет дисперсии, а показатель преломления находится из плазменной частоты в этой среде:

$$\varepsilon = 1 + \omega_p^2/\omega_0^2$$

подытожим: волна, падая на тело, действует на электрон в атоме, последний смещается на величину,

зависящую от частоты падающей волны, в результате атом приобретает дипольный момент, а тело - поляризацию, тем самым (я тут хз, что дальше, мы тупо находим $n$, а как из $\mathbf{D} = \varepsilon\mathbf{E}$ следует другая скорость?

тупо вытянули показатель преломления - супер, а почему?

см эл, я продолбал этот момент.)

### О случаях и границах применимости

(!??!?!?!? напишу потом нормально)

Для произвольной среды наша модель колебаний не применима, так как оказывается, что в атомах этой среды не одна, а много резонансных частот, вблизи каждой из которых могут происходить колебания. Более верная формула имеет вид:

$$\varepsilon = 1 + \frac{4\pi N e^2}{m} \sum_k \frac{f_k}{\omega_{0k}^2 - \omega^2 - 2i\Gamma_k \omega}$$

### Показатель преломления вещества для рентгеновских лучей???

### типы дисперсий

(пояснения позже)

Мы знаем, что фазовая скорость может быть больше или меньше фазовой скорости в среде

(см формулу 19.4.1, я из-за этого так думаю.

кстати, а как ей пользоваться?).

Однако мы будем характеризовать дисперсию по тому, как среда влияет именно на одну из частот $\omega$ нашей волны.

необходимо знать, каков для нее показатель преломления $n = \sqrt{\varepsilon\mu}$?

В случае вещественного $\varepsilon$, мы бы просто взяли корень, но так как мы работаем с комплексным, то при извлечении корня вообще говоря еще появится мнимая часть:

$$\sqrt{\varepsilon} = n + i\kappa.$$

*$n$ - показатель преломления, $\kappa$ - показатель поглощения*

При $\omega = 0$ получаем:

$$n_0 = n(0) = \sqrt{1 + \omega_p^2/\omega_0^2} > 1,$$

также при

$$\omega \to \infty, \qquad n \to 1$$

В общем случае мы хз, что тут делать, но за нас посчитали и построили график.

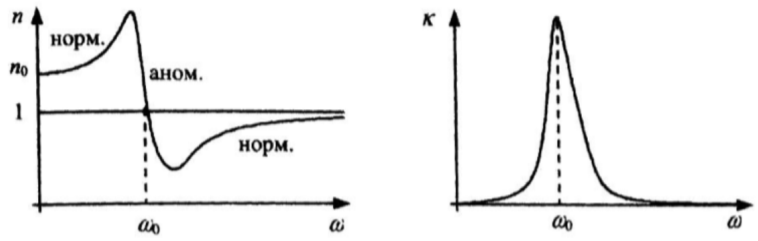
(а как посчитать???

место для тфкп теории.)

Про затухание нам уже (хз откуда) известно, что оно будет не нулевое вблизи $\omega_0$, так что посчитаем вблизи нее (после тфкп пойму, как последний переход сделан):

$$\kappa_{\max} \approx \operatorname{Im} \sqrt{\varepsilon(\omega_0)} \approx \operatorname{Im} \sqrt{1 + \frac{i\omega_p^2}{2\Gamma\omega_0}} \approx \frac{\omega_p}{2\sqrt{\Gamma\omega_0}}$$

Качественно это иллюстрируется так:




Видим, что есть качественно два типа дисперсий - назовем их *нормальной* и *аномальной дисперсией.*

(потом подумаю над связью их с групповой скоростью)

Затухание происходит по *закону Бугера*:

$$I = I_0 e^{-\alpha x},$$

(вывод полностью скипаю!)

здесь $\alpha$ - *коэффициент затухания.*

он связан с показателем затухания $\kappa$ соотношением:

$$\alpha = \frac{4\pi}{\lambda_0}\kappa, \quad \lambda_0 = \frac{2\pi c}{\omega}$$

А работаем мы так (без пояснений)

$$\varepsilon = 1 - \frac{\omega_p^2}{\omega^2}$$

поэтому

$$k^2 = \frac{\omega^2}{c^2}\varepsilon(\omega) \to \omega^2 = \omega_p^2 + k^2 c^2$$

а скорости:

$$v_\phi = \frac{\omega}{k} = \sqrt{\left(\frac{\omega_p}{k}\right)^2 + c^2}, v_{\text{гр}} = \frac{d\omega}{dk} = \frac{kc^2}{\sqrt{\omega_p^2 + k^2 c^2}} = \frac{kc^2}{\omega}$$

и

$$v_\phi v_{\text{гр}} = c^2$$

(место для задач и пояснений)

## 19.4.3 Plasma and Ionosphere

(????)

## 19.4.4 Advanced Theories of Dispersion (????)

если еще учитывать влияние электронов друг на друга, то нужно читать новые теории и думать дальше...

$\beta$ - поляризуемость.

зная ее, понимаем формулу лоренц-лоренца

$$\frac{4\pi}{3}N\beta = \frac{\varepsilon - 1}{\varepsilon + 2}$$

(место думать про это)

## 19.5 Other About Scattering

(мб тут многое будет, но пока об этом в дополнениях или других записях пишу, потому что вряд ли это важное для оптики. Стану умнее - буду создавать этот раздел уже.)

(?? see maybe electrodynamics about it???)

# 20 Effects and Models

## 20.1 Optics of Crystals

Тут будем рассматривать прохождение через кристаллы только плоские монохроматические электромагнитные волны. (пока я думаю, что это потому что с другими что-то сложное, нам итак хватает работы.

или там ничего не происходит и все?)

### 20.1.1 Crystals

(вообще там какая-то групповая еще большая теория, потом сюда мб ее суть укажу, пока не занимался ей.)

#### Теория

Кристаллы нам интересны благодаря их *анизотропии*, то есть в них не выполняется $\mathbf{D} = \varepsilon\mathbf{E}$, а вместо этого: $D_j = \sum\limits_k \varepsilon_{jk} E_k$.

здесь $\varepsilon_{jk}$ - *тензор диэлектрической проницаемости.*

оси, где он диагональный - *главные оси кристалла.*

в них: (а в другие оси как переход делается?)

$$D_x = \varepsilon_x E_x, D_y = \varepsilon_y E_y, D_z = \varepsilon_z E_z, \quad n_x = \sqrt{\varepsilon_x}, n_y = \sqrt{\varepsilon_y}, n_z = \sqrt{\varepsilon_z}$$

Если два показателя преломления совпадают - кристалл *одноосный.*В нем отличающаяся от других главная ось называется *оптической осью кристалла* . Они бывают *положительными* в случае, если $n_e > n_o$, так как в этом случае скорость обыкновенного луча более, чем необыкновенного (как бы все нормально),

а иначе говорят, что кристалл *отрицательный.* Если все 3 показателя преломления разные - кристалл *двуосный*,

но они сложные, так что они не будут нами рассматриваться.

Сам кристал имеет главную ось за счет потенциальных ям узлов решетки (кстати, а почему???).

в главных осях потенциальная энергия узла решетки имеет вид:

$$U = a_x x^2 + a_y y^2 + a_z z^2$$

(серьезно, я вообще хз, как потенциал связан с показателем преломления???)

(тут очень невнятный переход, я хз но из того, что выше следует, что:) поэтому $\boldsymbol{P} = \alpha_\perp \boldsymbol{E}_\perp + \alpha_\parallel \boldsymbol{E}_\parallel$, тем самым мы и получаем анизотропию, о которой говорилось выше: $\boldsymbol{D} = \varepsilon_\perp \boldsymbol{E}_\perp + \varepsilon_\parallel \boldsymbol{E}_\parallel$.

(у волны, которая попадает в кристалл не меняется длина волны, да?

типа фазовая ведь скорость меняется?

а потом меняется обратно при выходе?)

(угол между $D$ и $E$ определены кристаллом??? на входе **E** превращается в **D**, а потом от нее две волны, да?

или сперва отклонение, а потом уже раздвоение?)

## 20.1.2 Uniaxial Crystals and Double Refraction

**Теория**

Оказывается, чтобы в кристалле луч раздваивается или смотрит в одну сторону, а движется в другую (да?)! (черт, как бы это сказать?)

Ориентация волны в кристалле относительно его оси характеризуется *главным сечением* - плоскостью, содержащую оптическую ось кристалла и вектор $k$.

Ее следует сразу представлять, когда речь идет о распространении волны в нем.

В главном сечении сразу следует выделить угол между оптической осью и **k**, так как он нам в дальнейшем понадобится. Важность ее в том, что если вектор, перпендикулярен этой плоскости будет иметь обыкновенный показатель преломления,

причем других таких быть не может.

в кристалле поле удовлетворяет уравнения Максвелла:

$$\operatorname{rot} \boldsymbol{H} = \frac{1}{c}\frac{\partial \boldsymbol{D}}{\partial t}, \quad \operatorname{rot} \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t}$$

а так как наша волна в веществе имеет вид

$$\boldsymbol{E} = \boldsymbol{E}_0 e^{i(\omega t - \boldsymbol{kr})}, \quad \boldsymbol{H} = \boldsymbol{H}_0 e^{i(\omega t - \boldsymbol{kr})}, \quad \boldsymbol{D} = \boldsymbol{D}_0 e^{i(\omega t - \boldsymbol{kr})},$$

то при подстановки в уравнения Максвелла имеем:

$$\boldsymbol{D} = -\frac{c}{\omega}\boldsymbol{k} \times \boldsymbol{H}, \quad \boldsymbol{H} = \frac{c}{\omega}\boldsymbol{k} \times \boldsymbol{E}$$

Здесь мы получили важнейшие соотношения про отношения векторов друг относительно друга.

На этом этапе удивительно, но может быть, что вектор $\boldsymbol{D}$ не параллелен вектору $E$!

Важно вспомнить, что волна распространяется не туда, куда направлен $k$, а туда, куда направлен $\boldsymbol{S} = \frac{c}{4\pi}\boldsymbol{E} \times \boldsymbol{H}$ (потом нужно вписать это в первую главу, так как я там правда думал наоборот)

(что там со входом в кристалл, там получается, что **D** становится на место **E**, а второй - сразу же поворачивается относительно первого на угол, который зависит только от кристалла?

думаю, что наоборот происходит, но верно ли это???????)

есть уравнения, которые должны выполняться:

$$\boldsymbol{D} = -\frac{c}{\omega}\boldsymbol{k} \times \boldsymbol{H}, \quad \boldsymbol{H} = \frac{c}{\omega}\boldsymbol{k} \times \boldsymbol{E} \qquad \boldsymbol{D} = \varepsilon_\perp \boldsymbol{E}_\perp + \varepsilon_\parallel \boldsymbol{E}_\parallel$$

Дальше тут я не придумал, почему так, но решение может быть только в двух случаях: (интересно, а если вне главного сечения и не перпендикулярно ей, почему решений то нет?)

Если $\boldsymbol{D}$ перпендикулярен главному сечению - *обыкновенная волна* (потому что именно в перпендикулярном направлении к особенной оси обычный показатель преломления.)

Если $\boldsymbol{D}$ лежит в нем - *необыкновенная волна* (то есть именно у этой волны показатель преломления отличен от обыкновенного и будет как-то от $\varepsilon_\perp$ и $\varepsilon_\parallel$).

Для этой волны характерно, что она распространяется в главном сечение, но под углом к **k**.

Таким образом, зная направление поляризации для волны, входящей в кристалл, мы заменяем **E** на **D**, далее проецируем его на главное сечение и на нормаль к нему, в итоге пошли две волны.

В обыкновенной идет как и шла, но с новым направлением **D** (при выходе насколько я понимаю, **E** станет обратно таким же), а у необыкновенной направление напрямую определяется углом между **E** и **D**, а показатель преломления,

как будет ясно ниже - углом между оптической осью и **k** так что вне зависимости от направления поляризации при фиксированном направлении **k** будут одинаковые фазовые скорости (даже у волн с разными $\lambda$ будут одинаковые фазовые скорости, так что соответственно $\omega/k$ будет у разных волн одинаковое, а значит что вообще происходит???

если вправду длина волны не поменяется, то неужели это значит, что меняется все время чистота света при входе в кристалл???).

Это явление называется *двулучепреломлением.*

Прекрасная иллюстрация поможет не запутаться:

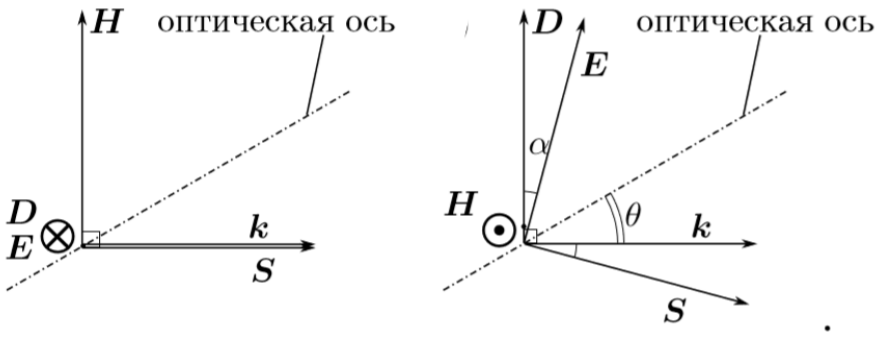



Figure 20.1: плоскость рисунка - главное сечение; а - обыкновенная, б - необыкновенная волна

Поэтому чтобы изменить показатель преломления у необыкновенной волны нужно менять только вектор $k$!!! Также видим, что поляризация у волн становится перпендикулярной друг другу.

(позже может быть добавлю вывод) Количественно фазовые скорости такие:

$$v_o = \frac{c}{\sqrt{\varepsilon_\perp}} = \frac{c}{n_o} \qquad v_e = \frac{c}{\sqrt{\varepsilon}} = \frac{c}{n(\theta)} \qquad \varepsilon = \frac{1}{\frac{\sin^2\theta}{\varepsilon_\parallel} + \frac{\cos^2\theta}{\varepsilon_\perp}}$$

Таким образом мы можем получать разности фаз при прохождении через кристалл.

(уже писал формулу, да?) Чаще всего мы работаем с малыми углами, в этом случае у необыкновенной волны $n_2 \approx n_o - (n_o - n_e)\,\theta^2$, так что разность фаз:

$$\Delta\varphi = \frac{2\pi}{\lambda}\ell\,(n_o - n_e)\,\theta^2$$

(и на что это повлияет??

они что, складываться будут?

в общем, я пойду про крист пластины писать, потом подумаю)

(а тут прямо взрыв мозга, я вообще хз, каким боком они интерферируют (что круги - это понятно) и дают еще крест, который при вращении еще и на негатив перейдет!!!!!!!!)

(это важная тема, я хз)

(нужно понять, как происходит их интерференция - и тут уже сразу про картину можно и сказать!!!)

**обзор обыкновенной и необыкновенной волны**

Обсудим их разницу

**распределение энергии по обыкновенной и необыкновенной волне**

## 20.1.3 Crystal Plates

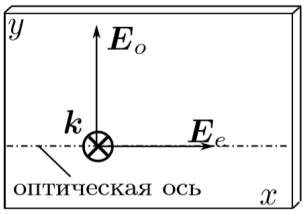
**простейшая теория**

Рассмотрим случай линейной поляризации.

Пусть вектор $k$ перпендикулярен оптической оси кристалла, например, когда волна попадает в тонкую пластинку, где ось, как у поляроида, смотрит в бок. Именно так и функционируют *кристаллические пластинки*.

В отличие от кристалла, тут поляризации показатель преломления для необыкновенной волны определяется только $\varepsilon_e$ (более отличий нет, да?).

(и тут необыкновенная не будет улетать от обыкновенной так как пластина тонкая, да?)




падение волны на пластинку

На примере кристаллической пластинки изучим подробно, как происходит прохождение волны через кристалл.

Таким образом, после прохождения пластинки толщиной $h$ появляется разность фаз

$$\Delta\varphi = kh\,(n_e - n_o)$$

А разность фаз напрямую влияет на тип поляризации.

(потом можно еще задуматься про это, пока вроде логично.

просто особенно не глубоко идея эта пока понята)

Есть пластинки $\lambda/2$ - в них сдвиг фаз происходит на $\pi$.

После прохода пластинки всегда снова образуется линейно поляризованная волна, только меняется плоскость поляризации.

(всегда?)

А в пластинках $\lambda/4$ - $\pi/2$.

поэтому всегда получаем круговую поляризацию.

Здесь полностью используем теорию поляризации, как сдвиг фаз влияет на тип.

Понятно, что эти соотношения у данной пластины для разных длин волн свои.

Кстати, чтобы не путаться, можно видеть разность фаз как $\Delta\varphi = \frac{2\pi}{\lambda} l$, где $l$ - разница оптических путей.

Так что можно вставлять $l = \lambda/2$ или $l = \lambda/4$ - и сразу получаем нужную разность фаз

Чаще всего на них падают монохроматические волны, которые общем случае всегда эллиптически поляризованы. В отличие от поляроидов, когда на выходе мы просто имели проекцию на ось поляроида (потом пойму, что там конкретно было, я пока что не доделал это),

(я не понимаю, это для круговой поляризации работает, или только для линейной?

скорее всего второе) здесь работаем так: проецируем **E** на ось пластинки и на перпендикуляр к ней.

(тут тоже хз, какая сдвигается из них?????) какая-то дальше сдвигается проекция на угол относительно второй - и получается, что между ними возникает разность фаз, так что в общем случае тип поляризации сменится.

Мы просто и один из векторов-проекций сдвигаем на угол, соответствующий типу пластинки, а далее находим векторную сумму проекций - результирующую амплитуду.

и как их мы применяем?

какие нюансы? потом все досмотрю. (и тут место задач)

**случай нелинейной поляризации**

(????)

## 20.2 Artificial Anisotropy (According to Landsberg)

(?????? I'll add it from his book later)

**Анизотропия, возникающая при деформациях**

**Двойное лучепреломление в электрическом поле (явление Керра)**

**Двойное лучепреломление в магнитном поле (явление Коттон-Мутона**

(?????)

## 20.3 Nonlinear Optics and Electro-Optical Effects (??)

(там по Ландсбергу еще есть это, пока не писал, итак по Ландау многое написано.)

### 20.3.1 Frequency Conversion in Nonlinear Media

(потом отработаю это, вроде тут простая идея, пока хз)

**Теория**

Изложенная в предыдущих главах теория распространения электромагнитных волн в диэлектрических средах основана на предположении о линейной связи индукции электрического поля **D** с его напряженностью **E**. Это приближение справедливо достаточной точностью, если (как это фактически имеет место) напряженность **E** мала по сравнению со значениями, характерными для внутриатомных полей. Но даже в этих условиях существующие малые нелинейные поправки в зависимости **D**(**E**) приводят к качественно новым эффектам и потому могут быть существенны.

Наиболее важной особенностью нелинейной среды является генерация в ней колебаний с новыми частотами. Так, если на такую среду падает монохроматическая волна с частотой $\omega_1$, то по мере ее распространения в среде генерируются волны с частотами $m\omega_1$ (где $m$ - целые числа); если первоначально имеется совокупность монохроматических сигналов с частотами $\omega_1$ и $\omega_2$, то с течением времени возникнут также и комбинационные частоты $m\omega_1 + m\omega_2$ и т. п.

Если среда бездиссипативна, то процесс преобразования частот подчиняется некоторым весьма общим соотношениям помимо очевидного условия сохранения суммарной энергии колебаний на всех частотах.

При этом предполагается, что нелинейность является слабой; смысл понятия этой слабости будет уточнен ниже.

Происхождение и смысл искомых соотношений наиболее ясен с квантовой точки зрения, из которой мы и будем исходить. Для упрощения рассуждений будем считать, что все частоты системы можно представить как линейные комбинации двух несоизмеримых основных частот $\omega_1$ и $\omega_2$ :

$$\omega_{mn} = m\omega_1 + n\omega_2$$

где $m, n$ - положительные или отрицательные целые числа.

Полную энергию излучения в среде можно представить как сумму энергий всех квантов:

$$\mathrm{U} = \sum_{m,n} N_{mn}\hbar\omega_{mn}$$

где $N_{mn}$ - число квантов с частотой $\omega_{mn}$. Суммирование производится по всем числам $m, n$, для которых $\omega_{mn} > 0$ (поскольку физическим смыслом обладают, конечно, только положительные частоты).

Процессы преобразования частот приводят к изменению чисел $N_{mn}$ со временем при сохранении полной энергии.

Поэтому

$$\frac{d\mathrm{U}}{dt} = \hbar\omega_1 \sum_{m,n} m\frac{dN_{mn}}{dt} + \hbar\omega_2 \sum_{m,n} n\frac{dN_{mn}}{dt} = 0$$

Но ввиду предположенной несоизмеримости $\omega_1$ и $\omega_2$ и целочисленности чисел квантов и их изменений это равенство требует обращения в нуль каждой из двух сумм в отдельности:

$$\sum_{m,n} m\frac{dN_{mn}}{dt} = 0, \quad \sum_{m,n} n\frac{dN_{mn}}{dt} = 0$$

Введем вместо чисел квантов соответствующие интенсивности - полные энергии $\overline{\mathrm{U}}_{mn}$, заключенные в излучении соответствующих частот:

$$\overline{\mathrm{U}}_{mn} = \hbar\omega_{mn}N_{mn}$$

Тогда соотношения $\sum\limits_{m,n} m\frac{dN_{mn}}{dt} = 0, \sum\limits_{m,n} n\frac{dN_{mn}}{dt} = 0$ примут вид

$$\sum_{m,n} \frac{m}{\omega_{mn}}\frac{d\overline{\mathrm{U}}_{mn}}{dt} = 0, \quad \sum_{m,n} \frac{n}{\omega_{mn}}\frac{d\overline{\mathrm{U}}_{mn}}{dt} = 0$$

Эти равенства и представляют собой искомые соотношения; они известны как теорема Мэнли-Роу (J.M. Manley, H.E. Rowe, 1956).

Здесь надо отметить обстоятельство, особенно ясное с точки зрения классической картины колебаний. Говоря об изменении $\mathrm{U}_{mn}$ со временем, мы имеем в виду только систематический ход этого изменения; другими словами, рассматривается энергия, усредненная по интервалам времени, большим по сравнению с периодами $1/\omega_1, 1/\omega_2$ (что и отмечено чертой над $\mathrm{U}_{mn}$. Именно здесь требуется слабость нелинейных эффектов: характерное время $\tau$ вызванного ими систематического нарастания возбуждающихся колебаний должно быть велико по сравнению с указанными периодами.

Только в таких условиях может иметь смысл рассмотрение временного хода величин, усредненных по интервалам $\Delta t$, таким, что

$$\frac{1}{\omega_1}, \frac{1}{\omega_2} \ll \Delta t \ll \tau.$$

Остается придать равенствам Мэнли-Роу более определенный вид, избавившись от ограничения $\omega_{mn} > 0$ при суммированиях Это легко сделать, заметив, что каждой паре чисел $m, n$, для которой $\omega_{mn} > 0$, отвечает пара чисел $-m, -n$, для которой частота отрицательна при том же $|\omega_{mn}|$. Положив по определению

$$\overline{\mathrm{U}}_{-m,-n} = \overline{\mathrm{U}}_{mn}$$

и распространив суммирование по всем целым числам от $-\infty$ до $\infty$, мы, следовательно, удвоим суммы, что не нарушит их равенства нулю.

После этого можно произвести еще одно упрощение.

В первом равенстве (107.4) разбиваем сумму по $m$ на две суммы - от 0 до $\infty$ и от $-\infty$ до 0, и во второй заменяем $m\ n \to -m, -n$; аналогичное преобразование во втором равенстве (107.4) производим в сумме по $n$. В результате получим окончательно:

$$\sum_{m=0}^{\infty}\sum_{n=-\infty}^{\infty}\frac{m}{m\omega_1+n\omega_2}\frac{d\overline{\mathrm{U}}_{mn}}{dt}=0,\quad \sum_{n=0}^{\infty}\sum_{m=-\infty}^{\infty}\frac{n}{m\omega_1+n\omega_2}\frac{d\overline{\mathrm{U}}_{mn}}{dt}=0$$

Обобщение теоремы на случай большего числа исходных несоизмеримых частот очевидно.

Конкретные свойства колебательной системы могут запрещать те или иные процессы преобразования частот, ведь суммирования производятся фактически лишь по разрешенным процессам. Так, в простейшем случае системы, допускающей лишь генерирование комбинационной частоты $\omega_1 + \omega_2$, числа $m, n$ пробегают значения 0,1 и мы находим

$$-\frac{1}{\omega_1}\frac{d\overline{\mathrm{U}}_{10}}{dt}=-\frac{1}{\omega_2}\frac{d\overline{\mathrm{U}}_{01}}{dt}=\frac{1}{\omega_1+\omega_2}\frac{d\overline{\mathrm{U}}_{11}}{dt}$$

Смысл этих равенств очевиден: убыль чисел квантов $\hbar\omega_1$ и $\hbar\omega_2$ равна прибыли числа генерируемых квантов $\hbar(\omega_1+\omega_2)$.

## 20.3.2 Nonlinear Permittivity

### Теория

При слабой нелинейности первая поправка к линейной зависимости $\mathbf{D}$ от $\mathbf{E}$ квадратична по полю. При наличии временной дисперсии она может быть представлена в произвольной анизотропной среде выражением

$$D_i^{(2)}(t)=\iint_0^{\infty} f_{i,kl}(\tau_1,\tau_2)\,E_k(t-\tau_1)\,E_l(t-\tau_2)\,d\tau_1 d\tau_2$$

аналогичным (77.3) (????). Разумеется, существование такого члена налагает определенные ограничения на допустимую симметрию среды; в частности, он отсутствует, если среда инвариантна относительно инверсии.

Хотя ниже мы будем рассматривать как типичный именно квадратичный по $\mathbf{E}$ член вида (108.1), необходимо отметить, что в квадратичном приближении электрическая индукция $\mathbf{D}$ может содержать также и билинейные по компонентам $\mathbf{E}$ и $\mathbf{H}$ и квадратичные по $\mathbf{H}$ члены; эти члены обычно играют меньшую роль и мы не будем их рассматривать. Мы не будем также обсуждать нелинейную зависимость магнитной индукции $\mathbf{B}$ от $\mathbf{H}$ ввиду аналогии с вопросом о зависимости $\mathbf{D}(\mathbf{E})$.

Введем величину

$$\varepsilon_{i,kl}(\omega_1,\omega_2):=\iint_0^{\infty} e^{i(\omega_1\tau_1+\omega_2\tau_2)} f_{i,kl}(\tau_1,\tau_2)\,d\tau_1 d\tau_2$$

которую естественно назвать нелинейной проницаемостьто второго порядка (по аналогии с линейной проницаемостью $\varepsilon_{ik}(\omega)$, определяемой выражением вида (77.5)). Она лишь множителем отличается от величин $\chi_{ikl} = \varepsilon_{ikl}/(4\pi)$, которые называют нелинейной восприимиивостью.

Ввиду симметрии, с которой входят $E_k$ и $E_l$ в определение $D_i^{(2)}(t)$, тензор $f_{i,kl}$ симметричен по индексам $kl$ при одновременной перестановке его аргументов: $f_{i,kl}(\tau_1, \tau_2) = f_{i,lk}(\tau_2, \tau_1)$. Поэтому такой же симметрией обладает и тензор $\varepsilon_{i,kl}$ :

$$\varepsilon_{i,kl}(\omega_1, \omega_2) = \varepsilon_{i,lk}(\omega_2, \omega_1)$$

В частности, при $\omega_1 = \omega_2$ тензор симметричен по последней паре индексов:

$$\varepsilon_{i,kl}(\omega, \omega) = \varepsilon_{i,lk}(\omega, \omega)$$

Кроме того, в силу вещественности функций $f_{i,kl}$ (в свою очередь следующей из определения (108.1) с вещественными $\mathbf{E}$ и $\mathbf{D}$). Имеем

$$\varepsilon_{i,kl}(-\omega_1, -\omega_2) = \varepsilon_{i,kl}^*(\omega_1, \omega_2)$$

Проницаемость (108.2) естественно появляется при рассмотрении монохроматических полей или их суперпозиций.

В нелинейных выражениях такие поля должны, конечно, быть представлены

в вещественном виде.

Так, если $\mathbf{E}$ - монохроматическое поле с одной частотой $\omega$, то надо писать $\mathbf{E}(t) = \mathrm{Re}\{\mathbf{E}_0 e^{-i\omega t}\}$, и его подстановка в (108.1) приводит к выражению

$$D_i^{(2)}(t) = \frac{1}{2}\,\mathrm{Re}\left\{\varepsilon_{i,kl}(\omega, \omega)e^{-2i\omega t}E_{0i}E_{0k} + \varepsilon_{i,kl}(\omega, -\omega)E_{0i}E_{0k}^*\right\}$$

Оно содержит колебания с удвоенной частотой (наряду с постоянным членом, отвечающим разности частот $\omega - \omega = 0$). В общем случае $\varepsilon_{i,kl}(\omega_1, \omega_2)$ описывает вклад в индукцию, пропорциональный $\exp(-i\omega_3 t), \omega_3 = \omega_1 + \omega_2$ Мы будем рассматривать ниже лишь бездиссипативные среды и будем называть их прозрачными, хотя они и не являются таковыми в буквальном смысле (для волн заданной частоты) ввиду возможности перехода энергии в другие частоты.

Будем также считать, что среда не обладает магнитной структурой.

Прежде всего покажем, что в таких условиях нелинейная проницаемость вещественна.

Это можно было бы увидеть непосредственно, если выразить компоненты тензора нелинейной восприимчивости через матричные элементы электрического дипольного взаимодействия среды с полем, играющего роль малого возмущения; восприимчивость второго порядка появляется в третьем приближении теории возмущений.

Но понять происхождение вещественности результата такого вычисления можно и без его фактического проведения.

Действительно, полную систему волновых функций, по которым вычисляются матричные

элементы, можно выбрать (для среды без магнитной структуры и потому инвариантной по отношению к обращению времени!) вещественной.

Веществен также и оператор взаимодействия поля с электрическим дипольным моментом среды.

Поэтому мнимые члены могли бы появиться лишь в результате обходов полюсов энергетических знаменателей теории возмущений.

Но отсутствие диссипации в среде означает, что ни одна из частот поля не совпадает с разностью энергетических уровней системы (или же, что вычеты в полюсах равны

нулю в силу тех или иных правил отбора); поэтому фактически обходить полюса не прихоДИТСЯ.

Прозрачность среды приводит также к появлению определенных соотношений симметрии для тензора $\varepsilon_{i,kl}$. И эти соотношения можно было бы усмотреть из конкретных выражений, полученных по теории возмущений.

Но и здесь можно прийти к требуемому результату более простым путем.

Для этого предположим, что поле в среде есть сумма трех почти монохроматических полей с несоизмеримыми частотами $\omega_1, \omega_2, \omega_3$

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 = \mathrm{Re}\left\{\mathbf{E}_{01}e^{-i\omega_1 t} + \mathbf{E}_{02}e^{-i\omega_2 t} + \mathbf{E}_{03}e^{-i\omega_3 t}\right\}, \quad (108.7)$$

причем $\omega_3 = \omega_1 + \omega_2$. Будем считать, что поля на частотах $\omega_1$ $\omega_2, \omega_3$ были созданы внешними источниками, впоследствии выключенными; под влиянием же слабой нелинейности среды их амплитуды $\mathbf{E}_{01}, \ldots$ становятся медленно меняющимися функциями времени.

Эта медленность позволяет писать уравнения Максвелла для полей каждой из основных частот в отдельности.

В свою очередь, из этих уравнений обычным образом следует закон сохранения энергии в виде

$$\frac{1}{4\pi}\left(\overline{\mathbf{E}_1\dot{\mathbf{D}}_1} + \overline{\mathbf{H}_1\dot{\mathbf{H}}_1}\right) + \mathrm{div}\,\frac{c}{4\pi}\left[\overline{\mathbf{E}_1\mathbf{H}_1}\right] = 0$$

(и аналогично для величин с индексами 2 и 3); $\mathbf{D}_1$ обозначает ту часть индукции $\mathbf{D}$, которая содержит множители $e^{\pm i\omega_1 t}$, а черта - усреднение по времени (нужное для дальнейшего).

При интегрировании по всему объему поля член с дивергенцией исчезает и остается

$$\frac{1}{4\pi}\int\left(\overline{\mathbf{E}_1\dot{\mathbf{D}}_1} + \overline{\mathbf{H}_1\dot{\mathbf{H}}_1}\right)dV = 0$$

Если теперь в явном виде выделить в $\mathbf{D}_1$ линейные и нелинейные по $\mathbf{E}$ члены.

.

$$\mathbf{D}_1 = \mathbf{D}_1^{(1)} + \mathbf{D}_1^{(2)}$$

то первые, вместе с $\mathbf{H}_1\dot{\mathbf{H}}_1$, дадут $d\overline{\mathrm{U}}_1/dt$ - изменение со временем энергии поля на частоте $\omega_1$. Это изменение, таким образом, определяется уравнением

$$\frac{d\overline{\mathrm{U}}_1}{dt} = -\frac{1}{4\pi}\int \mathrm{Re}\left\{\mathbf{E}_{01}\,\overline{e^{-i\omega_1 t}\}\,\dot{\mathbf{D}}_1^{(2)}}\right\}dV$$

Здесь производную $\partial\mathbf{D}_1^{(2)}/\partial t$ следует выразить через напряженность поля с помощью (108.1), (108.2). Усреднение по времени обращает в нуль все члены, кроме тех, в которых экспоненци- альные множители взаимно сокращаются.

Повторяя вычисления для остальных частот, находим окончательно:

$$\frac{d\overline{\mathrm{U}}_1}{dt} = -\frac{i\omega_1}{16\pi}\int \varepsilon_{k,il}\left(-\omega_3, \omega_2\right) E^*_{03i}E_{01k}E_{02l}dV + \text{К.с.}$$

$$\frac{d\overline{\mathrm{U}}_2}{dt} = -\frac{i\omega_2}{16\pi}\int \varepsilon_{i,ki}\left(\omega_1, -\omega_3\right) E^*_{03i}E_{01k}E_{02l}dV + \text{К.с.}$$

$$\frac{d\overline{\mathrm{U}}_3}{dt} = \frac{i\omega_3}{16\pi}\int \varepsilon_{i,kl}\left(\omega_1, \omega_2\right) E^*_{03i}E_{01k}E_{02l}dV + \text{К.с.}$$

где к. с. означает комплексно-сопряженное выражение. При вычислении использовано свойство (108.3).

Суммарное изменение энергии на всех трех частотах равно

$$\frac{d\overline{\mathrm{U}}}{dt} = \frac{d\overline{\mathrm{U}}_1}{dt} + \frac{d\overline{\mathrm{U}}_2}{dt} + \frac{d\overline{\mathrm{U}}_3}{dt}$$

В нелинейной среде эта сумма, вообще говоря, не должна точно обращаться в нуль ввиду возможности перехода энергии еще и в другие комбинационные частоты: $\omega_1 - \omega_2, \omega_3 + \omega_2$ и т. д. Но величина полей на частотах $\omega_1, \omega_2, \omega_3$, созданных внешними источниками, не имеет отношения к степени нелинейности; она не должна быть малой в противоположность полям на других частотах, появляющихся лишь в силу нелинейности среды.

Поэтому вклад последних в баланс энергии можно считать от- сутствующим и потребовать обращения суммы (108.10) в нуль.

Более того, поскольку такая среда представляет собой нелинейную систему всего с тремя частотами, можно применить к ней теорему Мэнли-Роу в ее простейшей форме (107.7). В используемых нами здесь обозначениях эти соотношения гласят:

$$\frac{1}{\omega_1}\frac{d\overline{\mathrm{U}}_1}{dt} + \frac{1}{\omega_3}\frac{d\overline{\mathrm{U}}_3}{dt} = 0, \quad \frac{1}{\omega_2}\frac{d\overline{\mathrm{U}}_2}{dt} + \frac{1}{\omega_3}\frac{d\overline{\mathrm{U}}_3}{dt} = 0.$$

Подставив сюда (108.9), найдем, что нелинейная проницаемость удовлетворяет следующим важным соотношениям симметрии

$$\varepsilon_{i,kl}\left(\omega_1, \omega_2\right) = \varepsilon_{k,il}\left(-\omega_3, \omega_2\right) = \varepsilon_{l,ki}\left(\omega_1, -\omega_3\right)$$

Выражаемая этими равенствами симметрия становится более наглядной, если приписать компонентам тензора еще и третий аргумент так, чтобы сумма всех трех была равна нулю:

$$\varepsilon_{i,kl}\left(-\omega_3; \omega_1, \omega_2\right) = \varepsilon_{k,il}\left(\omega_1; -\omega_3, \omega_2\right) = \\ = \varepsilon_{l,ki}\left(\omega_3; \omega_1, -\omega_3\right) = \varepsilon_{i,lk}\left(-\omega_3; \omega_2, \omega_1\right)$$

(последнее равенство $-(108.3)$).

Если условиться связывать три последовательных аргумента (частоты) с тремя последовательными индексами тензора, то можно произвольным образом переставлять эти индексы при условии одновременной такой же перестановки аргументов

Отметим, что само по себе требование отсутствия диссипации привело бы лишь к более слабому условию - равенству нулю суммы (108.10). т.е.

$$\omega_3\varepsilon_{i,kl}\left(\omega_1, \omega_2\right) - \omega_1\varepsilon_{k,il}\left(-\omega_3, \omega_2\right) - \omega_2\varepsilon_{l,ki}\left(\omega_1, -\omega_3\right) = 0.$$

Изложенный вывод неприменим непосредственно при $\omega_1 = \omega_2 \equiv \omega$, поскольку соотношения Мэнли- Роу сводятся при этом к одному лишь сохранению полной энергии.

Но о равенстве

$$\varepsilon_{i,kl}(\omega, \omega) = \varepsilon_{k,il}(-2\omega, \omega) = \varepsilon_{l,ki}(\omega, -2\omega)$$

можно заключить просто из соображений непрерывности при предельном переходе из (108.11).

Если обе частоты, $\omega_1$ и $\omega_2$, стремятся к нулю, тензор $\varepsilon_{i,kl}$ оказывается полностью симметричным.

Эта симметрия выражает собой просто тот факт, что в статическом случае индукция $\mathbf{D}$ может быть получена дифференцированием свободной энергии по $\mathbf{E}$:

$$D_i = -4\pi\frac{\partial\widetilde{F}}{\partial E_i}$$

так что

$$\frac{\partial D_i}{\partial E_k} = \frac{\partial D_k}{\partial E_i}$$

Отсюда следует симметричность тензора $\varepsilon_{i,kl}$ по индексам $ik$, а тем самым - по всем трем индексам.

Если же равна нулю лишь одна из частот, то соотношения (108.11) приводят к равенству

$$\varepsilon_{i,kl}(\omega, 0) = \varepsilon_{l,ki}(\omega, -\omega)$$

а также

$$\varepsilon_{i,kl}(\omega, 0) = \varepsilon_{k,il}(-\omega, 0) = \varepsilon_{k,il}(\omega, 0)$$

(в силу (108.5) вещественные функции $\varepsilon_{k,il}(\omega, 0)$ четны по $\omega$).

Тензор $\varepsilon_{i,kl}(\omega, 0)$ описывает линейный электрооптический эффект - изменение проницаемости кристалила под влиянием постоянного электрического поля, и, следовательно, совпадает с определенным в (100.4) тензором $\alpha_{ikl}$ :

$$\varepsilon_{i,kl}(\omega, 0) = \alpha_{ikl}(\omega)$$

как и должно быть, в силу (108.15) он симметричен по индексам ik.

Тензор же $\varepsilon_{l,ki}(\omega, -\omega)$ описывает другой эффект - появление в среде статической диэлектрической поляризации, пропорциональной квадрату приложенного слабого переменного периодического поля (ср.

второй член в (108.6)). Равенство (108.14) устанавливает, таким образом, свяЗь между этими двумя эффектамИ.

Аналогичными сооображениями, используя нелинейную восприимчивость, «перекрестную» между электрическими и магнитными величинами, можно было бы вновь прийти к связи между магнитооптическим эффектом Фарадея и намагничением среды вращающимся по направлению электрическим полем; это - связь, устанавливаемая формулами (101.15) и (101.25). Как уже было указано, для сред, инвариантных относительно пространственной инверсии, нелинейность второго порядка отсутствует.

В таких случаях нелинейные эффекты начинаются с кубических членов в разложении зависимости $\mathbf{D}(\mathbf{E})$. Соответствующая нелинейная проницаемость третьего порядка есть тензор четвертого ранга, зависяший от трех независимых частот:

$$\varepsilon_{i,klm}(\omega_1, \omega_2, \omega_3)$$

Его свойства симметрии полностью аналогичны свойствам тензора проницаемости второго порядка: если ввести еще и четвертую частоту

$$\omega_4 = \omega_1 + \omega_2 + \omega_3$$

и записать тензор в виде

$$\varepsilon_{i,klm}(-\omega_4; \omega_1, \omega_2, \omega_3)$$

то можно любым образом переставлять индексы вместе с такой же перестановкой четырех аргументов.

Нелинейность третьего порядка может быть существенна даже при наличии квадратичной нелинейности ввиду специфичности вызываемых ею эффектов.

**применения**

хз

## 20.3.3 Self-Focusing

**Теория**

Как уже говорилось, везде в этой главе напряженность электрического поля будет предполагаться достаточно малой, так что будут относительно малы соответствующие нелинейные члены в уравнениях.

В таких условиях представляют интерес те нелинейные эффекты, которые «накапливаются» с течением времени и за достаточно большое время приводят к существенному изменению характера явления.

В качестве первого примера такого рода в этом параграфе будут рассмотрены оптические эффекты, связанные с нелинейным изменением поля на частоте первичной волны.

**оптические эффекты, связанные с нелинейным изменением поля на частоте первичной волны**

Другими словами, рассматривается нелинейный вклад в $\mathbf{D}$ той же частоты $\omega$, которую имеет монохроматическое поле $\mathbf{E}$.

В квадратичных членах такой вклад отсутствует: они содержат только частоты $2\omega$ и 0.

Первый отличный от нуля эффект возникает от кубической нелинейности и содержится в членах вида $EEE^*$( частота $\omega+\omega-\omega=\omega$)

Ниже в этом параграфе будем считать среду изотропной (жидкость или газ).

Тогда интересующие нас члены третьего порядка в индукции имеют в общем случае вид

$$\mathbf{D}^{(3)}=\alpha(\omega)|\mathbf{E}|^2\mathbf{E}+\beta(\omega)\mathbf{E}^2\mathbf{E}^*$$

они содержат два независимых коэффициента; в прозрачной среде эти коэффициенты - вещественные четные функции частоты.

Такое число независимых коэффициентов находится в соответствии со свойствами симметрии тензора $\varepsilon_{i,klm}(-\omega;-\omega,\omega,\omega)$.

При указанных значениях аргументов этот тензор симметричен по парам индексов $ik$ и $lm$; в изотропной среде такой тензор имеет две независимые компоненты.

В пределе малых частот, как указано в предыдущем параграфе, тензор должен быть симметричен по всем индексам, т.е. в изотропной среде пропорционален комбинации

$$\delta_{ik}\delta_{lm}+\delta_{il}\delta_{km}+\delta_{im}\delta_{kl}$$

Это означает, что

$$\alpha(0)=2\beta(0)$$

Выражение (109.1) упрощается для линейно поляризованного поля E.

При такой поляризации комплексный вектор $\mathbf{E}$ сводится к вещественному вектору, умноженному на общий фазовый множитель; тогда выражения $|\mathbf{E}|^2\mathbf{E}$ и $\mathbf{E}^2\mathbf{E}^*$ совпадают и

$$\mathbf{D}^{(3)}=(\alpha+\beta)|\mathbf{E}|^2\mathbf{E}.$$

Такое же упрощение возникает при круговой поляризации поля $\mathbf{E}$; в этом случае $\mathbf{E}^2=0$ и (109.1) сводится к

$$\mathbf{D}^{(3)}=\alpha|\mathbf{E}|^2\mathbf{E}$$

В обоих случаях индукция поляризована так же, как и $\mathbf{E}$. В общем же случае эллиптической поляризации направления и отношения главных осей эллипсов $\mathbf{E}$ и $\mathbf{D}^{(3)}$ не совпадают.

$C_BA3$ b

$$\mathbf{D} = \mathbf{D}^{(1)} + \mathbf{D}^{(3)} = \varepsilon\mathbf{E} + \mathbf{D}^{(3)}$$

($\varepsilon(\omega)$ - обычная, линейная проницаемость) должна быть подставлена в уравнения Максвелла, которые при этом следует записать (исключив из них магнитное поле $\mathbf{H}$) в виде

$$\operatorname{rot}\operatorname{rot}\mathbf{E} + \tfrac{1}{c^2}\tfrac{\partial^2\mathbf{D}}{\partial t^2} = 0$$
$$\operatorname{div}\mathbf{D} = 0$$

Существенно, что эта нелинейная система уравнений допускает точное решение в виде монохроматической плоской волны

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{kr}-\omega t)}$$

с линейной или круговой поляризацией.

Действительно, для таких волн $|\mathbf{E}|^2 = |\mathbf{E}_0|^2$, так что формулы (109.3) или (109.4) имеют тот же характер, что и в линейном случае с проницаемостью, зависящей от амплитуды поля; поэтому можно брать вещественную часть после решения уравнений.

Связь $\mathbf{D}$ и $\mathbf{E}$ в этих случаях будем записывать в виде

$$\mathbf{D} = \left(\varepsilon + \frac{2c^2}{\omega^2}\eta\,|\mathbf{E}_0|^2\right)\mathbf{E}$$

введя удобное для дальнейшего обозначение: $\eta = \omega^2(\alpha+\beta)/(2c^2)$ для линейно поляризованной или $\eta = \omega^2\alpha/(2c^2)$ для поляризованной по кругу волныы.

Подстановка (109.8) в (109.6) дает $\operatorname{div}\mathbf{E} = 0$ - поле остается поперечным, как в линейной теории.

С учетом этого подстановка (109.7) в (109.5) приводит к дисперсионному соотношению

$$k^2 = \frac{\omega^2}{c^2}\varepsilon + 2\eta\,|\mathbf{E}_0|^2$$

Фазовая скорость $\omega/k$ оказывается зависящей теперь не только от частоты, но и от амплитуды волны.

Если $\eta > 0$, фазовая скорость убывает с увеличением амплитуды; такую среду называют фокусирующей (смысл этого названия выяснится ниже).

Если же $\eta < 0$, то фазовая скорость растет при увеличении амплитуды, и среду называют дефокусирующей.

Использование нелинейной связи (109.1) предполагает, конечно, лишь слабую нелинейность - члены более высоких порядков IOлжны быть малы по сравнению с членами $\mathbf{D}^{(3)}$.

Пои этом каче-

ственно новые явления могут возникнуть в результате «накопления» эффектов нелинейности за большие промежутки времени и на больших расстояниях.

Естественной постановкой вопроса является при этом рассмотрение почти монохроматической волны вида

$$\mathbf{E} = \mathbf{E}_0(t,\mathbf{r})e^{i(k_0x-\omega t)}$$

где $\mathbf{E}_0(t,\mathbf{r})$ - медленно меняюшаяся функция времени и координат (мало ее относительное изменение на интервалах времени $\sim 1/\omega$ и расстояниях $\sim 1/k_0$). Входяшие в фурье-разложение этого поля волновые векторы распределены в небольшом интервале значений

вокруг вектора $\mathbf{k}_0$, направленного вдоль оси $x$; его величину условимся считать связанной с $\omega$ равенством

$$k_0^2 = \frac{\omega^2}{c^2}\varepsilon(\omega)$$

отвечающим линейной теории.

Выведем уравнение для функции $\mathbf{E}_0(t, \mathbf{r})$

Прежде всего замечаем, что в уравнении (109.5) в члене

$$\operatorname{rot}\operatorname{rot}\mathbf{E} = \operatorname{grad}\operatorname{div}\mathbf{E} - \Delta\mathbf{E}$$

можно пренебречь величиной grad div $\mathbf{E}$. Действительно, в силу уравнения (109.6) дивергенция напряженности

$$\operatorname{div}\mathbf{E} \approx -\eta\frac{2c^2}{\varepsilon\omega^2}\mathbf{E}_0 \operatorname{grad}|\mathbf{E}_0|^2$$

т.е.

отлична от нуля только за счет производных от медленно меняющейся функции $\mathbf{E}_0$ и дополнительно мала в силу малости нелинейных членов; такими величинами мы пренебрегаем.

Таким образом, имеем

$$\operatorname{rot}\operatorname{rot}\mathbf{E} \approx -\Delta\mathbf{E} \approx \left[k_0^2\mathbf{E}_0 - 2ik_0\frac{\partial\mathbf{E}_0}{\partial x} - \Delta_\perp\mathbf{E}_0\right]e^{i(\mathbf{k}_0\mathbf{r}-\omega t)}$$

где $\Delta_\perp = \partial^2/\partial y^2 + \partial^2/\partial z^2$( опущен член со второй производной $\partial^2\mathbf{E}_0/\partial x^2$, не имеющий большого множителя $k_0$; поперечные же производные могут быть велики по сравнению с продольными) Вычисление $\partial^2\mathbf{D}^{(1)}/\partial t^2$ производится аналогично выводу формулы (80.10) и дает

$$\frac{1}{c^2}\frac{\partial^2\mathbf{D}^{(1)}}{\partial t^2} \approx -\frac{1}{c^2}\left(\omega^2\varepsilon(\omega)\mathbf{E}_0 + i\frac{\partial\left(\omega^2\varepsilon\right)}{\partial\omega}\frac{\partial\mathbf{E}_0}{\partial t}\right)e^{i(\mathbf{k}_0\mathbf{r}-\omega t)} =$$

$$= -\left(k_0^2\mathbf{E}_0 + 2i\frac{k_0}{u}\frac{\partial\mathbf{E}_0}{\partial t}\right)e^{i(\mathbf{k}_0\mathbf{r}-\omega t)}$$

где введена групповая скорость $u$ по определению

$$\frac{1}{u} = \frac{dk_0}{d\omega} = \frac{1}{c}\frac{\partial(\omega\sqrt{\varepsilon})}{\partial\omega}$$

В производной же от $\mathbf{D}^{(3)}$ достаточно оставить член

$$\frac{1}{c^2}\frac{\partial^2\mathbf{D}^{(3)}}{\partial t^2} = -2\eta(\omega)|\mathbf{E}_0|^2\mathbf{E}$$

пренебрегая здесь членом с малой производной $\partial\mathbf{E}_0/\partial t$. Подставив полученные выражения в (109.5), получим окончательно следующее уравнение:

$$ik_0\left(\frac{\partial}{\partial x} + \frac{1}{u}\frac{\partial}{\partial t}\right)\mathbf{E}_0 = -\frac{1}{2}\Delta_\perp\mathbf{E}_0 - \eta(\omega)|\mathbf{E}_0|^2\mathbf{E}_0$$

Комбинация производных в левой части равенства выражает собой тот факт, что возмущения амплитуды переносятся в направлении распространения волны с групIовой скоростью.

С помощью этого уравнения можно исследовать устойчивость неограниченной плоской волны, описываемой точным решением (109.7), (109.8)( В.

И.

Беспалов, В.

И.

Таланов, 1966).

Мы увидим, что в фокусирующей среде волна оказывается неустойчивой.

Согласно (109.9) в точном решении (109.7)

$$k \approx k_0 + \frac{\eta E_0^2}{k_0}$$

с $k_0$ из (109.11); в линейно поляризованной волне амплитуда $\mathbf{E}_0$ может быть определена как вещественный вектор.

Поэтому если записать волну (109.7) в виде (109.10), в последнем надо полоЖ ИТь

$$\mathbf{E}_0(x) = \mathbf{E}_0 \exp\left(ix\frac{\eta E_0^2}{k_0}\right)$$

Это выражение играет роль амплитуды невозмушенной волны.

Мы будем рассматривать стационарную задачу о пространственном развитии возмушений вдоль направления распространения волны.

Соответственно, амплитуду волны, подвергнутой малому возмущению, пишем как

$$\mathbf{E}_0(\mathbf{r}) = \{\mathbf{E}_0 + \delta\mathbf{E}(\mathbf{r})\} \exp\left(ix\frac{\eta E_0^2}{k_0}\right)$$

Будем считать, что $\delta\mathbf{E}$ направлено вдоль $\mathbf{E}_0$.

Подстановка (109.14) в (109.13) приводит к уравнению

$$ik_0\frac{\partial \delta E}{\partial x} = -\frac{1}{2}\Delta_\perp \delta E - \eta E_0^2\left(\delta E + \delta E^*\right)$$

Положим

$$\delta E = Ae^{i(\mathbf{qr}+\Gamma x)} + B^* e^{-i(\mathbf{qr}+\Gamma x)}$$

где q - вектор в плоскости уд.

Подставив это выражение в (109.15) и собрав по отдельности члены с $\exp\{\pm i(\mathbf{qr}+\Gamma x)\}$, по- лучим два уравнения

$$\left(\frac{q^2}{2} - \eta E_0^2 + k_0\Gamma\right)A - \eta E_0^2 B = 0$$
$$-\eta E_0^2 A + \left(\frac{q^2}{2} - \eta E_0^2 - k_0\Gamma\right)B = 0$$

Условие равенства нулю их определителя дает

$$\Gamma = \pm\frac{q}{2k_0}\sqrt{q^2 - 4\eta E_0^2}$$

При $\eta > 0$ и

$$q^2 < 4\eta E_0^2$$

Г мнимо, так что $\delta E$ из (109.16) содержит экспоненциально возрастающий член, т.е. волна неустойчива.

Отметим, что максимальный инкремент неустойчивости оказывается порядка величины нелинейной поправки к волновому вектору. Одно из проявлений этой неустойчивости - самофокусировка ограниченного по ширине светового пучка, распространяющегося в фокусирующей среде.

Происхождение этого явления связано с тем, что если амплитуда поля убывает от оси пучка к его периферии, то зависяшая от этой амплитуды диэлектрическая проницаемость среды тоже убывает (при $\eta > 0$) в том же направлении и среда ведет себя как фокусируюшая линза ( Г.

*А.* Аскарьяи, 1962).

Поведение пучка определяется игрой двух противоположных тенденций - такой фокусировкой и расширением пучка из-за дифракции.

Покажем, прежде всего, что эти тенденции могут взаимно компенсироваться в том смысле, что уравнение (109.13) допускает (при $\eta > 0$) решение в виде стационарного нерасширяющегося пучка.

Такое самоканалирование - специфический нелинейный эффект.

В линейной теории всякий ограниченный по сечению

пучок расходится из-за дифракции.

Мы ограничимся одномерным случаем, когда поле $\mathbf{E}$ зависит только от одной поперечной координаты $y$, будучи поляризовано вдоль оси $z$; волна распространяется вдоль оси $x$[1]).

В этом случае можно получить аналитическое решение задачи (В.И.

Таланов, 1965).

При этом мы отвлекаемся от того обстоятельства, что пучок бесконечной (в направлении оси $z$) ширины заведомо неустойчив, поскольку в нем возможны возмущения с малыми значениями $q_z$, неустойчивые согласно (109.17) Положим

$$E_{0z} = F(y)e^{ixx}$$

с малой величиной $x$, играющей роль поправки к волновому вектору $k_0$; функция $F(y)$ вещественна.

Подстановка в (109.13) дает для этой функции уравнение

$$\frac{1}{2}\frac{d^2F}{dy^2} = k_0 x F - \eta F^3$$

Оно имеет первый интеграл

$$\frac{1}{2}\left(\frac{dF}{dy}\right)^2 - k_0 x F^2 + \frac{\eta}{2}F^4 = \text{const}.$$

Нас интересует решение, в котором $F$ и $dF/dy$ стремятся к нулю при $|y| \to \infty$. Соответственно этому полагаем const $= 0$, после чего простое интегрирование дает

$$F = \left(\frac{2k_0 x}{\eta}\right)^{1/2} \frac{1}{\operatorname{ch}\left[(2k_0 x)^{1/2} y\right]}$$

(начало отсчета $y$ выбрано в центре пучка).

Ширина пучка по оси $y$ :

$$\delta \sim (k_0 x)^{-1/2} \sim \frac{1}{\sqrt{\eta}F(0)}$$

Поскольку протекающий по пучку поток энергии $W \sim E^2(0)\delta$, то $\delta$ пропорционально $1/W$ - пучок тем уже, чем больше переносимая им мощность.

Такой самоканалированный пучок представляет особый случай, когда фокусирующие свойства среды точно компенсиру- ются дифракцией.

Другие пучки будут либо расходиться, либо сходиться.

Напишем, прежде всего, качественный критерий самофокусировки для реального пучка ограниченного сечения.

Это можно сразу сделать исходя из условия неустойчивости (109.17).

В пучке с характерным радиусом $R$ возможны возмущения с поперечными к оси пучка длинами волн, меньшими $R$, т.е.

$cq \gtrsim 1/R$. Условие же (109.17) определяет верхнюю границу значений $q$, приводяших к неустойчивости.

Поэтому пучок будет неустойчив относительно фокусировки при

$$E_0^2 R^2 \eta \gtrsim 1$$

Переносимая вдоль пучка мощность определяется произведением $E_0^2 R^2$. Отметим, что критическое значение этой мощности, за которым начинается самофокусировка, не зависит от площади сечения пучка.

Оказывается возможным также установить точный (не по порядку величины) достаточный критерий самофокусировки пучка.

Для стационарного линейно поляризованного светового пучка, но без предварительных предположений о характере зависимости от $x$, уравнение для функции $E_0(x, \boldsymbol{\rho})$ имеет вид

$$ik_0 \frac{\partial E_0}{\partial x} = -\frac{1}{2}\Delta_\perp E_0 - \eta \left|E_0\right|^2 E_0$$

)$\boldsymbol{\rho}$ - двумерный радиус-вектор в плоскости $yz$; в этой же плоскости действуют дифференциальные операторы $\Delta_\perp$ и $\nabla_\perp$). .Тегко проверить, что из этого уравнения следует равенство

$$\frac{\partial \left|E_0\right|^2}{\partial x} + \mathrm{div}_\perp \mathbf{j} = 0$$

где

$$\mathbf{j} = \frac{i}{2k_0}\left(E_0 \nabla_\perp E_0^* - E_0^* \nabla_\perp E_0\right)$$

Отсюда, в свою очередь, следует «сохранение» (т.е.

независимость от $x$) интеграла

$$N = \int\limits_{-\infty}^{\infty} \left|E_0\right|^2 d^2\rho$$

Сохраняется также и интеграл

$$\mathrm{E} = \frac{1}{2k^2} \int\limits_{-\infty}^{\infty} \left\{\left|\nabla_\perp E_0\right|^2 - \eta \left|E_0\right|^4\right\} d^2\rho$$

в чем легко убедиться прямым дифференцированием по $x$ с использованием уравнения (109.22). Предполагается, конечно, что $E_0$ достаточно быстро убывает при $\rho \to \infty$, так что оба интеграла (равно как и интеграл (109.26) ниже) сходятся.

Покажем, что поведение пучка определяется знаком интеграла E: при $\mathrm{E} > 0$ пучок в среднем расходится, а при $\mathrm{E} < 0$ -

фокусируется.

Доказательство основано на простом уравнении, которое можно получить для среднего радиуса пучка $\bar{R}$, определенного согласно

$$R^2(x) = \frac{1}{N} \int\limits_{-\infty}^{\infty} \rho^2 \left|E_0\right|^2 d^2\rho$$

Для вывода этого уравнения пишем, используя (109.23):

$$\frac{d}{dx} \int \left|E_0\right|^2 \rho^2 d^2\rho = - \int \mathrm{div}_\perp \mathbf{j} \cdot \rho^2 d^2\rho = 2 \int \mathrm{j}\rho d^2\rho$$

Дифференцируя еще раз по $x$, подставив $\partial E_0/\partial x$ из (109.22) и интегрируя два раза по частям, получим в результате уравнение

$$N \frac{d^2}{dx^2} R^2 = 4\mathrm{E}$$

Отсюда

$$R^2(x) = \frac{2\mathrm{E}}{N} \left(x - x_0\right)^2 + R_0^2$$

где $x_0, R_0$ - постоянные.

Мы видим, что при $\mathrm{E} < 0$ на конечном расстоянии вдоль направления распространения достигается полная фокусировка пучка - его радиус $R$ обращается в нуль. Этот результат, полученный в рамках приближенного уравнения (109.13), не может иметь буквального физического смысла вблизи самого фокуса, где нарушаются предположения, сделанные при выводе уравнения.

Достаточно сказать, что при неограниченном увеличении плотности энергии поля при точной фокусировке уже нет оснований ограничиваться низшей степенью нелинейности - кубической.

Но существенна уже самая возможность самофокусировки пучка до такой степени, когда нелинейность перестает быть малой.

Подчеркнем, что установленный критерий носит достаточный, но не необходимый характер.

Пучок $\mathrm{cE} < 0$ заведомо фокусируется целиком, но расходимость пучка в среднем при $\mathrm{E} \geqslant 0$ не противоречит тому, что некоторая внутренняя его часть сфокусируется.

**применения**

?????

## 20.3.4 Strong Electromagnetic Waves

(а что в них такого?)

**Теория**

(добавлю наверное в запись про электромагнитные волны позже)

Возможность постановки рассмотренной в предыдущем параграфе задачи о генерации всего одной определенной гармоники была связана с наличием дисперсии.

Теперь мы рассмотрим обратный случай, когда во всем существенном интервале частот дисперсию можно считать отсутствующей, так что индукция $\mathbf{D}(t)$ в каждой точке определяется значением $\mathbf{E}(t)$ в тот же момент времени.

Мы будем считать среду изотропной; тогда направления **D** и E совпадают.

Нелинейность же в этом параграфе не предполагается малой, так что зависимость $D(E)$ - произвольная функция.

Пренебрежение поглощением и дисперсией имеет принципиальное значение в том отношении, что после этого из уравнений поля исчезают какие-либо параметры размерности частоты (или, тем самым, размерности длины).

Это обстоятельство делает возможным построение точного решения, обобщающего обычную одномерную плоскую волну линейного приближения.

**решение уравнений Максвелла для сильной волны**

Пусть волна распространяется в направлении оси $x$, электрическое поле направлено по оси $y$, а магнитное - по оси z ($E_y$ и $H_z$ обозначаем просто как $E$ и $H$).

Уравнения Максвелла

$$\operatorname{rot}\mathbf{H}=\frac{1}{c}\frac{\partial\mathbf{D}}{\partial t},\quad \operatorname{rot}\mathbf{E}=-\frac{1}{c}\frac{\partial\mathbf{H}}{\partial t}$$

принимают вид

$$-\frac{\partial H}{\partial x}=\frac{1}{c}\frac{\partial D}{\partial t}=\frac{\varepsilon}{c}\frac{\partial E}{\partial t},\quad -\frac{\partial E}{\partial x}=\frac{1}{c}\frac{\partial H}{\partial t},$$

где по определению

$$\varepsilon(E)=\frac{dD}{dE}$$

(при $E\to 0$ функция $\varepsilon(E)$ стремится к значению $\varepsilon_0$ - обычной диэлектрической проницаемости).

Ищем решение, в котором функции $E(t,x)$ и $H(t,x)$ могут быть выражены как функция одна другой: $H=H(E)$.

Тогда уравнения (111.1) можно переписать как

$$\frac{\varepsilon}{c}\frac{dE}{dt}+\frac{dH}{dE}\frac{\partial E}{\partial x}=0,\quad \frac{1}{c}\frac{dH}{dE}\frac{\partial E}{\partial t}+\frac{\partial E}{\partial x}=0$$

Для того чтобы эти два уравнения могли удовлетворяться отличными от нуля значениями неизвестных $\partial E/\partial t$ и $\partial E/\partial x$, должен быть равен нулю их определитель.

Это условие дает

$$\left(\frac{dH}{dE}\right)^2=\varepsilon(E)$$

откуда

$$H=\pm\int_0^E\sqrt{\varepsilon(E)}dE$$

Подставив теперь $dH/dE$ из (111.4) в одно из уравнений (111.3), имеем

$$-\frac{(\partial E/\partial t)_x}{(\partial E/\partial x)_t}=\left(\frac{\partial x}{\partial t}\right)_E=\pm\frac{c}{\sqrt{\varepsilon}}$$

Отсюда следует, что

$$x\mp\frac{c}{\sqrt{\varepsilon}}t$$

может быть произвольной функцией от $E$. Обозначив обратную функцию как $f$, имеем

$$E=f\left(x\mp\frac{c}{\sqrt{\varepsilon(E)}}t\right)$$

два знака здесь отвечают двум направлениям распространения волны.

После выбора функции $f$ формула (111.5) определяет в неявном виде зависимость $E(t,x)$. В слабых полях, когда можно положить $\varepsilon = \varepsilon_0$, (111.5) переходит в обычную плоскую волну с фазовой скоростью $c/\sqrt{\varepsilon_0}$.

Отметим, что полученное решение

сушествует только при $\varepsilon > 0$ - в соответствии с условием устойчивости (18.8)[1]).

По мере распространения волны заданный в начальный момент ее профиль искажается, поскольку разные участки бегут с различными скоростями.

Обычно $\varepsilon(E)$ убывает с ростом $E$ ( функция $\varepsilon(E)$ стремится к насыщению).

Тогда точки профиля с большими значениями $E$ бегут с большими скоростями, в результате чего увеличивается крутизна переднего фронта профиля (как это иллюстрирует рис.

60, на котором показана форма профиля в несколько последовательных моментов).

В некоторый момент возникает перегиб профиля, после чего он должен был бы стать неоднозначным.

В действительности, в этот момент в волне возникает электромагнитная ударная волна - разрыв величин $E$ и $H$. Граничные условия на разрыве имеют тот же вид (76.13), что и на любой движушейся поверхности.

Для плоской поперечной волны имеем:

$$H_2 - H_1 = \frac{v}{c}(D_2 - D_1)$$

сушествует только при $\varepsilon > 0$ - в соответствии с условием устойчивости (18.8)[1]).

По мере распространения волны заданный в начальный момент ее профиль искажается, поскольку разные участки бегут с различными скоростями.

Обычно $\varepsilon(E)$ убывает с ростом $E$ ( функция $\varepsilon(E)$ стремится к насыщению).

Тогда точки профиля с большими значениями $E$ бегут с большими скоростями, в результате чего увеличивается крутизна переднего фронта профиля (как это иллюстрирует рис.

60, на котором показана форма профиля в несколько последовательных моментов).

В некоторый момент возникает перегиб профиля, после чего он должен был бы стать неоднозначным.

В действительности, в этот момент в волне возникает электромагнитная ударная волна - разрыв величин $E$ и $H$. Граничные условия на разрыве имеют тот же вид (76.13), что и на любой движушейся поверхности.

Для плоской поперечной волны имеем:

$$H_2 - H_1 = \frac{v}{c}(D_2 - D_1)$$

$$E_2 - E_1 = \frac{v}{c}(H_2 - H_1)$$

где индексы 1 и 2 относятся к значениям величин соответственно впереди и позади фронта.

Перемножив эти два равенства, найдем для скорости ударной волны:

$$v^2 = c^2 \frac{E_2 - E_1}{D_2 - D_1}$$

В ударной волне происходит диссипация энергии.

Пусть $Q$ скорость диссипации, отнесенная к единице площади поверхности разрыва.

Для ее вычисления пишем закон сохранения энергии, примененный к цилиндрическому элементу объема среды, одно основание которого находится позади, а другое - впереди разрыва:

$$\frac{c}{4\pi}\left(E_2H_2 - E_1H_1\right) = v\left(U_2 - U_1\right) + Q$$

Слева стоит разность потоков энергии через оба основания, а справа - сумма скорости изменения внутренней энергии за счет

перемешения границы между областями 1 и 2 и диссипируемой в ней энергии.

Разность внутренних энергий (при неизменных плотности и температуре):

$$U_2 - U_1 = \frac{1}{4\pi}\int_{D_1}^{D_2} EdD + \frac{1}{8\pi}\left(H_2^2 - H_1^2\right)$$

Используя также равенства (111.6), (111.7), можно привести (111.8) к виду

$$Q = \frac{v}{4\pi}\left\{\frac{1}{2}\left(D_2 - D_1\right)\left(E_2 + E_1\right) - \int_{D_1}^{D_2} EdD\right\}$$

Если ударная волна слабая (т.е.

скачки величин в ней малы), то при вычислении $Q$ можно представить связь $D$ с $E$ в виде разложения

$$D(E) = D_1 + \varepsilon\left(E_1\right)\left(E - E_1\right) + \frac{1}{2}\varepsilon'\left(E_1\right)\left(E - E_1\right)^2$$

где $\varepsilon'(E) = d^2D/dE^2$. Простое вычисление приводит к результату

$$Q = -\frac{1}{48\pi}v\varepsilon'\left(E_1\right)\left(E_2 - E_1\right)^3$$

Таким образом, диссипация энергии в слабой электромагнитной ударной волне - третьего порядка по величине скачка поля в ней.

Поскольку должно быть $Q > 0$, то при $\varepsilon' < 0$ будет $E_2 > E_1$ - в соответствии с рис. 60.

Появление ударной волны нарушает применимость полученного решения: выражения (111.4),(111.5) для поля противоречат граничным условиям (111.6).

Существенно, однако, что волна остается приближенно (с точностью до величин второго порядка включительно) простой, пока ударную волну можно считать слабой.

С этой точностью выражение для скорости разрыва может быть представлено в виде

$$v = c\left[\varepsilon\left(\frac{E_1 + E_2}{2}\right)\right]^{-1/2}$$

В этом же приближении положение разрыва на профиле волны определяется условием равенства двух площадей между вертикальной прямой и штриховой кривой на рис. 60.

(там крч волна в цунами превращает, удивительно)

**определение отраженной волны**

Из вакуума на границу среды падает нормально плоская волна вида $E = f_i(t - x/c)$. Определить отраженную волну.

**Solution**

Поле в вакууме (полупространство $x < 0$) складывается из падающей и отраженной (индекс $r$) волн:

$$E = f_i\left(t - \frac{x}{c}\right) + f_r\left(t + \frac{x}{c}\right)$$
$$H = f_i\left(t - \frac{x}{c}\right) - f_r\left(t + \frac{x}{c}\right)$$

(в вакууме уравнения поля линейны и два решения можно складывать!).

В среде, $x > 0$, имеется лишь прошедшая волна, в которой

$$E = f_t\left(t - \frac{\sqrt{\varepsilon(E)}}{c}x\right), \quad H = \int_0^E \sqrt{\varepsilon} dE$$

Из условия непрерывности при $x = 0$ электрического поля имеем

$$f_t(t) = f_i(t) + f_r(t)$$

Условие же непрерывности $H$ на той же границе дает затем соотношение

$$f_i(t) - f_r(t) = \int_0^{f_i(t)} + f_r(t) \quad \sqrt{\varepsilon(E)} dE$$

которым и определяется, в неявном виде, функция $f_r$.

(????? конкретно потом попробую посмотреть, какой она будет по идее ничего особого не должно быть, такой же плоской, если падала плоская???)

## 20.3.5 Stimulated Raman Scattering

**обзор**

**итоговые формулы**

xз

**Theory**

К числу нелинейных эффектов третьего порядка относится влияние, оказываемое излучением некоторой частоты $\omega_1$ (волна накачки на распространение в той же среде волны другой частоты $\omega_2$. Эти эффекты заключены в нелинейной проницаемости

$$\varepsilon_{i,klm}\left(\omega_2, \omega_1, -\omega_1\right)$$

дающей вклад в индукцию с частотой $\omega_2$.

В изотропной среде индукция $\mathbf{D}_2$ на частоте $\omega_2$ с учетом указанного вклада дается выражением

$$\mathbf{D}_2 = \varepsilon_2 \mathbf{E}_2 + \alpha_2 \left(\mathbf{E}_1 \mathbf{E}_1^*\right) \mathbf{E}_2 + \beta_2 \mathbf{E}_1 \left(\mathbf{E}_1^* \mathbf{E}_2\right) + \Gamma_2 \mathbf{E}_1^* \left(\mathbf{E}_1 \mathbf{E}_2\right)$$

где

$$\mathbf{E}_1 = \mathbf{E}_{10} e^{i(\mathbf{k}_1 \mathbf{r} - \omega_1 t)}, \quad \mathbf{E}_2 = \mathbf{E}_{20} e^{i(\mathbf{k}_2 \mathbf{r} - \omega_2 t)}$$

В первом члене в (112.2)$\varepsilon_2 = \varepsilon_2\left(\omega_2\right)$ - обычная линейная проницаемость; в остальных членах $\alpha_2, \beta_2, \Gamma_2$ - три независимые компоненты тензора (112.1) (их число очевидно из самого способа построения выражения (112.2) из трех векторов $\mathbf{E}_1, \mathbf{E}_1^*, \mathbf{E}_2$).

Мы видим, что нелинейное воздействие поля $\mathbf{E}_1$ на поле с частотой $\omega_2$ может быть описано путем введения анизотропной диэлектрической проницаемости

$$\varepsilon_{2ik} = (\varepsilon_2 + \alpha_2 \mathbf{E}_1 \mathbf{E}_1^*)\,\delta_{ik} + \beta_2 E_{1i} E_{1k}^* + \Gamma_2 E_{1i}^* E_{1k}$$

В недиссипативной среде коэффициенты $\alpha_2, \beta_2, \Gamma_2$ (как и $\varepsilon_2$) вещественны и тензор (112.4) эрмитов. Как частный случай, при $\omega_1 \to 0$ и соответственно вещественном $\mathbf{E}_1$, он содержит в себе двойное преломление в статическом электрическом поле, описываемое формулой (100.1). При $\omega_1 \neq 0$ выражение (112.4) описывает также наведенную полем $\mathbf{E}_1$ гиротропию среды.

Сравнив (112.2)со(101.8), найдем вектор гирации (???)

$$\mathbf{g} = \frac{i}{2}\,(\beta_2 - \Gamma_2)\,[\mathbf{E}_1^* \mathbf{E}_1]$$

Он обращается в нуль, если поле $\mathbf{E}_1$ линейно поляризовано.

Более разнообразные явления могут происходить, если нелинейные взаимодействия поля со средой сопровождаются диссипацией. В таком случае коэффициенты $\alpha_2, \beta_2, \Gamma_2$ комплексны (линейную же проницаемость по-прежнему будем считать вещественной). Оказывается, что такая диссипация может приводить как к ослаблению, так и к усилению поля $\mathbf{E}_2$. В последнем случае говорят о вынужденном комбинационном рассеянии. (!?!!!!)

Вещественность линейных проницаемостей $\varepsilon(\omega_1), \varepsilon(\omega_2)$ означает, что на самих частотах $\omega_1$ и $\omega_2$ в среде нет диссипации: кванты $\hbar\omega_1$ и $\hbar\omega_2$ сами по себе не поглощаются средой.

Пусть в области частот, где среда способна к поглощению, лежит разность $\omega_1 - \omega_2$, но не сумма $\omega_1 + \omega_2$. Лиссипация осуществляется лишь путем превращения квантов большей энергии в кванты меньшей

энергии с отдачей освобождающегося избытка среде.

Таким образом, при $\omega_1 > \omega_2$ волна накачки усиливает волну меньшей частоты $\omega_2$. Усредненная по времени энергия, получаемая полем $\mathbf{E}_2$ (в единицу времени в единице объема) за счет слабых нелинейных эффектов, непосредственно дается взятым с обратным знаком выражением (96.5)

$$\frac{d\bar{U}_2}{dt} = -\frac{i\omega_2}{8\pi}\,(\varepsilon_{2ik}^* - \varepsilon_{2ki})\,E_{2i} E_{2k}^* = \\ = -\frac{\omega^2}{4\pi}\left\{\alpha_2''\,|\mathbf{E}_1|^2\,|\mathbf{E}_2|^2 + \beta_2''\,|\mathbf{E}_1\mathbf{E}_2|^2 + \Gamma_2''\,|\mathbf{E}_1\mathbf{E}_2^*|^2\right\}$$

(ср.

вывод формул (108.9)). Аналогичным выражением дается изменение энергии поля $\mathbf{E}_1$:

$$\frac{d\bar{U}_1}{dt} = -\frac{\omega_1}{4\pi}\left\{\alpha_1''\,|\mathbf{E}_1|^2\,|\mathbf{E}_2|^2 + \beta_1''\,|\mathbf{E}_1\mathbf{E}_2|^2 + \Gamma_1''\,|\mathbf{E}_1\mathbf{E}_2^*|^2\right\}$$

где $\alpha_1, \beta_1, \Gamma_1$ - независимые компоненты тензора проницаемости

$$\varepsilon_{i,klm}(\omega_1, \omega_2, -\omega_2)$$

описывающей влияние поля частоты $\omega_2$ на поле частоты $\omega_1$.

Соображения, подобные использованным в §107 при выводе теоремы Мэнли-Роу, позволяют утверждать, что

$$\frac{1}{\omega_1}\frac{d\bar{U}_1}{dt} = -\frac{1}{\omega_2}\frac{d\bar{U}_2}{dt}$$

- на каждый рождающийся квант $\hbar\omega_2$ приходится по одному исчезающему кванту $\hbar\omega_1$. Отсюда следует, что

$$\alpha_1'' = -\alpha_2'', \quad \beta_1'' = -\beta_2'', \quad \Gamma_1'' = -\Gamma_2''$$

Диссипируемая энергия определяется по убыли суммарной энергии обоих полей:

$$Q = -\frac{d\bar{U}_1}{dt} - \frac{d\bar{U}_2}{dt} = \frac{\omega_1 - \omega_2}{\omega_2}\frac{d\bar{U}_2}{dt}$$

При $\omega_1 > \omega_2$ из условия $Q > 0$ следует, что $d\bar{U}_2/dt > 0$ - волна меньшей частоты усиливается в соответствии со сказанным выше.

Условие положительности выражения (112.6) дается нера венствами

$$\alpha_2'' < 0, \quad \alpha_2'' + \beta_2'' < 0, \quad \alpha_2'' + \Gamma_2'' < 0, \quad \alpha_2'' + \beta_2'' + \Gamma_2'' < 0$$

Обратим внимание на то, что рассмотренный эффект не зависит от фазовых соотношений между полями.

Это свяЗано с тем, что поле волны накачки входит в уравнения в виде билинейных $\mathbf{E}_1$ и $\mathbf{E}_1^*$ выражений, из которых фазовые множители выпадают.

В конечном счете это приводит к тому, что длЯ усиления поля $\omega_2$ не требуется синхронизма полей - в противопОложность рассмотренным в §110 явлениям генерации гармоники и параметрического усиления сигнала.

Оказывается возможным свяЗать характеристики вынужденного комбинационного рассеяния с характеристиками обычного (спонтанного) рассеяния, о котором будет идти речь в гл. XV.

Соответствующие вычисления приведены в задаче к $118.

Приведенные выше соотношения справедливы, как уже было указано, если поглощение энергии средой происходит лишь на разностной частоте $\omega_1 - \omega_2$. Лругая ситуация имеет место, если в области поглощения лежит не разностная, а суммарная частоТа $\omega_1 + \omega_2$. В этом случае на каждый поглощенный квант $\hbar\omega_2$ поглощается также квант $\hbar\omega_1$, и среде отдается энергия, равная $\hbar(\omega_1 + \omega_2)$ ( двухфотионное поглощение).

Естественно, что в этом случае ослаЮтся волны обеих частот.

## 20.3.6 Second Harmonic Generation

### обзор

### итоговые формулы

хз

### количественная теория генерации второй гармоники

Изложим количественную теорию генерации второй гармоники, то есть возбуждения электромагнитного поля с частотой $2\omega$ полем частоты $\omega$.

Генерация второй гармоники - нелинейный эффект второго порядка. Он содержится в тензоре нелинейной восприимчивости

$$\varepsilon_{i,kl}(-2\omega;\omega,\omega)$$

и потому отсутствует в средах, допускающих пространственную инверсию.

Тензор $\varepsilon_{i,kl}(-2\omega;\omega,\omega)$ симметричен по индексам $kl$; его свойства симметрии в различных кристаллах - те же, что и у пьезоэлектрического тензора (§17). Мы будем считать среду непоглощающей, так что $\varepsilon_{i,kl}$ - вещественные величины.

Задача о генерации второй гармоники может быть поставлена следующим образом. Пусть на плоскую поверхность кристаллической среды падает плоская монохроматическая волна с частотой $\omega$. Наряду с отраженной и двумя преломленными (в двупреломляющем кристалле) волнами той же частоты возникают также отраженная и преломленные волны частоты $2\omega$. Волны этой частоты в кристалле представляют собой решение уравнений (!?!!??!?!)(109.5), (109.6), в которых нелинейный член в индукции, $\mathbf{D}^{(2)}$, должен быть выражен через поле основной волны. Амплитуды всех этих волн выражаются через амплитуду падающей волны с помощью граничных условий, на которых мы не будем останавливаться. Разумеется, амплитуда волн частоты $2\omega$ будет мала в меру малости нелинейной восприимчивости.

Преломленные волны распространяются далее в глубь кристалла как в неограниченной среде. По мере их распространения эффекты нелинейности накапливаются и интенсивность гармоник может достичь больших значений - происходит перекачка энергии из основной частоты в гармонику. Именно этот процесс и будет интересовать нас здесь.

Условия же на поверхности кристалла играют при этом лишь роль «начальных» условий, задающих некоторую малую, но отличную от нуля, амплитуду поля второй гармоники. Эти же условия задают (для данного направления падающей волны) волновые векторы первой $\mathbf{k}_1$ и второй $\mathbf{k}_2$ гармоник в кристалле.

Как будет видно из дальнейшего, эффективная перекачка энергии происходит, только если выполняется условие синхронизма основной волны и гармоники [2]):

$$\mathbf{k}_2 \approx 2\mathbf{k}$$

Подчеркнем, что для самой постановки задачи о генерации одной лишь второй гармоники принципиальное значение имеет наличие дисперсии. В отсутствие дисперсии условие (110.2) выполняется при преломлении автоматически вместе с аналогичными условиями и для более высоких гармоник ($\mathbf{k}_3 \approx 3\mathbf{k}_1$ и т. п.). При наличии дисперсии это уже не так и можно считать, что условие синхронизма, выполняясь для второй гармоники, не выполняется для других.

Подчеркнем, что уже само условие (110.2) фактически может быть выполнено, только если основная волна и гармоника относятся к различным поляризационным типам и соответственно, имеют различные законы дисперсии.

**Solution задачи о второй гармонике (???)**

Запишем поле в среде как суперпозицию двух волн:

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = \mathrm{Re}\left(\mathbf{e}_1 E_{10} e^{i(\mathbf{k}_1\mathbf{r}-\omega t)} + \mathbf{e}_2 E_{20} e^{i(\mathbf{k}_2\mathbf{r}-2\omega t)}\right]$$

причем, в силу условия $\mathbf{k}_2 \approx 2\mathbf{k}_1$,

$$\mathbf{k}_2 = 2\mathbf{k}_1 + \mathbf{q}$$

где $q \ll k_1$. Амплитуды волн представлены в виде произведений $\mathbf{E}_0 = \mathrm{e}E_0$, где e - единичный ($\mathrm{ee}^* = 1$) вектор поляризации. В линейном приближении эти амплитуды были бы постоянны, а учетом нелинейности они - медленно (т.е. мало на расстояниях $\sim 1/k_1$) меняющиеся функции координат.

Уравнения для амплитуд обеих волн получаются путем подстановки (110.3) в уравнения Максвелла (109.5),(109.6) и отделения в них членов с одинаковой зависимостью от времени.

Мы не станем подробно приводить здесь эти простые, но громоздкие вычисления и ограничимся лишь некоторыми принципиальными указаниями.

Мы будем искать решения уравнений, описывающие стационарную генерацию второй гармоники в кристалле бегушей основной волной; такие решения не зависят от времени.

При точном синхронизме ($\mathbf{q} = 0$) уравнения для амплитуд не содержали бы вовсе координат в явном виде; при неточном синхронизме координаты входят в уравнение только в комбинации qr (в множителе $\exp(-i\mathbf{qr})$). Выбрав направление вектора $q$ в качестве оси $z$, можно поэтому искать решения уравнений, зависящие только от $z$. Если исходить из указанной выше постановки вопроса о падении волны $\omega$ на поверхность кристалла, то задача Однородна в плоскостях, параллельных этой поверхности.

Поэтому ось $z$ ей перпендикулярна (перпендикулярность вектора $\mathbf{q}$ поверхности кристалла выполняется автоматически в силу граНИчных условий).

В линейном приближении волны в анизотропной (негиротропной) среде линейно поляризованы (см.

$97); для них е может быть определен как вещественный вектор, и именно эти значения для обеих волн будут подразумеваться ниже под $\mathbf{e}_1$ и $\mathbf{e}_2$.

Если разложить амплитуду $\mathbf{E}_0$ каждой волны по направлениям $\mathbf{e}, \mathbf{k}, [\mathbf{ek}]$, то составляющиие вдОль двух последних направлений ся как эффект нелинейности.

Составляюшщие же вдоль е приближенно совпадают с абсолютными величинами $E_0$ векторов $\mathbf{E}_0$.

Уравнения для них получаются умножением уравнения (109.5) на вектор $\mathbf{e}_1$ Или $\mathbf{e}_2$. Поскольку волны $cE_0 = \text{const}$ предполагаются точными решениями уравнений Максвелла в линейном приближении, то все линейные члены в уравнениях, не содержашие производных по $z$, взаимно сокращаются.

Члены же с составляющими $\mathbf{E}_0$ по направлениям $\mathbf{k}$ и $[\mathbf{ek}]$ (которые могли бы быть того же порядка величины, что и члены с производными $dE_0/dz$), как оказывается, выпадают при указанном умножении; это обстоятельство связано с ортогональностью индукции $\mathbf{D}$ направлениям как $\mathbf{k}$, так и $[\mathbf{ek}]$( см. (97.3)).

Ввиду предполагаемой медленности зависимости амплитуд от координат, в уравнениях можно пренебречь вторыми производными от $\mathbf{E}_0$ по $z$. Поэтому, например, имеющееся в уравнении для $\mathbf{E}_2$ (умноженном на $\mathbf{e}_2$) выражение

$$e_{2i}\left\{(\mathrm{rot}\,\mathrm{rot})_{ik} - \frac{4\omega^2}{c^2}\varepsilon_{ik}(2\omega)\right\} e_{2k}E_{20}e^{i\mathbf{k}_2\mathbf{r}}$$

с учетом всего сказанного приближенно сводится к

$$2i\mathbf{e}_2\left[\mathbf{k}_2\left[\mathbf{le}_2\right]\right]\frac{dE_{20}}{dz}$$

(где $\mathbf{l}$ - орт оси $z$), и аналогично для $\mathbf{E}_1$. Окончательные уравнения, получающиеся в результате описанных действий, гласят:

$$\alpha_2\frac{dE_{20}}{dz} = -i\eta e^{-iqz}E_{10}^2$$
$$\alpha_1\frac{dE_{10}^*}{dz} = i\eta e^{-iqz}E_{10}E_{20}^*$$

где введены обозначения:

$$\eta = \frac{\omega^2}{2c^2}\varepsilon_{i,kl}(\omega,\omega)e_{2i}e_{1k}e_{1l}$$
$$\alpha_1 = \mathbf{l}\left[\mathbf{e}_1\left[\mathbf{k}_1\mathbf{e}_1\right]\right],$$
$$\alpha_2 = \frac{1}{2}\mathbf{l}\left[\mathbf{e}_2\left[\mathbf{k}_2\mathbf{e}_2\right]\right] \approx \mathbf{l}\left[\mathbf{e}_2\left[\mathbf{k}_1\mathbf{e}_2\right]\right]$$

Умножив первое уравнение (110.5) на $E_{20}^*$, а второе - на $E_{10}$ и сложив их, найдем первый интеграл этой системы:

$$\alpha_1 |E_{10}|^2 + \alpha_2 |E_{20}|^2 = \text{const} \equiv P$$

Он выражает собой постоянство суммарного (в обеих волнах) потока энергии вдоль оси $z^2$).

От комплексных величин удобно перейти к вещественным абсолютным значениям и фазам величин $E_{10}$ и $E_{20}$. Для максимального упрощения уравнений введем новые неизвестные $\rho_1$, $\rho_2, \varphi_1, \varphi_2$ как безразмерные величины, определив их согласно

$$E_{10} = \sqrt{\frac{P}{\alpha_1}}\rho_1 e^{i\varphi_1}, \quad E_{20} = \sqrt{\frac{P}{\alpha_2}}\rho_2 e^{i\varphi_2}$$

Система уравнений (110.5) инвариантна относительно преобразования

$$\varphi_1 \to \varphi_1 + c, \quad \varphi_2 \to \varphi_2 + 2c.$$

Поэтому фактически в ней отделяются уравнения для функций $\rho_1, \rho_2$ и для инвариантной комбинации $2\varphi_1 - \varphi_2$, образующие вместе замкнутую систему уравнений:

$$\begin{aligned}
\frac{d\rho_1}{d\zeta} &= -\rho_1\rho_2 \sin\theta, \\
\frac{d\rho_2}{d\zeta} &= \rho_1^2 \sin\theta \\
\frac{d\theta}{d\zeta} &= -s - \left(2\rho_2 - \frac{\rho_1^2}{\rho_2}\right)\cos\theta
\end{aligned}$$

где

$$\theta = 2\varphi_1 - \varphi_2 - s\zeta$$

и введены безразмерная переменная

$$\zeta = z\eta\sqrt{\frac{P}{\alpha_1^2\alpha_2}}$$

и безразмерный параметр

$$s = \frac{q}{\eta}\sqrt{\frac{\alpha_1^2\alpha_2}{P}}$$

Первый интеграл (110.8) в этих переменных принимает вид

$$\rho_1^2 + \rho_2^2 = 1$$

Рассмотрим случай точного синхронизма: $q = 0$, т. е. $s = 0$. Тогда уравнения (110.10),(110.11) имеют еще один первый интеграл:

$$\rho_1^2\rho_2 \cos\theta = \text{const} \equiv \delta$$

(причем постоянная $\delta^2 \leqslant 4/27$; это следует, как легко убедиться, из равенств (110.15),(110.16) в силу условия $|\cos\theta| \leqslant 1$). С использованием обоих этих интегралов решение уравнений (110.10) приводится к квадратурам - к эллиптическому интегралу

$$\zeta = \pm\frac{1}{2}\int\limits_{\rho_2^2(0)}^{\rho_2^2(\zeta)} \frac{du}{[u(1-u)^2 - \delta^2]^{1/2}};$$

выбор знака перед интегралом определяется знаком начального ( при $\zeta = 0$) значения $\sin\theta$. Тут кубическое уравнение

$$u(1-u)^2 - \delta^2 = 0$$

имеет (при $\delta^2 \leqslant 4/27$) три вещественных положительных корня, из которых два меньше единицы; обозначим последние как $\rho_a^2$ и $\rho_b^2$, причем $\rho_a^2 < \rho_b^{21}$). Определяемая формулой (110.17) функция $\rho_2^2(\zeta)$ периодически меняется в пределах между этими значениями с периодом, равным интегралу

$$\int_{\rho_a^2}^{\rho_b^2} \frac{du}{[u(1-u)^2 - \delta^2]^{1/2}}$$

Аналогичным образом меняется и функция $\rho_1^2(\zeta) = 1 - \rho_2^2(\zeta)$, причем, когда одна из них максимальна, другая минимальна.

Величину $\rho_a^2$ надо отождествить со значением интенсивности второй гармоники, $\rho_2^2(0)$, задаваемой граничными условиями на поверхности кристалла ($z = 0$). Мы видим, что в направлении оси $z$ в глубь кристалла происходит периодическая перекачка энергии из основной частоты во вторую гармонику и обратно.

При уменьшении $\rho_2(0)$ период этого процесса возрастает, и при $\rho_2(0) \to 0$ стремится к бесконечности по логарифмическому закону.

Предельному значению $\rho_2(0) = \rho_a = 0$ отвечает решение

$$\rho_1 = \frac{1}{\operatorname{ch}\zeta}, \quad \rho_2 = \operatorname{th}\zeta$$

получающееся из (117.17) при $\delta = 0$ элементарным интегрированием; в нем амплитуда второй гармоники монотонно возрастает и асимптотически при $\zeta \to \infty$ вся энергия переходит из основной частоты в гармонику.

Рассмотрим теперь обратный случай, когда амплитуда $\rho_2$ остается везде малой по сравнению с $\rho_1$. Мы увидим, что этот случай отвечает значительной расстройке синхронизма волн.

При $\rho_2 \ll \rho_1$ можно, в первом приближении, считать $\rho_1$ постоянным ($\rho_1 = \rho_1(0)$), а уравнения для $\rho_2$ и $\theta$ записать в виде

$$\frac{d\rho_2}{d\zeta} = \rho_1^2(0)\sin\theta, \quad \frac{d\theta}{d\zeta} = -s + \frac{\rho_1^2(0)}{\rho_2}\cos\theta.$$

Решение этих уравнений, равное нулю в начальной точке $\zeta = 0$ :

$$\rho_2(\zeta) = \frac{2}{s}\rho_1^2(0)\sin\frac{s\zeta}{2}, \quad \theta = \frac{\pi}{2} - \frac{s\zeta}{2}$$

Эти формулы определяют изменение поля на протяжении одного интервала $0 \leqslant \zeta \leqslant 2\pi/s$ (т. е. $0 \leqslant x \leqslant 2\pi/q$), после чего процесс повторяется периодически. Условие $\rho_2 \ll \rho_1$ означает, что должно быть $\rho_1^2(0)/s \ll \rho_1(0)$, т.е. $s \gg \rho_1(0)$ или

$$qz_0 \gg 1, \quad z_0 \sim \frac{1}{\eta\rho_1(0)\sqrt{P}}$$

Это - условие сравнительно большой расстройки синхронизма.

Вообще, величина параметра $q$ определяет, какой эффект ограничивает генерацию гармоники (т.е.

рост амплитуды $\rho_2$) - линейный эффект нарушения синхронизма при $qz_0 \gg 1$ или нелинейные эффекты при $qz_0 \ll 1^2$).

**Об обратном явлении (??)**

Рассмотренные уравнения описывают и обратное явление: усиление (его называют параметрическим) слабого сигнала частоты $\omega$ в поле интенсивного излучения частоты $2\omega$; этот процесс, рассмотрением которого мы здесь ограничимся, представляет собой простейший случай более общего явления - усиления сигналов с различными частотами $\omega_2$ и $\omega_1 - \omega_2$ в поле интенсивной волны с частотой $\omega_1$ (С. А. Ахманов, Р.В. Хохлов, 1962; R.H. Kingston, 1962).

Прежде всего, подчеркнем следующее отличие этого процесса от генерации второй гармоники. Последняя могла начинаться с нулевой интенсивности гармоники. Возможность же усиления основной частоты во всяком случае требует хотя бы малой ее начальной интенсивности: если в начальной точке $\rho_1(0) = 0$, то Так останется везде; из уравнений (110.10) видно, что вместе с $\rho_1$ обращаются в нуль также и производные всех порядков от $\rho_1$ И $\rho_2$.

Снова рассмотрим случай точного синхронизма и, сверх того, пусть начальное значение фазовой переменной $\theta(0) = -\pi/2$; при точном синхронизме это значение будет сохраняться. Тогда параметр $\delta = 0$ в силу равенства $\cos\theta = 0$, хотя начальные значения $\rho_1$ и $\rho_2$ отличны от нуля. Решение уравнений (110.10) в этом случае есть

$$\rho_1 = \frac{1}{\operatorname{ch}(\zeta - \zeta_0)}, \quad \rho_2 = -\operatorname{th}(\zeta - \zeta_0)$$

где $\zeta_0 > 0$ - постоянная.

При большом значении этой постоянной начальное значение $\rho_1(0) = 1/\mathrm{ch}\zeta_0$ мало.

Мы видим, что вдоль оси $z$ в глубь кристалла происходит усиление волны основной частоты за счет интенсивности гармоники.

Последняя убывает до нуля ( при $\zeta = \zeta_0$), а затем снова возрастает, пока асимптотически вся интенсивность не окажется сосредоточенной в ней.

## 20.3.7 Phase-Matching Conditions (?)

(мб это выше уже обсуждалось, пока не знаю, потом подумаю.)

## 20.3.8 Optical Rectification

(в общефизе прописаны основы, от туда можно начать выписку)
??

## 20.3.9 Interference of Polarized Waves

Интересно, что интерференция никогда не наблюдается, если волны поляризованы во взаимно перпендикулярных плоскостях.

прямо очень важная глава.

складываем амплитуды!!!

скоро займусь этим.

я просто уже забил на это все, а ответ мб будет тут.

**в параллельных лучах**

**в сходящихся лучах**

тема очень интересная, я уже на этот вопрос отвечал, когда сдавал эффект поккельса, сюда еще добавлю разные красивые картинки, но это займет целый шаг. зато выглядеть

будет прямо отлично.

### 20.3.10 Pockels Effect

в данном случае потенциал имеет зависимости более высокого порядка, нежели второй. и о чем он?

**Суть**

**Модель**

**О применениях**

### 20.3.11 Kerr Effect

**Суть**

**Модель**

**О применениях**

### 20.3.12 Faraday Effect

пока не касался.

**Суть**

**Модель**

**О применениях**

# Part VII
# Optical Phenomena in Nature

## 21 Some Phenomena

### 21.1 Optics in Everyday Life

(discuss all effects which can be seen on daily basics)

### 21.2 Effects with Photography

(??? why this and that is seen on the picture???)

### 21.3 Light in Atmosphere

(??? explain, how it works???)

### 21.4 Other Phenomena

#### 21.4.1 Color of the sky

про это люди любят узнавать.

**Что нужно поменять, чтобы небо стало красным?**

(с формулами решу задачу)

#### 21.4.2 Rainbow

Подробно обсудим, откуда она

#### 21.4.3 Double Rainbow

Подробно обсудим, откуда она

**Описание радуги**

**Описание двойной радуги**

(мб когда-то порешаю, что есть она?)

#### 21.4.4 Optical Halo

??

# Part VIII
# Optical Equipment and Experiments

Очень часто приходится иметь дело с той или иной оптической установкой, так что опишу подробно установки, которые мне известны

## 22 Basic Laboratory Experiments

(I'll add soon from theory below!)

### 22.1 Experiments about Geometrical Theory and Modeling of Optical Systems

#### 22.1.1 The Study of Centered Optical Systems

Цель работы: изучить методы определения фокусных расстояний линз и сложных оптических систем; определить характеристики оптической системы, составленной из тонких линз; изучить недостатки реальных линз - сферическую и хроматическую аберрации.

В работе используются: оптическая скамья с набором рейтеров, положительные и отрицательные линзы, экран, осветитель с ирисовой диафрагмой, зрительная труба, светофильтры, кольцевые диафрагмы, линейка.

**I. Определение фокусного расстояния тонкой собирающей линзы и сложных оптических систем по методу Аббе**

Измерение фокусного расстояния по методу Аббе основано на определении поперечного увеличения для нескольких (не менее двух) различных положений предмета, находящегося на оптической оси исследуемой оптической системы. На рис. 1 представлена соответствующая схема эксперимента.

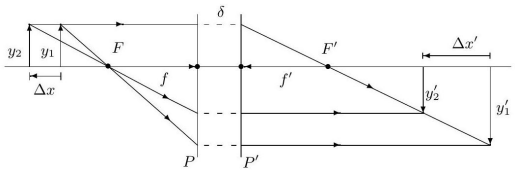



Рис. 1. Измерение фокусного расстояния оптической системы по методу Аббе

С учётом определения (1.15) и формул (1.8), (1.18) фокусное расстояние системы можно выразить через положения предмета и соответствующие увеличения следующим образом:

$$f = \frac{\Delta x}{\Delta\left(y/y'\right)} = -\frac{\Delta x'}{\Delta\left(y'/y\right)} \tag{1}$$

Здесь $\Delta x = x_2 - x_1$ - смещение предмета, $\Delta x' = x'_2 - x'_1$ - соответствующее ему перемещение изображения, $\Delta\left(y'/y\right) = y_2/y'_2 - y_1/y'_1$ - приращение поперечного увеличения, а $\Delta\left(y/y'\right)$ - приращение величины,
обратной поперечному увеличению. Для повышения точности измерений следует выбирать такие смещения $\Delta x$, чтобы увеличения заметно отличались друг от друга. С

целью уменьшения случайной ошибки, возникающей при фокусировке изображения, измерения следует проводить несколько раз, усредняя полученные данные.

Основное достоинство метода Аббе состоит в том, что фокусное расстояние сложной системы или линзы, как это видно из рисунка, может быть получено при неизвестном расстоянии $\delta$ между главными плоскостями $P$ и $P'$.

## II. Определение фокусного расстояния собирающих линз и сложных оптических систем по методу Бесселя

Схема метода Бесселя для случая, когда $n = n'$ и $f' = -f$, представлена на рис. 2. Она основана на том, что при заданном расстоянии $L$ между предметом и экраном уравнение (1.24) представляет собой квадратное уравнение относительно расстояния $s$ от главной плоскости пространства предметов до предмета ($s < 0$):

$$-\frac{1}{s} + \frac{1}{L - \delta + s} = \frac{1}{f} \tag{2}$$

имеющее при условии $L > 4f + \delta$ решения $s_1$ и $s_2$, показанные на рис. 2, где $\delta$ - расстояние между главными плоскостями системы (линзы).

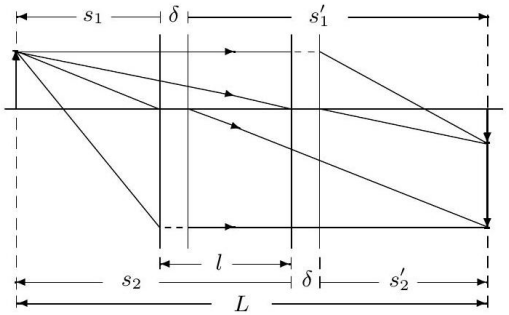



Рис. 2. Измерение фокусного расстояния оптической системы по методу Бесселя

С учётом симметрии и направлений измерения расстояний, положения предметов определяются соотношениями $s_2' = -s_1$ и $s_1' = -s_2$. Для расстояния $L$ между предметом и экраном и расстояния $\ell$ между двумя положениями системы (линзы) получаем: $L - \delta = s_1' - s_1$,

$$\ell = -s_2 + s_1 = s_1 + s_1'. \text{ Отсюда следует, что}$$

$$s_1 = -\frac{1}{2}(L - \delta - \ell), \quad s_1' = \frac{1}{2}(L - \delta + \ell) \tag{3}$$

Подставляя (3) в формулу (1.22), после несложных преобразований находим выражение

$$f = \frac{(L - \delta)^2 - \ell^2}{4(L - \delta)} \tag{4}$$

Если выполняется условие $|\delta| \ll L$, то с относительной погрешностью

$$\varepsilon_f = \frac{L^2 + \ell^2}{L^2 - \ell^2}\frac{|\delta|}{L} \tag{5}$$

формула (4) может быть представлена в более простом виде

$$f = \frac{L^2 - \ell^2}{4L} \tag{6}$$

Для определения фокусного расстояния $f$ достаточно, таким образом, измерить расстояние $L$ между предметом и экраном и расстояние $\ell$ между двумя положениями системы, при которых на экране видны чёткие изображения. Допускаемая при этом относительная погрешность результата измерений (без учёта случайных ошибок) даже для толстых линз и некоторых сложных оптических систем может быть мала.

Если выполнить измерения для двух пар величин $L$ и $\ell$, то из двух уравнений вида (4) можно найти обе характеристики системы: $f$ и $\delta$.

При выводе формул (2) - (5) по умолчанию предполагалось, что расстояние между главными плоскостями системы положительно $\delta > 0$, однако полученные результаты, очевидно, справедливы и для $\delta \leqslant 0$.

### III. Определение фокусного расстояния тонкой собирающей линзы

Если собирающую линзу можно считать тонкой, т. е. принять, что главные плоскости линзы совпадают, то её фокусное расстояние может быть измерено, например, следующими способами (в дополнение к изложенным выше).

Способ 1. Воспользуемся формулой тонкой линзы (1.37):

$$-\frac{1}{a} + \frac{1}{a'} = \frac{1}{f} \tag{7}$$

Здесь $a$ и $a'$ - расстояния предмета и его изображения от плоскости линзы, причём $a < 0, a' > 0$, если предмет находится слева, а его изображение - справа от линзы; фокусное расстояние $f > 0$.

Для измерения фокусного расстояния достаточно провести измерения по стандартной схеме «предмет - линза - экран» при произвольном расстоянии между предметом и экраном с различными значениями $a$ при увеличенном и при уменьшенном изображении. Относительная погрешность определения фокусного расстояния линзы по этому методу порядка $d/f$, где $d$ - толщина линзы. Сравнение среднеквадратичного отклонения измеренных значений $f$ от среднего $d$ позволит сделать вывод о том, нужны ли многократные измерения.

Способ 2. Фокусное расстояние тонкой собирающей линзы можно определить с помощью зрительной трубы, настроенной на бесконечность, то есть на параллельный пучок лучей. Разместив между предметом и зрительной трубой положительную линзу и перемещая её вдоль оси системы, можно найти резкое изображение предмета в окуляре зрительной трубы. При этом расстояние от середины линзы до предмета равно фокусному расстоянию тонкой линзы. Для толстой линзы зрительная труба позволяет определить только положение главного фокуса.

### IV. Определение фокусного расстояния тонкой рассеивающей линзы

Способ 1. Определение фокусного расстояния рассеивающей линзы затруднено тем, что изображение предмета получается мнимым (при действительном источнике) и поэтому не может быть получено на экране. Эту трудность легко обойти с помощью вспомогательной собирающей линзы.

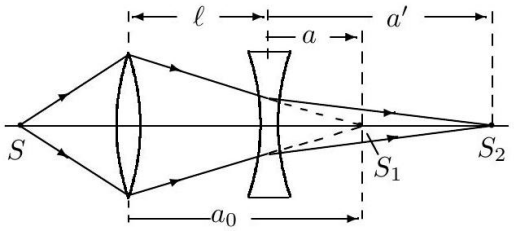



Рис. 3. Измерение фокусного расстояния рассеивающей линзы

Сначала с помощью собирающей линзы получают на экране действительное изображение предмета $S$ (точка $S_1$ на рис. 3). Затем на пути лучей, выходящих из собирающей линзы, располагают исследуемую рассеивающую линзу и, отодвигая экран, получают чёткое изображение предмета на экране, образованное двумя линзами. Точка $S_1$ пересечения сходящихся лучей играет по отношению к рассеивающей
линзе роль мнимого источника. Изображение источника переместится теперь в точку $S_2$.

Определив расстояния $a = a_0 - \ell > 0$ и $a' > 0$, рассчитывают фокусное расстояние рассеивающей линзы по формуле (7), которое, естественно, должно получиться отрицательным.

Способ 2. Если расстояние $a$ на рис. 3 совпадает с модулем фокусного расстояния рассеивающей линзы, то изображение $S_2$ перемещается в бесконечность, то есть лучи выходят из линзы параллельным пучком.

Параллельность пучка можно установить с помощью зрительной трубы, настроенной на бесконечность. Зная расстояние от первой линзы до точки $S_1$ и расстояние между линзами, нетрудно определить фокусное расстояние тонкой рассеивающей линзы. Для толстой отрицательной линзы этот метод позволяет определить только положение главного фокуса.

## V. Определение положения главных и фокальных плоскостей сложной оптической системы

Для нахождения главных плоскостей системы недостаточно знать фокусное расстояние, нужно определить ещё положения главных фокусов. Это можно сделать при помощи зрительной трубы, настроенной на бесконечность. Отложив от главных фокусов отрезки, равные фокусному расстоянию, можно найти положения главных плоскостей системы. При этом необходимо учитывать возможность различного взаимного расположения кардинальных точек (плоскостей) сложной системы.

Все характеристики оптической системы, состоящей из двух тонких линз, можно рассчитать (см. (1.38)-(1.40)), если известны фокусные расстояния $f_1, f_2$ каждой линзы и оптический интервал $\Delta$ или расстояние между центрами линз $\ell_{12}$.

Экспериментальная установка. Оптическая скамья с осветителем, транспарант, набор линз, экран и зрительная труба позволяют определить параметры оптических систем всеми описанными способами. Все оптические элементы устанавливаются на скамье при помощи рейтеров.

Важную роль играет правильная центрировка элементов системы. Проходя через плохо отцентрированную систему, лучи света могут отклониться и пройти мимо экрана или глаза наблюдателя. Центрировать линзы следует как по высоте, так и в поперечном направлении, для чего линзы установлены на поперечных салазках.

## ЗАДАНИЕ

### A. Определение фокусных расстояний тонких линз и характеристик сложной оптической системы

В этом упражнении предлагается определить фокусные расстояния тонких собирающих и рассеивающих линз, рассчитать их оптическую силу, а также определить фокусное расстояние и положения главных и фокальных плоскостей оптической системы из двух тонких линз.

**Внимание! Не следует касаться поверхности линз пальцами, потому что пылинки и жирные пятна на стекле рассеивают световые лучи и ухудшают чёткость изображения.**

### I. Центрировка элементов оптической системы

1. Из имеющегося комплекта отберите собирающие линзы. Для этого, держа линзу в одной руке, получите на ладони другой руки изображение любого удалённого объекта (окна, лампочки) и оцените на глаз фокусное расстояние линзы. Линза, которая не даёт действительного изображения, - рассеивающая. Запишите приближённые значения фокусных расстояний собирающих линз.

2. Соберите и отцентрируйте установку. Для этого на одном конце оптической скамьи установите осветитель, на котором в качестве предмета закреплён транспарант с изображением ориентированной вверх шкалы или другого объекта известного размера, например, стрелки (↑). Отрегулируйте высоту экрана так, чтобы его центр совпадал с центром яркого круглого пятна от осветителя.

3. Проведите центрировку собирающей линзы. Для этого отодвиньте экран от осветителя и разместите в промежутке рейтер с собирающей линзой №1.

Оптические оси линз устанавливаются параллельно ребру оптической скамьи на глаз. Легко убедиться на опыте, что при небольших увеличениях, получаемых в настоящей работе, такая установка является достаточной.

Передвигая линзу и экран вдоль скамьи, добейтесь чёткого изображения предмета на экране. Закрепите рейтеры. Смещая линзу с помощью поперечных салазок и по высоте, приведите центр изображения к центру экрана.

4. Остальные линзы можно отцентрировать позднее. Для центрировки рассеивающих линз следует воспользоваться уже отцентрированной положительной линзой, расположив её впереди отрицательной. Дальнейшие действия продумайте самостоятельно.

### II. Определение фокусных расстояний тонких линз при помощи экрана

Измерения проводятся для одной собирающей и одной рассеивающей линзы.

5. Для проведения измерения фокусного расстояния собирающей линзы №1 по методу Аббе (рис. 1) установите её на оптической скамье между осветителем с траспарантом и экраном. Перемещая осветитель вдоль скамьи, получите на экране резкое изображение предмета при двух различных положениях осветителя и соответственно экрана. Размеры изображений для повышения точности измерений должны заметно отличаться друг от друга. Зарегистрируйте перемещения предмета $\Delta x$ и изображения $\Delta x'$ и соответствующие двум положениям предмета

размеры изображений $y_1'$ и $y_2'$. Рассчитайте фокусное расстояние линзы по формулам (1), оцените точность результата.

6. Для измерения фокусного расстояния линзы №1 по методу Бесселя или по формуле тонкой линзы (по указанию преподавателя) расположите экран за линзой на расстоянии $L > 4f$ от предмета (рис. 2). Перемещая линзу вдоль скамьи, получите увеличенное и уменьшенное изображения предмета на экране.

Измерьте расстояния от средней плоскости линзы до предмета $a$ и до изображения $a'$ для обоих положений линзы. Средняя плоскость линзы и положение предмета отмечены на оправах проточками. Измерьте также расстояние $L$ от предмета до экрана и перемещение $\ell$ линзы (рейтера).

По результатам нескольких измерений с различными расстояниями $L$ определите среднее значение фокусного расстояния, используя формулу (6) или (7). Оцените среднеквадратичную случайную погрешность $\sigma_{\text{отд}}$ однократного измерения $f$ и сравните её с толщиной линзы. Сделайте вывод о том, нужны ли многократные измерения в этом эксперименте.

7. Для определения фокусного расстояния тонкой рассеивающей линзы используйте вспомогательную собирающую линзу. Сначала с помощью короткофокусной собирающей линзы №3 получите на экране увеличенное изображение предмета и измерьте расстояние от линзы до экрана ($a_0$ на рис. 3). Затем между собирающей линзой и экраном разместите рассеивающую линзу и, отодвигая экран от линзы, найдите действительное изображение предмета, образованное системой линз. Измерьте расстояние $a'$ от рассеивающей линзы до экрана и расстояние между линзами $\ell$.

Рассчитайте величину $a$ и определите фокусное расстояние рассеивающей линзы с помощью формулы (7). При вычислении фокусного расстояния нужно приписать величинам $a$ и $a'$ правильные знаки.

## III. Определение фокусных расстояний тонких линз с помощью зрительной трубы

Измерения проводятся для двух собирающих и одной рассеивающей линзы.

8. Для определения фокусных расстояний линз с помощью зрительной трубы необходимо настроить трубу на бесконечность. Эту настройку проще всего осуществить, наведя трубу на удалённый объект (например, на окно в конце длинного коридора). Предварительно вращением глазной линзы окуляра трубы настройтесь на резкое видение окулярной шкалы (если окулярная шкала отсутствует - совместите штрих на глазной линзе с точкой на тубусе). Не следует настраивать трубу на предмет, расположенный за оконным стеклом, так как оконное стекло часто оказывается недостаточно плоским.

9. Поставьте собирающую линзу №1 на расстоянии от предмета, примерно равном фокусному расстоянию. На небольшом расстоянии от линзы закрепите трубу, настроенную на бесконечность, и отцентрируйте её по высоте. Передвигая линзу вдоль скамьи и, если необходимо, перемещая линзу с помощью поперечных салазок, получите в окуляре зрительной трубы чёткое изображение предмета. При этом расстояние между предметом и серединой тонкой линзы (между проточками на оправах) равно фокусному расстоянию.

Поверните линзу другой стороной к источнику и повторите измерение её фокусного расстояния. По результатам измерений сделайте вывод, можно ли считать линзу тонкой.

10. Измерьте фокусное расстояние собирающей линзы №2 при помощи зрительной трубы.

11. Для определения фокусного расстояния тонкой рассеивающей линзы используйте схему, изображённую на рис. 3. Сначала получите на экране увеличенное изображение предмета при помощи короткофокусной положительной линзы. Измерьте расстояние $a_0$ между линзой и экраном.

Разместите сразу за экраном трубу, настроенную на бесконечность, и закрепите её. Уберите экран и поставьте на его место исследуемую рассеивающую линзу. Отцентрируйте световой пучок с помощью листа бумаги. Перемещая рассеивающую линзу, найдите в окуляре зрительной трубы резкое изображение предмета.

Измерив расстояние $\ell$ между линзами, рассчитайте фокусное расстояние рассеивающей линзы в пространстве предметов $f = \ell - a_0$.

Поверните рассеивающую линзу другой стороной к источнику и повторите измерения. Оцените, можно ли считать линзу тонкой.

12. Составьте таблицу результатов экспериментального определения фокусных расстояний линз, включив в неё дополнительно колонку значений оптической силы каждой линзы.

**IV. Определение фокусного расстояния и положения главных и фокальных плоскостей сложной оптической системы**

13. Соберите модель оптической системы, установив в центре оптической скамьи две тонкие собирающие линзы. Для определения фокусного расстояния системы по методу Аббе (рис. 1) расположите экран на дальнем конце скамьи. При проведении эксперимента следует учитывать, что согласно формулам (1.8), (1.14), (1.15) поперечное увеличение системы

$$\frac{y'}{y} = \frac{f_{1\Sigma}}{x - F_{1\Sigma}}$$

может принимать разный знак в зависимости от знака фокусного расстояния и взаимного расположения предмета и переднего фокуса системы. Соответственно изображение может быть как прямым, так и обратным.

Перемещая осветитель с транспарантом вдоль скамьи, получите на экране резкое изображение предмета. Измерьте расстояние от предмета до первой линзы и величину изображения $y'_1$.

Отодвиньте осветитель на некоторое расстояние от прежнего положения и, передвигая экран, вновь получите резкое изображение предмета. Для повышения точности измерений размеры изображений $y'_1$ и $y'_2$ должны заметно отличаться друг от друга. Зарегистрируйте перемещения предмета $\Delta x$ и изображения $\Delta x'$ и соответствующие двум положениям предмета размеры изображений $y'_1$ и $y'_2$. Рассчитайте фокусное расстояние системы по формулам (1).

14. Для нахождения положения главных фокусов системы закрепите зрительную трубу за второй линзой, подвиньте осветитель к первой линзе и отцентрируйте систему с помощью листа бумаги.

Медленно отодвигая осветитель от системы, найдите резкое изображение предмета в окуляре зрительной трубы. Если необходима более резкая фокусировка, задиафрагмируйте первую линзу и подберите подходящую яркость пучка. Определите положение главного фокуса системы $F_{1\Sigma}$ в пространстве предметов относительно первой линзы, измерив расстояние от этой линзы до предмета.

15. Поменяйте линзы местами и повторите измерения п. 3, сохранив неизменным расстояние между линзами. Определите расстояние от предмета до ближней линзы и соответственно найдите положение главного
фокуса системы $F_{2\Sigma}$ в пространстве изображений относительно второй линзы в исходном состоянии системы.

16. На миллиметровой бумаге постройте в масштабе чертёж оптической системы. Укажите на нём положения фокусов каждой из линз, а также главных фокусов $F_{1\Sigma}$ и $F_{2\Sigma}$ системы.

Постройте ход луча, вошедшего в систему слева параллельно главной оптической оси, последовательно через каждую из линз. Точка пересечения луча, вышедшего из системы (или его продолжения), с оптической осью определяет положение главного фокуса системы $F_{2\Sigma}$, а точка его пересечения с продолжением входящего луча определяет положение главной плоскости системы $P_2$ в пространстве изображений. Расстояние точки $H_{2\Sigma}$ пересечения задней главной плоскости с оптической осью от заднего главного фокуса $F_{2\Sigma}$ равно фокусному расстоянию системы $f_{2\Sigma}$ в пространстве изображений (рис. 1.2, рис. 1).

Аналогичное построение сделайте для луча, вошедшего в систему справа, из которого найдите положения главной плоскости $P_1$, главной точки $H_{1\Sigma}$, а также фокусное расстояние $f_{1\Sigma}$ системы в пространстве предметов.

17. По экспериментальным данным и построенным чертежам определите основные характеристики исследованной оптической системы: координаты главных и фокальных точек в пространствах предметов и изображений и фокусное расстояние системы.

18. Рассчитайте те же основные характеристики системы по формулам (1.38). Сведите экспериментальные, расчётно-графические и теоретические результаты в таблицу, сравните их и предложите свои версии причин замеченных расхождений.

### Б. Основные аберрации оптических систем*

В этом упражнении предлагается исследовать аберрации плосковыпуклой линзы: набор кольцевых диафрагм позволяет исследовать продольную сферическую аберрацию (зависимость фокусного расстояния от радиуса кольцевого пучка, падающего на линзу), а набор светофильтров - хроматическую аберрацию (зависимость фокусного расстояния от длины волны). Расчётные формулы и схемы экспериментов приведены в п. 5 введения к разделу.

**Внимание! Не следует касаться поверхности линз пальцами, потому что пылинки и жирные пятна на стекле рассеивают световые лучи и ухудшают чёткость изображения.**

---

0

- Выполняется по указанию преподавателя.

**I. Сферическая аберрация**

1. Для качественного наблюдения сферической аберрации расположите осветитель и экран на дальних концах скамьи. Установите плосковыпуклую линзу №3, которая может перемещаться относительно рейтера в продольном направлении, на расстоянии $a_1$ от предмета, чуть большем фокусного расстояния. Линза должна быть обращена измерительной шкалой к наблюдателю. Наденьте на линзу диафрагму диаметром $2h = 1$ см. Перемещая линзу, получите на удалённом экране резкое изображение ирисовой диафрагмы осветителя.

Установите диафрагму максимального диаметра ($2h = 4$ см) и убедитесь, что при неизменном расстоянии $a_1$ расстояние $a_2$ от линзы до изображения заметно изменилось.

2. Для измерения аберраций поставьте вместо экрана за линзой зрительную трубу. Установите линзу строго в нулевое положение с учётом нониуса. Используя зрительную трубу, получите параллельный пучок параксиальных лучей от линзы с диафрагмой минимального диаметра и запишите отсчёт по нониусной шкале. Увеличивая диаметр диафрагмы и подстраиваясь к новому положению фокуса при помощи нониусного винта, регистрируйте соответствующие отсчёты по нониусной шкале.

3. По результатам измерений постройте график функции $-s\,(h^2)$, который согласно (1.50) должен представлять собой прямую линию. Экстраполируя эту прямую к значениям $h = r$ и $h = 0$, рассчитайте продольную сферическую аберрацию линзы $\delta s(r) = s(r) - s(0)$, а по наклону прямой найдите показатель преломления стекла линзы $n$.

**II. Хроматическая аберрация**

4. Используя зрительную трубу и три светофильтра (красный, жёлтый и синий), найдите фокусные расстояния $f_F$ и $f_C$, определив по нониусной шкале положения плосковыпуклой линзы №3, соответствующие резкому изображению диафрагмы. Фильтры следует располагать в параллельном пучке (за окуляром зрительной трубы), чтобы они не изменяли хода лучей. Фокусное расстояние $f_D$ линзы № 3, измеренное с жёлтым светофильтром, указано на оправе линзы.

5. Рассчитайте хроматическую аберрацию по формуле (1.54) и число Аббе по формуле (1.56). По табличным данным определите сорт стекла, из которого сделана линза №3.

**Контрольные вопросы**

1. Покажите, что если расстояние между предметом и экраном превышает $4f$, то изображение на экране может быть получено при двух различных положениях линзы.

2. Опишите методику измерения фокусных расстояний отрицательных линз.

3. Дайте определения главных фокусов, фокусных расстояний и главных плоскостей сложной центрированной оптической системы.

4. Определите графически положение главных плоскостей системы, состоящей из двух тонких линз - собирающей и рассеивающей, - с известными фокусными расстояниями и расстоянием между ними.

5. Дайте определения сферической и хроматической аберраций.

6. Объясните, почему сферическая аберрация зависит от того, какой стороной плосковыпуклая линза обращена к источнику.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§9 – 16.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 71-79, 81, 82, 86.

3. Дитчберн Р. Физическая оптика. - М.: Наука, 1965. Гл. 7.

4. Борн М., Вольф Э. Основы оптики. - М.: Наука, 1973. Гл. 3-5.

## 22.1.2 Modeling of Optical Devices and Determining Their Magnification

Цель работы: изучить модели зрительных труб (астрономической трубы Кеплера и земной трубы Галилея) и микроскопа, определить их увеличения.

В работе используются: оптическая скамья, набор линз, экран, осветитель со шкалой, зрительная труба, диафрагма, линейка.

В работе предлагается измерить фокусные расстояния линз, смоделировать трубу Кеплера, трубу Галилея, микроскоп (см. пп. 4.2, 4.3 введения к разделу) и определить их увеличения.

Экспериментальная установка. Набор линз, осветитель, экран, зрительная труба, необходимые для моделирования оптических приборов, устанавливаются при помощи рейтеров на оптической скамье. Предметом служит миллиметровая сетка, нанесённая на матовое стекло осветителя.

Центрирование линз. При юстировке любых оптических приборов важно правильно центрировать входящие в систему линзы. Проходя через плохо отцентрированную систему линз, лучи света отклоняются
в сторону и могут вообще не доходить до глаза наблюдателя. Центрировать линзы следует как по высоте, так и в поперечном направлении (для чего линзы крепятся на поперечных салазках).

Юстировка коллиматора. При составлении моделей телескопических систем необходимо иметь удалённый объект. В качестве такого объекта обычно используется бесконечно удалённое изображение предмета (шкалы осветителя), установленного в фокальной плоскости положительной линзы. Лучи, выходящие из одной точки предмета, пройдя через линзу, образуют параллельный пучок. Устройство такого рода называется коллиматором. Для юстировки коллиматора удобно использовать вспомогательную зрительную трубу, предварительно настроенную на бесконечность. Передвигая линзу коллиматора вдоль скамьи, добиваются появления резкого изображения предмета в окуляре зрительной трубы.

Измерение фокусных расстояний линз. Для того чтобы моделировать оптические инструменты, нужно знать фокусные расстояния линз, которые могут быть использованы в качестве объектива или окуляра модели. Фокусные расстояния положительных линз проще всего найти с помощью вспомогательной зрительной трубы, установленной на бесконечность (другие методы измерения фокусных расстояний можно найти в описании работы 4.1.1). Работа выполняется так же, как при юстировке коллиматора.

При определении фокусного расстояния отрицательной линзы предметом служит изображение сетки, которое даёт вспомогательная положительная линза.

## ЗАДАНИЕ

**Внимание! Не следует касаться поверхности линз пальцами, потому что царапины, пылинки или жирные пятна на стекле рассеивают световые лучи и ухудшают чёткость изображения.**

### I. Центрировка элементов оптической системы

1. Из имеющегося набора отберите собирающие линзы; для этого, держа линзу в одной руке, получите на ладони другой изображение любого удалённого объекта (окна, лампочки) и оцените на глаз фокусное расстояние. Линза, которая не даёт действительного изображения, рассеивающая. Запишите номера линз и приближённые значения фокусных расстояний.

2. Соберите и отцентрируйте установку. Для этого на одном конце оптической скамьи установите предмет - осветитель с сеткой - и вплотную к нему - экран на рейтере. При юстировке системы следует соблюдать следующие простые правила: все винты рейтеров на скамье и винты поперечных салазок должны находиться с одной стороны оптической скамьи (со стороны экспериментатора). После всякого перемещения рейтеров вдоль скамьи они должны надёжно закрепляться винтами.

Отрегулируйте высоту экрана так, чтобы его центр совпадал с центром яркого круглого пятна от осветителя. Эта операция должна выполняться при плотно затянутом винте, закрепляющем рейтер на оптической скамье.

Для перемещения рейтера вдоль скамьи винт следует только слегка ослабить, чтобы не допустить перпендикулярного скамье смещения центра. Перемещая рейтер, прижимайте его к скамье со стороны, противоположной винту.

3. Отодвиньте экран от осветителя и разместите в промежутке рейтер с одной из собирающих линз.

Передвигая линзу и экран вдоль скамьи, добейтесь чёткого изображения края диафрагмы осветителя или миллиметровой шкалы на экране. Закрепите рейтеры. Перемещая линзу поперёк оптической оси, приведите центр изображения к центру экрана.

Оптические оси линз устанавливаются параллельно ребру оптической скамьи на глаз.

4. Таким же образом центрируются остальные собирающие линзы. Для центрировки рассеивающих линз воспользуйтесь уже отцентрированной положительной линзой, расположив её впереди отрицательной. Способ центрировки рассеивающих линз продумайте самостоятельно.

### II. Определение фокусных расстояний тонких линз с помощью зрительной трубы

5. Для определения фокусных расстояний линз с помощью зрительной трубы (рис. 1) необходимо настроить трубу на бесконечность. Эту настройку проще всего осуществить наведя трубу на удалённый объект (например, на окно в конце длинного коридора). Предварительно вращением глазной линзы окуляра трубы настройтесь на резкое видение окулярной шкалы. Не следует настраивать трубу на предметы, расположенные за оконным стеклом, так как оконное стекло часто оказывается недостаточно плоским.

6. Поставьте собирающую линзу на расстоянии от предмета, примерно равном фокусному. На небольшом расстоянии от линзы закрепите трубу, настроенную на бесконечность (рис. 1), и отцентрируйте её по высо те. Диафрагма диаметром $d = 1$ см, надетая на ближнюю к осветителю линзу, уменьшит сферические аберрации и повысит чёткость изображения.

Передвигая линзу вдоль скамьи, получите в окуляре зрительной трубы изображение предмета - миллиметровой сетки. При этом расстояние между предметом и серединой тонкой линзы (между проточками на оправах) равно её фокусному расстоянию.

Поверните линзу другой стороной к источнику и повторите

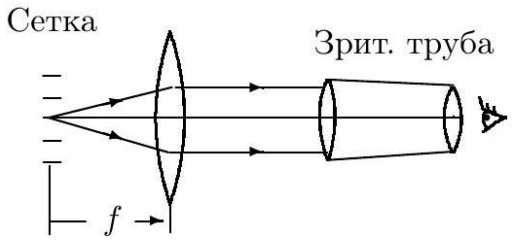



Рис. 1. Определение фокусного расстояния собирающей линзы измерения фокусного расстояния. По результатам измерений сделайте вывод, можно ли считать линзу тонкой.

7. Измерьте фокусные расстояния остальных положительных линз при помощи зрительной трубы.

8. Для определения фокусного расстояния тонкой рассеивающей линзы сначала получите на экране увеличенное изображение сетки при помощи одной короткофокусной собирающей линзы. Измерьте расстояние $a_0$ между линзой и экраном (удобно взять $a_0 \approx 30$ cm).

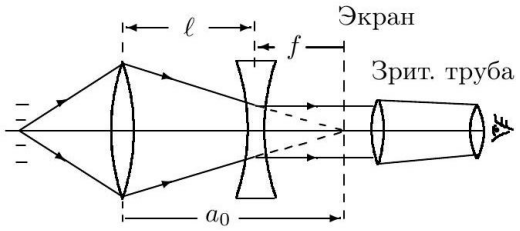



Рис. 2. Определение фокусного расстояния рассеивающей линзы

9. Разместите сразу за экраном трубу, настроенную на бесконечность, и закрепите её. Уберите экран и поставьте на его место исследуемую рассеивающую линзу (рис. 2). Отцентрируйте световой пучок с помощью листа бумаги. Перемещая рассеивающую линзу, найдите в окуляре зрительной трубы резкое изображение сетки. Если изображение недостаточно чёткое, задиафрагмируйте ближнюю к осветителю линзу и заново проведите настройку с экраном. Подберите оптимальную яркость источника.

Измерив расстояние между линзами $\ell$, рассчитайте фокусное расстояние рассеивающей линзы $f = \ell - a_0$.

10. Поверните рассеивающую линзу другой стороной к источнику и повторите измерения.

**III. Телескоп Кеплера**

11. Из имеющегося набора отберите две собирающие линзы для создания модели зрительной трубы Кеплера с увеличением 2-3 (рис. 3). В качестве коллиматора используйте линзу с фокусным расстоянием $f \approx 15 \div 20$ см (при этих условиях размер изображения предмета не будет превышать размера поля зрения вспомогательной зрительной трубы).

12. Настройте коллиматор при помощи вспомогательной зрительной трубы так же, как вы делали это при измерении фокусных расстояний. Чёткость изображения повысится, если надеть на коллиматор диафрагму диаметром 1 см и уменьшить яркость осветителя.

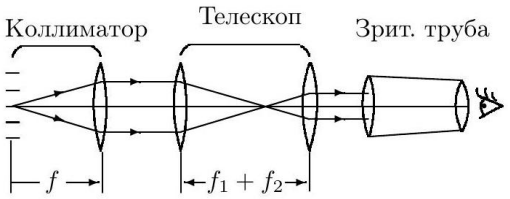



Рис. 3. Определение увеличения телескопа Кеплера

Для последующих расчётов увеличения телескопа определите размер изображения $h_1$ одного миллиметра шкалы осветителя в делениях окулярной шкалы зрительной трубы. Очевидно, $h_1 = k \operatorname{tg} \alpha_1 \approx k\alpha_1$, где $k$ - некоторый коэффициент, характеризующий увеличение зрительной трубы, $\alpha_1$ - угловой размер изображения миллиметрового деления шкалы осветителя, наблюдаемого через коллиматор.

13. Соберите модель телескопа: линзу с максимальным фокусным расстоянием - объектив телескопа - расположите почти вплотную к линзе коллиматора, окуляр - на расстоянии, примерно равном сумме фокусных расстояний обеих линз трубы.

14. Закрепите вспомогательную зрительную трубу за окуляром телескопа и отцентрируйте световое пятно при помощи листа бумаги. Слегка перемещая окуляр телескопа вдоль оптической скамьи, получите изображение сетки в объективе вспомогательной трубы.

Измерьте расстояние между объективом и окуляром телескопа и сравните его с суммой фокусных расстояний.

15. Рассчитайте увеличение исследуемой модели телескопа через отношение передних фокусных расстояний линз $f_1, f_2$ согласно формуле (1.48):

$$N_T = -\frac{f_1}{f_2} \tag{1}$$

16. Определите увеличение телескопа через отношение углов, под которыми объект виден через телескоп и без него (рис. 1.8а и формулы (1.40) введения к разделу). Для этого найдите размер $h_2$ изображения миллиметрового деления шкалы осветителя в делениях окулярной шкалы вспомогательной трубы при наблюдении по схеме рис. 3. Очевидно, что $h_2 = k \operatorname{tg} \alpha_2 \approx k\alpha_2$, где $\alpha_2$ - угловой размер изображения миллиметрового деления шкалы при наблюдении через телескоп. Сравнив $h_2$ с величиной $h_1$, полученной в п. 12 без телескопа, рассчитайте искомое увеличение по формуле

$$N_T = \frac{\alpha_2}{\alpha_1} = -\frac{h_2}{h_1} \tag{2}$$

где учтено различие знаков углов лучей на входе и выходе из трубы Кеплера (рис. 1.8а).

17. Определите увеличение телескопа, сравнив диаметр оправы его объектива и диаметр изображения этой оправы в окуляре (см. рис. 1.86 и формулу (1.49)). Для этого отодвиньте вспомогательную трубу и расположите экран за окуляром телескопа. Снимите диафрагму с коллиматора и убедитесь, что световое пятно полностью освещает объектив телескопа и проходит через окуляр. Отодвигая экран от окуляра, получите на нём чёткое изображение оправы объектива. Поднеся к объективу какой-нибудь предмет (например, край линейки), убедитесь, что наблюдается именно изображение оправы объектива. Измерьте диаметр объектива $D_1$ и диаметр его изображения $D_2$. Определите увеличение телескопа по формуле (1.49).

18. Сравните результаты измерения увеличения телескопа Кеплера тремя рассмотренными выше способами и объясните причины возможных расхождений.

**IV. Труба Галилея**

19. Соберите модель трубы Галилея, поставив в модели трубы Кеплера вместо собирающей окулярной линзы рассеивающую линзу на расстоянии от объектива, равном разности модулей фокусных расстояний объектива и окуляра. При таком расположении линз труба Галилея представляет собой телескопическую систему, которая может быть исследована на уже освоенном вами стенде.

20. Выполните настройки системы и проведите измерения, аналогичные тем, которые были проделаны для телескопа Кеплера по пп. 14-16.

21. Сравните результаты измерений увеличения трубы Галилея в телескопическом режиме, полученные разными способами, между собой и с результатом расчёта по формуле (1).

*V*. **Модель микроскопа**

22. Для создания модели микроскопа с увеличением $N_M = 5$ отберите самые короткофокусные собирающие линзы из набора. Рассчитайте необходимые оптический интервал $\Delta$ и длину тубуса $\ell_{12}$ по формулам (см. 1.47) :

$$N_M = N_1 N_2, \quad N_1 = -\frac{\Delta}{f_1}, \quad N_2 = \frac{L}{f_2}, \quad \Delta = \ell_{12} - f_1 - f_2 \tag{3}$$

где $N_1, N_2$ - увеличения объектива и окуляра, $f_1, f_2$ - положительные передние фокусные расстояния линз, $L = 25$ см - расстояние наилучшего зрения.

23. Расположите объектив и окуляр на соответствующем расстоянии $\ell_{12}$ друг от друга (рис. 4) и закрепите рейтеры. Сфокусируйте модель микроскопа на сетку осветителя. Для этого перемещайте осветитель вдоль оптической скамьи до тех пор, пока в окуляре микроскопа не появится отчётливое увеличенное изображение сетки.

24. Расположите за окуляром моде-

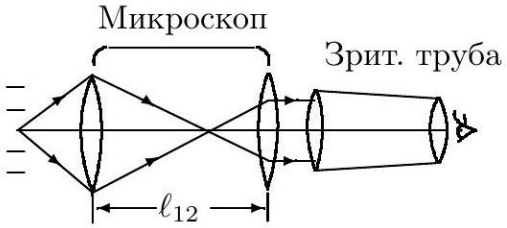



Рис. 4. Модель микроскопа ли микроскопа зрительную трубу, настроенную на бесконечность. Слегка перемещая осветитель, получите в поле зрения трубы изображение миллиметровой шкалы осветителя. Чёткость изображения повысится, если надеть на объектив микроскопа диафрагму диаметром 1 см и уменьшить яркость осветителя.

25. Для экспериментального определения увеличения микроскопа измерьте величину изображения $h_2$ миллиметрового деления предметной шкалы в делениях окулярной шкалы зрительной трубы. Используя результат аналогичных измерений ($h_1$ в п. 12) с коллиматорной линзой, фокус $f$ которой известен, рассчитайте увеличение микроскопа по формуле

$$N_M = -\frac{h_2}{h_1}\frac{L}{f} \tag{4}$$

Сравните результат с теоретическим расчётом по формулам (3).

26. Оцените погрешности в определении увеличения всех моделей оптических приборов.

**Контрольные вопросы**

1. С какой целью при исследовании моделей телескопов и микроскопа используется дополнительная зрительная труба?
2. Чему равны поперечное, угловое и продольное увеличения телескопа, если передние фокусные расстояния его объектива и окуляра равны соответственно $f_1 = 40$ см и $f_2 = -2$ см?
3. При каком условии угловое увеличение микроскопа равно поперечному?
4. Выведите формулу (4) и объясните соответствующий метод измерения увеличения микроскопа.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §24.
2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 91, 92.
3* Дитчберн Р. Физическая оптика. - М.: Наука, 1965. Гл. 7.
4. Борн М., Волъф Э. Основы оптики. - М.: Наука, 1973. Гл. 3-5.

## 22.1.3 Abbe Refractometer

Цель работы: измерить показатели преломления твёрдых и жидких тел в монохроматическом свете.

В работе используются: технический рефрактометр Аббе; осветитель; набор стеклянных образцов; жидкости с неизвестными показателями преломления (глицерин, этиловый спирт); монохлорнафталин; дистиллированная вода.

Рефрактометрия. Показатели преломления жидких и твёрдых тел могут измеряться с большой точностью. При данной температуре и для данной длины волны они являются важнейшими постоянными, характеризующими вещество. Измерение показателей преломления может быть использовано для исследования веществ - соответствующий раздел науки носит название рефрактометрии. Обычно измерения проводятся для $D$-линии натрия (дублет со средней длиной волны 589,3 нм).

В основе рефрактометрического метода исследования лежит формула Лоренц-Лорентца, связывающая показатель преломления $n$ изотропного вещества с числом молекул $N$ в единице объёма и поляризуемостью [1]$\alpha$ молекул вещества:

$$\frac{n^2-1}{n^2+2} = \frac{4\pi}{3} N\alpha \tag{1}$$

Величину

$$r = \frac{1}{\rho}\frac{n^2-1}{n^2+2} \tag{2}$$

где $\rho$ - плотность вещества, называют удельной рефракцией. Согласно формуле Лоренц-Лорентца (1), удельная рефракция чистого химического вещества зависит только от молекулярных характеристик (массы молекулы $m_0$ и поляризуемости $\alpha$) и равна

$$r = \frac{4\pi}{3}\frac{\alpha}{m_0} = \text{ const.}$$

Иными словами, она должна быть константой (для света с заданной длиной волны), не зависящей от состояния, в котором вещество находится. [2]

Предполагая, что оптическое поведение молекул каждого компонента смеси практически не зависит от присутствия других компонентов, можно записать удельную рефракцию смеси веществ через удельные рефракции компонетов $r_1, r_2, \ldots$ и их массовые доли (отношения масс компонентов к массе смеси) $c_1, c_2, \ldots$:

$$r = c_1 r_1 + c_2 r_2 + \ldots \tag{3}$$

Опыт показывает, что данное соотношение обычно хорошо выполняется на практике.

Также, согласно опыту, во многих случаях аналогичное эмпирическое правило (аддитивность рефракиий) может быть применено и для расчёта рефракции сложного химического соединения, если известны рефракции составляющих его элементов.

Для каждого элемента удобно ввести понятие атомной рефракции $R$ как произведение удельной рефракции $r$ данного элемента на его атомную массу $A$:

[1] Напомним, что поляризуемостью частицы $\alpha$ называют коэффициент пропорциональности между электрическим полем $E$, в которое она помещена, и возникающим при этом дипольным моментом частицы $p = \alpha E$.

2 Хотя формула Лоренц-Лорентца имеет ограниченную область применимости (см., например, [1, 3])), указанное свойство удельной рефракции часто хорошо выполняется на практике. Например, для воздуха при изменении давления на несколько сотен атмосфер удельная рефракция изменяется не более чем на $0,1\%$. Нередко удельная рефракция остаётся практически постоянной даже при изменении агрегатного состояния вещества (например, при замерзании воды).

$$R = Ar \tag{4}$$

и аналогично для химического соединения - понятие молекулярной рефракции $R_{\text{M}}$:

$$R_M = Mr = \frac{M}{\rho}\frac{n^2-1}{n^2+2} = \frac{4\pi}{3}N_A\alpha \tag{5}$$

где $M$ - молекулярная масса соединения, $N_A$ - постоянная Авогадро. Тогда, если принять, что для молекулярной рефракции выполняется свойство аддитивности, можно записать её как сумму рефракций атомов, составляющих молекулу:

$$R_M = q_1A_1r_1 + q_2A_2r_2 + \ldots = q_1R_1 + q_1R_2 + \ldots \tag{6}$$

где $q_1, q_2, \ldots$ - числа атомов элементов, входящих в состав молекулы. Аддитивность молекулярной рефракции означает, что взаимодействие отдельных атомов с полем световой волны не зависит от присутствия других атомов, входящих в состав той же молекулы. Если взаимное влияние атомов не мало, аддитивность нарушается.

В нашей работе исследуются рефракции соединений из углерода, водорода и кислорода: воды ($H_2O$), глицерина ($C_3H_8O_3$) и этилового спирта ($C_2H_6O$). Если измерить показатели преломления этих соединений, предполагая аддитивность молекулярной рефракции, можно вычислить показатель преломления любого другого соединения из углерода, водорода и кислорода (например, метилового спирта). Для этого нужно сначала вычислить по формуле (5) молекулярные рефракции воды, глицерина и этилового спирта (по измеренным значениям показателей преломления), а затем рассчитать атомные рефракции $R_{\text{C}}, R_{\text{H}}$ и $R_{\text{O}}$, используя соотношения

$$\begin{aligned} R_{\text{H}_2\text{O}} &= 2R_{\text{H}} + R_{\text{O}} \\ R_{\text{C}_3\text{H}_8\text{O}_3} &= 3R_{\text{C}} + 8R_{\text{H}} + 3R_{\text{O}} \\ R_{\text{C}_2\text{H}_6\text{O}} &= 2R_{\text{C}} + 6R_{\text{H}} + R_{\text{O}} \end{aligned} \tag{7}$$

По найденным значениям атомных рефракций $R_{\text{C}}, R_{\text{H}}$ и $R_{\text{O}}$ можно найти молекулярную рефракцию любого вещества, состоящего из этих атомов, и (зная плотность) определить ожидаемую величину его показателя преломления.

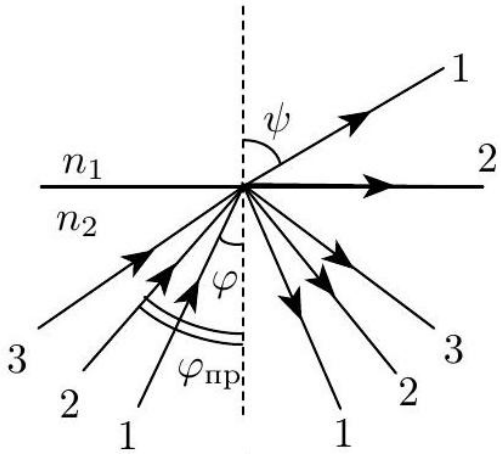
Принцип работы рефрактометра Аббе. Технический рефрактометр Аббе служит для быстрого (и сравнительно грубого) измерения показателей преломления жидких и твёрдых тел. Устройство рефрактометра Аббе основано на явлении полного внутреннего отражения.

Пусть луч света падает на границу раздела двух сред (рис. 1а) со стороны оптически более плотной среды ($n = n_2$). Для углов падения $\varphi$, меньших предельного $\varphi_{\text{пр}}$, свет частично проникает в оптически менее плотную среду ($n = n_1$), и частично отражается (луч 1 на рис. 1а). При $\varphi_{\text{пр}} < \varphi < 90^\circ$ преломлённый луч отсутствует, и наступает полное отражение (луч 3 на рис. 1а).

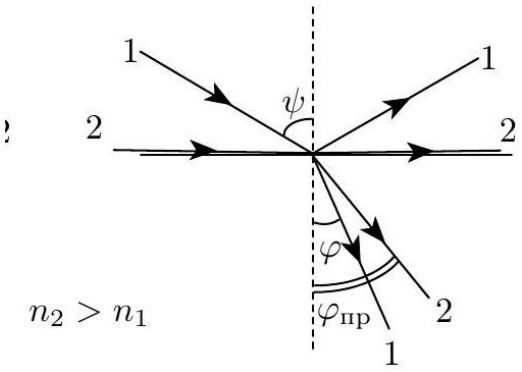
Предельный угол $\varphi_{\text{пр}}$ соответствует углу преломления $\psi = 90^\circ$ (луч 2 на рис. 1а); следовательно,

$$\sin\varphi_{\text{Ip}} = \frac{n_1}{n_2} \tag{8}$$

Зная показатель преломления одной из сред и определяя на опыте предельный угол, можно с помощью (8) найти показатель преломления второй среды.




а)




б)

Рис. 1. Предельный угол полного внутреннего отражения (а) и предельный угол преломления (б)

Пусть свет падает на границу раздела со стороны оптически менее плотной среды (рис. 1б). В зависимости от угла падения луч во второй среде может составлять с нормалью углы, расположенные в интервале от нуля до $\varphi_{\text{пр}}$ (луч 1 на рис. 1б). Предельный угол преломления $\varphi_{\text{пр}}$ соответствует углу падения $\psi = 90°$ (скользящий луч 2 на рис. 1б). Легко видеть, что величина предельного угла и в этом случае определяется формулой (8).

При измерениях показателя преломления с помощью рефрактометра Аббе можно пользоваться как методом полного внутреннего отражения, так и методом скользящего луча.

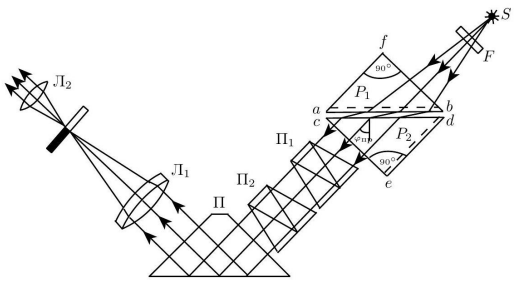



Рис. 2. Ход лучей в рефрактометре при измерении показателя преломления жидкости методом скользящего луча

Оптическая схема рефрактометра Аббе и ход лучей при измерении показателя преломления жидкости по методу скользящего луча показаны на рис. 2. Основной частью рефрактометра являются две стеклянные прямоугольные призмы $P_1$ и $P_2$,

изготовленные из стекла с большим показателем преломления. В разрезе призмы имеют вид прямоугольных треугольников, обращённых друг к другу гипотенузами; зазор между призмами составляет около 0,1 мм и служит для помещения исследуемой жидкости. Свет проникает в призму $P_1$ через грань $bf$ и попадает в жидкость через матовую грань $ab$. Свет, рассеянный матовой поверхностью, проходит слой жидкости и под всевозможными углами падает на грань $cd$ призмы $P_2$.

Скользящему лучу в жидкости ($\psi = 90°$) соответствует предельный угол преломления $\varphi_{\text{пр}}$ . Преломлённые лучи с углами больше $\varphi_{\text{пр}}$ не возникают. В связи с этим угол выхода лучей из грани ce может изменяться лишь в некотором интервале.

Если свет, выходящий из грани $ce$, пропустить через собирающую линзу $_1$, то в её фокальной плоскости наблюдается резкая граница света и темноты. Граница рассматривается с помощью линзы $_2$. Линзы $_1$ и $\Omega_2$ образуют зрительную трубу, установленную на бесконечность. В их общей фокальной плоскости находится изображение шкалы вели чин показателя преломления и указатели (нить и перекрестие). В поле зрения окуляра $\Omega_2$ трубы одновременно можно увидеть только часть изображения шкалы и часть поля сфокусированных лучей, выходящих из призмы $P_2$. Вращая систему призм $P_1$ и $P_2$ и, следовательно, изменяя наклон предельного пучка лучей относительно оси зрительной трубы, можно добиться, чтобы граница света и тени оказалась в поле зрения окуляра $\Omega_2$ и совпала с положением указателя. При вращении системы призм поворачивается и шкала показателей преломления, установленная на пластине, жёстко связанной с системой призм $P_1$ и $P_2$ (на рис. 2 шкала не показана). Значение показателя преломления жидкости отсчитывается по шкале на уровне резкой границы света и тени.

Если источник света $S$ не является монохроматическим, то наблюдаемая в окуляре трубы граница света и темноты часто оказывается размытой и окрашенной из-за дисперсии показателя преломления исследуемого вещества (т. е. из-за зависимости $n$ от длины волны $\lambda$). Для того чтобы получить и в этом случае резкое изображение границы, на пути лучей, выходящих из призмы $P_2$, помещают компенсатор с переменной дисперсией. Компенсатор содержит две одинаковые дисперсионные призмы Амичи (призмы $\Pi_1$ и $\Pi_2$ на рис. 2), каждая из которых состоит из трёх склеенных призм, обладающих различными показателями преломления и различной дисперсией. Призмы рассчитываются так, чтобы монохроматический луч с длиной волны $\lambda_D = 589,3\text{Н}$ (среднее значение длины волны жёлтого дублета натрия) не испытывал отклонения. Лучи с другими длинами волн отклоняются в ту или иную сторону. Если положение призм соответствует рис. 2, то дисперсия двух призм равна удвоенной дисперсии каждой из них. При повороте одной из призм Амичи на 180° относительно другой (вокруг оптической оси) полная дисперсия компенсатора оказывается равной нулю, так как дисперсия одной из призм скомпенсирована дисперсией другой. В зависимости от взаимной ориентации призм дисперсия компенсатора изменяется в пределах от нуля до удвоенного значения дисперсии одной призмы.

Для поворота призм друг относительно друга служат специальная рукоятка и система конических шестерен, с помощью которых призмы одновременно поворачиваются в противоположных направлениях. Вращая ручку компенсатора, следует добиваться того, чтобы граница света и тени в поле зрения стала достаточно резкой. Положение границы при этом соответствует длине волны $\lambda_D$, для которой обычно и приводятся значения показателя преломления.

В некоторых случаях, когда дисперсия исследуемого вещества особенно велика, диапазон компенсатора оказывается недостаточным и
чёткой границы получить не удаётся. В этом случае рекомендуется устанавливать перед осветителем жёлтый светофильтр.

Применяемая в рефрактометре Аббе поворотная призма П (призма Дове) позволяет сделать прибор более компактным.

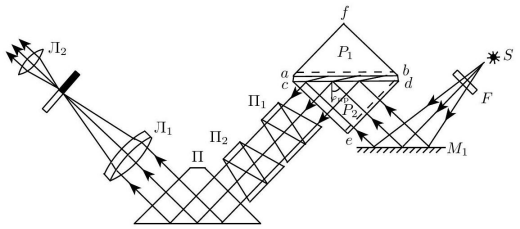



Рис. 3. Ход лучей в рефрактометре при измерении показателя преломления жидкости методом полного внутреннего отражения

На рис. 3 показан ход лучей в рефрактометре при работе по методу полного внутреннего отражения. В этом случае свет от источника $S$ после отражения от зеркала $M_1$ падает на матовую грань $ed$ призмы $P_2$ (в методе скользящего луча эта поверхность закрывается металлической шторкой). После рассеяния на грани ed свет падает на границу раздела стекло-жидкость под всевозможными углами. При $\varphi > \varphi_{\text{пр}}$ наступает полное внутреннее отражение, при $\varphi < \varphi_{\text{пр}}$ свет отражается частично. В поле зрения трубы наблюдается граница света и полутени.

Так как условия, определяющие величину предельного угла в методе скользящего луча и в методе полного внутреннего отражения, совпадают, положение линии раздела в обоих случаях также оказывается одинаковым.

Заметим, что, в отличие от метода скользящего луча, метод полного внутреннего отражения позволяет измерять показатели преломления непрозрачных веществ.

Рефрактометр Аббе можно использовать и для измерения показателей преломления твёрдых тел. И в этом случае применимы как метод полного внутреннего отражения, так и метод скользящего луча. Исследуемый образец должен иметь плоскую полирован ную поверхность; этой поверхностью он прижимается к гипотенузе $cd$ призмы $P_2$ (призма $P_1$ при этом отклоняется в сторону).

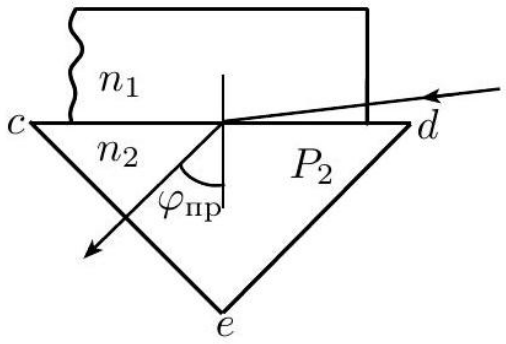



Рис. 4. Измерение

показателя преломления твёрдых тел Для обеспечения оптического контакта в зазор между соприкасающимися поверхностями вводится тонкий слой жидкости, показатель преломления $n$ которой удовлетворяет условию: $n_1 \leqslant n \leqslant n_2$, где $n_1$ - показатель преломления исследуемого образца, а $n_2$ - показатель преломления призмы $P_2$. При выполнении этого условия наличие слоя жидкости не искажает результатов измерения. (Студентам предлагается самостоятельно разобраться в этом вопросе.) Обычно для создания оптического контакта используют монобромнафталин, показатель преломления которого для жёлтых линий натрия равен $n_0 = 1,66$.

При работе по методу скользящего луча образец должен иметь полированную боковую поверхность, сквозь которую в него проникает свет (рис. 4).

Работа с прибором. На рис. 5 представлен внешний вид рефрактометра Аббе ИРФ-22. Измерение показателя преломления проводят в следующем порядке. Откинув призму $P_1$, с помощью пипетки помещают каплю исследуемой жидкости на поверхность призмы $P_2$. Возвратив призму $P_1$ в исходное положение, направляют свет от осветителя на грань $bf$ верхней призмы $P_1$ (при измерениях по методу скользящего луча) или с помощью зеркала $M_1$ - на грань $ed$ нижней призмы $P_2$ (при измерениях по методу полного внутреннего отражения). В первом случае шторка $L$ должна закрывать грань ed призмы $P_2$, а во втором её необходимо отвести в сторону.

С помощью зеркала $M_2$ направляют свет от окна в лаборатории (или от осветительной лампы) на шкалу показателей преломления, находящуюся за матовым стеклом в корпусе прибора. Яркость освещения шкалы подбирают, наблюдая её изображение через окуляр $_2$. Вращая блок призм $P_1$ и $P_2$ с помощью ручки $H_1$, приводят границу света и темноты (или света и полутени) в поле зрения окуляра; вращая ручку $H_2$ компенсатора, добиваются того, чтобы эта граница была резкой. После этого по шкале определяют величину показателя преломления.

Следует иметь в виду, что рабочие поверхности призм $P_1$ и $P_2$ легко повредить, поэтому при работе с ними необходима осторожность. Чтобы получать правильные значения показателей преломления, необходимо перед каждым опытом тщательно очищать рабочие поверхности

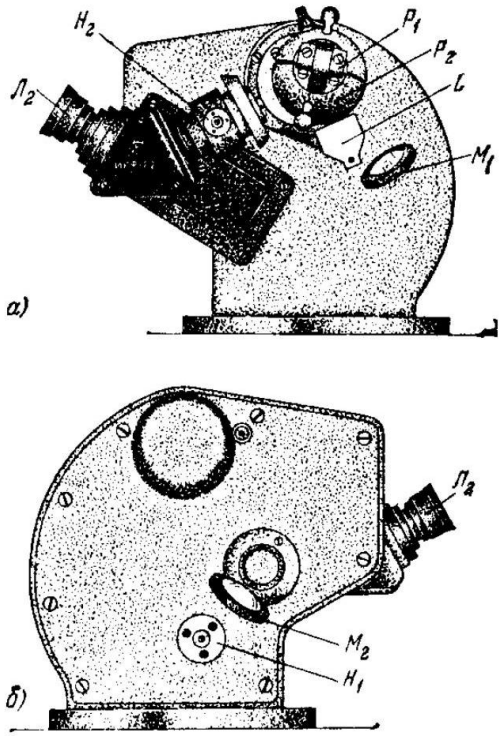



Рис. 5. Рефрактометр Аббе

призм. Приступая к измерениям, необходимо прежде всего убедиться в правильной работе прибора. Такую проверку проще всего выполнить, измерив показатель преломления вещества с известным показателем преломления. Для этого к каждому прибору придаётся эталонная стеклянная пластинка с известным $n_D$. В правильности работы прибора можно убедиться и в отсутствие такой пластинки, измеряя, например, показатель преломления дистиллированной воды, для которой (при $t = 20°\text{C}$) $n_D = 1,33291$. Если измерение даёт другой результат, следует определить поправку к шкале.

Абсолютная погрешность прибора по паспорту $\pm 2 \cdot 10^{-4}$. Область измерения в проходящем свете $1,3 \div 1,7$, в отражённом свете $1,3 \div 1,57$.

## ЗАДАНИЕ

В работе предлагается измерить показатели преломления нескольких стёкол и жидкостей.

1. Перед каждым измерением тщательно протрите рабочую поверхность призмы чистой сухой салфеткой, затем нанесите на поверхность каплю дистиллированной воды и снова протрите призму.

2. Проделайте серию контрольных измерений показателя преломления дистиллированной воды (или эталонной стеклянной пластины) и убедитесь в правильной работе прибора. Определите систематическую погрешность прибора и оцените случайную погрешность, возникающую при измерении (однократном) показателя преломления.

3. Измерьте показатели преломления $2 - 3$ стеклянных образцов (по указанию преподавателя), используя как метод скользящего луча, так и метод полного внутреннего отражения. Во избежание возникновения царапин на измерительной призме стеклянный образец кладётся на ровный тонкий слой монохлорнафталина, покрывающий всю поверхность призмы. Монохлорнафталин капают из пипетки на призму прибора и придавливают стеклянным образцом до возникновения ровного по всему полю наблюдения слоя жидкости.

4. Для каждого образца проведите измерения не менее 5 -ти раз каждым методом и оцените погрешность каждого метода. Сравните результат с паспортной погрешностью.

5. Аналогичным образом измерьте показатели преломления глицерина и этилового спирта.

6. По экспериментальным значениям показателей преломления вычислите молекулярные рефракции воды $R_{\mathrm{H_2O}}$, глицерина $R_{\mathrm{C_3H_8O_3}}$ и этилового спирта $R_{\mathrm{C_2H_6O}}$, а также поляризуемости $\alpha$ молекул этих веществ.

7. С помощью системы уравнений (7) вычислите атомные рефракции углерода $R_{\mathrm{C}}$, водорода $R_{\mathrm{H}}$ и кислорода $R_{\mathrm{O}}$.

8. Предполагая справедливым правило аддитивности, вычислите молекулярную рефракцию метилового спирта $R_{\mathrm{CH_4O}}$ и с помощью формулы (5) найдите его показатель преломления. Сравните вычисленное значение с табличным.

9. Используя вычисленные значения рефракций для воды $R_{\mathrm{H_2O}}$ и углерода $R_{\mathrm{C}}$, рассчитайте показатели преломления льда и алмаза. Сравните полученные значения с табличными.

### Контрольные вопросы

1. Что называется полным внутренним отражением?

2. Для чего грани $ab$ и $ed$ призм $P_1$ и $P_2$ делают матовыми?

3. При каких условиях слой жидкости между исследуемым твёрдым образцом и призмой $P_2$ не влияет на измеренное значение показателя преломления?

4. Что называется атомной и молекулярной рефракцией? Сформулируйте правило аддитивности для рефракций.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §84.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 137, 156.

3* Борн М., Волъф Э. Основы оптики. - М.: Наука, 1973. 2.3.

4. Дитчберн Р. Физическая оптика. - М.: Наука, 1965. Гл. 14, 15.

## 22.2 Experiments about Interference and Coherence

### 22.2.1 Newton's Rings

Цель работы: познакомиться с явлением интерференции в тонких плёнках (полосы равной толщины) на примере колец Ньютона и с методикой интерференционных измерений кривизны стеклянной поверхности.

В работе используются: измерительный микроскоп с опак-иллюминатором; плосковыпуклая линза; пластинка из чёрного стекла; ртутная лампа ДРШ; щель; линзы; призма прямого зрения; объектная шкала.

В нашей установке кольца Ньютона образуются при интерференции световых волн, отражённых от границ тонкой воздушной прослойки, заключённой между выпуклой поверхностью линзы и плоской стеклянной пластинкой.

Линии постоянной разности хода представляют собой концентрические кольца с центром в точке соприкосновения. Радиусы тёмных и светлых колец определяются формулами (2.71) и (2.72) соответственно. Для протяжённого источника линии равной толщины локализованы на поверхности линзы, если пластинка лежит на линзе, и вблизи поверхности линзы, если линза лежит на пластинке, как в нашем случае. Наблюдение ведётся в отражённом свете.

Экспериментальная установка. Схема экспериментальной установки приведена на рис. 1. Опыт выполняется с помощью измерительного микроскопа. На столике микроскопа помещается держатель с пластинкой чёрного стекла. На пластинке лежит исследуемая линза.

Источником света служит ртутная лампа, находящаяся в защитном кожухе. Для получения монохроматического света применяется призменный монохроматор, состоящий из конденсора $K$, коллиматора (щель $S$ и объектив $O$) и призмы прямого зрения П. Эти устройства с помощью рейтеров располагаются на оптической скамье. Свет от монохроматора попадает на опак-иллюминатор (ОИ), расположенный между окуляром и объективом микроскопа - специальное устройство для освещения объекта при работе в отражённом свете. Внутри опакиллюминатора находится полупрозрачная пластинка $P$, наклоненная под углом 45° к оптической оси микроскопа. Свет частично отражается от этой пластинки, проходит через объектив микроскопа и попадает на исследуемый объект. Пластинка может поворачиваться вокруг горизонтальной оси $x$, а опак-иллюминатор - вокруг вертикальной оси.

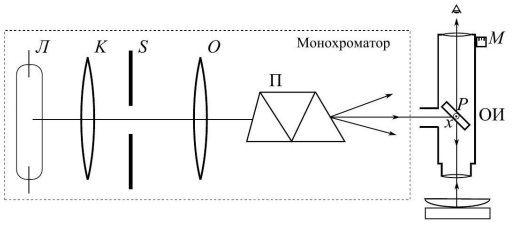


Рис. 1. Схема установки для наблюдения колец Ньютона

Столик микроскопа может перемещаться в двух взаимно перпендикулярных направлениях с помощью винтов препаратоводителя. Отсчетный крест окулярной шкалы перемещается перпендикулярно оптической оси микроскопа с помощью микрометрического винта.

Оптическая схема монохроматора позволяет получить в плоскости входного окна опак-иллюминатора достаточно хорошо разделённые линии спектра ртутной лампы. Изображение щели $S$ фокусируется на поверхность линзы объективом микроскопа, и в том же месте находится плоскость наблюдения микроскопа, т. е. точка источника и точка наблюдения интерференции совпадают. Картина интерференции как и в случае расположения пластинки сверху, так и в данном случае не зависит от коэффициента преломления линзы и определяется величиной зазора между нижней поверхностью линзы и стеклянной пластинкой.

Рекомендуется сначала настроить микроскоп на кольца Ньютона в белом свете (свете ртутной лампы), а затем после выделения монохроматором зелёной линии провести измерения в монохроматическом свете.

Настройка микроскопа. Включив ртутную лампу, располагают перед окном лампы конденсор, а остальные элементы схемы временно убирают с оптической скамьи. Осветив входное окно опак-иллюминатора, располагают исследуемую линзу на столике микроскопа.

Глядя на линзу сбоку невооружённым глазом, поворачивают опакиллюминатор вокруг вертикальной оси, а полупрозрачную пластинку $P$ - вокруг горизонтальной оси (рис. 1), пока на оправе линзы (или
листе бумаги, положенном на линзу) не появится яркое световое пятно. Перемещением тубуса микроскопа вдоль вертикальной оси фокусируют пятно на оправу линзы; при помощи препаратоводителей подводят точку соприкосновения линзы к оптической оси микроскопа.

Глядя в окуляр, вращением пластинки $P$ подводят световое пятно к центру поля зрения. Слегка перемещая тубус микроскопа по вертикали, добиваются появления чётких колец Ньютона в поле зрения микроскопа. Крест нитей должен проходить через середину тёмного пятна.

Настройка монохроматора. С помощью конденсора освещают щель $S$ светом лампы Л. Располагают объектив $O$ на расстоянии от щели, близком к фокусному, и размещают в промежутке между объективом и микроскопом призму прямого зрения. При этом в плоскости опак-иллюминатора появляются достаточно хорошо разделённые линии спектра ртутной лампы. Переход от одной линии к другой осуществляется путём перемещения микроскопа, яркость линий меняется при перемещении конденсора поперёк оптической оси при помощи специального винта. При измерении диаметров колец удобно использовать яркую зелёную линию ртути.

Передвигая микроскоп как целое и поворачивая его вокруг вертикальной оси, добиваются наибольшей освещённости поля зрения. Перемещение объектива $O$ вдоль оптической оси также влияет на освещённость поля зрения.

Определение радиуса кривизны линзы. Для определения радиуса кривизны линзы измеряют диаметры колец: устанавливают перекрестие на середину какого-либо достаточно удалённого от центра, но ещё отчётливо видимого тёмного кольца и снимают отсчёт по окулярной шкале: целые деления отсчитываются слева от риски, проходящей через окулярную шкалу, десятые и сотые доли деления - по окулярному микрометрическому винту $M$.

Перемещая перекрестие, последовательно устанавливают его на середины тёмных колец и записывают соответствующие показания окулярной шкалы и микрометра. После

прохождения через центральное пятно продолжают измерения, записывая возрастающие номера колец и координаты их диаметров. Для устранения ошибок, возникающих изза люфта в винте, перекрестие всегда следует подводить к кольцу с одной стороны. Цену одного деления окулярной шкалы определяют сравнивая её с изображением эталонной (объектной) шкалы. По разности отсчётов определяют диаметры, а затем и радиусы тёмных колец. Аналогичная серия измерений выполняется для светлых колец Ньютона.

При обработке результатов измерений удобно пользоваться графическим методом. Проще всего построить графики зависимости $r_m^2$ и $(r'_m)^2$ от номера $m$ кольца. Согласно теории (формулы (2.71) и (2.72)) эти графики должны представлять собой прямые линии, первая из которых проходит через начало координат. Зная длину волны $\lambda$, по наклону прямых можно рассчитать радиус $R$ кривизны линзы. По тому, насколько близко экспериментальные точки группируются вдоль прямой, можно судить о качестве аппаратуры и величине ошибок. Деформация линзы и стеклянной пластинки в месте их соприкосновения может приводить к отступлению от прямой зависимости при малых $m$.

Наблюдение «биений». При освещении системы светом, содержащим две спектральные компоненты, наблюдается характерная картина биений: чёткость интерференционных колец периодически изменяется. Это объясняется наложением двух систем интерференционных колец, возникающих для разных длин волн $\lambda_1$ и $\lambda_2$. Чёткие кольца в результирующей картине образуются при наложении светлых колец на светлые и тёмных на тёмные. Размытые кольца получаются при наложении светлых колец одной картины на тёмные кольца другой.

Нетрудно рассчитать период возникающих биений. Пусть в промежутке между двумя центрами соседних чётких участков укладывается $\Delta m$ колец для спектральной линии с длиной волны $\lambda_1$. Тогда в этом промежутке должно располагаться $(\Delta m + 1)$ колец для спектральной линии с длиной волны $\lambda_2$ (при $\lambda_2 < \lambda_1$). Получение окончательной формулы для $\Delta m$ мы предоставляем читателю.

Для освещения входного окна опак-иллюминатора сразу двумя спектральными линиями (например, жёлтой и зелёной) можно расфокусировать монохроматор, смещая объектив $O$ и призму.

Если смешать две линии не удаётся, то, убрав призму прямого зрения, можно объективом $O$ сфокусировать на окно опак-иллюминатора белый свет ртутной лампы. Результаты измерений не изменятся, так как остальные линии в спектре ртутной лампы заметно слабее жёлтой и зелёной.

Калибровка окулярной шкалы. Для определения цены деления окулярной шкалы сверху на линзу кладут калиброванную объектную шкалу. Плавно поднимая тубус, находят изображение миллиметровой шкалы и совмещают его с окулярной шкалой.

Объектная шкала размером 1 мм разбита на 100 делений. Используя всё поле зрения микроскопа, отмечают, какие из самых дальних штрихов объектной шкалы лучше всего совпадают со штрихами окулярной шкалы. Можно использовать для калибровки окулярный микрометр, совмещая перекрестие с началом и концом объектной шкалы.

## ЗАДАНИЕ

В работе предлагается, измерив диаметры колец Ньютона, определить радиус кривизны линзы; исследовать картину биений и рассчитать разность длин волн между жёлтой и зелёной спектральными линиями ртути.

При выполнении работы руководствуйтесь дополнительным описанием, расположенным на установке.

**I. Измерения**

1. Включите ртутную лампу и настройте микроскоп на кольца Ньютона в белом свете.
2. Настройте монохроматор, фокусируя на входном окне опакиллюминатора изображение зелёной линии ртути.
3. Вращая окулярный микрометрический винт, убедитесь, что перекрестие проходит через центр тёмного пятна и что поле зрения освещено симметрично слева и справа от центра.

Определите координаты диаметров тёмных и светлых колец (удобно координаты диаметров соответствующих колец записывать друг под другом).

Оцените диаметр пятна соприкосновения линзы со стеклянной пластинкой.

4. Проведите наблюдение биений для жёлтой и зелёной линий; на месте оцените разность длин волн и сопоставьте результат с табличным.
5. Прокалибруйте окулярную шкалу, используя эталонную объектную шкалу.

**II. Обработка результатов**

6. Рассчитайте цену деления окулярной шкалы и оцените погрешность результата.
7. Найдите разность длин волн жёлтой и зелёной линий ртути, полученную при наблюдении биений. Сравните результат с табличным.
8. Рассчитайте радиусы тёмных и светлых колец и постройте на одном листе графики зависимости $r_m^2$ и $(r'_m)^2$ от номера кольца $m$. Укажите на графике границы тёмного пятна. Какой из графиков должен проходить через начало координат?

Деформация линзы и стеклянной пластинки в месте их соприкосновения может приводить при малых $m$ к отступлению от формул (2.71) и (2.72). Меньше всего искажены деформацией кольца с большими номерами. При наличии искажений прямая, проведённая через точки, соответствующие слишком малым $m$, не пройдёт через начало координат.

9. Зная длину волны $\lambda$, по наклону прямых (с учётом калибровки) рассчитайте радиус $R$ кривизны линзы и оцените погрешность результата.

**Контрольные вопросы**

1. Почему в центре колец Ньютона в отражённом свете расположено тёмное пятно?
2. Как расположены кольца Ньютона в проходящем свете?
3. Почему ширина колец убывает с увеличением порядка?
4. Почему при расчёте интерференционной картины не учитывают отражение от передней (плоской) поверхности линзы?
5. Выведите формулу, связывающую период биений, выраженный в числе колец, с длинами волн спектральных линий.
6. Почему видность колец увеличивается при диафрагмировании пучка перед опакиллюминатором?

7* Почему интерференционные кольца больших номеров кажутся размытыми?

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. IV; VI, 26.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. III.

3* Дитчбёрн Р. Физическая оптика. - М.: Наука, 1965. Гл. V.
4* Захаръевский А.Н. Интерферометры. - М.: Оборонгиз, 1952. 7.

## 22.2.2 Jamin Interferometer

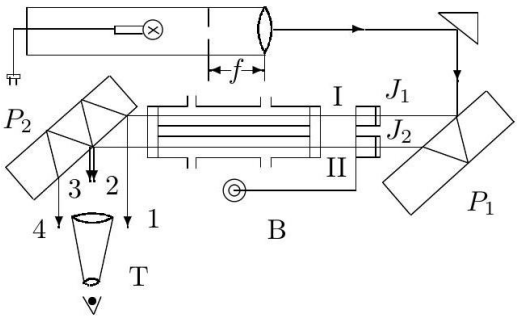
Цель работы: знакомство с техникой интерференционных измерений показателей преломления газов с помощью интерферометра Жамена.

В работе используются: интерферометр Жамена; газовая кювета; осветитель; зрительная труба; сильфон; баллон с углекислым газом; манометр; краны; светофильтр.

Принципиальная схема интерферометра Жамена описана в п. 5.2 введения к разделу.

Экспериментальная установка. В нашем интерферометре свет от лампы накаливания проходит коллиматорный объектив и поворотную призму, а затем падает на пластинку $P_1$ под углом 45° к ней (рис. 1). Пластины $P_1$ и $P_2$ закреплены на панели, ниже которой имеются два установочных винта, позволяющих в небольших пределах поворачивать зеркала. При этом пластина $P_1$ может поворачиваться вокруг горизонтальной оси (изменение ширины полос), а пластина $P_2$ - вокруг вертикальной оси (изменение положения полос). Между пластинами на пути лучей I и II расположена кювета длины $\ell$, состоящая из двух одинаковых камер, закрытых с торцов плоскопараллельными стеклянными пластинками.

В одну из камер вводится исследуемый газ, а вторая заполнена воздухом при атмосферном давлении. При этом разность хода $\Delta$, вызванная разностью показателей преломления газов $\delta n$, приводит к сдвигу интерференционных полос:




$$\Delta = \delta n \cdot \ell \tag{1}$$

Рис. 1. Экспериментальная установка

Сдвиг на одну полосу соответствует дополнительной разности хода $\Delta = \lambda$. Определив число полос $m$, на которое сместилась картина, можно рассчитать

$$\delta n = \frac{\Delta}{\ell} = m\frac{\lambda}{\ell} \tag{2}$$

На пути лучей I и II расположен компенсатор Жамена, состоящий из двух одинаковых плоскопараллельных стеклянных пластинок $J_1$ и $J_2$ (рис. 1). Если обе пластинки установлены под одинаковым углом к лучам, то и оптическая длина пути в них для обоих лучей оказывается одинаковой. Поворот одной из пластинок вокруг горизонтальной оси вызывает увеличение или уменьшение оптической длины пути

соответствующего луча. Это позволяет скомпенсировать разность хода, возникающую в камерах. Для точного отсчёта угла поворота одна из пластинок снабжена рычагом, конец которого смещается при помощи микрометрического винта В. Пластинки компенсатора ставятся под углом 45° к горизонтали, что позволяет использовать линейную экстраполяцию при измерениях. Смещение полос можно наблюдать через зрительную трубу Т. Расчёт величины смещения приведён в п. 5.2.

Интерферометр Жамена можно применять для измерения небольших изменений показателей преломления жидкостей или газов, а также для определения примесей различных газов в воздухе (например, для измерения концентрации рудничного газа в шахте).

Показатель преломления $n$ исследуемого газа определяется путём сравнения с воздухом при атмосферном давлении:

$$n = n_{\text{возд}} + \frac{\Delta}{\ell}. \tag{3}$$

Для определения величины $\Delta$ компенсатор следует прокалибровать.

Юстировка интерферометра. Перед началом работы камеру кюветы продувают воздухом, чтобы удалить из неё остатки углекислого газа.

Включают осветитель и с помощью экрана проверяют ход лучей I и II между зеркалами $P_1$ и $P_2$: оба луча должны проходить через камеры кюветы и пластинки компенсатора, установленные параллельно друг другу, и падать на пластину $P_2$. На выходе из пластины $P_2$ должны быть видны три пятна (рис. 1): крайние соответствуют лучам 1 и 4, среднее - лучам 2 и 3.

Для получения интерференционных полос в поле зрения трубы необходимо, чтобы ребро двугранного угла, образованного плоскостями пластин $P_1$ и $P_2$, было приблизительно горизонтальным. К такому расположению можно прийти путём вращения пластины $P_2$ относительно вертикальной оси. Следует иметь в виду, что используемые в интерферометрах в качестве зеркал стеклянные пластины не всегда оказываются достаточно хорошо изготовленными. Это приводит к некоторым особенностям в расположении интерференционных полос. В частности, полосы могут оказаться несколько наклонёнными к горизонтали, и этот наклон полос не удаётся устранить поворотом пластинки $P_1$ вокруг горизонтальной оси.

Далее установочным винтом пластинки $P_1$ регулируют ширину полос.

Калибровка компенсатора. Отъюстировав интерферометр, с помощью установочного винта пластинки $P_2$ совмещают нулевую полосу с перекрестием нитей в окуляре зрительной трубы.

Замечают нулевое деление микрометрического винта компенсатора. Вращая винт компенсатора, последовательно совмещают с перекрестием первую, вторую и т.д. полосы и записывают отсчёты. Градуировку следует проводить выделяя узкий интервал длин волн, для чего на пути лучей из осветителя устанавливается светофильтр. Результаты изображаются на градуировочном графике $z = f(m)$, где $z$ - отсчёт по шкале компенсатора. График позволит определять разность хода лучей в интерферометре по шкале компенсатора.

## ЗАДАНИЕ

В работе предлагается отъюстировать интерферометр; прокалибровать компенсатор, используя светофильтр; исследовать смещение интерференционных полос а) при изменении давления воздуха в одной из камер, б) при заполнении одной из камер углекислым газом при атмосферном давлении. По результатам измерений

рассчитываются показатели преломления воздуха и углекислого газа при нормальных условиях и средняя поляризуемость молекул воздуха.

### I. Юстировка прибора

1. Ознакомьтесь с устройством газовой системы (см. описание экспериментальной установки). Уравняйте давление в обеих камерах кюветы: первую соедините с атмосферой, открыв краны $К_1$ и $К_2$, а вторую (с открытым концом) продуйте с помощью груши Г, чтобы удалить из неё остатки углекислого газа (кран $К_0$ в положении 3).

2. Убедитесь, что пластинки $J_1$ и $J_2$ компенсатора установлены параллельно друг другу (на глаз).

3. Включите осветитель и с помощью экрана (листа белой бумаги) убедитесь, что оба луча I и II проходят через входные окна кюветы и попадают на пластину $P_2$. Световые пятна на второй пластине должны быть правильной круглой формы; если пятна имеют несимметричную форму, обратитесь за помощью к лаборанту.

Рассматривая глазом лучи, отражённые от пластины $P_2$, убедитесь, что видны три пятна: ближе к камере расположено пятно 1 (слабое - луч отражается от внешних поверхностей пластин), дальше всех от камеры - пятно 4 (яркое, т.к. оба раза луч отражается от зеркальных поверхностей) и центральное пятно средней яркости, образованное наложением лучей 2 и 3.

Поочерёдно закрывая лучи I и II, убедитесь в том, что яркость среднего пятна меняется незначительно.

4. Установите зрительную трубу так, чтобы лучи 2 и 3 попали в объектив. Настройте окуляр зрительной трубы на чёткое видение измерительного креста.

5. Настройте интерференционную картину: для этого медленно поворачивая пластину $P_2$ вокруг вертикальной оси с помощью установочного винта, добейтесь появления интерференционных полос в поле зрения трубы; убедитесь, что при этом наклон полос практически не изменяется. Ориентация полос в пространстве существенно зависит от качества (параллельности) пластин.

Установочным винтом пластинки $P_1$ задайте ширину полос порядка 1/10 поля зрения. Во всех дальнейших опытах установочные винты пластинок $P_1$ и $P_2$ трогать не следует.

### II. Калибровка компенсатора

6. Установите начало отсчёта, совместив перекрестие окуляра с нулевой полосой с помощью винта компенсатора. Из-за дисперсии стекла чёткой нулевой полосы нет, но за нуль можно принять один из самых сильных максимумов.

Повторите настройку нуля несколько раз, вращая винт в одну сторону, чтобы исключить люфт.

7. Прокалибруйте компенсатор в единицах $\lambda$, выделив узкий интервал длин волн с помощью светофильтра. Для этого наденьте на оправу окуляра красный светофильтр и, последовательно совмещая первую, вторую, ... $m$-ю подвижные полосы с перекрестием, запишите соответствующие отсчёты $z_m$ по микрометрическому винту компенсатора. При смещении на одну полосу разность хода меняется на длину волны.

При калибровке используйте все полосы, наблюдаемые в окуляре сверху и снизу от нулевой полосы (со светофильтром их больше двадцати).

8. Запишите длину кюветы $\ell$, указанную на установке, а также длину волны $\lambda$ и полосу пропускания светофильтра, указанные на его оправе.

### III. Зависимость $\delta n$ от $P$ для воздуха

9. Убедитесь, что давление воздуха в обеих камерах кюветы атмосферное (краны $K_1$ и $K_2$ открыты); установите шток сильфона приблизительно на середину его длины и отсоедините первую камеру от атмосферы, перекрыв кран $K_1$.

10. Изменяя давление с помощью сильфона и совмещая нулевую полосу с перекрестием, снимите зависимость показаний компенсатора $z$ от перепада давлений $\Delta P$. Если давление «плывёт», фиксируйте величину $\Delta P$ сразу после совмещения перекрестия с нулевой полосой.

Давление следует изменять в обе стороны от атмосферного в пределах рабочей области манометра ($\pm$1000 мм вод. ст.).

### IV. Сравнение показателей преломления воздуха и углекислого газа при атмосферном давлении

11. Соедините первую камеру кюветы с атмосферой, открыв кран $K_1$. Заполните углекислым газом камеру с открытым концом. Для этого $3-4$ раза плавно, чтобы избежать резкого изменения температуры газа при расширении, переведите кран $K_0$ из положения 1 в положение 2.

12. Совместите нулевую полосу с перекрестием. Слишком малое смещение картины ($<$ 25 полос для камеры длиной 10 см) означает, что камера заполнена смесью углекислого газа с воздухом. Если повторная прокачка камеры не помогает, обратитесь за помощью к лаборанту.

Снимите зависимость показаний компенсатора от времени, раз в минуту возвращая нулевую полосу к перекрестию, и оцените время установления равновесия. Повторите измерения, стараясь заполнять кювету как можно более плавно.

13. Определите температуру $T$ и давление $P$ по показаниям лабораторного термометра и барометра.

### Обработка результатов

1. Постройте калибровочный график $z_m = f(m)$ - зависимость отсчёта по компенсатору от номера совмещённой полосы.

2. Постройте график $z = f(\Delta P)$ (от +1000 до -1000 мм $H_2O$). Определите угол наклона прямой; с помощью калибровочного графика и формулы (2) перейдите от делений компенсатора $\Delta z$ к величине $\delta n$; рассчитайте сначала среднюю поляризуемость молекулы воздуха, а затем - показатель преломления воздуха в условиях опыта.

Пересчитайте показатель преломления к нормальным условиям и сравните результат с табличным.

3. Молекулу газа можно представить как металлический шарик в однородном электрическом поле. Оцените радиус молекулы азота по результатам измерений.

4. Рассчитайте показатель преломления $n$ для углекислого газа в условиях опыта по формуле (3), взяв показатель преломления воздуха, рассчитанный по результатам эксперимента.

Пересчитайте $n$ к нормальным условиям и сравните результат с табличным.

5. Оцените экспериментальные погрешности.

6. Оцените интервал $\delta n$, доступный для измерений, исходя из возможностей компенсатора: минимальная величина $\delta n$, доступная для измерений, определяется точностью компенсатора, максимальная - диапазоном его работы.

**Контрольные вопросы**

1. Почему пластинки $P_1$ и $P_2$ интерферометра Жамена должны быть достаточно толстыми (не менее $2-3$ см)?

2. Могут ли эти пластинки сильно отличаться по толщине?

3. Объясните, почему увеличивается число наблюдаемых интерференционных полос, если на пути лучей из осветителя установлен светофильтр.

4. Покажите, что ширина наблюдаемых интерференционных полос пропорциональна отношению $\lambda/\beta$.

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. IV, гл. VI, 25; гл.VII, §28.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. III.

3* Захаръевский А.Н. Интерферометры. - М.: Оборонгиз, 1952. 10.

## 22.2.3 Rayleigh Interferometer

Цель работы: ознакомление с интерференцией на двух щелях, устройством и принципом действия интерферометра Релея и с его применением для измерения показателей преломления газов.

В работе используются: технический интерферометр ИТР-1, светофильтр, баллон с углекислым газом, сильфон, манометр, краны.

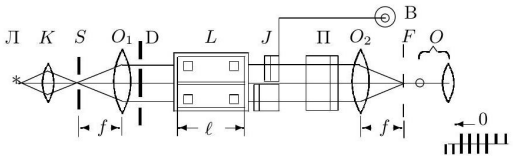
Принципиальная схема устройства интерферометра Релея описана в п. 5.1 введения к разделу.

Экспериментальная установка. Интерферометр Релея - прибор для измерения разности показателей преломления - основан на явлении дифракции света на двух параллельных щелях. Схема прибора представлена на рис. 1 в вертикальной и горизонтальной проекциях. Лампа накаливания Л с помощью конденсора $K$ ярко освещает узкую входную щель $S$, расположенную в фокусе объектива $O_1$ (фокусное расстояние $f$). Коллиматор, состоящий из щели $S$ и объектива $O_1$, посылает параллельный пучок на диафрагму $D$ с двумя вертикальными щелями (расстояние между щелями $d$). Свет после двойной щели проходит кювету $L$, состоящую из двух одинаковых стеклянных камер, в которые вводятся исследуемые газы (в нашей

установке $-CO_2$ или воздух). Кювета занимает только верхнюю часть пространства между объективами $O_1$ и $O_2$, длина кюветы $\ell$. За кюветой расположены две стеклянные пластинки $J$ (компенсатор Жамена, см. ниже) и пластинка .

Интерференционная картина (картина дифракции на двух щелях), наблюдаемая в фокальной плоскости $F$ объектива $O_2$, представляет собой две системы равноотстоящих полос, параллельных щелям: верхняя

а)




б)

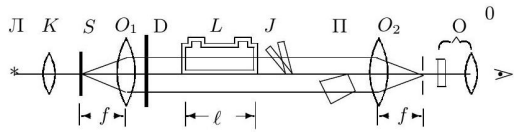



Рис. 1. Устройство интерферометра Релея:

а) вид сверху; б) вид сбоку

(подвижная) образована лучами, прошедшими через кювету, нижняя (неподвижная) - лучами, прошедшими под кюветой. Обе системы интерференционных полос разграничены при помощи пластины П тонкой разделительной линией. Для наблюдения двух систем полос в окуляре применена цилиндрическая линза диаметром 2,2 мм, ось которой расположена вертикально. Вторая («глазная») линза окуляра - обычная сферическая. Она служит для подстройки чёткости картины под глаз наблюдателя.

При малых дифракционных углах $\varphi = \lambda/d$ расстояние между соседними светлыми (или тёмными) полосами $\delta y$ зависит от длины волны $\lambda$, фокусного расстояния $f$ объектива $O_2$ и расстояния между дифракционными щелями $d$:

$$\delta y = f\frac{\lambda}{d} \tag{1}$$

В техническом интерферометре ИТР-1, который используется в нашей работе, $f \simeq 20$ см, $d \simeq 1,5$ см, и $\delta y$ оказывается порядка $10^{-3}$ см. Для наблюдения таких мелких интерференционных полос требуется окуляр с большим увеличением ($\gamma \simeq 150^{\times}$). Короткофокусная цилиндрическая линза окуляра О сильно растягивает интерференционную картину по горизонтали, не меняя её вертикальных размеров и тем самым мало ослабляя освещённость полос. Изображение светящейся точки в фокальной плоскости объектива $O_2$ при рассматривании через цилиндрическую линзу имеет вид светлой вертикальной линии, длина которой определяется диаметром объектива. Поэтому распределение освещённости в нижней части светлой линии зависит от действия нижней ча сти объектива, а в верхней части линии - от верхней части объектива. Таким образом, наблюдатель видит две системы полос: верхняя образована лучами, прошедшими через кюветы, нижняя - лучами, прошедшими под кюветами.

При заполнении камер газами с одинаковым показателем преломления $n$ обе системы полос совпадают. Оптическая разность хода $\Delta = \delta n \cdot l$, возникающая при прохождении света через камеры с разными газами $\delta n = n_2 - n_1$, ведёт к поперечному смещению верхней дифракционной картины относительно неподвижной нижней. Смещение на одну полосу

соответствует дополнительной разности хода $\Delta = \lambda$. Просчитав число полос $m$ между центрами обеих картин, можно рассчитать

$$\delta n = \frac{\Delta}{\ell} = m\frac{\lambda}{\ell} \tag{2}$$

Для точного измерения разности хода используется компенсатор Жамена ($J$ на рис. 1) - устройство, которое позволяет вернуть подвижную систему полос к первоначальному положению, т. е. вновь совместить обе системы полос. В установке компенсатор Жамена расположен за кюветой. Он состоит из двух одинаковых плоскопараллельных стеклянных пластинок, установленных на пути лучей под углом 45° к горизонтали. Вращение одной из пластин вокруг горизонтальной оси, перпендикулярной оси системы, вызывает увеличение или уменьшение оптической длины пути соответствующего луча. Ось вращения снабжена рычагом, конец которого смещается при помощи микрометрического винта В.

Интерферометр Релея можно применять для измерения небольших изменений показателей преломления жидкостей или газов, а также для определения примесей различных газов в воздухе (например, для измерения концентрации рудничного газа в шахте). Показатель преломления $n$ исследуемого газа определяется путём сравнения с воздухом при атмосферном давлении:

$$n = n_{\text{возд}} + \frac{\Delta}{\ell} \tag{3}$$

Для определения величины $\Delta$ компенсатор следует прокалибровать.

## ЗАДАНИЕ

В работе предлагается исследовать изменение показателя преломления воздуха при изменении давления и определить разность показателей преломления воздуха и углекислоты при атмосферном давлении. По результатам измерений рассчитываются показатели преломления воздуха и углекислого газа при нормальных условиях.

### I. Подготовка к работе

1. Включите осветитель интерферометра в сеть и убедитесь, что в поле зрения окуляра видны две системы интерференционных полос. Если картина недостаточно яркая - обратитесь за помощью к лаборанту.

2. Ознакомьтесь с устройством газовой системы (см. описание экспериментальной установки). Уравняйте давление в обеих камерах кюветы: первую соедините с атмосферой, открыв краны $К_1$ и $К_2$, а вторую (с открытым концом) продуйте с помощью груши Г, чтобы удалить из неё остатки углекислого газа (кран $К_0$ в положении 3).

### II. Калибровка компенсатора

3. Уравняв давление в камерах, подождите 2-3 минуты, пока выровняются температуры. Установите начало отсчёта, совместив с помощью компенсатора обе системы полос. Установка нуля в белом свете - совмещение центральных полос - облегчается тем, что боковые полосы окрашены. Совместите (приблизительно) боковые полосы с симметричной окраской, а затем (как можно точнее) - центральные. Один оборот винта (100 делений) соответствует одному делению вертикальной шкалы компенсатора. Для удобства отсчёта используется лупа.

Повторите опыт несколько раз, вращая винт в одну сторону, чтобы исключить люфт.

4. Прокалибруйте компенсатор в единицах $\lambda$, выделив узкий интервал длин волн с помощью светофильтра. При калибровке используйте все полосы, наблюдаемые в окуляре слева и справа от нулевой полосы (со светофильтром их больше двадцати).

5. Запишите длину кюветы $\ell$, указанную на установке, а также длину волны $\lambda$ и полосу пропускания светофильтра, указанные на его оправе.

**III. Зависимость $\Delta n$ от $P$ для воздуха**

6. Убедитесь, что давление воздуха в обеих камерах кюветы атмосферное (см. п. 2). Установите сильфон в среднее положение и отсоедините первую камеру от атмосферы, перекрыв кран $K_1$.

7. Изменяя давление с помощью сильфона и совмещая нулевые полосы, снимите зависимость показаний компенсатора $z$ от перепада давлений $\Delta P$. Если давление «плывёт», фиксируйте величину $\Delta P$ сразу после совмещения нулевых полос.

Давление следует изменять в обе стороны от атмосферного в пределах рабочей области манометра ($\pm$1000 мм вод. ст.).

IV. Сравнение показателей преломления воздуха и углекислого газа при атмосферном давлении

8. Соедините первую камеру кюветы с атмосферой, открыв кран $K_1$, и отключите манометр, закрыв кран $K_2$. Заполните углекислым газом камеру с открытым концом. Для этого 3-4 раза плавно, чтобы избежать резкого изменения температуры газа при расширении, переведите кран $K_0$ из положения 1 в положение 2.

9. Совместите нулевые полосы. Небольшое смещение подвижной картины ($< 25$ полос для камеры длиной 10 см) означает, что камера заполнена смесью углекислого газа с воздухом. Если повторная прокачка камеры не помогает, обратитесь за помощью к лаборанту.

Снимите зависимость равновесного положения компенсатора от времени, раз в минуту совмещая нулевые полосы, и оцените время установления равновесия. Повторите измерения, стараясь заполнять кювету как можно более плавно.

10. Определите температуру $T$ и давление $P$ по показаниям лабораторного термометра и барометра.

**V. Обработка результатов**

1. Постройте калибровочный график, отложив по оси абсцисс номер совмещённой полосы $m$, а по оси ординат величину $z$ (отсчёт по компенсатору). Проверьте, сохраняется ли линейность графика вдали от нулевого положения компенсатора.

2. Постройте график зависимости показателя преломления $\delta n$ от разности давлений $\Delta P$ (от +1000 до -1000 мм вод. ст.). Величину $\delta n$ рассчитайте по формуле (2) с помощью калибровочного графика.

По углу наклона рассчитайте среднюю поляризуемость молекул воздуха, а затем - показатель преломления воздуха в условиях опыта.

Пересчитайте показатель преломления к нормальным условиям и сравните результат с табличным.

3. Рассчитайте показатель преломления для углекислого газа в условиях опыта по формуле (3), взяв показатель преломления воздуха, рассчитанный по результатам эксперимента.

Пересчитайте $n\,(CO_2)$ к нормальным условиям и сравните результат с табличным.

4. Оцените экспериментальные погрешности.

5. Оцените интервал $\delta n$, доступный для измерений, исходя из возможностей компенсатора: минимальная величина $\delta n$, доступная для измерений, определяется точностью компенсатора, максимальная - диапазоном его работы.

**Контрольные вопросы**

1. Объясните, почему при поочерёдном перекрывании щелей $D$ наблюдаемая дифракционная картина не смещается.

2. Почему ширина щели $S$ коллиматора должна быть достаточно малой? Оцените максимальный размер этой щели, если известны фокус коллиматора и расстояние между щелями.

3. Будет ли наблюдаться чёткая нулевая полоса, если исследуемое вещество обладает заметной дисперсией $n = n(\lambda)$? Сделайте соответствующие оценки.

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. IX, 44, 45.

2. Бутиков Е.И. Оптика. - М.: Высшая школа, 1986. 5.6.

3. Горелик Г.С. Колебания и волны. - М.: Физматгиз, 1959. Гл. IX.

4. Кингсеп А.С., Ложшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 7, §7.8.

5* Захаръевский А.Н. Интерферометры. - М.: Оборонгиз, 1952. 11, 12.

## 22.2.4 Michelson Interferometer

Цель работы: Изучение двухлучевой интерференции, определение длины волны, проверка эффекта Доплера.

В работе используются: интерферометр Майкельсона с подвижным зеркалом, лазер, фотоумножитель, частотомер, линзы.

Интерферометр Майкельсона находит применение в спектрометрах с высоким разрешением, для абсолютных и относительных измерений длин с точностью 0,005 мкм. Принципиальная схема интерферометра описана в п. 5.3 введения к разделу.

Оптическая схема интерферометра приведена на рис. 1. Источником света служит лазер ЛГ. Лазер излучает узкий пучок света, который фокусируется линзой $_1$. В фокусе этой линзы возникает точечный источник света $S$. Сферическая световая волна от источника $S$ падает на делительный кубик ДК и делится его диагональной гранью на две волны - отражённую 1 и проходящую 2. Волна 1 отражается от зеркала $_1$, возвращается к кубику, частично проходит сквозь него и попадает на экран Э. Волна 2 отражается от зеркала $3_2$, частично отражается от кубика и также попадает на экран.

Световые волны 1 и 2
испускаются одним источником $S$, и они когерентны между собой. Эти волны создают на экране Э интерференционную картину. Для увеличения масштаба интерференционной картины может быть использована линза $\Omega_2$.

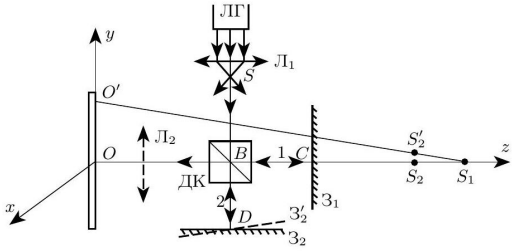



Рис. 1. Схема интерферометра

Зеркало $З_1$ установлено перпендикулярно падающему лучу. Оно может перемещаться вдоль луча. Это зеркало в дальнейшем будет называться подвижным. Зеркало $З_2$ вдоль направления падающего луча не перемещается. Его, однако, можно наклонять по отношению к лучу.

Экспериментальная установка. Схема экспериментальной установки приведена на рис. 2. Источником света служит гелий-неоновый лазер ЛГН-203. Его излучение обладает большой длиной когерентности, что позволяет получать хорошо различимую глазом интерференционную картину при разности хода в десятки сантиметров. Неподвижное зеркало $З_2$, поворачивается микрометрическими винтами $М_г$ (относительно горизонтальной) и $М_В$ (относительно вертикальной оси). Зеркало $З_1$ установлено перпендикулярно падающему лучу. Оно может передвигаться вдоль луча с помощью микрометрического винта, соединённого с двигателем Дв через муфту и редуктор РД, позволяющий менять скорость движения зеркала. Двигатель питается от сети через блок питания пульта управления. Концевые контакты $К_1$ и $К_2$ меняют направление движения зеркала на обратное. Включение лазера и двигателя производится с пульта управления. Сигнальные лампочки указывают, в какую сторону движется зеркало.

Интерференционная картина наблюдается на экране Э. Она может быть увеличена с помощью линзы $\Omega_2$. В этом случае на экране в увеличенном масштабе воспроизводится интерференционная картина, ко-

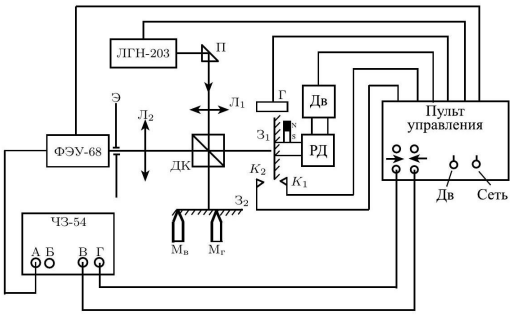



Рис. 2. Схема экспериментальной установки

торая создаётся перед линзой в плоскости, сопряжённой экрану. Линза закреплена на съёмном столике, её фокусное расстояние 4,3 см.

Для регистрации изменения интенсивности света используется фотоэлектронный умножитель ФЭУ-68, установленный непосредственно за экраном. Свет на окно ФЭУ попадает через небольшое отверстие в центре экрана. Для питания ФЭУ используется высоковольтный выпрямитель. Выпрямитель включается тумблером «Сеть» на пульте управления.

Периодическое изменение интенсивности света, возникающее при движении зеркала $_1$, приводит к такому же изменению сигнала ФЭУ. Число периодов изменения интенсивности света пересчитывается частотомером Ч3-54. Частотомер может работать в одном из трёх режиМОВ.

1. Он может измерять число импульсов, поступающих на его входы (А или Б) за некоторый промежуток времени (его продолжительность определяется поступлением сигналов на управляющие входы В и Г).

2. С помощью частотомера можно измерять промежутки времени. Для таких измерений в прибор встроен кварцевый генератор. Частотомер измеряет время, прошедшее между поступлением сигналов на его
управляющие входы, подсчитывая соответствующее число импульсов кварцевого генератора.

3. Наконец, частотомер может измерять частоту сигнала, поступающего на его вход, сравнивая число периодов исследуемого сигнала с числом импульсов кварцевого генератора.

Для получения управляющих сигналов используется геркон Г (герметичный магнитоуправляемый контакт). Схема работает следующим образом. На отсчётной головке микрометрического винта зеркала $3_1$ закреплён небольшой магнит. Головка вращается вместе с винтом. После срабатывания концевого контакта $К_1$ зеркало начинает двигаться к экрану. При приближении магнита к геркону вырабатывается управляющий сигнал, который подаётся через схему пульта управления на вход В частотомера. Частотомер начинает счёт импульсов. Сигнал на окончание счёта подаётся на вход Г после того, как с помощью геркона зарегистрировано 32 оборота ходового винта. После срабатывания концевого контакта $К_2$ зеркало начинает движение от экрана. На этом участке движения счёт импульсов не производится. Один оборот микрометрического винта приводит к перемещению зеркала на 1 мм. Таким образом, полное перемещение зеркала $3_1$ составляет $L = 32$ мм.

## ЗАДАНИЕ

Работа содержит несколько упражнений.

В первом упражнении исследуется интерференционная картина, возникающая при различных положениях зеркала $3_2$, проверяется формула (2.69).

Во втором упражнении измеряется длина волны лазерного излучения путём пересчёта интерференционных колец, пропадающих в центре экрана при движении зеркала $3_1$. Такие измерения не очень точны, но обладают тем преимуществом, что не требуют сложных расчётов.

В третьем упражнении исследуется эффект Доплера.

Перед началом работы следует отъюстировать установку.

### I. Юстировка системы

1. Включите блок питания установки (тумблер слева от частотомера).

2. Убедитесь, что луч от поворотной призмы (П на рис. 2) идёт параллельно столу на высоте $100 \pm 2$ мм (при этом линза $\Omega_1$ и делительный кубик сняты).

3. Установите оправу с зеркалом $_2$ перпендикулярно лучу поворотом оправы в зеркале (грубо): для этого освободите фиксирующие (жёлтые) винты, микровинтами $М_г$ и $М_в$ установите плоскость правильно и снова зажмите фиксирующие винты.

Поочерёдно вращая винты $М_г, М_в$ зеркала $S_2$, заведите (точно) отражённый луч на поворотную призму.

4. Установите делительный кубик в центре системы и определите его положение относительно вертикали: луч от поворотной призмы, отразившись от полупрозрачной грани кубика, должен попасть на центр подвижного зеркала $3_1$, а прямой луч, отразившись от зеркала $3_2$, - на центр экрана. При этом пучки должны проходить через центр кубика, а яркость пятен, отражённых от подвижного и неподвижного зеркал, на экране должна быть примерно одинаковой. Если яркость пятен на экране заметно отличается, поверните кубик относительно вертикали на 90° и ещё раз проведите настройку. Центр экрана (вход в ФЭУ) должен находиться на той же высоте, что и луч от поворотной призмы до зеркала $_2$ (около 10 см).

5. Поочерёдно вращая винты столика разделительного кубика, совместите оба луча в центре экрана.

Полезно убедиться в высокой чувствительности оптической системы к механическим деформациям. Последите за тем, как «дышит» интерференционная картина на экране, если надавить пальцем на металлическую станину интерферометра: полосы равного наклона двигаются по дифракционному пятну лазера на экране.

6. Поставьте линзу $\Omega_1$ между поворотной призмой и кубиком. Убедитесь, что лазерный луч проходит через центр линзы (все пятна на зеркалах и на экране должны остаться на одной высоте). При необходимости совместите центр линзы с осью пучка.

Перемещая линзу поперёк луча, совместите центр интерференционной картины с отверстием ФЭУ на экране. Прикрепите линзу к столу скобой с винтом.

**II. Исследование интерференционной картины**

1. Чтобы увеличить интерференционную картину, установите между экраном и кубиком столик с линзой $_2$. Перемещая линзу, добейтесь того, чтобы центр колец снова совпал с отверстием в экране.

2. Слегка проворачивая вручную муфту двигателя, добейтесь того, чтобы в центре экрана оказалось тёмное пятно. Приложите к экрану лист бумаги, на котором проведён крест из взаимно перпендикулярных линий. Совместите центр креста с центром колец, наметьте положение 5-6-ти первых тёмных колец и измерьте их диаметры.

3. Убедитесь в справедливости формулы (2.69). Для этого постройте график зависимости квадрата радиуса кольца от его номера. По наклону полученной прямой найдите величину $L(L-a)/m_0$. Сравните найденное число с результатом теоретического расчёта. Согласно рис. 1, $L = OS_1 = BS_1 + OB, L - a = OS_2 = BS_2 + OB, BS_1 = 2BC + SB$, $BS_2 = 2BD + SB$. Проведите расчёт, принимая $SB = 5$ см, $BC = 26$ см (для среднего положения зеркала), $BD = 22$ см, расстояние от делительного кубика до линзы $\Omega_2 - 16,5$ см, расстояние от линзы до экрана - 15 см, фокусное расстояние линзы $\Omega_2 - 4,3$ см.

[01] Из-за недостаточной механической развязки измерить радиусы колец затруднительно. Упражнения 2 и 3 легко выполняются, если сфотографировать интерференционную картину и провести измерения по фотографии.

4. Получите на экране картину вертикальных полос. Для этого уведите центр пятна в сторону по горизонтали ($O \to O'$ на рис. 1), сместив микрометр М на $0,08 - 0,09$ мм от исходного положения. Измерьте ширину полосы в центре экрана (в т. $O$) и расстояние $O'O$ на экране. Зная $OS_1$, оцените угол поворота зеркала и смещение изображения источника света ($S_2S_2'$). Угол поворота зеркала вдвое меньше угла $OO'/OS_1$.

5. Вновь получите картину колец. Снимите столик с линзой $_2$.

**III. Измерение длины волны лазерного излучения**

1. Совместите центр интерференционной картины с отверстием в экране (входом «ФЭУ»).

2. Включите тумблер «Сеть» на частотомере, при этом должны загореться лампы индикаторного табло.

3. Установите частотомер в режим счёта импульсов. Для этого переключатель рода работы поставьте в положение «Интервал А/В-Г», кнопка «блок» отжата, ручка «время индикации» - в среднем положении, кнопка ручной работы с изображением руки отжата (автоматический режим).

Положения ручек на блоке частотомера «измеритель интервалов времени» отмечены.

4. Переместите рычаг регулятора скорости двигателя в положение 2 (регулятор разрешается переключать ТОЛЬКО при выключенном двигателе).

Включите двигатель тумблерами «Сеть» и «Дв» на пульте управления; при этом загорится индикатор, указывающий направление движения («←» или « →) и зеркало начнёт движение от ФЭУ. Проследите за работой схемы управления двигателем при $2 - 3$ изменениях направления движения зеркала. Если вы обнаружите сбои в работе аппаратуры, обратитесь к преподавателю или лаборанту.

5. Определите по табло частотомера число колец $N$, прошедших через центр экрана (через вход «ФЭУ») за время движения зеркала. В выбранном режиме работы частотомер включается при отходе зеркала от контакта $К_1$ и выключается после 32 -го оборота ходового винта. Перемещение зеркала $L = 32$ мм. Запуск частотомера происходит автоматически: после вспышки индикатора «» (движение к экрану) зажигается лампочка селектора частотомера и начинается отсчёт.

Измерения числа интерференционных полос $N$, проходящих через центр экрана, проведите не менее 10 раз. Отключите двигатель.

По результатам этих измерений найдите длину волны лазерного излучения и оцените ошибку. Сравните полученный результат с паспортным: $\lambda = 6328$Å.

[0]2 Возможные сбои: 1) счётчик числа колец не включается; 2) счётчик не выключается после прохождения 32 -х см (не гаснет кнопка «счёт» на частотомере); в этом случае следует выключить и снова включить двигатель тумблером «Дв».

### IV. Исследование эффекта Доплера

Измерения сводятся к определению частоты колебаний интенсивности света на входе ФЭУ при движении зеркала и определению скорости этого движения. Для проверки результатов измерений используется формула (2.70).

При измерении частоты частотомер работает в режиме измерения частоты, при измерении скорости - в режиме измерения времени (скорость рассчитывается по времени, которое требуется для перемещения зеркала на расстояние $L = 32$ мм).

1. При измерении времени передвижения зеркала переключатель «Род работы» ставится в положение «интерв. В-Г», переключатель «Метки времени» - в положение 1mS (т.е. время движения зеркала измеряется в миллисекундах).

Чтобы измерить доплеровское смещение частоты, переключатель «Род работы» устанавливается в положение «Частота А », переключатель «Время счёта-множитель» - в положение $10^3$, кнопка ручной работы отжата. Вращением ручки «Время индикации» подбирается удобное для наблюдения время, в течение которого измеренное значение частоты высвечивается на табло ($\sim 2 - 3$ с).

2. Установите максимальную скорость передвижения зеркала (рычаг регулятора скорости - в положение I). Включите двигатель.

3. При выбранном значении скорости сначала измерьте время передвижения зеркала на расстояние $L$, затем доплеровскую частоту - частоту изменения яркости света, падающего на ФЭУ (число полос, проходящих через ФЭУ за 1 с). Повторите измерения столько раз, сколько необходимо, чтобы погрешность измерения этих величин не превышала 2%.

4. Повторите измерения ещё для $2 - 3$-х значений скорости передвижения зеркала (переключать регулятор ТОЛЬКО при выключенном двигателе!).

5. При остановленном двигателе измерьте частоту дрожания картины. Сравните с изменением доплеровской частоты во время движения (скорость двигателя флуктуирует).

6. Постройте график доплеровской частоты в зависимости от скорости передвижения зеркала. Используйте полученные результаты для определения длины световой волны. Совпадает ли найденное вами значение с длиной волны, определённой во втором упражнении?

7. Выключите тумблер «Сеть» частотомера. В тот момент, когда зеркало находится приблизительно посередине между контактами $K_1$ и $K_2$, выключите двигатель (тумблеры «Дв» и «Сеть» на пульте управления). Установите рычаг переключения скорости редуктора в положение 1.

Выключите напряжение на блоке питания.

### Контрольные вопросы

1. С какой скоростью движется картина интерференционных полос при перемещении подвижного зеркала?

2. Чем ограничена наибольшая и наименьшая скорости движения зеркала, которые могут быть измерены в нашей установке по доплеровскому смещению частоты?

3. Чем определяется точность в измерении длины волны лазера? Как эта точность зависит от скорости перемещения зеркала?

4. Какую форму имеют интерференционные полосы при произвольном расположении источников $S_1$ и $S_2$?

**Литература**

1. Калитеевский Н.И. Волновая оптика. М.: Высшая школа, 1978. 5, 6, 7, 3.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 28, 133.

3. Борн М., Волъф Э. Основы оптики. - М.: Наука, 1973. 7.5.4.

## 22.2.5 Coherence of Light

Цель работы: оценка длины волны излучения полупроводникового лазера, определение зависимости радиуса когерентности от размера источника, оценка спектрального диапазона чувствительности зрительной системы экспериментатора.

В работе используются: лазер, лампы накаливания с блоком питания, объектив, оптические щели, микроскоп.

Источник света. Источником света в лабораторной установке является лампа накаливания с вольфрамовой спиралью, нагретой примерно до 2000 К. Это тепловой источник, в котором распределение мощности излучения по длинам волн зависит от температуры спирали. Излучение нагретого вольфрама может приближенно быть описано моделью равновесного излучения абсолютно черного тела. Спектральный состав излучения в этой модели описывается известной формулой Планка (вывод формулы можно найти в курсах квантовой механики):

$$u_\lambda = \frac{8\pi hc}{\lambda^5} \cdot \frac{1}{e^{\frac{hc}{\lambda k_\mathrm{B} T}} - 1}$$

где $u_\lambda$ есть объемная плотность энергии излучения, отнесенная к единичному интервалу длин волн [1] и имеет размерность Дж/м [4], $h$ - постоянная Планка, $c$ - скорость света, $k$ - постоянная Больцмана, $T$ - температура источника. На рис. 1 приведена зависимость спектральной плотности излучения абсолютно черного тела при температуре 2000 К. График приведен в относительных единицах и нормирован на значение $u_\lambda$ в максимуме, который наблюдается на длине волны $\lambda_{\max} \approx 1,45$ мкм (инфракрасный диапазон). Следствием формулы Планка является уменьшение $\lambda_{\max}$ с ростом температуры, которое описывается законом Вина:

$$\lambda_{\max} = b/T$$

где константа $b \approx 0,2\frac{hc}{k_\mathrm{B}} \approx 0,2898$ см $\cdot$ К.

Как видно из рис. 1, характерная ширина [2] $\Delta\lambda_T$ спектра излучения вольфрамовой нити достаточно велика, так что $\Delta\lambda_T \gtrsim \lambda_{\max}$. Согласно

[1] То есть для нахождения энергии единицы объёма излучения в интервале длин волн $(\lambda_1, \lambda_2)$ необходимо вычислить интеграл $\int\limits_{\lambda_1}^{\lambda_2} u_\lambda d\lambda$.

[2] Ширину спектра $\Delta\lambda$ обычно определяют как диапазон длин волн, в котором плотность энергии отличается от максимального значения не более чем вдвое.

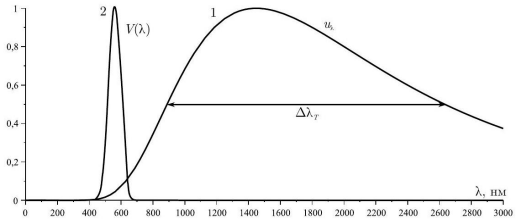



Рис. 1. Нормированная спектральная плотность равновесного теплового излучения для температуры $T = 2000$ K(1) и график относительной спектральной чувствительности зрительной системы человека (2)

формулам (2.54) мы не сможем получить даже одной интерференционной полосы от такого источника.

Однако если уменьшить ширину спектрального диапазона $\Delta\lambda$, в котором ведется наблюдение (например, пропустить тепловое излучение через светофильтр или регистрировать его приемником с узкой спектральной чувствительностью), то длина когерентности увеличится. В наших опытах таким приемником является глаз экспериментатора. На рис. 1 приведена зависимость от длины волны относительной спектральной чувствительности зрительной системы человека $V(\lambda)$. Эта чувствительность характеризует дневное зрение и является стандартом, принятым в 1924 г. Международной комиссией по освещению после многочисленных специальных исследований. Максимальная чувствительность, на которую нормирована кривая, приходится на длину волны 555 нм. Приведенная кривая чувствительности характеризует «в среднем» людей с нормальным зрением, индивидуальная же кривая для конкретного экспериментатора может быть иной. Тем не менее можно принять, что чувствительность зрительной системы экспериментатора превышает уровень 0,5 в диапазоне длин волн 510-610 нм и использовать этот диапазон для оценок того, что можно уверенно наблюдать глазом. Максимальный порядок интерференции и число наблюдаемых интерференционных полос для этого диапазона согласно (2.54):

$$m = \frac{\lambda}{\Delta\lambda} \approx \frac{555}{610 - 510} \approx 5 \div 6$$

длина и время когерентности

$$\ell_{\text{ког}} = \lambda\frac{\lambda}{\Delta\lambda} \approx 3, \quad \tau_{\text{ког}} = \frac{\ell_{\text{ког}}}{c} \approx 10^{-14}\text{с}$$

Из сопоставления кривых 1 и 2 на рис. 1 видно, что только несколько процентов мощности излучения электрической лампочки воспринимается зрительной системой человека.

Описание интерферометра. Для решения задачи измерения радиуса когерентности электромагнитного поля в некоторой плоскости можно использовать набор экранов с разными расстояниями между двумя отверстиями. Мы считаем, что контраст интерференционной картины, возникающей в дальней зоне при дифракции света на двух отверстиях, не зависит от координат этих отверстий, а определяется только расстоянием между ними (условие однородности статистических характеристик поля). Меняя экраны в выбранной плоскости так, чтобы расстояние между отверстиями постепенно увеличивалось, следует найти минимальное расстояние $\rho_0$ (2.62), при котором интерференционная картина пропадает. В лабораторной работе используется интерферометр, специально сконструированный для визуального наблюдения распределения интенсивности по времени задержки (2.48) и расстоянию между отверстиями.

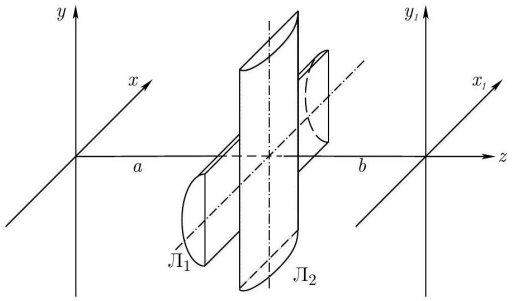



Рис. 2. Принципиальная схема интерферометра

Принципиальная оптическая схема интерферометра приведена на рис. 2. Экран с отверстиями располагается в плоскости $(x, y), (x_1, y_1)$ - плоскость наблюдения, $z$ - оптическая ось системы. Изображение формируется объективом, состоящим из двух тонких вплотную приложен ных друг к другу цилиндрических линз $\Omega_1$ (ось цилиндра горизонтальна), $_2$ (ось цилиндра вертикальна). Пусть в экране имеется одно прямоугольное отверстие (щель с резкими границами), фокусные расстояния линз одинаковы $f_1 = f_2 = f$ и вдвое превосходят расстояния от экрана до объектива a и от объектива до плоскости наблюдения $b$, т. е. $a = b = 2f$. В этом случае при малых углах дифракции света на отверстии две цилиндрические линзы эквивалентны одной сферической с фокусным расстоянием $f$. В плоскости наблюдения возникает изображение щели с увеличением $b/a = 1$. Линза $_1$ (ось цилиндра горизонтальна) обеспечивает резкие горизонтальные границы изображения, линза $\Omega_2$ (ось цилиндра вертикальна) - резкие вертикальные границы (рис. 3а). Заменим линзу $\Omega_2$ на линзу с большим фокусным расстоянием $f_2 = b$. В этом случае плоскость наблюдения является фокальной для этой линзы. Если на экран падает плоская монохроматическая световая волна, то в плоскости наблюдении возникает распределение интенсивности света, соответствующее дифракции Фраунгофера на щели (рис. 3б). Это распределение развернуто вдоль горизонтальной оси, ширина главного максимума обратно пропорциональна ширине щели, горизонтальные границы остаются резкими и на прежнем месте (действие линзы $\Omega_1$). Рассмотрим еще один пример изображения. Пусть в плоскости $(x, y)$ помещен экран с двумя парами одинаковых квадратных отверстий $A_1, B_1$ и $A_2, B_2$. Центры отверстий лежат на биссектрисах координатных углов в нижней полуплоскости, их координаты $(\pm\frac{h}{2}, -\frac{h}{2})$ и $(\pm h, -h)$ (рис. 4а). Соответствующее изображение в плоскостях $(x, y)$ приведено на рис. 4.



а)



б)

Рис. 3. Изображения прямоугольной щели: а) $a = b = 2f_1 = 2f_2$, б) $a = b = 2f_1 = f_2$

Линза $\Omega_1$ формирует две яркие полосы (изображение перевернутое). Внутри каждой полосы линза $\Omega_2$ формирует картину дифракции Фраунгофера на двух отверстиях $A_1, B_1$

и $A_2, B_2$. Вертикальные координаты полос $\frac{h}{2}b/a$ (расстояние между отверстиями $\rho = h$) и $hb/a$ (расстояние между отверстиями $\rho = 2h$). Ширина главного дифракционного максимума вдоль горизонтальной оси одинакова для обеих

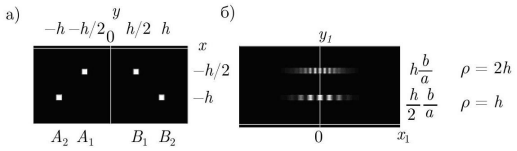



Рис. 4. Экран с двумя парами отверстий (а), изображение в плоскости наблюдения (б)

полос и определяется размером отверстий, фокусным расстоянием линзы $\Omega_2$ и длиной волны света. Число полос в главном дифракционном максимуме в нашем примере ($\rho = h$ и $\rho = 2h$) отличается в два раза, так как оно прямо пропорционально расстоянию между отверстиями $\rho$ вдоль горизонтальной оси $x$.

Ясно, что плавного увеличения расстояния между отверстиями $\rho$ можно добиться, проделав большое количество пар отверстий $A, B$. Для этой цели во входной плоскости интерферометра размещают экран с двумя взаимно перпендикулярными щелями (рис. 5). Возникающее для креста щелей изображение приведено на рис. 6. Можно сказать, что горизонтальные яркие полоски, показанные на рис. 4б, тесно примыкают друг к другу, расстояние между интерференционными полосами уменьшается с увеличением вертикальной координаты $y_1$.

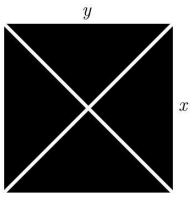



Рис. 5. Экран интерферометра

В лабораторной установке объектив интерферометра собран не из двух цилиндрических линз, а из одной сферической (фокусное расстояние $F$) и одной цилиндрической (фокусное расстояние $f$, ось цилиндра горизонтальна). Для тонких линз, вплотную приложенных друг к другу, настройка объектива должна удовлетворять условиям

$$F = b, \quad \frac{1}{a} + \frac{1}{b} = \frac{1}{F} + \frac{1}{f}$$

Объектив с двумя фокусными расстояниями позволяет получить изображение дифракционной картины в координатах угол дифрак-

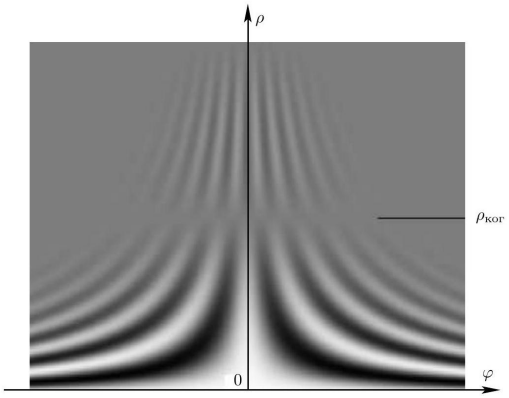


Рис. 6. Изображение в плоскости $(x_1, y_1)$. Переменные: угол дифракции (время задержки), расстояние между отверстиями в экране

ции $\varphi$ (развертка изображения вдоль горизонтальной оси $x_1$) и расстояние между малыми отверстиями $\rho$ (развертка вдоль вертикальной оси $y_1$). Формулы пересчета координат имеют вид

$$\varphi = \frac{x_1}{b}, \quad \rho = 2\frac{a}{b}\,|y_1| \tag{1}$$

Для длины волны $\lambda_0$ угловое расстояние между интерференционными полосами $\Delta\varphi = \lambda/\rho$, линейное расстояние

$$\Delta x_1 = b\Delta\varphi_1 = \frac{\lambda_0 b^2}{2a\,|y_1|}$$

т.е. в переменных $y_1, x_1$ интерференционные полосы представляют собой гиперболы. Число наблюдаемых цветных полос для теплового источника определяется относительной чувствительностью зрительной системы наблюдателя. В наших опытах это число равно $5 \div 6$. Если на входную плоскость интерферометра (экран, рис. 5) падает параллельный пучок света, то распределение интенсивности излучения в плоско сти наблюдения вдоль горизонтальной прямой описывается выражением (2.48), так как кривая чувствительности (рис. 1, кривая 2) является практически гауссовой. Время задержки, используемое в (2.48), равно $\tau = \rho\varphi/c$.

В опытах между крестом и электрической лампочкой размещается вертикальная оптическая щель. Если считать щель квазимонохроматическим источником случайного поля, то для расчета видности интерференционной картины можно использовать распределение интенсивности (2.64), где радиус когерентности $\rho_0$ определяется угловым размером щели (2.62). Если глаз не различает интерференционные полосы на некоторой высоте изображения $y_1$, можно, используя второе равенство (1), рассчитать радиус когерентности поля во входной плоскости интерферометра.

Используя (2.48), (2.62), (2.64) и приведенные выше формулы пересчета масштабов, можно выписать относительное отклонение интенсивности света наблюдаемого изображения $I$ от фона $I_0$:

$$\frac{I - I_0}{I_0} = \exp\left[-\left(\frac{2\pi\Delta\lambda a x_1 y_1}{\lambda_0^2 b^2}\right)^2\right] \frac{\sin\left(2\pi\frac{a}{b}\frac{y_1 d}{\lambda_0 L}\right)}{2\pi\frac{a}{b}\frac{y_1 d}{\lambda_0 L}} \cos\left(4\pi\frac{a x_1 y_1}{b^2\lambda_0}\right)$$

где длина волны $\lambda_0 \approx 560$ нм соответствует максимуму, а $\Delta\lambda \approx 120$ нм - ширине по уровню $1/$ е чувствительности зрительной системы человека, $b/a$ - увеличение объектива интерферометра, $d$ - ширина вертикальной щели, $L$ - расстояние от этой щели до экрана с крестом щелей. Первый сомножитель определяется кривой относительной чувствительности зрительной системы человека или спектром регистрируемого излучения. Второй сомножитель связан с пространственной когерентностью случайного поля, испытавшего дифракцию на щели. Третий сомножитель описывает интерференционные полосы-гиперболы.

## ЗАДАНИЕ

В работе предлагается собрать и отъюстировать интерферометр, оценить длину волны излучения полупроводникового лазера, определить зависимость радиуса когерентности поля световой волны от размера источника, оценить ширину относительной чувствительности зрительной системы экспериментатора.

Элементы интерферометра располагаются на металлической плите оптического стола и крепятся к плите с помощью прижимных пластин и болтов. При фиксации на столе положения щелей, линз, микроскопа и т. д. возникает их неконтролируемое смещение и наклон относительно

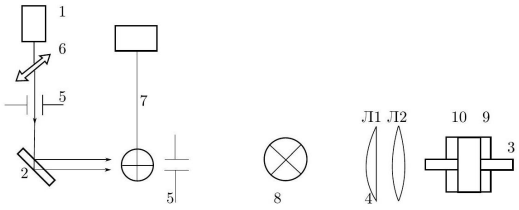



Рис. 7. Элементы лабораторной установки

первоначального выбранного положения. Крепление элементов интерферометра следует проводить в несколько этапов, каждый раз выправляя смещение элементов. Сборка интерферометра - трудоемкий процесс, требующий внимания, аккуратности и взаимопонимания. Один студент устанавливает и крепит элементы установки, другой контролирует качество сборки по возникающим изображениям и управляет этим процессом. Расположение элементов интерферометра для различных опытов приведено на рис. 7.

**I. Измерения**

Прямое неослабленное лазерное излучение опасно для зрения! Положение лазерных лучей контролируется по экрану.

1. Установка оптической оси системы. На столе устанавливается полупроводниковый лазер 1, стеклянная пластина 2, микроскоп 3. Стеклянная пластина ослабляет излучение лазера примерно в 20 раз и формирует два пучка света для наблюдения интерференционной картины. Микроскоп закреплен на двухкоординатном столике 9. Установите верхнюю пластину столика без смещения относительно нижней. Микроскоп должен быть полностью вставлен в его оправу 10. Получите резкое изображение окулярной шкалы микроскопа, саму шкалу установите в горизонтальное положение, вращая микроскоп вокруг его оптической оси. Пластина и микроскоп устанавливаются вблизи углов оптического стола. Включите лазер. Ось системы устанавливается по лазерным лучам.

Высота оси над поверхностью стола 20-25 см, положение оси по вертикали контролируется линейкой. Грубая установка оси осуществляется перемещением и поворотом рейтера пластины, тонкая - юстировочными винтами ее оправы.

Лучи лазера должны попадать в объектив микроскопа, а за окуляром на экране должны наблюдаться два ярких световых пятна. Такого положения можно добиться перемещением рейтера микроскопа в горизонтальном направлении, поднимая или опуская двухкоординатный столик и поворачивая его вокруг вертикальной оси. Закрепите на оптическом столе лазер, пластину, микроскоп.

2. Установка объектива интерферометра. Объектив 4 составлен из цилиндрической линзы $_1$ (ось цилиндра горизонтальна) и сферической линзы $\Omega_2$. Линзы объекива не идеальные, а толстые и имеют значительную хроматическую аберрацию. Объектив устанавливается сферической линзой к микроскопу. Если верхняя пластина координатного столика микроскопа не смещена относительно нижней, а сам микроскоп полностью вставлен в его оправу, то расстояние между осями

стержней, которыми крепятся столик и объектив, должно составлять $470 \pm 3$ мм. Лучи лазера должны проходить вблизи центра цилиндрической линзы перпендикулярно ее поверхности. За микроскопом на экране должна наблюдаться яркая вертикальная полоса.

3. Оценка длины волны излучения лазера. Перекройте лазерный пучок поляроидом 6. Между поляроидом и стеклянной пластинкой установите вертикально щель 5 пластинами к лазеру. Щель должна пропускать центральную часть лазерного пучка. Установите максимальный размер щели, а поляроидом 6 максимально ослабьте излучение лазера. Теперь возникающую дифракционную картину можно наблюдать через микроскоп глазом. Изображение имеет вид огибающей, показанной на рис. 3б, внутри которой расположены интерференционные полосы. Подберите удобные для наблюдения ширину щели и ослабление поляроидом лазерного излучения. Измерьте по окулярной шкале микроскопа расстояние между интерференционными полосами. Расстояние между лазерными пучками за пластиной 2 измерьте линейкой.

4. Определение предметной плоскости обдектива. Установите окулярную шкалу микроскопа вертикально. Положение плоскости на оси системы, в которую следует поместить щели в виде креста 8, определяется по изображению другого транспаранта - оптической щели 5. Если оптическая щель строго горизонтальна и расположена в предметной плоскости объектива, то в микроскопе наблюдается резкое изображение ее горизонтальных границ. Установите ширину оптической щели 1 мм, ориентация щели должна быть горизонтальной, ее пластинки повернуты к
лазерным лучам, расстояние от щели до объектива по осям стержней крепления оправ около 45 см. Горизонтальное и вертикальное положения щели установите по лазеру. Перед щелью установите электрическую лампочку 7, включите ее питание. Перемещая щель вдоль оптической оси, определите ее положение, при котором изображение щели наиболее резкое. Уточните поворот щели относительно оптической оси интерферометра (ее горизонтальное положение). Еще раз найдите оптимальное расположение щели. Эту процедуру можно повторить несколько раз. Объектив с цилиндрической линзой лишен аксиальной симметрии и весьма чувствителен к повороту щели. При правильной настройке горизонтальные границы щели должны быть резкими без изменения их цвета. Измерьте размер изображения щели и определите увеличение объектива $b/a$. Измерьте расстояние между осями стержней оправ щели и объектива. По этому расстоянию вместо оптической щели установите оправу с крестом щелей. Ширина щелей около 100 мкм. Угол между щелями и вертикалью 45°. Сместите электрическую лампу и по лазерным лучам совместите центр щелей с оптической осью объектива. Проверьте расстояние между оправами щелей и объектива, закрепите рейтер щелей.

5. Измерение радиуса когерентности и максимального порядка интерфереииии. Установите перед экраном с крестом щелей 8 оптическую щель 5 (рис. 7) на расстоянии около 100 см от экрана. Оптическая щель ориентирована вертикально, пластинами к лазеру. По лазерным лучам совместите центр щели с оптической осью интерферометра. Те перь можно выключить полупроводниковый лазер. Полностью откройте оптическую щель. Перед щелью установите электрическую лампочку и включите ее на полную мощность. Перемещением лампы добейтесь симметричного освещения креста щелей интерферометра (на экране со щелями хорошо видна яркая вертикальная полоса света). Наблюдайте возникающее

изображение интерференционной картины и одновременно плавно закрывайте оптическую щель. Если свет перестанет попадать на крест щелей при уменьшении ширины оптической щели, то немного сместите лампу. В результате этого этапа настройки в микроскопе должна наблюдаться картина, похожая на приведенную на рис. 6. Закрепите на столе лампу.

Далее следует процедура тонкой настройки интерферометра. Качество возникающего изображения интерференционной картины заметно зависит от отклонения от вертикали оптической щели (ориентации источника света). Установите ширину щели около 0,3 мм. Очень плавно и медленно поворачивая щель вокруг оптической оси интерферометра (при ослабленных крепежных винтах ее оправы), зафиксируйте ее по ложение, при котором интерференционные полосы пропадают, затем с малым контрастом возникают при увеличении вертикальной координаты изображения (рис. 6). Вертикальная ориентация полосы нулевого порядка интерференции достигается небольшим поворотом креста щелей вокруг оптической оси интерферометра. Симметричность интерференционной картины относительно вертикальной оси - малым поворотом объектива относительно вертикальной оси (в объективе имеется небольшой перекос типа клина). Зарисуйте схему интерферометра, на схеме укажите расстояния между его элементами. Измерьте расстояние $L$ от оптической щели 5 до креста щелей.

6. Определите максимальный порядок интерференции, который различает ваш глаз. Для $4-5$ значений ширины $d$ оптической щели 5 измерьте радиус когерентности поля световой волны во входной плоскости интерферометра. Для этого по шкале микроскопа измеряется вертикальная координата исчезновения полос и проводится расчет по формуле (1).

Замечание. Возможно, что плохое качество интерференционной картины связано с нарушением центрировки интерферометра, возникшей при установке элементов. Следует окончательно проверить прохождение лазерного луча по центру системы. Рекомендуется начинать проверку с микроскопа, затем поочередно уточнять положение остальных элементов, закончив измерительной щелью. Как правило, этого достаточно. Горизонтальные темные неубираемые линии в поле зрения микроскопа вызваны перекрытием щелей креста пылью, которая убирается тонким листом бумаги.

## II. Обработка результатов

1. Оцените длину волны излучения полупроводникового лазера. Фокусное расстояние объектива можно принять равным $F \approx 33$ см.

2. Оцените спектральную ширину относительной чувствительности вашей зрительной системы. Длину волны для максимальной чувствительности $\lambda_{\max}$ примите равной стандартной $\lambda_{\max} \approx 560$ нм. Сравните полученный результат со сделанной выше оценкой $\Delta\lambda \approx 100 \div 120$ нм. Оцените эквивалентную длину когерентности, время когерентности, отношение времени когерентности к периоду колебаний электромагнитного поля с длиной волны $\lambda_{\max}$.

3. Для проверки теоретической зависимости $\rho_0 = \frac{\lambda}{d}L$, связывающей между собой радиус когерентности поля с угловым размером щели $d/L$, постройте соответствующий график по полученным экспериментальным данным. Переменные следует выбрать так, чтобы график имел вид прямой. По графику оцените длину волны максимальной чувствительно сти зрительной системы, сравните результат со стандартным значением $\lambda_{\max} \approx 560$НМ.

**Контрольные вопросы**

1. Как связаны между собой функции когерентности и контрастности изображений?

2. Как измерить длину когерентности, радиус когерентности?

3. На рис. 6 при переходе через $\rho_{\text{ког}}$ максимумы интерференционных полос меняются на минимумы. Почему?

4. Каков КПД электрической лампы накаливания, используемой для освещения?

5. Как зависит радиус когерентности от ширины оптической щели и расстояния от щели до входной плоскости интерферометра?

6. Как изменится число наблюдаемых интерференционных полос, если между глазом и микроскопом установить оптический фильтр?

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§28, 30 – 32.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 17, 21, 22.

## 22.2.6 Diffraction of Coherent Radiation on a Rough Surface

Цель работы: получение и исследование изображений стохастических дифракционных полей (спеклов), приобретение навыков обработки двумерных цифровых сигналов.

В работе используются: гелий-неоновый лазер, матовая пластинка, поляроид, цифровая камера, компьютер.

Дифракционные явления в учебных курсах оптики рассматривают, как правило, на примере регулярных структур или одинаковых элементов - отверстий, решёток и т. п. Однако дифракционные картины возникают и при когерентном освещении случайно-неоднородных объектов, таких как шероховатая поверхность. В этом случае интерференционная картина образуется при сложении взаимно когерентных волн, имеющих случайные сдвиги фаз. На такой картине обычно наблюдаются светлые пятна (их называют спеклами от англ. speckle - крапинка, пятнышко), разделенные темными участками. В некоторых прикладных задачах, например в лазерной локации, исследуются спеклкартины, возникающие при отражении света от шероховатой поверхности. Статистический анализ случайной картины дифракции позволяет получать информацию о статистических свойствах объектов и излучения.

Принципиальная схема установки для получения спеклов приведена на рис. 1.

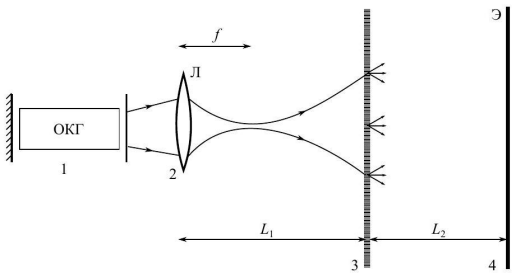


Рис. 1. Схема получения спекл-картины: 1 - гелий-неоновый лазер, 2 линза, 3 - рассеиватель, 4 - экран или фотокамера

Источником излучения служит He − Ne-лазер с длиной волны излучения 0,63 мкм. Пучок излучения, формируемый линзой, рассеивается на матовой стеклянной пластинке, на экране наблюдается спекл-картина рассеяния. Характерный вид картины приведен на рис. 5 и 6. В простых экспериментах можно исследовать зависимость спекл-картины от поперечного размера лазерного луча на рассеивателе: чем «острее» фокусировка, тем более крупные спеклы возникают на экране, и, наоборот, увеличение диаметра луча приводит к уменьшению размеров спеклов. При увеличении расстояния $L_2$ между рассеивателем и экраном спеклы увеличиваются в размере.

Перемещение рассеивателя поперек лазерного луча приводит к движению спеклов на экране. Направление движения спеклов может совпадать с направлением перемещения матовой пластинки или быть противоположным в зависимости от положения плоскости рассеивателя относительно точки фокусировки луча (т.е. в зависимости от соотношения между расстояниями $f$ и $L_1$ на рис. 1). При фокусировке ла зерного луча на рассеивателе характер движения спеклов оказывается совершенно своеобразным: спеклы хаотически возникают и исчезают, наподобие пузырей в кипящей жидкости, без какого-либо заметного смещения, этот эффект называется «кипением» спеклов.

Картина спеклов возникает в результате интерференции когерентных волн, рассеянных отдельными неровностями поверхности. Фаза колебаний в волне, рассеянной отдельной неровностью, зависит как от её положения на освещенной площадке (от её координат $x, y$), так и от «высоты» $h$ данной неровности относительно «подстилающей» плоскости. Форма рассеивающей поверхности $h(x, y)$ является случайной функцией координат $x$ и $y$, поэтому фаза колебаний световой волны в данной точке экрана является случайной функцией координат. В некоторые точки экрана волны, рассеянные разными неровностями матовой поверхности, приходят с «благоприятными» фазовыми соотношениями и образуют интерференционный максимум - яркое пятнышко-спекл. В других точках волны, интерферируя, гасят друг друга и образуют темные области между спеклами. Если матовая пластинка неподвижна, то фазовые соотношения между рассеянными волнами, попадающими в различные точки экрана, сохраняются во времени в силу высокой когерентности лазерного излучения, и картина спеклов стационарна.

Статистические характеристики рассеивателя и спекл-картины. На рис. 2 показан рельеф поверхности рассеивателя - матовой стеклянной пластинки (нем. matt - тусклый, блеклый), одна сторона которой подвергнута травлению плавиковой кислотой. Светлые и темные пятна на рис. 2 показывают отклонения шероховатой поверхности от некоторой средней плоскости в ту или другую сторону. На рис. 3 приведен график зависимости отклонения толщины пластинки $h$ от средней толщины вдоль некоторой прямой, принятой за ось $x$. Чтобы подчеркнуть неровность шероховатой поверхности, масштаб отклонения $h$ увеличен примерно в десять раз по сравнению с масштабом по координате $x$.

Из качественного описания ясно, что вид спекл-картины определяется возмущениями фронта световой волны, вносимыми шероховатой рассеивающей поверхностью. Возмущения волнового фронта создаются случайной функцией $h(x, y)$, поэтому для анализа этих возмущений необходимо использовать статистические методы.

Функция корреляции. Как известно из курса теории вероятностей, для описания случайных величин используют понятия среднего значения и дисперсии. В нашем случае случайной величиной является толщина пластинки, ее среднее значение характеризует толщину пластин-

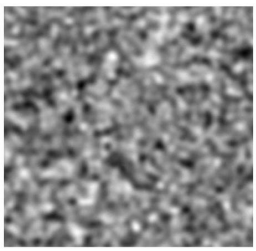


Рис. 2. Рельеф матовой поверхности. Светлые и темные области изображения отвечают выпуклостям и впадинам соответственно

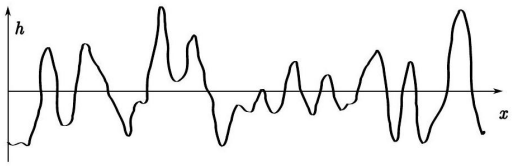



Рис. 3. Профиль поверхности рассеивателя

ки «в целом» и соответствует прямой $\langle h \rangle = 0$ на рис. 3. Дисперсия $D_h = \langle h^2 \rangle - \langle h \rangle^2$, а также связанная с ней величина, называемая среднеквадратичным отклонением $\sigma_h = \sqrt{D_h}$, характеризует отклонения случайной величины от среднего значения. Величина $\sigma_h$ определяет масштаб высоты зубцов и глубины провалов на кривой, приведенной на рис. 3. Величина среднеквадратичного отклонения также описывает матовую пластинку «в целом». Однако для анализа картины спеклов нам понадобится еще информация о том, насколько быстро изменяется случайная функция $h(x, y)$ при переходе от одной точки поверхности к другой. Для характеристики этого свойства в теории вероятностей вводят функцию, которая называется корреляционной. В рассматриваемом случае матовой пластинки случайной функцией является высота неровности $h$ в точке поверхности, задаваемой радиус-вектором $\boldsymbol{r}_i$. Согласно определению, корреляционной функцией $\Psi$ является функция

$$\Psi_h(\boldsymbol{\rho}) = \langle h(\boldsymbol{r}) h(\boldsymbol{r} + \boldsymbol{\rho}) \rangle - \langle h(\boldsymbol{r}) \rangle^2 \tag{1}$$

где $\boldsymbol{\rho}$ - смещение относительно точки $\boldsymbol{r}$, скобки $\langle .. \rangle$ - нение по всем точкам поверхности. Функция показывает, насколько в среднем изменяется величина $h$ при смещении из данной точки на расстояние $\boldsymbol{\rho}$, другими словами, насколько значения $h(\boldsymbol{r})$ и $h(\boldsymbol{r} + \boldsymbol{\rho})$ скоррелированы между собой. Для однородной пластинки функция $\Psi$ не зависит от выбора точки $\boldsymbol{r}$, а определяется только величиной смещения $\rho$, т.е. $\Psi_h = \Psi_h(\rho)$. Общие соображения позволяют представить качественно вид функции $\Psi_h(\rho)$. При $\rho = 0$ значения $h(\boldsymbol{r})$ и $h(\boldsymbol{r} + \boldsymbol{\rho})$ одинаковы, т. е. их корреляция наибольшая, и функция $\Psi_h(\rho)$ достига-

ет максимума (в этом случае $\Psi_h(\rho) = \langle h^2(\boldsymbol{r}) \rangle - \langle h(\boldsymbol{r}) \rangle^2 \equiv \sigma_h^2$, согласно определению дисперсии). При больших значениях $\rho$ в силу случайного характера функции $h$ значения $h(\boldsymbol{r})$ и $h(\boldsymbol{r}+\boldsymbol{\rho})$ оказываются независимыми и функция $\Psi(\rho) \to 0$. Таким образом, функция $\Psi_h(\rho)$ должна достигать максимального значения при $\rho = 0$ и уменьшаться с ростом $\rho$.

Обычно хорошим приближением для корреляционной функции $\Psi_h(\rho)$ является функция Гаусса:

$$\Psi_h(\rho) = \sigma_h^2 e^{-\frac{\rho^2}{\rho_h^2}} \tag{2}$$

где $\rho_h$ - характерное расстояние, называемое радиусом корреляции, на котором функция корреляции уменьшается в $e$ раз. Радиус корреляции $\rho_h$ - это масштаб, характеризующий своего рода «пространственную память» случайной величины $h(\boldsymbol{r})$.

По своему физическому смыслу $\rho_h$ можно трактовать как средний размер неоднородностей матовой пластинки. Для справки отметим, что в исследуемых образцах матовых стекол $\rho_h \approx 5 \div 10$ мкм.

Световая волна, проходя сквозь матовую поверхность, меняет амплитуду и фазу своих колебаний. Однако удовлетворительным приближением является рассмотрение матовой поверхности как чисто фазового транспаранта, т.е. случайной фазовой решетки. Случайная составляющая $\varphi$ изменения фазы прошедшей световой волны определяется оптической разностью хода, пропорциональной случайным вариациям толщины пластины $h$:

$$\varphi = \frac{2\pi}{\lambda}(n-1)h \tag{3}$$

где $n \approx 1,5$ - показатель преломления стекла. Фронт световой волны, приобретя набег фазы (3) после прохождения матовой пластинки,
становится случайным. Наблюдаемые спеклы представляют собой визуализацию этого случайного поля, и нам необходимо установить связь статистических свойств этого поля со статистическими свойствами рассеивающей пластинки. Отметим свойство, следующее из соотношения (3): высота $h$ отсчитывается от среднего уровня $\langle h \rangle = 0$, поэтому и $\langle \varphi \rangle = 0$. Кроме того, функции корреляции фазы $\Psi_\varphi$ и высоты $\Psi_h(2)$ совпадают с точностью до постоянного множителя:

$$\Psi_\varphi(\rho) = \langle \varphi(\boldsymbol{r}+\boldsymbol{\rho})\varphi(\boldsymbol{r})\rangle = \sigma_\varphi^2 e^{-\frac{\rho^2}{\rho_h^2}}$$

Для небольших смещений $\rho$

$$\Psi_\varphi(\rho) \approx \sigma_\varphi^2\left(1-\rho^2/\rho_h^2\right), \quad \sigma_\varphi^2 = \left(\frac{2\pi}{\lambda}\right)^2 (n-1)^2\sigma_h^2 \tag{4}$$

Рассмотрим прохождение плоской монохроматической волны, фронт которой параллелен поверхности $\langle h \rangle = 0$, через плоскую шероховатую пластинку. Непосредственно за фазовым транспарантом комплексная амплитуда поля равна $E(\boldsymbol{r}) = E_0 e^{i\varphi(\boldsymbol{r})}$, где $E_0$ - амплитуда падающей волны. Напомним, что областью определения фазы $\varphi$, входящей в комплексную амплитуду $e^{i\varphi}$, является отрезок $\{-\pi, \pi\}$. В матовых стеклах характерный размер неровностей превышает длину световой волны ($\sigma_h > \lambda$), из этого на основании соотношения (3) заключаем, что случайное изменение фазы колебаний выходит за диапазон $\{-\pi, \pi\}$ (т.е. $\sigma_\varphi > \pi$). На этом основании в дальнейших расчетах комплексных амплитуд мы будем считать фазу $\varphi$ случайной величиной, равномерно распределенной на отрезке $\{-\pi, \pi\}$ и имеющей нулевое среднее значение. В этом случае равны нулю и средние значения синуса и косинуса фазы:

$$\left\langle e^{i\varphi(\boldsymbol{r})}\right\rangle = \langle \cos\varphi(\boldsymbol{r})\rangle = \langle \sin\varphi(\boldsymbol{r})\rangle = 0$$

Эксперименты подтверждают умозаключение о том, что $\left\langle e^{i\varphi(r)}\right\rangle = 0$. Следовательно, $\langle E(\boldsymbol{r})\rangle = 0$. Если бы это предположение было неверно и $\langle e^{i\varphi}\rangle \neq 0$ (малая модуляция фазы), то на экране наблюдалось бы сечение лазерного пучка, но в наших опытах этого эффекта нет.

Получим функцию корреляции фазовых множителей $e^{i\varphi}$, необходимую для расчета картины спеклов. Для комплексных случайных величин $\xi(r)$ с нулевым средним значением $\langle \xi(r)\rangle = 0$ определение функции корреляции (1) преобразуется к виду

$$\Psi_\xi(\rho) = \langle \xi(\boldsymbol{r}+\boldsymbol{\rho})\xi^*(\boldsymbol{r})\rangle \tag{5}$$

В нашем случае $\xi(r) = e^{i\varphi(r)}$, поэтому с использованием (5) получаем

$$\Psi_\varphi(\rho) = \left\langle e^{i[\varphi(\boldsymbol{r}+\boldsymbol{\rho})-\varphi(\boldsymbol{r})]} \right\rangle$$

В силу того что случайные величины $h(r)$ и $\varphi(r)$ связаны между собой соотношением (3), корреляционная функция $\Psi_\varphi(\rho)$ должна иметь гауссову форму со своим радиусом корреляции $\rho_\varphi$:

$$\left\langle e^{i[\varphi(\boldsymbol{r}+\boldsymbol{\rho})-\varphi(\boldsymbol{r})]} \right\rangle = e^{-\frac{\rho^2}{\rho_\varphi^2}}. \tag{6}$$

Оценим радиус корреляции $\rho_\varphi$. Полагая смещение $\rho$ малым, разложим в ряд левую и правую части равенства (6) до членов $\rho^2$ включительно. Используя (4), получаем

$$\begin{aligned}
&\left\langle e^{i[\varphi(\boldsymbol{r}+\boldsymbol{\rho})-\varphi(\boldsymbol{r})]} \right\rangle \\
&\approx 1 + i[\langle\varphi(\boldsymbol{r}+\boldsymbol{\rho})\rangle - \langle\varphi(\boldsymbol{r})\rangle] - \frac{1}{2}\left[\left\langle\varphi^2(\boldsymbol{r}+\boldsymbol{\rho}) + \varphi^2(\boldsymbol{r})\right\rangle - 2\sigma_\varphi^2\left(1 - \frac{\rho^2}{\rho_h^2}\right)\right] = \\
&= 1 - \frac{1}{2}\left[2\sigma_\varphi^2 - 2\sigma_\varphi^2\left(1 - \frac{\rho^2}{\rho_h^2}\right)\right] = 1 - \sigma_\varphi^2\frac{\rho^2}{\rho_h^2} = 1 - \frac{\rho^2}{\rho_\varphi^2}
\end{aligned}$$

Искомый радиус корреляции

$$\rho_\varphi = \frac{\rho_h}{\sigma_\varphi} = \frac{\rho_h}{\sigma_h}\frac{\lambda}{2\pi(n-1)} \tag{7}$$

Из соотношения (7) видно, что радиус корреляции для фазового множителя меньше радиуса корреляции для неровностей. Как упоминалось выше, для матового стекла $(\lambda/\sigma_h) < 1, n \approx 1,5$ и оценка по формуле (7) дает $\rho_\varphi < \rho_h/\pi$.

Для характеристики матовой поверхности используют так называемый параметр шероховатости $\zeta$, равный отношению средней высоты неровности к ее усредненному поперечному размеру. Используя статистические характеристики, можно записать:

$$\zeta = \frac{\sigma_h}{\rho_h} \tag{8}$$

Применяемая в работе матовая пластинка имеет параметр шероховатости $\zeta \sim 0,1$, длина волны гелий-неонового лазера $\lambda = 0,63$ мкм, оценка по формуле (7) дает $\rho_\varphi \approx \frac{10}{\pi}\lambda \approx 3\lambda \approx 2$ мкм.

Дифракция излучения лазера на шероховатой поверхности. Гелий-неоновый лазер генерирует слабо расходящийся пучок света на длине волны $\lambda = 0,63$ мкм. Можно принять, что на выходе из лазера пучок имеет плоский волновой фронт, а распределение интенсивности света в пучке описывается гауссовой функцией:

$$I_L(\boldsymbol{r}) = |E_L(\boldsymbol{r})|^2 = I_0 e^{-\frac{r^2}{\rho_L^2}} \tag{9}$$

где $\rho_L$ - поперечное расстояние от оси лазерного луча, на котором интенсивность уменьшается в  раз по сравнению с максимальной. На выходе лазера $\rho_L \approx 100$ мкм. Как известно, при распространении гауссова пучка его поперечный размер $\rho_L$ увеличивается, но форма (9) сохраняется. Линза, показанная на рис. 1, изменяет размер пучка, но не меняет его гауссовой формы. Матовая пластинка, стоящая на пути лазерного луча, вносит поперечную неоднородность в световое поле и значительно увеличивает угловую расходимость пучка. Когерентное лазерное излучение преобразуется матовой поверхностью в поперечно-некогерентное.

Изучая картину спеклов, мы наблюдаем на экране распределение интенсивности света, рассеянного матовой пластинкой. Наша задача состоит в том, чтобы на основании

анализа статистических свойств распределения интенсивности получить информацию о статистических свойствах рассеивателя.

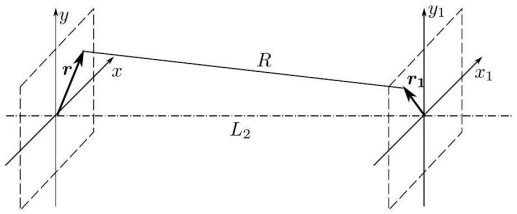



Рис. 4. К расчету разности хода: $\boldsymbol{r}$ - радиус-вектор точки в плоскости рассеивателя, $\boldsymbol{r}_1$ - в плоскости экрана или фотокамеры, $L_2$ - расстояние вдоль оси $z$ между матовой пластинкой и экраном

Пусть на матовую поверхность нормально падает излучение лазера с комплексной амплитудой $E_L(\boldsymbol{r})$. Непосредственно за этой поверхностью комплексная амплитуда равна $E_L(\boldsymbol{r})e^{i\varphi(\boldsymbol{r})}$ (изменением амплитуды поля мы пренебрегаем). Площадки $d\boldsymbol{r} = dxdy$ являются источниками вторичных сферических волн, которые распространяются к плоскости наблюдения и суммируются в интересующей нас точке с координатой $\boldsymbol{r}_1$ (рис. 4). Расстояние $L_2$ между матовым стеклом и плоскостью наблюдения $L_2 \gg r_1, r_2$, поэтому воспользуемся приближением Френеля:

$$R = \sqrt{L_2^2 + |\boldsymbol{r} - \boldsymbol{r}_1|^2} \approx L_2 + \frac{1}{2}\frac{|\boldsymbol{r} - \boldsymbol{r}_1|^2}{L_2} = L_2 + \frac{r^2 + r_1^2}{2L_2} - \frac{(\boldsymbol{r}\cdot\boldsymbol{r}_1)}{L_2}$$

С физической точки зрения использование приближения Френеля означает замену сферических волн параболическими. В этом приближении комплексная амплитуда поля $E(\boldsymbol{r}_1)$ в плоскости наблюдения с точностью до постоянного множителя может быть записана как

$$E(\boldsymbol{r}_1) \propto e^{i\frac{\pi r_1^2}{\lambda L_2}} \int\limits_{-\infty}^{\infty} E_L(\boldsymbol{r}) e^{i\varphi(\boldsymbol{r})} e^{i\frac{\pi r^2}{\lambda L_2}} e^{-i\frac{2\pi}{\lambda L_2}(\boldsymbol{r}\boldsymbol{r}_1)} d\boldsymbol{r}$$

Для дальнейшего расчета статистических характеристик спекл-картин нам понадобится проанализировать функцию вида $|\langle E(\boldsymbol{r}_1) E^*(\boldsymbol{r}_1')\rangle|, \boldsymbol{r}_1' = \boldsymbol{r}_1 + \boldsymbol{\rho}$, где $\boldsymbol{\rho}$ - смещение, введённое в (1). Нужная нам функция имеет вид

$$|\langle E(\boldsymbol{r}_1) E^*(\boldsymbol{r}_1')\rangle| \propto$$
$$\propto \left| \int\limits_{-\infty}^{\infty} \int\limits_{L} (\boldsymbol{r}) E_L(\boldsymbol{r}') \left\langle e^{i[\varphi(\boldsymbol{r}) - \varphi(\boldsymbol{r}')]} \right\rangle e^{i\frac{\pi}{\lambda L_2}\left(r^2 - \boldsymbol{r}'^2\right)} e^{-\frac{2\pi i}{\lambda L_2}\left[(\boldsymbol{r}\boldsymbol{r}_1) - \left(\boldsymbol{r}'\boldsymbol{r}_1'\right)\right]} d\boldsymbol{r} d\boldsymbol{r}' \right|$$

Подынтегральная функция сильно зависит от разности аргументов $|\boldsymbol{r}_1' - \boldsymbol{r}_1| = |\boldsymbol{\rho}|$. Действительно, наиболее быстро с увеличением смещения $\rho$ меняется функция корреляции фазового множителя (6):

$$\left\langle e^{i[\varphi(\boldsymbol{r}) - \varphi(\boldsymbol{r}')]} \right\rangle = e^{-\frac{\rho^2}{\rho_\varphi^2}}$$

где $\rho_\varphi \sim 2$ мкм. При смещении на $\rho_\varphi$ амплитуды и фазы полей $E_L(\boldsymbol{r})$ и $E_L^*(\boldsymbol{r}')$ практически совпадают (поскольку $\rho_L \sim 10^2$ мкм, поэтому можно считать, что (см. (9)) $E_L(\boldsymbol{r})E_L^*(\boldsymbol{r}') = |E_L(\boldsymbol{r})|^2 = I_0 e^{-r^2/\rho_L^2}\Big)$. Аналогично для квадратичного фазового множителя при $L_2 > 1$ см имеем оценку

$$\frac{\pi}{\lambda L_2}\left(r^2-r'^2\right)=\frac{\pi\left(r+r'\right)\left(r-r'\right)}{\lambda L_2}\sim\frac{2\pi\rho_L\rho_\varphi}{\lambda L_2}\sim\frac{\pi}{30}\ll\pi$$

и мы пренебрежём этим множителем в дальнейших выкладках. С учётом приведённых соображений модуль функции корреляции поля про порционален

$$\left|\left\langle E\left(\boldsymbol{r}_1\right)E^*\left(\boldsymbol{r}'_1\right)\right\rangle\right|\propto\left|\int_{-\infty}^{+\infty}e^{-\frac{r^2}{\rho_L^2}}e^{-\frac{\left(\boldsymbol{r}-\boldsymbol{r}'\right)^2}{\rho_\varphi^2}}e^{-\frac{2\pi i}{\lambda L_2}\left[\left(\boldsymbol{r}\boldsymbol{r}_1\right)-\left(\boldsymbol{r}'\boldsymbol{r}'_1\right)\right]}d\boldsymbol{r}d\boldsymbol{r}'\right| \tag{10}$$

Отметим, что в (10) площадка рассеивателя $d\boldsymbol{r}$ является источником плоских волн $e^{-2\pi i(\boldsymbol{r}_1\boldsymbol{r})/\lambda L_2}$, т. е. выполняются условия дифракции Фраунгофера, или, другими словами, изображения спекл-картин наблюдаются в дальней зоне. В дальней зоне дифракционная картина приобретает устойчивую структуру и угловое распределение интенсивности излучения перестаёт зависеть от координаты $z$. Полезно провести наблюдение спекл-структуры при изменении расстояния от рассеивателя до экрана и убедиться в том, что при увеличении координаты $z$ расстояния между спеклами и их линейные размеры будут увеличиваться, но исчезновения и появления новых спеклов не будет. Угловые координаты спеклов $\theta_i=r_{1i}/L_2\,(\theta_i\ll 1)$ при этом остаются неизменными.

Полученное выражение (10) показывает, что корреляционные свойства картины спеклов в дальней зоне определяются распределением не амплитуды, а интенсивности излучения $I_L(r)$ на рассеивателе. Соотношение (10) позволяет анализировать статистические свойства пучка излучения и статистические свойства картины спеклов.

Индикатриса рассеяния. На рис. 5 приведена фотография практически всей спекл-картины, полученной при минимальном удалении фотокамеры от матового стекла ($L_2\approx 15$ мм). Видно, что яркость картины спеклов максимальна в центре и убывает в радиальном направлении. Распределение яркости в картине спеклов обусловлено угловой зависимостью интенсивности рассеянного матовой пластинкой излучения. Кривая, показывающая распределение относительной интенсивности рассеянного света по направлениям (углам), носит название индикатрисы ($\oint p.$ indicatrice - указывающая).

Для конкретной спекл-картины (рис. 5) индикатриса представляет собой сильно изрезанную кривую, максимумы которой соответствуют ярким спеклам. Если провести усреднение по многим участкам матовой пластины, т. е. наложить друг на друга множество картин, подобных приведенной на рис. 5, то распределение интенсивности будет представляться гладкой кривой, спадающей от центра вдоль радиуса, которая будет представлять собой среднюю индикатрису. Грубую оценку для масштаба $\rho_i$, на котором изменяется «сглаженная» индикатриса, можно получить из угловой расходимости рассеянного излучения на основании общих представлений о дифракционных явлениях. Излуче-

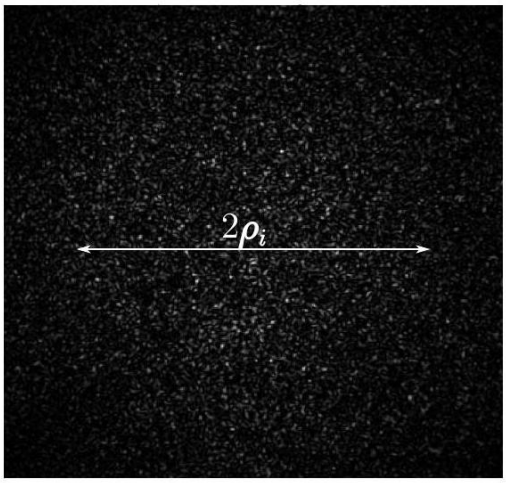



Рис. 5. Общий вид спекл-картины (мелкие спеклы)

ние лазера, падающее на матовую поверхность, имеет угловую расходимость порядка $\theta_L \sim \lambda/\rho_L \sim 0,6 \cdot 10^{-2}$. Для оценки расходимости излучения после прохождения матовой пластины воспользуемся соотношением $\theta_\varphi \sim \lambda/\rho_\varphi$. При $\rho_\varphi \sim 2$ мкм получаем $\theta_\varphi \sim 0,3$, откуда видно, что $\theta_\varphi \gg \theta_L$. Из простых геометрических соображений следует $\rho_i \sim L_2 \frac{\lambda}{\rho_\varphi}$. Более аккуратный расчет (в (10) следует положить $\boldsymbol{r}_1 = \boldsymbol{r}_1'$) показывает, что индикатриса представляет собой гауссову функцию вида

$$\left\langle |E(r_1)|^2 \right\rangle \propto e^{-r_1^2/\rho_i^2}$$

где

$$\rho_i = \frac{\lambda L_2}{\pi \rho_\varphi} \tag{11}$$

- радиус индикатрисы (эта величина отмечена на рис. 5).

Та же запись в угловых переменных дает

$$\theta = \frac{r_1}{L_2}, \quad \theta_i = \frac{\lambda}{\pi \rho_\varphi}, \quad \left\langle |E(\theta)|^2 \right\rangle \sim e^{-\theta^2/\theta_i^2}$$

Анализ статистических свойств рассеянного лазерного излучения (определение величины $\rho_i$) позволяет найти параметр шероховатости матовой поверхности $\zeta$. Используя соотношения (7), (8) и (11), получаем

$$\zeta = \frac{\rho_i}{2L_2 (n_\lambda - 1)} \tag{12}$$

Функции корреляции интенсивности спекл-картины. Крупные спеклы получаются при удалении фотокамеры от матовой пластинки на расстояние $L_2 \approx 25 \div 80$ мм, фотография крупных спеклов приведена на рис. 6. «Крупными» будем называть спеклы, внутри которых заметно распределение яркости, отражающее распределение интенсивности рассеянного излучения. Изучение статистики распределения интенсивности в крупных спеклах позволяет судить о характерном поперечном размере лазерного пучка в месте установки матового стекла.

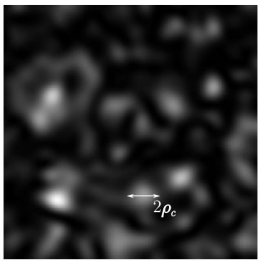



Рис. 6. Крупномасштабное изображение спеклов

Обработка фотографий спекл-картин при различном удалении $L_2$ матовой пластины от фотокамеры показывает, что функция корреляции интенсивности хорошо описывается гауссовой кривой:

$$\Psi_I(\rho) \propto e^{-\rho^2/\rho_c^2} \tag{13}$$

где $\rho_c$ - характерный радиус функции корреляции, физический смысл которого можно трактовать как средний размер спекла. Найдем зависимость радиуса корреляции $\rho_c$ от условий проведения опытов. Для этого вначале оценим функцию корреляции поля $E(\boldsymbol{r})$. Заметим, что эта функция является однородной (т. е. не зависит от выбора координат точек $\boldsymbol{r}$ и $\boldsymbol{r}'$), зависит только от величины смещения $\boldsymbol{\rho} = \boldsymbol{r} - \boldsymbol{r}'$, и, положив $\boldsymbol{r}' = 0$ и $\boldsymbol{r} = \boldsymbol{\rho}$, получим из (10)

$$|\Psi_E(\rho)| \propto \left| \int_{-\infty}^{+\infty} I_L(\boldsymbol{r}) e^{-\frac{2\pi i}{\lambda L_2}(\boldsymbol{\rho}\boldsymbol{r})} d\boldsymbol{r} \right|$$

Радиус корреляции фазового множителя $\rho_\varphi$ в соотношении (10) теперь определяет только нормировочную константу или среднюю интенсивность наблюдаемой спекл-картины. Используя выражение (9) для интенсивности лазерного излучения, получаем

$$|\Psi_E(\rho)| \propto e^{-\frac{\pi^2 \rho_L^2}{\lambda^2 L_2^2}\rho^2} \tag{14}$$

Таким образом, две функции корреляции - экспериментальная (для интенсивности) и теоретическая (для поля) являются гауссовыми. В теории случайных полей доказывается, что функции корреляции интенсивности и поля связаны так же, как интенсивность и поле, т. е.

$$\Psi_I(\rho) = |\Psi_E(\rho)|^2 = e^{-\frac{2\pi^2 \rho_L^2}{\lambda^2 L_2^2}\rho^2}$$

Отсюда следует, что искомый радиус корреляции спекл-картины равен

$$\rho_c = \frac{\lambda L_2}{\sqrt{2}\pi \rho_L} \tag{15}$$

Сделаем важное замечание. Корректное описание случайных полей возможно только статистическими методами. Тем не менее полученные оценки радиуса индикатрисы $\rho_i$ и радиуса корреляции $\rho_c$ хорошо согласуются с выводами классической оптики. Углы дифракции $\theta_i = \rho_i/L_2$ и $\theta_c = \rho_c/L_2$ определяются длиной волны излучения и характерными размерами объектов, в данном случае - радиусами корреляции фазового множителя и лазерного пучка. Из (11) и (15) следует:

$$\theta_i = \frac{\lambda}{\pi \rho_\varphi} \sim \frac{\lambda}{2\rho_\varphi}, \quad \theta_c = \frac{\lambda}{\sqrt{2}\pi \rho_L} \sim \frac{\lambda}{2\rho_L}$$

Функция распределения интенсивности спекл-картины. Интенсивность в картине спеклов представляет собой случайную величину. Одной из характеристик случайной величины является функция плотности распределения вероятности. В нашем случае знание функции распределения интенсивности в спекл-картинах необходимо для правильного выбора режима работы фотокамеры: яркости, контрастности и выдержки.

Рассмотрим формирование случайного поля световой волны в плоскости наблюдения. В каждой точке этой плоскости комплексную амплитуду поля E можно представить себе как большую сумму полей, рассеянных различными неровностями матовой пластинки. Суммирование таких полей удобно представить в форме векторной диаграммы

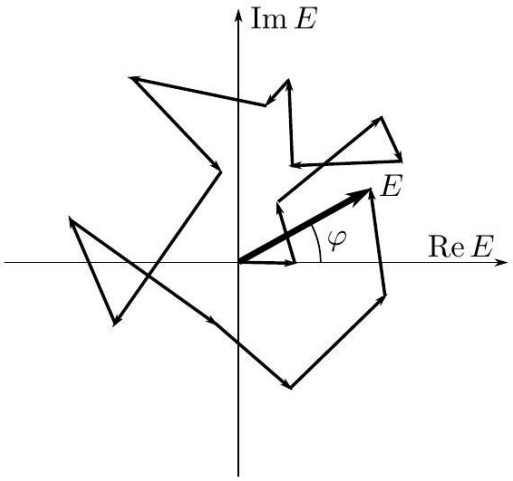



Рис. 7. Векторная диаграмма для вычисления комплексной амплитуды поля в выбранной точке спекл-картины

(рис. 7). Длина отдельного вектора $E_n$ зависит от размера неровности, её формы и т. д. Угол поворота вектора определяется относительной фазовой задержкой, вносимой неровностью, и разностью хода от матовой пластины до плоскости наблюдения. Число таких векторов велико: для $\rho_\varphi \sim 1$ мкм и $\rho_L \sim 10^2$ мкм их количество $(\rho_L/\rho_\varphi)^2 \sim 10^4$.

Как отмечалось ранее, средние значения суммарного вектора поля, его действительной и мнимой частей равны нулю:

$$\langle \boldsymbol{E} \rangle = 0, \quad \langle \mathrm{Re}\, E \rangle = \langle |E| \cos\varphi \rangle = 0, \quad \langle \mathrm{Im}\, E \rangle = \langle |E| \sin\varphi \rangle = 0$$

Рассмотрим действительную $E_R = |E| \cos\varphi$ и мнимую $E_I = |E| \sin\varphi$ части вектора $\boldsymbol{E}$. Величина $dE_R dE_I$ есть элемент площади на комплексной плоскости, изображенной на рис. 7.

Первоначально найдем вероятность отклонения комплексной амплитуды поля $\boldsymbol{E}$ от её среднего значения, т. е. от нуля. Ответ следует из центральной предельной теоремы теории вероятностей: если некоторая случайная величина равна большой сумме других случайных величин, то искомая плотность вероятности имеет нормальный (гауссов) закон распределения. В нашем случае вероятность того, что конец вектора $\boldsymbol{E}$ (рис. 8) лежит в области от $E_R$ до $E_R + dE_R$, от $E_I$ до $E_I + dE_I$, пропорциональна $e^{-\left(E_R^2+E_I^2\right)/I_0} dE_R dE_I$, где $I_0$ - некоторая константа Вероятность того, что амплитуда суммарного поля спекл-картины лежит в интервале от $|E|$ до $|E| + d|E|$:

$$W(|E|)d|E| \propto |E| e^{-|E|^2/I_0} d|E|$$

(это распределение называется распределением Релея). Поскольку интенсивность света $I = |E|^2$, то искомая функция распределения вероятности для спекл-картины имеет вид [1]:

$$dP(I) = W(I)dI = \frac{1}{I_0} e^{-I/I_0} dI \tag{16}$$

Константа $I_0$ в выражении (16) имеет смысл среднего значения интенсивности. Распределение (16) называется экспоненциальным; такого рода распределения случайных величин часто встречаются в физике: такому распределению подчиняется длина свободного пробега молекул в газе, а также энергия частиц в равновесной системе (распределение Больцмана). Особенностью экспоненциального распределения является то, что $\langle I \rangle = I_0$, а среднеквадратичное отклонение также равно $I_0$:

$$\sigma_I = \sqrt{\langle I^2 \rangle - \langle I \rangle^2} = I_0. \tag{17}$$

В справедливости приведенных соотношений легко убедиться непосредственным вычислением, используя (16).

Регистрация и обработка изображений спеклов. Для регистрации картины спеклов используется цифровая фотокамера, для обработки изображений применяются компьютерные программы «Индикатриса» и «Корреляция», написанные в среде Mathcad. Фотокамера имеет светочувствительную матрицу, содержащую 480 фоточувствительных элементов - сэнселей (от англ. sensel = sensor element) - по вертикали (нумерация $0 \ldots 479$) и 640 сенселей по горизонтали (нумерация $0 \ldots ....639$). Расстояние между чувствительными элементами $d = 8,5$ мкм. Таким образом, получаемое изображение состоит из $480 \times 640$ пикселей (англ. pixel = picture element). Интенсивность сигнала в каждом пикселе квантуется в 8-битной шкале и записывается в виде целого числа от 0 до 255 ($2^8 = 256$ уровней интенсивности). Если световой сигнал превосходит максимальный, то ему присваивается значение 255 (режим насыщения). Наиболее точные результаты обработки изображений достигаются в том случае, когда используется весь динамический диапазон фотокамеры, т. е. количество чувствительных элементов, находящихся в режиме насыщения (значение интенсивности 255), минимально. Для контроля режима работы фотокамеры используется режим вывода гистограмм интенсивности на экран монитора компьютера.

На рис. 8 показаны гистограммы сигналов фотокамеры при фотографировании индикатрисы рассеяния (рис. 5). По оси абсцисс отложена величина сигнала (целое число от 0 до 255), по оси ординат - количество пикселей с соответствующей интенсивностью. Оптимальной является ситуация, когда гистограмма в логарифмическом представлении (рис. 8б) максимально приближается к прямой (в соответствии с (16)). Приемлемые результаты получаются, когда в режиме насыщения находится небольшое количество пикселей ($\sim 10^2$). Линейность амплитудной характеристики изображения достигается варьированием режима съемки: яркости, контраста, выдержки (значению выдержки «1» соответствует реальное время 1/30

[0]1 Подобная функция распределения возникает в термодинамике при описании тепловых скоростей молекул идеального газа. Скорость молекулы газа $\boldsymbol{v}$ можно представить как сумму большого числа скачков скоростей $\Delta\boldsymbol{v}$, возникающих при взаимных столкновениях молекул. Векторная диаграмма, показанная на рис. 7, соответствует этому представлению для двумерного распределения по скорости, где каждый элементарный вектор изображает скачок скорости при соударении. Для двумерного распределения средние значения проекций скоростей $\langle v_x \rangle = \langle v_y \rangle = 0$. Вероятность того, что после большого числа столкновений вектор $\boldsymbol{v}$ имеет проекции от $v_x$ до $v_x + dv_x$, от $v_y$ до $v_y + dv_y$, определяется распределением Максвелла $\propto \exp\left[-\left(v_x^2 + v_y^2\right)/2\sigma_v^2\right] dv_x dv_y$, т. е. имеет гауссову форму. Для газа среднеквадратическое отклонение $\sigma_v^2 = \left\langle v^2 \right\rangle$ определяется массой молекулы и температурой. Вероятность того, что модуль скорости частицы газа лежит в интервале от $v$ до $v + dv$, пропорциональна $\propto \exp\left[-v^2/2\sigma_v^2\right] \cdot 2\pi v dv$, что аналогично (16). Формы функций распределения для комплексной амплитуды поля и скоростей газа основаны на центральной предельной теореме, поэтому совпадают и распределения по интенсивности поля (16) и по энергиям молекул идеального газа.

секунды.). Кроме того, между лазером и линзой можно установить поляроид и с его помощью регулировать интенсивность света. Установка параметров работы камеры производится при фотографировании мелких спеклов (индикатриса рассеяния, рис. 5), так как количество спеклов в этом случае велико, и экспериментальное распределение интенсивности удовлетворительно соответствует теоретическому распределению (16). Рекомендуемые значения параметров указаны на установке.

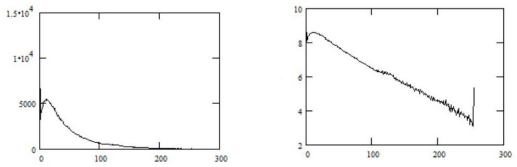



Рис. 8. Гистограмма (а) и ее логарифм (б). Среднее значение сигнала 45, среднеквадратическое отклонение 44. Сигнал превышает порог 255 в 200 пикселях (0, 07% от общего числа)

Для обработки фотографий мелких спеклов и определения радиуса индикатрисы служит программа «Индикатриса». Физически установить камеру так, чтобы центр спекл-картины совпадал с центром фоточувствительной матрицы, весьма сложно. Поэтому центр изображения определяется программными средствами, подобно тому, как вычисляется центр масс пластины с переменной плотностью, если вместо распределения масс использовать распределение интенсивностей по пикселям. Затем программа выделяет из всего изображения квадрат около найденного центра. На рис. 5 приведено изображение мелких спеклов внутри выделенного квадрата, а на рис. 9 приведено сечение сигналов вдоль диагонали квадрата. Программа строит гауссовскую поверхность, которая имеет минимальное среднеквадратическое отклонение от матрицы зарегистрированных сигналов.

Вычисления проводятся с безразмерными величинами, и их результаты выдаются в безразмерных числах (пикселях). Для получения правильного значения радиуса индикатрисы $\rho_i$ в размерных единицах необходимо учесть, что расстояние между пикселями составляет $d = 8, 5$ мкм.

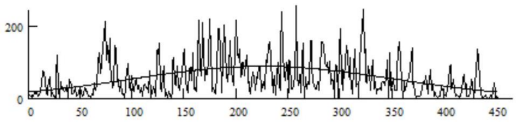



Рис. 9. Развертка сигнала фотокамеры и сечение поверхности средней интенсивности. По оси абсцисс отложен номер пикселя

Определение радиуса корреляции спекл-картины позволяет найти характерный размер лазерного пучка в месте установки матового стекла. Необходимые для этого вычисления реализует программа «Корреляция». Записанная матрица сигналов фотокамеры программно смещается относительно самой себя на малое число пикселей, и при каждом смещении проводятся вычисления корреляционной функции. Смещения проводятся для всех направлений, полученная поверхность аппроксимируется наилучшей гауссовой поверхностью. В работе предлагается определить радиус корреляции спекл-картины и соответственно характерный размер лазерного пучка на матовом стекле для четырех различных значений $L_2$.

Определения и формулы, используемые в работе. Проведенный выше теоретический анализ включает в себя следующие основные моменты.

1. Характеристики матового стекла:

$\sigma_h$ - среднеквадратическое отклонение толщины стеклянной пластины от среднего значения.

$\sigma_\varphi = \frac{2\pi}{\lambda}(n_\lambda - 1)\sigma_h$ - среднеквадратическое отклонение случайной фазы световой волны.

$\rho_h$ - радиус корреляции случайных неровностей матовой поверхности.

$\zeta = \frac{\sigma_h}{\rho_h}$ - параметр шероховатости.

$\left\langle e^{i\varphi(\boldsymbol{r}+\boldsymbol{\rho})} e^{-i\varphi(\boldsymbol{r})} \right\rangle = e^{-\frac{\rho^2}{\rho_\varphi}}$ - функция корреляции случайного фазового множителя.

$\rho_\varphi = \frac{\rho_h}{\sigma_\varphi} = \frac{\rho_h}{\sigma_h}\frac{\lambda}{2\pi(n_\lambda - 1)} = \frac{\lambda}{2\pi(n_\lambda - 1)\zeta}$ - радиус корреляции случайного фазового множителя.

2. Индикатриса рассеяния:

$\langle I \rangle \propto e^{-\rho^2/\rho_i^2}$ - форма индикатрисы.

$\rho_i = \frac{\lambda L_2}{\pi \rho_\varphi} = 2L_2(n_\lambda - 1)\zeta$ - радиус индикатрисы.

3. Статистические характеристики интенсивности спекл-картины: $I(r) \sim e^{-r^2/\rho_L^2}$ - форма интенсивности лазерного луча.

$\Psi_I(\boldsymbol{\rho}) = e^{-\rho^2/\rho_c^2}$ - безразмерная функция корреляции интенсивности.

$\rho_c = \frac{\lambda L_2}{\sqrt{2}\pi\rho_L}$ - радиус корреляции интенсивности спекл-картины.

$dN = \frac{N}{I_0} e^{-I/I_0} dI$ - гистограмма сигналов фотокамеры.

$I_0 = \langle I \rangle = \sqrt{\langle I^2 \rangle - \langle I \rangle^2}$ - среднее значение и среднеквадратическое отклонение интенсивности (экспоненциальное распределение).

## ЗАДАНИЕ

В работе предлагается провести качественные наблюдения спеклкартин, получить набор их фотографий, обработать фотографии с помощью компьютера и количественно оценить статистические характеристики спеклов.

### I. Качественные наблюдения

1. Включите лазер. Поместите в луч лазера матовое стекло, за стеклом установите экран. Перемещая экран вдоль оптической скамьи от рассеивателя, проследите за движением спеклов. Убедитесь в применимости
приближения Фраунгофера. Получите изображение индикатрисы рассеяния, см. рис. 5, измерьте необходимые расстояния и оцените параметр шероховатости матового стекла. Получите изображение крупных спеклов, см. рис. 6, оцените диаметр лазерного пучка.

2. На расстоянии 50 см от лазера установите линзу с фокусным расстоянием $f \sim 55$ мм. Контролируя положение лазерного пучка на удаленном экране, проведите центровку линзы. В некоторых прикладных задачах исследуются спекл-картины, возникающие при отражении света от шероховатой поверхности. Проведите наблюдение спекл-картины, возникающие при отражении расширенного пучка лазера от бумажного экрана. При перемещении глаза наблюдателя освещенный участок экрана начинает мерцать.

3. Установите матовое стекло между линзой и её фокальной плоскостью ($f = 55$ мм). Перемещая стекло при помощи установочного винта, отметьте связь направления движения стекла и спеклов. Проделайте такой же опыт, поместив стекло за фокальной плоскостью. Объясните различие направлений смещения стекла и спеклов. Аккуратно перемещая стекло к фокальной плоскости линзы, получите изображение «кипящих» спеклов. По размеру одиночного спекла оцените поперечный диаметр лазерного пучка вблизи фокальной плоскости линзы.

**II. Фотографии спеклов**

4. Установите между линзой и лазером поляроид. Матовое стекло установите на расстоянии 75 мм от линзы рассеивающей поверхностью от лазера, камеру - на расстояние 20 мм от матовой поверхности рассеивателя (последовательность операций указана в инструкции «Регистрация и обработка изображений»). Необходимо максимально ослабить поляроидом прошедший пучок.

5. Используя инструкцию, получите изображение индикатрисы рассеяния (мелких спеклов). Яркость изображения подбирается выдержкой и поляроидом.

Используя инструкцию, перейдите в режим фотографирования, сделайте фотографию и получите гистограмму оцифрованного сигнала. Форма гистограммы (см. рис. 8) характеризует правильность выбора параметров работы камеры. В режиме насыщения должны работать $\sim 10^2$ пикселей, что регулируется выдержкой и поляроидом. Сделайте и проанализируйте несколько пробных фотографий.

6. Сфотографируйте и запомните картину мелких спеклов. В названии файлов рекомендуется вводить расстояние от линзы до матового стекла и от стекла до матрицы, например, 75_20.

Получите 3 фотографии крупных спеклов, перемещая камеру от стекла шагом 20 мм (файлы 75_40, 75_60, 75_80). Форма гистограммы сигналов для крупных спектров не обязательно должна соответствовать приведенной на рис. 8 (мал объем выборки), контролируйте только количество пикселей в режиме насыщения.

7. Получите фотографию индикатрисы и 3 фотографии крупных спеклов для расстояния между линзой и матовым стеклом 85 мм.

**III. Обработка фотографий**

В работе используются две программы «Индикатриса» и «Корреляция», написанные в среде Mathcad.

8. Откройте программу «Индикатриса» и обработайте с её помощью полученные Вами фотографии мелких спеклов для двух расстояний линзаматовое стекло. Введите имя фотографии и расстояние линза-стекло $L_1$. Зарисуйте в рабочей тетради гистограмму в логарифмическом масштабе, запишите среднее значение и среднеквадратическое отклонение, число пикселей с сигналом 255.

Программа вычисляет среднеквадратические ошибки отклонения сигнала набора гауссовских поверхностей с различными радиусами индикатрисы рассеяния. Обратите внимание на обязательное наличие минимума ошибки для некоторого радиуса индикатрисы. Запишите выбранное значение радиуса в пикселях, миллиметрах, найденный параметр шероховатости матового стекла. Для радиуса корреляции высоты шероховатой поверхности $\rho_h \sim 10$ мкм оцените дисперсию самих шероховатостей $\sigma_h$ и дисперсию фазы $\sigma_\varphi$. Закройте программу «Индикатриса».

9. Откройте программу «Корреляция». При помощи этой программы обработайте 4 фотографии для расстояний матовое стекло-линза $L_1 = 75$ мм, затем $L_1 = 85$ мм, всего 8 фотографий. Введите имя файла. Программа позволяет развернуть сигнал матрицы по выбранному столбцу (пиксель с номером)0,0) находится в левом верхнем углу изображения). Рекомендуется посмотреть форму случайного сигнала.

Программа вычисляет функцию корреляции интенсивности спеклкартины и находит наилучший радиус корреляции, аппроксимируя эту функцию гауссовской поверхностью (по минимуму ошибки аппроксимации). Обратите внимание, что используемая функция корреляции нормирована на единицу при нулевом смещении изображений. В программе приведены сводные таблицы результатов измерений и расчётов. Величина расстояния линза-стекло $L_1$ или стекло-камера $L_2$ измеряются в мм, радиус корреляции в пикселях. После обработки каждой фотографии внесите результаты в таблицу и в рабочую тетрадь. Про верьте линейность зависимости радиуса корреляции от расстояния матовое стекло-камера. В программе введена ошибка расчётов радиуса корреляции 0,5 пикселя. Занесите графики в рабочую тетрадь, запишите полученные значения радиусов лазерного пучка по уровню интенсивности 1/ e. Закройте программу «Корреляция».

**Контрольные вопросы**

1. Как возникает спекл-структура и чем она отличается от дифракции Фраунгофера на мелкой сетке?
2. При каких условиях будет наблюдаться спекл-картина при отражении от шероховатой поверхности?
3. Как зависит радиус корреляции фазового множителя от характеристик матовой поверхности, показателя преломления материала?
4. Чем определяется длина вектора $|E_n|$ и угол его поворота на рис. 7?
5. Что такое радиус индикатрисы рассеяния и от чего он зависит?
6. Как использовать гистограмму камеры для её настройки?
7. Наблюдалось ли в опыте нарушение условия $\langle e^{i\varphi} \rangle = 0$?

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. С. 433-436.
2. Ахманов С.А., Дьяков Ю.Е., Чиркин А.С. Введение в статистическую радиофизику и оптику. - М.: Наука, 1981 г. - С. 320-326.

## 22.3 Experiments about Diffraction. Fourier Optics

### 22.3.1 Study of Light Diffraction

Цель работы: исследовать явления дифракции Френеля и Фраунгофера на щели, изучить влияние дифракции на разрешающую способность оптических инструментов.

В работе используются: оптическая скамья, ртутная лампа, монохроматор, щели с регулируемой шириной, рамка с вертикальной нитью, двойная щель, микроскоп на поперечных салазках с микрометрическим винтом, зрительная труба.

Перед выполнением работы необходимо ознакомиться с теорией дифракции (см. п. 1 введения к разделу).

## A. Fresnel Diffraction on a Slit

Схема установки для наблюдения дифракции Френеля на щели представлена на рис. 1. Световые лучи освещают щель $S_2$ и испытывают на ней дифракцию. Дифракционная картина рассматривается с помощью микроскопа М, сфокусированного на некоторую плоскость наблюдения П.

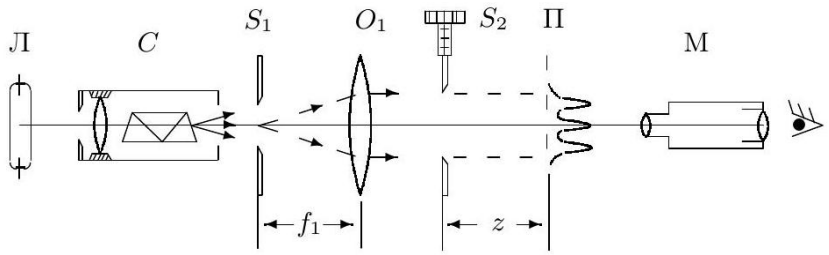



Рис. 1. Схема установки для наблюдения дифракции Френеля

Щель $S_2$ освещается параллельным пучком монохроматического света с помощью коллиматора, образованного объективом $O_1$, и щелью $S_1$, находящейся в его фокусе. На щель $S_1$ сфокусировано изображение спектральной линии, выделенной из спектра ртутной лампы Л при помощи простого монохроматора $C$, в котором используется призма прямого зрения.

Распределение интенсивности света в плоскости наблюдения П проще всего рассчитывать с помощью зон Френеля (для щели их иногда называют зонами Шустера). При освещении щели $S_2$ параллельным
пучком лучей (плоская волна) зоны Френеля представляют собой полоски, параллельные краям щели (рис. 2). Результирующая амплитуда в точке наблюдения определяется суперпозицией колебаний от тех зон Френеля, которые не перекрыты створками щели. Графическое определение результирующей амплитуды производится с помощью векторной диаграммы - спирали Корню. Суммарная ширина $n$ зон Френеля (Шустера) определяется соотношением (3.20):

$$\xi_n = \sqrt{zn\lambda} \tag{1}$$

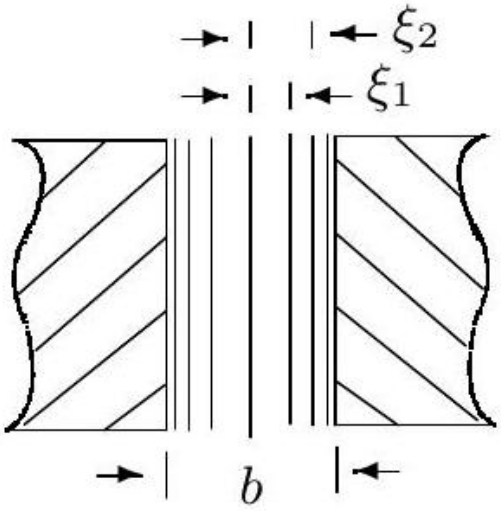
где $z$ - расстояние от щели до плоскости наблюдения (рис. 1), а $\lambda$ - длина волны.

Вид наблюдаемой дифракционной картины на щели шириной $b$ определяется волновъм $na$ раметром (3.1)

$$p = \frac{\sqrt{z\lambda}}{b}$$

Также используют так называемое число Френеля:

$$C = \frac{b^2}{z\lambda} = \frac{1}{p^2}$$




• это, как ясно из (1), полное число открытых

Рис. 2. Зоны Френеля в плоскости щели зон Френеля на всей ширине щели.

Дифракционная картина отсутствует вблизи щели при $p \ll 1$ ($C \gg 1$, т.е. на щели укладывается огромное число зон), а распределение интенсивности света за щелью можно приближённо получить с помощью законов геометрической оптики. Дифракционная картина в этом случае наблюдается только в узкой области на границе света и тени у краёв экрана.

При небольшом удалении от щели (или изменении ширины щели $S_2$) эти две группы дифракционных полос перемещаются практически независимо друг от друга. Каждая из этих групп образует картину дифракции Френеля на краю экрана. Распределение интенсивности при дифракции света на краю экрана может быть найдено с помощью спирали Корню (см. п. 1.4.2 введения к разделу).

При дальнейшем увеличении расстояния $z$ (или уменьшении ширины щели $S_2$) обе системы дифракционных полос постепенно сближаются и, наконец, при $C \gtrsim 1$ накладываются друг на друга. Распределение интенсивности в плоскости наблюдения в этом случае определяется числом зон Френеля, укладывающихся на полуширине щели $b/2$. Если
это число равно $n$, то в поле зрения наблюдается $m = n - 1$ тёмных полос. Таким образом, по виду дифракционной картины можно оценить число зон Френеля на полуширине щели.

## ЗАДАНИЕ

В этом упражнении предлагается исследовать дифракцию Френеля на узкой щели, на краю экрана, на тонкой нити.

### I. Подготовка приборов к работе

1. Соберите схему согласно рис. 1. Для этого включите ртутную лампу, поставьте входную щель $S_1$ за монохроматором и сфокусируйте на неё зелёную линию спектра. Откройте щель пошире и с помощью экрана или листа белой бумаги проверьте, что луч идёт вдоль оптической скамьи.

2. Для освещения параллельньм пучком щели $S_2$ (щель с микрометрическим винтом!) установите линзу $O_1$ на расстоянии от щели $S_1$, близком к фокусному (оно указано

на оправе линзы). Точная настройка пучка на параллельность осуществляется с помощью зрительной трубы, настроенной на бесконечность.

3. Для настройки зрительной трубы на бесконечность сначала настройте её окуляр: поворотом глазной линзы добейтесь появления в поле зрения чёткого изображения окулярной шкалы или креста; затем вращением винта кремальеры найдите чёткое изображение удалённого объекта (окно в конце коридора). Запомните показания винта кремальеры, чтобы исключить случайный сбой настройки.

4. Поставьте зрительную трубу за линзой $O_1$. Убедитесь, что пучок света проходит через центр линзы и попадает на центр объектива зрительной трубы.

Слегка перемещая линзу $O_1$ вдоль оси системы, найдите в окуляре зрительной трубы резкое изображение входной щели $S_1$. Закрепите входную щель и линзу на оптической скамье.

5. Определите нуль микрометрического винта щели $S_2$: глядя сквозь щель на окно или лампу накаливания, определите момент её открытия. Повторите эту процедуру несколько раз, вращая винт всё медленнее, и зафиксируйте новое значение нуля. Максимальная ширина щели 0,4 мм, один оборот винта - 100 дел - соответствует 0,1 мм (цена деления - 0,001 мм). Установите ширину щели $0,2 - 0,3$ мм.

Поставьте щель $S_2$ за линзой $O_1$ и закрепите.

6. Сфокусируйте микроскоп на щель $S_2$. Для этого сначала, глядя в окуляр микроскопа, убедитесь, что окулярная шкала видна чётко. Если
нет - подстройте окуляр под свой глаз: вытащив окуляр из тубуса, направьте его на любой яркий объект (лампу, окно) и поворотом глазной линзы (перемещением её вдоль оси окуляра) добейтесь чёткого видения окулярной шкалы. Настроив окуляр, поставьте его на место.

Предметная плоскость микроскопа расположена практически в фокусе его объектива (обычно это $2-4$ см). Поставьте микроскоп за щелью $S_2$ так, чтобы указатель продольного перемещения салазок микроскопа был расположен со стороны окуляра. Убедитесь с помощью экрана, что световой пучок попадает на объектив.

Перемещая микроскоп вдоль оптической оси, найдите в окуляре резкое изображение щели $S_2$. Проверьте, что при небольшом удалении микроскопа от щели на ярком фоне изображения щели (сначала около краёв) появляются узкие тёмные полосы, количество которых уменьшается по мере удаления микроскопа (дифракция в ближней волновой зоне).

7. Методом последовательных приближений попытайтесь увеличить контрастность картинъ. Она зависит от ширины щели $S_1$ (при уменьшении щели растёт контрастность, но падает яркость), от центрировки линзы $O_1$, от равномерности освещения щелей и их параллельности.

**II. Измерения**

8. Добившись наибольшей чёткости дифракционной картины, снова найдите резкое изображение щели (чёткие края без дифракционных полос). Запишите начальное положение микроскопа - координату по шкале продольной линейки, расположенной на оптической скамье.

Если тубус микроскопа закреплён на подставке с продольной шкалой, зафиксируйте начальное положение микроскопа по этой шкале.

9. Постепенно отодвигая микроскоп от щели $S_2$, заметьте по шкале положение микроскопа, при котором на фоне щели видна одна тёмная полоса. Смещение микроскопа от первоначального положения (от резкого изображении щели) даёт величину $z$ - расстояние от щели до плоскости наблюдения [см. рис. 1 и формулу (1)]. Если это расстояние меньше двух см, следует увеличить ширину щели $S_2$ и снова поработать над контрастностью картины (см. п. 7).

Приближая микроскоп к щели, снимите зависимость координаты микроскопа от числа $n$ наблюдаемых тёмных полос. При хорошей настройке можно увидеть $4 - 5$ полос.

10. Измерьте ширину $b$ щели $S_2$, используя микрометрический винт поперечных салазок микроскопа (один оборот винта - 100 дел. - 1 мм). Освещённость изображения увеличится, если осветить щель настольной лампой.

Сравните результат с показаниями микрометрического винта щели $S_2$. В случае заметного расхождения (причиной может быть люфт) откройте щель пошире и снова сравните результаты измерения ширины. Влияние люфта можно исключить, если при измерениях вращать винт в одну сторону.

### III. Качественные наблюдения

Вновь сфокусируйте микроскоп на щель. При небольшом удалении микроскопа от щели у её краёв появляются узкие частые полосы. Это дифракиия на краю экрана. Обратите внимание, что возле границы щели расположена самая яркая светлая полоса.

11. Закрепите микроскоп на оптической скамье и проследите за изменением дифракционной картины при уменьшении ширины щели $S_2$. Опишите результат.

12. Для исследования дифракиии Френеля на препятствии поставьте вместо щели $S_2$ рамку с тонкой вертикальной нитью. Настройте микроскоп на резкое изображение нити. Убедитесь, что при удалении микроскопа от нити на её фоне всегда наблюдается чётное число тёмных дифракционных полос (светлый центр).

### IV. Обработка результатов

13. Сравните размер зон Френеля с измеренной шириной $b$ щели $S_2$. Для этого свяжите число тёмных полос $m$ в поле зрения с числом зон Френеля $n$ на полуширине щели, рассчитайте величину $2\xi_n$ по формуле (1) и постройте график $2\xi_n = f(n)$. Отложите на графике величину $b$. Длина волны зелёной линии ртути $\lambda = 5461\text{Å}$.

14. Нарисуйте качественный график распределения интенсивности на границе света и тени при дифракции на краю экрана. С помощью спирали Корню сравните интенсивности на краю экрана и вдали от него.

15. Объясните с помощью зон Френеля, как связано видимое число тёмных полос с шириной щели.

### Б. Дифракция Фраунгофера на щели

На значительном удалении от щели, когда выполнено условие $C \ll 1$ (то есть ширина щели становится значительно меньше ширины первой зоны Френеля, $b \ll \sqrt{\lambda z}$, изображение щели размывается и возникает дифракционная картина, называемая дифракцией Фраунгофера (см. п. 1.5.2 введения).

Дифракцию Френеля и Фраунгофера можно наблюдать на одной и той же установке (рис. 1). Однако при обычных размерах установки дифракция Фраунгофера возникает только при очень узких щелях.

Например, при $z \approx 20 - 40$ см и $\lambda \approx 5 \cdot 10^{-5}$ см получаем $b \ll 0,3$ мм. Поскольку работать с такими тонкими щелями неудобно, для наблюдения дифракции Фраунгофера к схеме, изображённой на рис. 1, добавляется объектив $O_2$ (рис. 3).

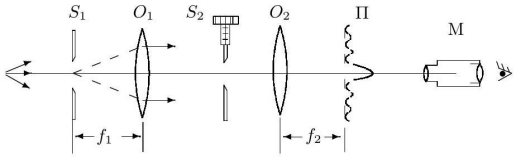



Рис. 3. Схема установки для наблюдения дифракции Фраунгофера на щели

Дифракционная картина наблюдается здесь в фокальной плоскости объектива $O_2$. Каждому значению угла $\theta$ соответствует в этой плоскости точка, отстоящая от оптической оси на расстоянии

$$x = f_2 \operatorname{tg} \theta \approx f_2 \theta \tag{2}$$

Поскольку объектив не вносит дополнительной разности хода между интерферирующими лучами (таутохронизм), в его фокальной плоскости наблюдается неискажённая дифракционная картина Фраунгофера. Эта картина соответствует бесконечно удалённой плоскости наблюдения.

В центре поля зрения наблюдается дифракционный максимум (светлая полоса). При малых углах $\theta$ положение минимумов (тёмных полос) определяется, согласно формуле (3.27), соотношением

$$\theta_m = m \frac{\lambda}{b} \tag{3}$$

Расстояние $x_m$ от тёмной полосы до оптической оси объектива $O_2$ пропорционально фокусному расстоянию $f_2$. Из (2) и (3) следует

$$x_m = m \frac{\lambda}{b} f_2 \tag{4}$$

Из (4) видно, что при малых углах минимумы эквидистантны, а расстояния $\delta x$ между минимумами обратно пропорциональны ширине $b$ щели $S_2$.

### ЗАДАНИЕ

В этом упражнении предлагается исследовать дифракцию Фраунгофера на щели и проследить, как влияют изменение ширины щели и её смещение на характер дифракционной картины.

### V. Настройка установки

1. Для перехода из ближней волновой зоны (дифракция Френеля) в дальнюю (дифракция Фраунгофера) к установке, собранной по рис. 1, достаточно добавить линзу $O_2$ (рис. 3). Если параллельность пучка за линзой $O_1$ нарушилась, воспользуйтесь рекомендациями, изложенными в пп. А2-А4.

2. Поставьте линзу $O_2$ между щелью $S_2$ и микроскопом (чем короче установка, тем ярче будет картина).

Настройте микроскоп на фокальную плоскость П линзы (фокус линзы указан на её оправе). Для этого временно снимите со скамьи щель $S_2$ и убедитесь, что свет проходит через центры линз и попадает на объектив микроскопа. Перемещая микроскоп вдоль скамьи, найдите резкое изображение щели $S_1$.

Уменьшая ширину щели $S_1$, проверьте, что микроскоп настроен на края щели, а не на цветную полосу, которую даёт монохроматор. Закрепите микроскоп и линзы на скамье.

3. Поместите щель $S_2$ между линзами и подберите её ширину так, чтобы в поле зрения микроскопа появилась дифракционная картина.

В отличие от френелевой дифракционной картины фраунгоферова занимает всё поле зрения микроскопа. Добейтесь наибольшей контрастности картины (см. п. А7).

### VI. Измерения

4. Измерьте с помощью винта поперечного перемещения микроскопа координаты $X_m$ нескольких дифракционных минимумов (от $-m$ до $+m$). Запишите ширину $b$ щели $S_2$ и фокусное расстояние линзы $O_2$.

### VII. Качественные наблюдения

5. Убедитесь, что смещение щели $S_2$ в боковом направлении не приводит к сдвигу дифракционной картины. Объясните явление.

6. Проследите, как изменяется масштаб дифракционной картины при уменьшении ширины щели $S_2$.

### VIII. Обработка результатов

7. Постройте график, откладывая по горизонтали номер минимума $m$, а по вертикали - его координату $X_m$ (от $-m$ до $+m$). По углу наклона прямой определите среднее расстояние $\Delta X$ между соседними минимумами; рассчитайте ширину щели $b$ по формуле (4) и сравните с измеренной. Оцените расхождение результатов.

## В. Дифракция Фраунгофера на двух щелях

Для наблюдения дифракции Фраунгофера на двух щелях в установке (рис. 3) следует заменить щель $S_2$ экраном Э с двумя щелями (рис. 4). При этом для оценки влияния ширины входной щели на чёткость дифракционной картины вместо входной щели $S_1$ следует поставить щель с микрометрическим винтом. Два дифракционных изображения входной щели, одно из которых образовано лучами, прошедшими через левую, а другое - через правую щели, накладываются друг на друга.

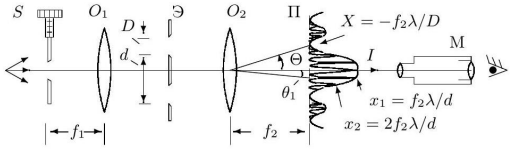



Рис. 4. Схема установки для наблюдения дифракции Фраунгофера на двух щелях

Если входная щель достаточно узка, то дифракционная картина в плоскости П (рис. 4) подобна той, что получалась при дифракции на одной щели (рис. 3), однако теперь вся картина испещрена рядом дополнительных узких полос (см. п. 1.5.3 введения к разделу).

Как ясно из формулы (3.29), угловая координата $\theta_m$ интерференционного максимума $m$-го порядка определяется соотношением

$$\theta_m = m\frac{\lambda}{d} \tag{5}$$

где $d$ - расстояние между щелями. Линейное расстояние $\delta x$ между соседними интерференционными полосами в плоскости П равно, поэтому

$$\delta x = f_2\frac{\lambda}{d} \tag{6}$$

На рис. 4 показано распределение интенсивности в фокальной плоскости объектива $O_2$. Штриховой линией (в увеличенном масштабе) изображено распределение интенсивности при дифракции света на одиночной щели. Нетрудно оценить число $n$ интерференционных полос, укладывающихся в области центрального дифракционного максимума. Согласно (4) полная ширина главного максимума равна $2f_2\lambda/b$, где $b$ - ширина щели, отсюда

$$n = \frac{2\lambda f_2}{b}\frac{1}{\delta x} = \frac{2d}{b} \tag{7}$$

При дифракции света на двух щелях чёткая система интерференционных полос наблюдается только при достаточно узкой ширине входной щели $S$, которую можно рассматривать как протяжённый источник света размером $b$. Для наблюдения интерференции необходимо, чтобы расстояние $d$ между щелями не превышало радиуса когерентности (см. раздел II, п. 4.3):

$$d \leqslant \frac{\lambda}{b}f_1. \tag{8}$$

Здесь $b$ - ширина входной щели $S$ и, следовательно, $b/f_1$ - её угловая ширина. Таким образом, по размытию интерференционной картины можно оценить размер источника. Этот метод используется в звёздном интерферометре при измерении угловых размеров звёзд.

## ЗАДАНИЕ

В этом упражнении предлагается исследовать картину дифракции на двух щелях и оценить влияние размеров источника на чёткость картины.

### IX. Настройка и измерения

1. Не перемещая (!) линз и микроскопа, замените в установке (рис. 3) входную щель $S_1$ щелью с микрометрическим винтом и, слегка передвигая её вдоль скамьи, найдите в микроскопе резкое изображение новой входной щели. Закрепите щель.

Поставьте между линзами экран Э с двойной щелью. В области главного дифракционного максимума должна появиться система равноотстоящих тёмных и светлых полос (рис. 4). Центрировкой системы и подбором ширины щели $S$ добейтесь наибольшей чёткости дифракционной картины.

2. Определите с помощью микрометрического винта поперечных салазок микроскопа координаты самых удалённых друг от друга тёмных полос внутри центрального максимума и просчитайте число светлых промежутков между ними. Измерьте ширину центрального максимума.

3. Исследуйте влияние пространственной когерентности на видность интерференционной картины. Для этого, расширяя входную щель $S$, подберите такую ширину щели $b_0$, при которой наступает первое исчезновение интерференционных полос, и запишите эту величину.

Убедитесь, что при дальнейшем увеличении входной щели картина вновь появляется, но она заметно менее контрастна. Определите соответствующую ширину входной щели и ширину, при которой картина наиболее контрастна.

4. Запишите фокусные расстояния обеих линз.

5. Метод измерения параметров двойной щели, необходимых для расчётов, изложен в п. ГЗ.

**X. Обработка результатов**

6. Определите расстояние $\delta x$ между минимумами по результатам измерений, рассчитайте величину $d$ по формуле (6) и сравните с измеренной.

Рассчитайте число полос внутри главного максимума по формуле (7) и сравните с экспериментальным.

7. Сравните измеренную ширину $b_0$ щели $S$ с расчётом по формуле (10).

## Г. Влияние дифракции на разрешающую способность оптического инструмента

Установка, представленная на рис. 3, позволяет исследовать влияние дифракции на разрешающую способность оптических инструментов.

Как уже было выяснено, линзы $O_1$ и $O_2$ в отсутствие щели $S_2$ создают в плоскости П изображение щели $S_1$, и это изображение рассматривается в микроскоп М. Таким образом, нашу установку можно рассматривать как оптический инструмент, предназначенный для получения изображения предмета. При этом коллиматор (щель $S_1$ и объектив $O_1$) является моделью далёкого предмета, а объектив $O_2$ и микроскоп М составляют зрительную трубу, наведённую на этот предмет.

Щель $S_2$, установленная непосредственно перед объективом $O_2$, позволяет изменять эффективный размер объектива и, следовательно, разрешающую способность оптической системы.

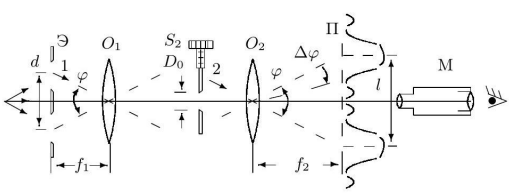


Рис. 5. Схема установки для исследования разрешающей способности оптического инструмента

Поместим вместо щели $S_1$ экран Э с двумя узкими щелями, расстояние между которыми равно $d$ (рис. 5). Тогда расстояние $\ell$ между изображениями щелей в плоскости П равно

$$\ell = \varphi f_2 = d \cdot \frac{f_2}{f_1} \tag{9}$$

а ширина каждого изображения

$$\delta x \approx \frac{\lambda}{b} f_2$$

определяется дифракцией света на щели $S_2$. Когда полуширина дифракционного изображения превышает расстояние между изображениями, то по виду дифракционной картины трудно определить, представляет собой источник двойную или одиночную щель.

Условия, при которых ещё можно различить, имеем мы дело с одной или двумя щелями, для разных наблюдателей различны. Для того чтобы исключить связанный с этим произвол, пользуются обычно критерием Рэлея (см. п. 2.9.3 введения к разделу), который приблизительно соответствует возможностям визуального наблюдения: изображения считаются различимыми, когда максимум одного дифракционного пятна совпадает с минимумом другого, а в условиях нашей задачи - когда полуширина дифракционного изображения $\delta x$ совпадает с расстоянием $\ell$ между изображениями отдельных щелей:

$$\delta x \sim \ell \quad \rightarrow \quad \frac{\lambda}{b} \sim \frac{d}{f_1} \tag{10}$$

**ЗАДАНИЕ**

В этом упражнении предлагается исследовать влияние размера диафрагмы, ограничивающей поперечный размер пучка света, на чёткость изображения объекта.

XI. Настройка и измерения

1. Соберите схему согласно рис. 5. Для этого в предыдущей схеме (рис. 4), не меняя (!) положения линз и микроскопа, вместо щели $S$ поставьте двойную щель и, перемещая её вдоль оси, получите в поле зрения микроскопа чёткое, симметричное изображение двойного источника.

2. Поставьте между линзами щель $S_2$ и, уменьшая её ширину, наблюдайте за ухудшением качества изображения. Подберите ширину щели $S_2$ так, чтобы изображения обеих щелей почти сливались, но всё-таки ещё воспринимались раздельно. Запишите показания микрометрического винта щели $S_2$ (ширину $b_0$).

3. Поставьте двойную щель перед микроскопом и измерьте с помощью микрометрического винта поперечных салазок микроскопа расстояние $d$ между щелями и ширину каждой щели $b$ (сделайте чертёж щели и запишите координату каждой из 4-х вертикалей). Используйте подсветку лампой.

XII. Обработка результатов

4. Для проверки справедливости критерия Рэлея сравните измеренную ширину $b_0$ щели $S_2$ с расчётом по формуле (10): $b_0 = \lambda f_1/d$.

**Контрольные вопросы**

1. Объясните, почему при дифракции Френеля на круглом отверстии в центре картины может быть как тёмное, так и светлое пятно, а при дифракции на препятствии - круглом экране - пятно всегда светлое (пятно Пуассона).

2. Объясните, почему смещение щели $S_2$ в боковом направлении в установке на рис. 3 не приводит к сдвигу дифракционной картины. Будет ли смещаться картина дифракции при смещении щели $S_2$ вдоль оси установки?

3. Получите формулу для распределения интенсивности при дифракции Фраунгофера на двух щелях.

4. Выведите формулу (8). На примере размытия интерференционной картины при широком источнике объясните, что такое пространственная когерентность.

5. Как связана видность интерференционной картины с пространственной когерентностью источника? Чем отличаются графики видности при дискретном и непрерывном изменении параметров источника? То же для временной когерентности.

6. Сформулируйте теорему Бабине.

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 8.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. VIII, 33, 36-38; гл. IX, §§39 – 41, 44, 45.

3. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. III, 28; гл. IV, 39-46.

## 22.3.2 Diffraction of light by an ultrasonic wave in a liquid

Цель работы: изучение дифракции света на синусоидальной акустической решётке и наблюдение фазовой решётки методом тёмного поля.

В работе используются: оптическая скамья, осветитель, два длиннофокусных объектива, кювета с жидкостью, кварцевый излучатель с микрометрическим винтом, генератор ультразвуковой частоты, линза, вертикальная нить на рейтере, микроскоп.

В работе изучается дифракция света на фазовой решётке. Фазовая решётка создаётся в жидкости ультразвуковыми волнами и наблюдается методом тёмного поля. Перед выполнением работы необходимо ознакомиться с теорией дифракции на фазовых структурах (см. п. 2.2, 2.8.2 введения к разделу).

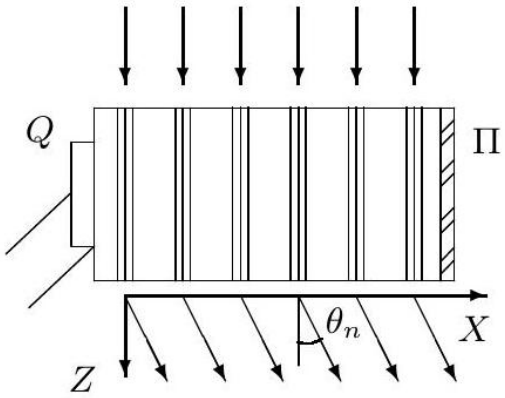



Рис. 1. Дифракция световых волн на акустической решётке

При прохождении ультразвуковой (УЗ) волны через жидкость в ней возникают периодические оптические неоднородности, обусловленные разницей значений коэффициента преломления в областях сжатия и разрежения. Эти периодические неоднородности играют роль своеобразной дифракционной решётки для проходящего сквозь жидкость света.

Пусть УЗ-волна распространяется вдоль оси $X$ (рис. 1) в жидкости, налитой в стеклянную кювету. В направлении оси $Z$ сквозь жидкость проходит световая волна, испытывающая дифракцию на акустической решётке. Поскольку скорость света значительно больше скорости звука, акустическую решётку можно считать неподвижной. Вызванное ультразвуком возмущение показателя преломления жидкости в нашем случае очень мало. При этом естественно сделать предположение (справедливость которого мы потом исследуем теоретически и экспериментально), что акустическую решётку можно рассматривать как тонкий фазовый экран.

При небольших амплитудах звуковой волны показатель преломления жидкости $n$ меняется по закону

$$n = n_0(1 + m\cos\Omega x) \tag{1}$$

где $\Omega$ - волновое число для УЗ-волны ($\Omega = 2\pi/\Lambda$), $\Lambda$ - длина УЗволны, $m$ - глубина модуляции показателя преломления, определяемая интенсивностью ультразвуковой волны ($m \ll 1$).

Пусть фаза световых колебаний на передней поверхности жидкости равна нулю. Тогда на задней поверхности (т.е. в плоскости $z = 0$) она равна

$$\varphi = knL = \varphi_0(1 + m\cos\Omega x) \tag{2}$$

где $L$ - толщина слоя жидкости в кювете, $k$ - волновое число для света ($k = 2\pi/\lambda$), $\lambda$ - длина световой волны, $\varphi_0 = kn_0L$. Таким образом, в плоскости $z = 0$ фаза световых колебаний является периодической функцией координаты $x$, иными словами - УЗ-волна в жидкости создаёт фазовую дифракционную решётку.

Можно сформулировать качественный критерий, при выполнении которого можно считать акустическую решётку чисто фазовой, т.е. рассматривать её как тонкий фазовый экран. Для нашей задачи условие тонкого транспаранта (3.66) можно записать в виде

$$m \ll \frac{\Lambda}{L}\sqrt{\frac{\lambda}{L}} \tag{3}$$

Таким образом, чисто фазовая акустическая решётка реализуется лишь на достаточно слабой УЗ-волне. При повышении мощности ультразвука акустическая волна начинает работать как сложная амплитуднофазовая решётка.

Проведённое в п. 2.2.2 рассмотрение дифракции на фазовой решётке справедливо только в случае слабой фазовой модуляции. В общем случае после прохождения через кювету световое поле представляет совокупность не трёх, а большого числа плоских волн, распространяющихся под углами, определяемыми условием

$$\Lambda \sin\theta_m = m\lambda \quad (m = 0, \pm1, \pm2, \ldots) \tag{4}$$

Каждая из этих волн соответствует одному из максимумов в дифракционной картине Фраунгофера.

Определяя на опыте положение дифракционных максимумов различного порядка, можно по формуле (4) найти длину $\Lambda$ УЗ-волны и вычислить скорость $v$ распространения ультразвуковых волн в жидкости, если известна частота $\nu$ колебаний кварцевого излучателя:

$$v = \Lambda\nu \tag{5}$$

Рассматриваемая теория применима как для бегущих, так и для стоячих ультразвуковых волн. Стоячие УЗ-волны образуются при наложении волны, идущей от излучателя, и волны, отражённой от задней
стенки кюветы. Если же заднюю стенку кюветы покрыть слоем пористой резины (слой П на рис. 1), то волна от неё не отражается, и в кювете образуется практически чистая бегущая волна. Следует иметь в виду, что в стоячей волне амплитуда изменения давления (а следовательно, и коэффициента преломления) больше, чем в бегущей волне, создаваемой тем же излучателем. В связи с этим дифракционная картина в первом случае содержит большее число максимумов.

Экспериментальная установка. Для наблюдения дифракции света на УЗ-волнах на оптической скамье собирается установка, изображённая на рис. 2.

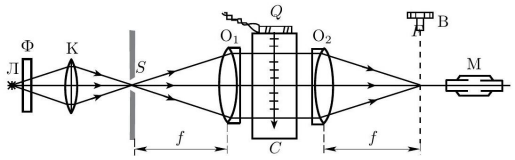



Рис. 2. Схема наблюдения дифракции на акустической решётке

Источник света Л через светофильтр Ф и конденсор К освещает щель $S$, которая расположена в фокусе объектива $O_1$. Выходящий из объектива параллельный пучок света проходит через кювету $C$ перпендикулярно направлению распространения УЗ-волн. Эти волны возбуждаются в жидкости пьезокварцевой пластинкой $Q$, прикреплённой к стенке кюветы. На кварцевую пластинку подаётся напряжение ультразвуковой частоты от генератора (на рис. 2 не показан). В фокальной плоскости второго объектива $O_2$ образуется дифракционная картина, наблюдаемая при помощи микроскопа М. При этом обязательно применяют монохроматическое излучение (красный светофильтр).

Чёткость дифракционных полос зависит от ряда факторов, например, от ширины щели $S$, от её наклона по отношению к вертикали, от угла наклона кюветы к падающему лучу и т.д.

Длина $\Lambda$ ультразвуковой волны определяется с помощью (4); в силу малости углов $\theta_m$ окончательное выражение может быть представлено в виде

$$l_m = mf\frac{\lambda}{\Lambda} \tag{6}$$

где $l_m$ - измеренное на опыте линейное расстояние между $m$-м и нулевым максимумами, а $f$ - фокусное расстояние объектива $O_2$.

Наблюдение оптических неоднородностей, создаваемых ультразвуковыми волнами в жидкости, методом тёмного поля. Попробуем теперь получить видимое изображение фазовой акустической решётки. Для этого, прежде всего, необходимо получить в поле зрения микроскопа изображение задней плоскости (считая по ходу световых лучей) кюветы. Это достигается с помощью вспомогательной положительной линзы О, которую располагают на оптической скамье за фокальной плоскостью объектива $O_2$ (рис. 3).

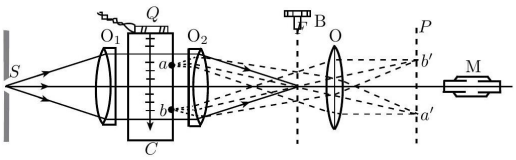



Рис. 3. Наблюдение акустической решётки методом тёмного поля

Перемещая микроскоп вдоль оптической скамьи, фокусируют его на плоскость $P$, где расположено чёткое изображение $a'b'$ какого-либо предмета $ab$, вплотную прижатого к стенке кюветы. Можно ли теперь увидеть в микроскоп УЗ-волну? Ясно, что чисто фазовая акустическая решётка является невидимой (если, конечно, выполнено условие (3)).

Для наблюдения пространственной структуры фазовой решётки можно использовать методы фазового контраста или тёмного поля (см. п. 2.8.2). Фазовую структуру можно сделать видимой при изменении фазы колебаний в центральном дифракционном максимуме на $\pm\pi/2$. Такой метод наблюдения носит название метода фазового контраста. В настоящей работе используется другой способ получения видимого изображения решётки - метод тёмного поля, основанный на устранении центрального дифракционного максимума с помощью специального экрана. Как нетрудно показать, в поле зрения микроскопа будут наблюдаться чередующиеся светлые и тёмные полосы, причём расстояние между тёмными полосами соответствует смещению в плоскости кюветы на $\Lambda/2$. Таким образом, должно наблюдаться характерное для метода тёмного поля удвоение числа деталей рассматриваемой структуры.

Этот опыт можно проводить только со стоячими волнами, т.к. в случае бегущей волны визуальное наблюдение оказывается невозможным: глаз не успевает следить за быстро перемещающейся волной.

## A. Установка с вертикальной щелью

### ЗАДАНИЕ

В работе предлагается измерить координаты дифракционных полос, образующихся при дифракции света на акустической решётке, а также определить период этой решётки методом тёмного поля. По результатам измерений рассчитывается скорость ультразвука в воде. Все измерения ведутся на стоячей волне.

### I. Определение скорости ультразвука по дифракционной картине

1. Соберите схему согласно рис. 2. Для получения параллельного пучка установите щель $S$ на фокусном расстоянии от главной плоскости объектива $O_1$ (положение главной плоскости и фокусное расстояние объектива указаны на установке).

2. Ярко осветите щель с помощью конденсора. Предварительную настройку можно вести с зелёным фильтром.

С помощью листа бумаги найдите резкое изображение щели $S$ в фокальной плоскости $F$ объектива $O_2$. Перемещая микроскоп вдоль оптической скамьи, сфокусируйте его на резкое изображение щели. Предметная плоскость микроскопа расположена на расстоянии $2-4$ см от его объектива.

Если окулярная шкала видна нерезко, следует вынуть окуляр из тубуса и, направив на любой освещённый объект, вращением глазной линзы подстроить его под свой глаз. Вернув окуляр на место, снова найдите резкое изображение щели.

3. Включите генератор УЗ-частоты и частотомер. Плавно изменяя частоту генератора в диапазоне 1-1,5 МГц, получите в поле зрения микроскопа систему дифракционных полос.

4. Замените широкополосный зелёный фильтр красным. Изменяя ширину щели $S$, её наклон и положение конденсора, добейтесь оптимальных условий наблюдения дифракционных полос.

5. Проследите, как изменяется число дифракционных полос при изменении мощности ультразвука.

6. Найдите чёткую дифракционную картину и, перемещая излучатель с помощью микрометрического винта, оцените по порядку величины длину УЗ-волны как удвоенное расстояние между наиболее чёткими дифракционными картинами.

7. Измерьте положения $x_m$ шести-восьми дифракционных максимумов с помощью поперечного микрометрического винта микроскопа. Один оборот винта)100 дел.) соответствует перемещению в 1 мм.

Не рекомендуется использовать для этих измерений окулярную шкалу, т.к. цена деления её неизвестна, а калибровка шкалы с помощью винта поперечного смещения микроскопа даёт большую погрешность.

Определите рабочую частоту по показаниям частотомера.

8. Повторите измерения п. 7 для $3-4$-х частот в одном из диапазонов: вблизи одного, 4 -х или 6 -ти МГц. Отключите сигнал генератора, не выключая его из сети.

9. Запишите фокусное расстояние объектива $O_2$ и полосу пропускания светофильтра, указанные на установке.

10. Для каждой частоты постройте график зависимости координаты $x_m$ от порядка $m$ и по наклону прямой определите расстояние между соседними полосами $l_m/m = \Delta x_m/\Delta m$. Рассчитайте длину $\Lambda$ УЗ-волны по формуле (6).

11. Рассчитайте скорость ультразвука в воде по формуле (5). Оцените погрешность эксперимента и сравните результат с табличным.

**II. Определение скорости ультразвука методом тёмного поля**

12. Для перехода к методу тёмного поля (рис. 3), не смещая микроскоп, поставьте рейтер с вертикальной нитью в фокальной плоскости $F$ объектива $O_2$ и, перемещая рейтер вдоль оптической скамьи, найдите резкое изображение нити. Оно должно совпадать с резким изображением щели.

Не смещая рейтер с нитью, отодвиньте микроскоп и поставьте дополнительную линзу между нитью и микроскопом.

13. К задней (по ходу луча) стенке кюветы прижмите стеклянную пластинку с миллиметровыми делениями. Откройте пошире входную щель. С помощью листа бумаги найдите плоскость, в которой располагается резкое изображение линейки, созданное двумя линзами. Передвигая микроскоп, сфокусируйте его на изображение пластинки.

14. Определите цену деления окулярной шкалы в условиях опыта. Для этого совместите самые дальние из хорошо видимых в поле зрения миллиметровых делений пластинки с делениями окулярной шкалы и запишите количество тех и других делений.

Вместо калибровки окулярной шкалы можно найти соответствие между делениями поперечного микрометрического винта и линейки в кювете.

Проведённая калибровка соответствует определённому положению микроскопа, поэтому не следует перемещать микроскоп вдоль скамьи до конца эксперимента.

15. Уберите пластинку из кюветы и уменьшите ширину входной щели. Включите генератор и попытайтесь увидеть звуковую решётку. Если решётка видна без применения метода тёмного поля (при открытом нулевом максимуме), значит, она не чисто фазовая, а сложная амплитудно-фазовая. В этом случае приведённая выше теория плохо применима. Уменьшая мощность излучателя, добейтесь исчезновения видимого изображения решётки в поле зрения микроскопа.

16. Закройте центральный дифракционный максимум вертикальной нитью. Удобно устанавливать нить в отсутствие УЗ-сигнала. В этом случае нет никаких максимумов, кроме нулевого (изображения щели), так что правильно установленная нить должна полностью затемнять поле зрения (отсюда название метод тёмного поля).

17. Включите генератор и найдите изображение акустической решётки.

18. Измерьте длину УЗ-волны в воде. Для этого с помощью окулярной шкалы (или микрометрического винта) измерьте расстояние между самыми дальними из хорошо видимых в поле зрения тёмных (или светлых) полос и просчитайте число промежутков между ними.

19. В заключение работы проделайте качественные эксперименты. Перемещая нить, закрывайте последовательно максимумы первого, второго и т.д. порядков и наблюдайте за изменениями картины звукового поля.

Убрав нить, постепенно закрывайте дифракционные максимумы с помощью края экрана (лезвия бритвы).

Объясните изменения в картине звукового поля.

20. Рассчитайте длину УЗ-волны $\Lambda$ с учётом удвоения числа деталей, наблюдаемых методом тёмного поля. Определите скорость ультразвука в воде и оцените погрешность эксперимента.

### Б. Установка с горизонтальной щелью

Источник света Л (рис. 4) с помощью конденсора К проецируется на входную (коллиматорную) щель $S$. Входная щель ориентирована горизонтально и прикрыта красным светофильтром Ф. Коллиматорный объектив $O_1$ посылает параллельный пучок на кювету с водой $C$.

Излучатель $Q$, погружённый в кювету, создаёт УЗ-волну. Вертикальное перемещение излучателя осуществляется винтом I (рис. 5), тонкая подача - лимбом II. При определённых положениях излучателя волна становится стоячей.

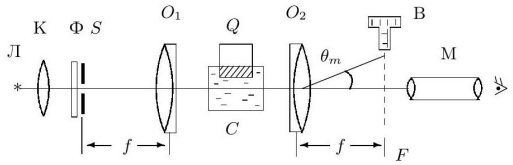



Рис. 4. Схема для наблюдения дифракции на акустической решётке

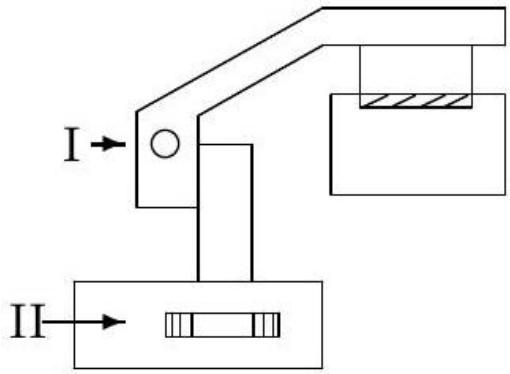



Рис. 5. Устройство для вертикального перемещения излучателя

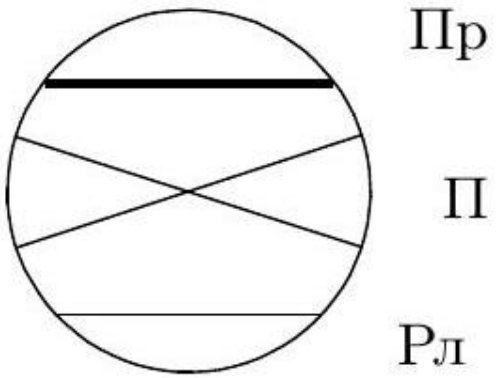



Рис. 6. Проволока Пр, перекрестие П и реперная линия Рл в фокальной плоскости объектива $O_2$

Параллельный пучок света, дифрагируя на стоячей звуковой волне, образует дифракционную картину в фокальной плоскости $F$ (рис. 4) камерного объектива $O_2$. Картину можно наблюдать в микроскоп М.

Дифракционные полосы ориентированы горизонтально. Расстояние между ними можно измерить с помощью микрометрического винта В. Винт передвигает размещённые на стекле (рис. 6) в плоскости $F$ перекрестие П, тонкую реперную линию Рл и толстую проволоку Пр, которая используется в методе тёмного поля.

## ЗАДАНИЕ

В работе предлагается измерить координаты полос, образующихся при дифракции света на акустической решётке, а также определить период этой решётки методом тёмного поля. По результатам измерений рассчитывается скорость ультразвука в воде. Все измерения ведутся на стоячей волне.

III. Определение скорости ультразвука по дифракционной картине

1. а) Соберите схему согласно рис. 4. Включите осветитель. Максимально откройте входную щель.
   б) Поместив лист бумаги между коллиматором и кюветой, убедитесь, что световое пятно равномерно освещено; затем проверьте освещённость пятна на выходе из прибора. В случае необходимости отцентрируйте источник света и конденсор (см. Техническое описание - ТО, «Юстировка», пп. 1 и 2).

в) Настройте микроскоп и отсчётное устройство (см. ТО, пп. 3 и 4).
Установите рабочую ширину щели 20-30 мкм.

2. Получите дифракционную картину (см. ТО, п. 5). Картина видна наиболее чётко, когда в кювете образуется стоячая УЗ-волна.

Перемещая излучатель с помощью лимба (рис. 5), оцените по порядку величины длину УЗ-волны как удвоенное расстояние между наиболее чёткими дифракционными картинами. Цена деления лимба 10 мкм, один оборот - 50 делений. Максимальное перемещение излучателя - 2 мм.

3. Определите положения дифракционных полос. Это удобно делать в следующем порядке:

а) увидев дифракционную картину, установите частоту генератора на ближайшее фиксированное значение (например, 1,2; 1, 3 МГц и т. д. для удобства отсчёта);
б) вращением лимба II (рис. 5) добейтесь наилучшей картины: в поле зрения должно быть 8-10 дифракционных полос;
в) с помощью перекрестия П и микрометрического винта, установленного на выходе из прибора, определите координату $Y$ (в делениях винта) каждой светлой полосы (окулярная шкала микроскопа в этом упражнении не используется). Запишите соответствующую частоту.

4. Повторите измерения п. 3 для трёх-четырёх фиксированных частот, лежащих в одном из диапазонов (см. рис. 2-ТО).

5. Для подготовки установки к следующему упражнению, глядя в окуляр, закройте проволочкой (Пр на рис. 6) центральный максимум (изображение щели) и запишите показания винта, перемещающего проволочку. Это облегчит Вам настройку методом тёмного поля.

6. Постройте на одном листе графики $Y = Y(m)$ (от $-m$ до $+m$). Для каждой частоты определите по наклону прямой расстояние между соседними полосами (цена деления микрометрического винта - 4 мкм). Зная фокусное расстояние объектива $O_2$($f = 28$ см) и полосу пропускания красного фильтра ($\lambda = 6400 \pm 200$Å), рассчитайте длину УЗ-волны $\Lambda$ по формуле (6). Рассчитайте скорость звука для каждой частоты по формуле (5) и среднюю скорость.

7. Оцените погрешность эксперимента. Сравните результат с табличным.

### IV. Определение скорости ультразвука методом тёмного поля

8. Для перехода к методу тёмного поля отодвиньте микроскоп от щели и разместите в промежутке между ними дополнительную линзу (рис. 3).

Поднимите излучатель над кюветой, опустите в воду квадратную сетку (сторона квадрата - 1 мм) и прижмите её к задней по ходу луча стенке кюветы.

Расширьте входную щель, чтобы увеличить освещённость поля зрения микроскопа. Найдите изображение сетки на листе бумаги (в плоскости $P$) и убедитесь, что свет попадает на объектив микроскопа. Передвигая микроскоп вдоль оптической оси, настройтесь на резкое изображение сетки в плоскости $P$. Центрируя линзу О по высоте и вокруг горизонтальной и вертикальной осей, добейтесь того, чтобы чётко были видны как горизонтальные, так и вертикальные штрихи сетки. Закрепите микроскоп.

Окулярная шкала микроскопа ориентирована вертикально. Используя как можно большую часть поля зрения микроскопа, зафиксируйте координаты совпадающих штрихов окулярной шкалы и сетки.

Проведённая калибровка соответствует определённому положению микроскопа, поэтому не следует перемещать микроскоп вдоль скамьи до конца эксперимента.

Рассчитайте цену малого деления окулярной шкалы в этом эксперименте.

9. Для наблюдения акустической решётки установите рабочую ширину щели (20-30 мкм). Уберите калибровочную сетку из кюветы, опустите туда излучатель и, варьируя частоту, постарайтесь увидеть звуковую решётку в микроскоп. Если решётка видна при открытом центральном максимуме, значит, она сложная амплитудно-фазовая. Уменьшая мощность ультразвука, добейтесь исчезновения видимого изображения решётки.

10. Закройте нулевой дифракционный максимум проволочкой (см. упр. I, п. 5). Это можно сделать, глядя мимо микроскопа на стекло в фокальной плоскости объектива $O_2$ (рис. 3). При этом проволочка закрывает изображение щели, и поле зрения микроскопа затемняется.

11. Меняя частоту, наблюдайте акустическую решётку. Убедитесь, что при удалении проволочки с главного максимума решётка не видна.

12. Определение периода звуковой решётки удобно вести при фиксированных частотах, настраивая чёткую картину перемещением излучателя (см. упр. I, п. 3).

Зафиксируйте с помощью окулярной шкалы микроскопа координаты первой и последней из хорошо видимых в поле зрения тёмных полос и количество светлых промежутков между ними.

13. Повторите измерения п. 12 для 6 -8-ми частот внутри одной из рабочих полос УЗ-излучателя (рис. 2-ТО).

14. Закрывая проволочкой последовательно 1 -й, 2-й, -1 -й, -2 -й максимумы, наблюдайте за изменением картины. Объясните явление.

15. Для каждой частоты рассчитайте длину УЗ-волны $\Lambda$ с учётом удвоения числа наблюдаемых полос.

16. Постройте график $\Lambda = F(1/\nu)$ и определите по наклону прямой скорость ультразвука в воде.

Оцените погрешность эксперимента, сравните результаты с теоретическими.

**Контрольные вопросы**

1. Покажите, что период акустической решётки совпадает с длиной бегущей ультразвуковой волны как в случае бегущих, так и в случае стоячих волн. Покажите, что формула (4) справедлива для любой периодической решётки, независимо от того, является ли она амплитудной или фазовой.

2* Как изменяется во времени интенсивность света в дифракционных максимумах разного порядка при дифракции света на бегущих и на стоячих ультразвуковых волнах?

3. В чём отличие метода тёмного поля от метода фазового контраста?

4. Сформулируйте экспериментальный критерий того, что акустическая решётка является чисто фазовой.

**Литература**

1. Ландсбер» Г.С. Оптика. - М.: Физматлит, 2003. Гл. X, 56.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. IV, 59.

3. Кингсеп А.С., Локшин Г.Р., Ольхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. §8.8.

**Работа 4.3.3**

## 22.3.3 Investigation of Microscope Resolving Power by the Abbe Method

Цель работы: определение дифракционного предела разрешения объектива микроскопа методом Аббе.

В работе используются: лазер; кассета с набором сеток разного периода; линзы; щель с микрометрическим винтом; оптический стол с набором рейтеров и крепёжных винтов; экран; линейка.

Перед выполнением работы необходимо ознакомиться со спектральным методом задачи дифракции (см. п. 2 введения к разделу).

Всякая оптическая система, предназначенная для получения изображений, имеет конечный предел разрешения, т. е. ограниченную возможность раздельного наблюдения близко расположенных предметов. Принципиальной причиной, ограничивающей предел разрешения, является дифракция световых волн. Разрешающей способностью оптического прибора называют минимальное расстояние $\ell_{\min}$ между двумя точками в пространстве предметов, которое прибор может разрешить. При визуальном наблюдении изображения в качестве критерия разрешения применяют так называемый критерий Рэлея (см. п. 2.9 введения).

Для иммерсионного микроскопа (объект находится в иммерсионной среде - жидкости с показателем преломления $n$) разрешающая способность объектива при некогерентном освещении

$$\ell_{\min} \approx \frac{0,61\lambda}{n \sin A}$$

где $A$ - апертурный угол объектива микроскопа (см. рис. 1), т. е. угол между оптической осью и лучом, направленным из центра объекта в край линзы (напомним, что при наблюдении в микроскоп объект занимает небольшой участок, располагающийся вблизи оптической оси объектива). Вывод этой формулы предоставляем читателю.

Рассмотрим теперь когерентно освещённый объект, наблюдаемый в микроскоп. Схема образования изображения в объективе микроскопа представлена на рис. 1. Для простоты рассмотрим случай, когда предметом является периодическая структура (дифракционная решётка), освещаемая параллельным пучком лучей. При наблюдении в микроскоп предмет располагается вблизи переднего фокуса объектива. При освещении решётки волнами, наклонными к оси, с углом наклона чуть

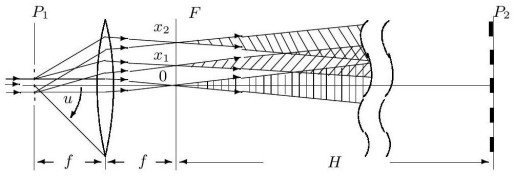



Рис. 1. Образование изображения в объективе микроскопа. $P_1$ - плоскость предмета, $F$ - задняя фокальная плоскость объектива, $P_2$ - плоскость, сопряжённая с предметной плоскостью. В плоскости $P_2$ световые пучки сильно перекрываются

меньшим апертуры $A$ волны нулевого порядка сфокусируются на край диафрагмы. Для получения изображения достаточно, чтобы на противоположные края сфокусировались волны 1-го порядка, т. е. угол между волнами 0 и 1 порядка должен быть равен $2A$ (см. п. 2.10). Согласно формуле (3.82), минимальное разрешаемое объективом расстояние определяется условием

$$\ell_{\min} \approx \frac{\lambda}{\sin A} \approx \frac{\lambda}{D/2f} \tag{1}$$

где $D$ - диаметр диафрагмы. При этом диафрагма, расположенная симметрично, пропускает нулевой и $\pm 1$ дифракционные максимумы.

В нашей работе применяется двумерная решётка - сетка. Её можно рассматривать как две скрещенные (перпендикулярные друг к другу) решётки. Узкий пучок монохроматического света, пройдя через решётку с вертикальными штрихами, даёт совокупность максимумов, расположенных вдоль горизонтальной линии. Световой пучок, соответствующий каждому максимуму, проходя через вторую решётку, распадается на новую совокупность световых пучков, дающих максимумы вдоль вертикальной линии. Главные максимумы возникают тогда, когда одновременно выполняются условия:

$$d \sin \theta_x = m_x \lambda, \quad d \sin \theta_y = m_y \lambda \tag{2}$$

где $m_x$ и $m_y$ - целые числа, характеризующие порядки дифракционных максимумов, $\theta_x$ и $\theta_y$ - направления на главные дифракционные максимумы в горизонтальной и вертикальной плоскостях соответственно.

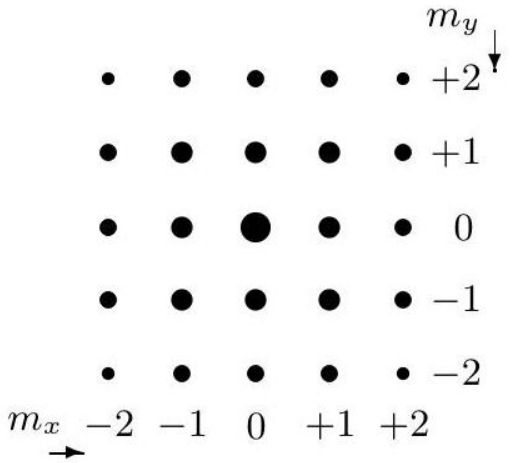



Рис. 2. Дифракция Фраунгофера на двумерной решётке (сетке). Максимумы изображены кружками, размеры которых характеризуют интенсивности

Максимумы, удовлетворяющие условию $\theta_x, \theta_y < u$, создают в задней фокальной плоскости $F$ объектива картину дифракции Фраунгофера (рис. 2) - первичное изображение.

Если теперь поместить в фокальной плоскости вертикальную щель так, чтобы через неё проходили дифракционные максимумы с $m_x = 0$ и $m_y = 0, \pm 1, \pm 2, \ldots$, то в плоскости $P_2$ получится изображение решётки с горизонтально расположенными штрихами. Если, наоборот, пропустить максимумы с $m_y = 0$ и $m_x = 0, \pm 1, \pm 2, \ldots$, то в $P_2$ получится изображение решётки с вертикальными штрихами. Таким образом можно продемонстрировать явление пространственной филътраиии выделение различных структур в изображении (см. п. 2.8 введения к разделу).

Экспериментальная установка. Схема модели проекционного микроскопа приведена на рис. 3. Предметом служат сетки, расположенные в кассете. Смена сеток осуществляется поворотом внешнего кольца кассеты.

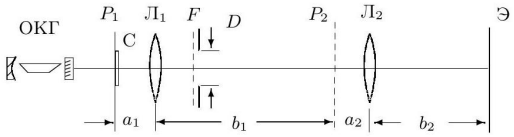



Рис. 3. Схема экспериментальной установки - модель проекционного микроскопа

Излучение лазера (ОКГ) почти перпендикулярно падает на сетку С, установленную вблизи фокальной плоскости линзы $\Omega_1$ - объектива
микроскопа. Обычно и объектив, и окуляр микроскопа - короткофокусные линзы (1-3 см). В нашей модели линза $\Omega_1$ выбирается достаточно длиннофокусной ($f \approx 10$ cm), т. к. размер первичного изображения в фокальной плоскости $F$ должен быть не слишком малым, чтобы дополнительными диафрагмами можно было влиять на вторичное изображение в плоскости $P_2$. Вторичное изображение из плоскости $P_2$ проецируется на экран Э линзой $\Omega_2$ (короткофокусной, чтобы изображение на экране было крупнее). Во избежание микротравм глаза от излучения лазера не следует использовать эту линзу традиционным образом как окуляр микроскопа.

Изображение сетки периодически повторяется - репродуиируется - в пространстве между сеткой и первой линзой, поэтому для того, чтобы среди множества репродуцированных изображений сетки можно было выделить её геометрическое изображение, на одну из сеток наложена тонкая проволочка, т. е. непериодический объект, изображение которого не репродуцируется.

В фокальной плоскости $F$ могут быть установлены диафрагмы щелевая или ирисовая (отверстие с переменным диаметром) и различного рода маски (препятствия).

Как видно из соотношения (1), минимально разрешимый шаг решётки или сетки определяется апертурным углом $A$ объектива. Обычно апертура микроскопа меняется при помощи ирисовой диафрагмы на объективе (на линзе $\Omega_1$ такая диафрагма есть), но в наших условиях удобнее располагать щелевую диафрагму в плоскости $F$. Имея набор сеток с различными периодами $d$ и изменяя апертурный угол объектива с помощью щелевой диафрагмы, можно экспериментально проверить соотношение (1).

В нашей работе период сеток рассчитывается двумя способами: в первом способе (по дифракции Фраунгофера) расстояние между дифракционными максимумами на экране измеряется при помощи линейки, а затем по формулам (2) определяется её период; во втором способе период определяется по увеличенному (с помощью модели микроскопа) изображению сетки на экране.

С помощью откалиброванных таким образом сеток определяется разрешающая способность микроскопа. Для этого в задней фокальной плоскости $F$ объектива устанавливается щелевая диафрагма с микрометрическим винтом и подбирается её минимальный размер, при котором ещё видно изображение сетки на экране (щель пропускает максимумы с $m = 0, \pm 1$). По размеру диафрагмы и фокусному расстоянию объектива рассчитывается апертурный угол $u$ и проверяется соотношение (1).

Для выполнения последней части работы ширина вспомогательной щели и угол наклона к оси системы подбираются так, чтобы на экране вместо изображения сетки получалось изображение решётки, расположенной наклонно, т. е. осуществлялась пространственная фильтраиил. Если сетку и щель поменять местами, то соответствующим подбором сетки можно «рассечь» первичное изображение так, что изображение щели на экране будет многократно повторяться - мультиплицироваться.

## ЗАДАНИЕ

В работе предлагается определить периоды сеток сначала по их спектру на удалённом экране, затем по увеличенному с помощью модели микроскопа изображению сеток на экране и, наконец, по результатам измерения разрешающей способности микроскопа; наблюдать явления саморепродукции, пространственной фильтрации и мультиплицирования.

**Внимание! Попадание прямого лазерного луча на сетчатку глаза крайне опасно!**

### I. Определение периода решёток по их пространственному спектру

1. Включите в сеть блок питания лазера. Из-за переотражения в выходных стеклах лазера на его выходе может быть несколько лучей. Следите за лучом с максимальной яркостью. Яркость зелёного лазера можно ослабить с помощью светофильтра.

2. Закрепите кассету с двумерными решётками (сетками) вблизи выходного окна лазера. В каждой кассете 6 различных решёток.

Вращая наружное кольцо кассеты, пронаблюдайте на удалённом экране дифракционные картины для разных сеток.

Для определения расстояния между соседними дифракционными максимумами измерьте расстояние между удалёнными друг от друга максимумами (горизонтальными или вертикальными) и число промежутков между ними.

Проведите измерения для шести разных сеток и укажите их номера.

3. Измерьте расстояние $H$ от сетки до экрана.

Запишите длину волны лазера, указанную на установке.

II. Определение периода решёток по изображению, увеличенному с помощью модели микроскопа

4. Соберите модель проекционного микроскопа (рис. 3). При настройке добивайтесь центрировки системы: световой пучок, прошедший одну
линзу, и пучок, прошедший обе линзы, должны попадать в то же место экрана, куда падал луч от лазера в отсутствие линз.

Длиннофокусная линза $\Omega_1 (f \approx 10$ см) должна отстоять от сетки на расстоянии, чуть большем фокусного, при этом изображение сетки в плоскости $P_2$ получится на расстоянии, в несколько раз превышающем фокусное. Найдите это изображение на листе бумаги, затем с помощью короткофокусной линзы $_2$ получите резкое изображение сетки на экране.

Для того чтобы выделить изображение сетки, соответствующее законам геометрической оптики (для устранения влияния саморепродукции), найдите резкое изображение проволочки, расположенной на одной из сеток.

5. Определите расстояния $a_1, b_1, a_2, b_2$ (рис. 3), учитывая следующие факторы: положение центров линз указаны на оправах проточками; расстояние $a_2 \simeq f_2$, поэтому положение промежуточного изображения можно не искать, а измерять сразу $(b_1 + a_2)$ - «длину тубуса» микроскопа.

Так как для одной линзы увеличение равно $b/a$, то для всей системы оно составляет $b_1 b_2 / (a_1 a_2)$.

Измерьте периоды изображений сеток на экране.

**III. Определение периодов решёток по оценке разрешающей способности микроскопа**

6. Поместите щелевую диафрагму с микрометрическим винтом в фокальную плоскость $F$ линзы $_1$. Определите для каждой решётки минимальный размер диафрагмы $D$, при котором на экране ещё видно изображение сетки (при меньших размерах щели изображение выглядит как одномерная решётка).

**IV. Пространственная фильтрация и мультиплицирование**

7. Проделайте качественный опыт по пространственной фильтрации. Подберите сетку средних размеров с достаточно крупным вторичным изображением. Ширину щели подберите так, чтобы она свободно пропускала максимум нулевого порядка и не пропускала максимумы первого порядка, расположенные в поперечном направлении. Поворачивая щель относительно оси системы, получите изображения решёток при различных ориентациях щели: для вертикального положения щели, когда она пропускает только дифракционные максимумы $(0, m_x)$, для горизонтального положения, когда проходят только дифракционные максимумы $(m_y, 0)$, и для наклонного положения под углом 45°, когда пропускаются максимумы с $m_x = m_y$. В этом случае на экране видна наклонная решётка - периодическая структура, которой нет в исходном объекте.

Объясните наблюдаемые явления.

8. Для наблюдения явления мультиплицирования поменяйте местами сетку C и щель $D$ (рис. 3): сначала, не трогая линз, получите на экране резкое изображение щели, а затем в фокальной плоскости $F$ объектива поставьте кассету с сетками, которые будут «рассекать» фурье-образ щели.

Подберите такую ширину входной щели $D$, чтобы на экране можно было наблюдать мультиплицированное изображение для всех сеток. Чем уже щель, тем шире её фурье-образ и тем легче «рассечь» его сетками.

Проследите, как влияет на изображение уменьшение ширины щели при неизменной сетке и что происходит при смене сеток при постоянной ширине щели. Объясните явления.

**V. Обработка результатов**

1. По измерениям спектров определите дифракционные углы $\theta_x, \theta_y$ и рассчитайте периоды решёток.

2. По измерениям увеличенных с помощью микроскопа изображений сеток рассчитайте их периоды и сравните с результатами, полученными по дифракционной картине.

3. По измерениям со щелью рассчитайте по формуле (1) минимальное расстояние (период решётки $d$), разрешаемое микроскопом, и сравните с результатами предыдущих измерений.

4. Для проверки теории Аббе постройте график зависимости $d = f(1/D)$, взяв периоды сеток, определённые по спектру. Проверьте справедливость формулы (1).

5. Зарисуйте и опишите результаты качественных опытов по пространственной фильтрации и мультиплицированию.

**Контрольные вопросы**

1. Какому условию должно удовлетворять расстояние между сеткой и экраном, чтобы на экране наблюдалась дифракционная картина Фраунгофера?

2. Почему при наблюдении дифракции на удалённом экране дифракционные максимумы в случае крупных сеток перекрываются?

3. Почему предмет располагается вблизи передней фокальной плоскости объектива микроскопа?

4. Почему при смещении сетки в поперечном направлении первичное изображение не меняется, а вторичное смещается?

5. Какой пространственный фильтр следует расположить в задней фокальной плоскости объектива, чтобы получить в плоскости $P_2$ изображение сетки с уменьшенными в два раза линейными размерами ячеек?

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. XIV, 92; гл. XV, 96, 97.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. IV, 56, 57.

3. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 8, §8.10.

4. Дитчберн Р. Физическая оптика. - М.: Наука, 1965. Гл. 8, 8.19-8.29.

## 22.3.4 Fourier Transform Method in Optics

Цель работы: исследование особенностей применения пространственного преобразования Фурье для анализа дифракционных явлений.

В работе используются: гелий-неоновый лазер, кассета с набором сеток разного периода, щель с микрометрическим винтом, линзы, экран, линейка.

Анализ сложного волнового поля во многих случаях целесообразно проводить, разлагая его на простейшие составляющие, например, представляя его в виде разложения по плоским волнам. При этом оказывается, что если мы рассматриваем поле, полученное после прохождения плоской монохроматической волны через предмет или транспарант (изображение предмета на фотоплёнке или стеклянной пластинке) с функцией пропускания $t(x)$, то разложение по плоским волнам соответствует преобразованию Фурье от этой функции. Если за предметом поставить линзу, то каждая плоская волна сфокусируется в свою точку в задней фокальной плоскости линзы. Таким образом, картина, наблюдаемая в фокальной плоскости линзы, даёт нам представление о спектре плоских волн падающего на линзу волнового поля. Поэтому можно утверждать, что с помощью линзы в оптике осуществляется пространственное преобразование Фурье (подробнее см. п. 2 введения к разделу).

Экспериментальная установка. Схема установки представлена на рис. 1. Щель переменной ширины $D$, снабжённая микрометрическим винтом В, освещается параллельным пучком света, излучаемым лазером (радиус кривизны фронта волны велик по сравнению с фокусными расстояниями используемых в схеме линз).

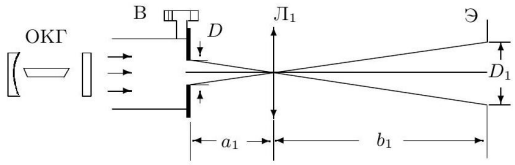



Рис. 1. Схема для определения ширины щели с помощью линзы

Увеличенное изображение щели с помощью линзы $_1$ проецируется на экран Э. Величина изображения $D_1$ зависит от расстояний от линзы до предмета $-a_1$ и до изображения $-b_1$, т. е. от увеличения Г системы:

$$\Gamma = \frac{D_1}{D} = \frac{b_1}{a_1} \tag{1}$$

Изображение спектра щели образуется в задней фокальной плоскости Ф линзы $\Omega_1$. Размещая в плоскости Ф двумерные решётки-сетки, можно влиять на первичное изображение и получать мультиплицированное изображение щели.

Убрав линзу, можно наблюдать на экране спектр щели (рис. 2), а если заменить щель решёткой - спектр решётки. Крупные решётки дают на экране очень мелкую картину спектра, которую трудно промерить. В этом случае используют две линзы (рис. 3): первая (длиннофокусная) формирует первичное изображение - спектр, вторая (короткофокусная) - проецирует на экран увеличенное изображение спектра.

## ЗАДАНИЕ

В работе предлагается: А) определить размеры щели сначала по увеличенному с помощью линзы изображению, затем - по спектру на экране; Б) определить периоды сеток сначала по спектру, затем по увеличенному изображению спектра; В) исследовать изображение щели, мультиплицированное с помощью сеток; Г) проследить влияние щелевой диафрагмы, расположенной в фурье-плоскости, на изображение сетки.

**Внимание! Попадание прямого лазерного луча на сетчатку глаза крайне опасно!**

### А. Определение ширины щели

#### I. Определение ширины щели по увеличенному изображению

1. Включите в сеть блок питания лазера. Обратите внимание на распределение интенсивности излучения лазера на экране: на его сложную структуру, обусловленную возбуждением различных типов колебаний в резонаторе лазера.

2. Установите тубус со щелью вплотную к выходному окну лазера (рис. 1). Настройку системы следует вести, наблюдая за пятном света на листе бумаги.

3. С помощью короткофокусной линзы $\Omega_1$ ($F_1 \simeq 3-4$ см) получите на экране Я увеличенное изображение щели. Фокусы линз указаны на оправах.

4. Определите начало отсчёта ширины щели по её открытию, т. к. ноль может быть сбит. Цена деления винта - 10 мкм.

5. Меняя ширину щели от 50 до 500 мкм (5-50 делений от нового нуля), снимите зависимость размера изображения $D_1$ от ширины щели $b$. Изменение ширины щели следует вести в сторону увеличения, чтобы исключить влияние люфта (свободного хода винта).

6. Измерьте расстояния $a_1$ и $b_1$ для определения увеличения Г системы. Погрешность этих измерений велика (особенно для малого расстояния $a_1$), поэтому, измерив дополнительно $L = a_1 + b_1$ и зная $F_1$, можно по формуле линзы рассчитать $a_1$ и $b_1$ и сравнить с измеренными. Для удалённого экрана расчёт даёт обычно $a_1 \simeq F_1$.

7. Зная увеличение линзы и размер изображения, рассчитайте по формуле (1) ширину входной щели $D_{\text{л}}$ («л» - с применением линзы).

#### II. Определение ширины щели по её спектру

8. Получите на удалённом экране спектр щели (рис. 2). Меняя ширину щели, проследите за изменением спектра на экране и оцените интервал, для которого можно наблюдать и измерять спектр.

9. Измерьте ширину спектра для самой маленькой щели. Для большей точности следует измерять расстояние $X$ между удалёнными от центра минимумами, расположенными симметрично относительно центра картины, и отмечать порядок минимума $m$ (например, 1 -й, 2 -й, 3 -й, ..., начиная от центра).

Проведите серию измерений $X(m)$, меняя ширину щели в тех же пределах, что и в п. 5.

Измерьте расстояние $L$ от щели до экрана.

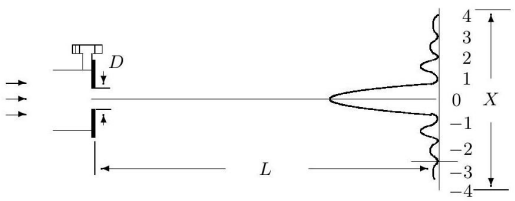



Рис. 2. Схема для определения ширины щели по спектру

10. По результатам измерений спектра рассчитайте ширину щели $b_{\text{с}}^{(\text{«с»}}$ по спектру), используя соотношения

$$\Delta X = \frac{X}{2m} = \frac{\lambda}{b_{\text{с}}} L \tag{2}$$

Длина волны $\text{He} - \text{Ne}$-лазера $\lambda = 6328\text{Å}$.

Постройте на одном листе графики $b_{\text{л}} = f(b)$ и $b_c = f(b)$.

Обратите внимание, что по спектру можно определить размер малой щели, который не может быть достаточно точно определён с помощью линзы, т. к. размер изображения щели на экране $2 - 3$ мм.

11. Если поднести волос с выходному окну лазера, то по дифракционной картине на экране можно определить его диаметр. Проведите измерения и расчёт так же, как для щели.

### Б. Определение периода решёток

III. Определение периода по спектру на удалённом экране

12. Поставьте кассету с двумерными решётками (сетками) вплотную к выходному окну лазера. В окошке под отверстием с сеткой виден № сетки. Вращением наружного кольца кассеты можно менять сетки.

13. Для каждой сетки измерьте расстояние $X$ между $m$-ми максимумами и отметьте $m$ - порядок максимума.

Измерьте расстояние $L$ от кассеты до экрана.

14. Рассчитайте расстояния $\Delta X$ между соседними максимумами и определите период каждой решётки $d_{\text{с}} = f($ №$)$, используя соотношения

$$\Delta X = \frac{X}{2m} = \frac{\lambda}{d_{\text{с}}} L \tag{3}$$

Для крупных решёток спектры промерить не удаётся - они слишком узки. Их можно увеличить с помощью системы линз.

IV. Определение периода решёток по увеличенному изображению спектра

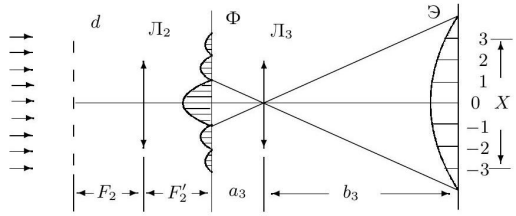


Рис. 3. Схема определения периода решётки по увеличенному изображению спектра

15. Линзу $\Omega_2$ с максимальным фокусом ($F_2 \simeq 10$ см) поставьте на расстоянии $\simeq F_2$ от кассеты. В плоскости Ф линза $\Omega_2$ даёт фурье-образ сетки её спектр, а короткофокусная линза $\Omega_3\,(F_3 \simeq 2,5$ см) создаёт на экране увеличенное изображение этого спектра.

Так как экран достаточно удалён ($b_3 \gg a_3$), то практически $a_3 = F_3$, и расстояние между линзами $\simeq F_2 + F_3$.

16. Измерьте $X$ и $m$ для всех сеток, где это возможно.

17. Зная увеличение линзы $\Omega_3\,(\Gamma_3 = b_3/a_3)$, можно рассчитать расстояние между максимумами $\Delta x$ в плоскости Ф, а затем период сетки $d$:

$$\Delta x = \frac{\Delta X}{\Gamma_3} = \frac{\lambda}{d_\pi} F_2. \tag{4}$$

B. Пространственное преобразование спектров

**V. Мультиплицирование**

18. Снова поставьте тубус со щелью к окну лазера (рис. 4) и найдите на экране резкое изображение щели с помощью линзы $\Omega_2$ ($F_2 \simeq 10$ см).

В фокальной плоскости Ф линзы $_2$ поставьте кассету с сетками, которые будут «рассекать» фурье-образ щели - осуществлять пространственную фильтрацию.

19. Подберите такую ширину входной щели $D$, чтобы на экране можно было наблюдать мультиплицированное изображение для всех сеток. Чем уже щель, тем шире её фурье-образ и тем легче рассечь его сетками.

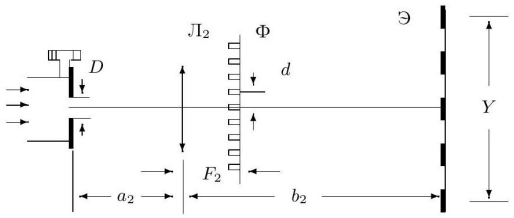



Рис. 4. Схема для наблюдения мультиплицирования

20. Снимите зависимость $Y$ (расстояние между удалёнными изображениями щели) и $K$ (число промежутков между изображениями) от № (номер сетки) для фиксированной ширины входной щели.

Запишите величины $D$ и $F_2$. Измерьте расстояния $a_2$ и $b_2$ для расчёта увеличения $\Gamma_2$.

21. Рассчитайте периоды $\Delta y$ «фиктивных» решёток, которые дали бы такую же периодичность на экране: $\Delta y = \Delta Y/\Gamma_2$, где $\Delta Y = Y/K$.

Постройте график $\Delta y = f\,(1/d_c)$, где $d_c$ - периоды решёток, определённые по спектру. Зависимость должна быть линейной, поскольку

$$\frac{\lambda}{\Delta y} F_2 = d_\mathrm{c} \tag{5}$$

22. Зная размер щели $D$, постройте в масштабе первичное изображение (спектр щели) и отложите на нём величины $d$( №) - периоды самой плотной и самой редкой из использованных решёток.

VI. Влияние щелевой диафрагмы на изображение сетки (пространственная фильтрация)

23. Поставьте на место щели (рис. 4) кассету с сетками и сфокусируйте на экран изображение сетки. Убедитесь, что изображение остаётся резким при смене сеток. Если нет - добейтесь этого (устраните саморепродукцию), слегка перемещая линзу $\Omega_2$.

Поставьте в плоскости Ф вертикальную щель и проследите за изменением изображения на экране при сужении щели. Проделайте то же для щели, ориентированной горизонтально и под углом 45° к вертикали. Объясните явление.

**Контрольные вопросы**

1. Как связаны между собой изображения, наблюдаемые в плоскостях $P_1, \Phi, P_2$?
2. Что общего между спектром одиночной щели и спектром периодической последовательности щелей такой же ширины?
3. Как связаны между собой ширина щели и ширина её спектра?
4. В работе наблюдается картина дифракции Фраунгофера на сетке. Какие параметры сетки можно определить по этой картине?
5. Что общего и чем отличаются спектры, образующиеся при дифракции Фраунгофера на щели и на волосе?

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Гл. 1, Гл. 8.
2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. IV, 52.

**Работа 4.3.5**

## 22.3.5 Study of a Hologram

Цель работы: изучить свойства голограмм точечного источника и объёмного предмета.

В работе используются: гелий-неоновый лазер, голограммы, набор линз, предметная шкала, экран, линейка.

В настоящей работе определяются характеристики голограммы точечного источника и исследуется голограмма трёхмерного объекта. Получение и исследование голограмм возможно при наличии источников с большой длиной когерентности, поэтому в работе применяется лазерное излучение. Перед выполнением работы необходимо ознакомиться с теорией голографического изображения (см. п. 3 введения к разделу). С более подробной

теорией голографического изображения, учитывающей конечную толщину голограммы, можно познакомиться в приложении к данной работе (с. 256).

Экспериментальная установка. В работе исследуются голограммы точечного источника и трёхмерного объекта. Эти голограммы записаны на фотопластинках типа ЛОИ-2, имеющих высокое разрешение: $\approx 5000$ линий /мм. При записи был использован He-Ne-лазер с длиной
волны излучения $\lambda = 632,8$ нм. Такой же лазер используется при восстановлении изображения.

В работе требуется определить расстояние $d$ между голограммой и точечным источником света. Это расстояние можно рассчитать по формуле (3.88), если измерить радиусы $\rho$ нескольких колец. Но размеры колец в нашей голограмме слишком малы для непосредственных

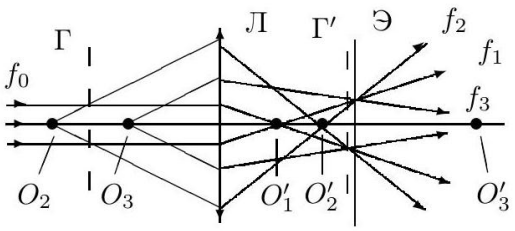



Рис. 1. Схема установки: Г - голограмма точечного источника измерений ($d \approx 1$ см, $\lambda \approx 0,6, \rho \approx 0,1$m), поэтому требуется получить увеличенное изображение голограммы. Это изображение получается при помощи короткофокусной линзы на экране (рис. 1).

Согласно формуле (3.85) при просвечивании голограммы Г (рис. 1) точечного источника плоской волной с амплитудой $f_0 = \text{const}$ на выходе имеется три волны: плоская с амплитудой $f_1 = \text{const}$, расходящаяся сферическая волна $f_2 \propto e^{ikr}$, отвечающая мнимому изображению $O_2$, и сходящаяся сферическая волна $f_3 \propto e^{-ikr}$, отвечающая действительному изображению $O_3$. После прохождения линзы Л волна $f_1$ собирается в фокусе линзы в точке $O_1'$, волны $f_2$ и $f_3$ фокусируются соответственно в точках $O_2'$ и $O_3'$. Изображение, возникающее на экране Э, можно рассматривать как результат интерференции сферических волн от трёх точечных источников: $O_1', O_2'$ и $O_3'$. Поэтому картина концентрических колец возникает на экране не только тогда, когда на нём образуется изображение Г′ голограммы, но и при многих других положениях линзы. Это легко проверить экспериментально.

Таким образом, чтобы определить радиусы колец, следует убедиться, что на экране действительно возникло изображение голограммы, то есть что плоскости Г и Г′ сопряжены. Для этого в плоскость Г вместо голограммы помещают прозрачную предметную шкалу, на которую нанесён тонкий крест с делениями. В этом случае увеличенное изображение креста получается при единственном положении линзы для определённого расстояния между предметом и экраном.

Предметная шкала и голограмма точечного источника смонтированы на платформе в единую плоскую кассету транспарантов (шкала в отверстии № 1, голограмма - в отверстии № 2). Такая же предметная шкала, закреплённая в отдельной оправе, используется в качестве предмета при определении фокусного расстояния голографической линзы.

Кроме голограммы точечного источника в работе исследуется голограмма объёмного предмета. Предмет представляет собой горизонтально расположенную миллиметровую линейку, за которой расположен вертикальный металлический стержень. При записи голограммы предмет располагался на расстоянии 10 см от пластинки. Фотопластинка была обращена к предмету той стороной, на которую нанесена эмульсия. Голограмма установлена в оправе вертикально и может поворачиваться вокруг вертикальной оси.

Для измерения угла поворота служит транспортир, закреплённый под голограммой в горизонтальной плоскости.

Источником света служит He-Ne-лазер с диаметром луча $< 1$ мм.

В опытах с объёмной голограммой требуются более широкие световые пучки. Для расширения пучка используются две линзы: короткофокусная - с фокусным расстоянием $< 2$ мм и длиннофокусная - с фокусным расстоянием $\approx 8$ см. Собранный из этих линз расширитель создаёт пучок диаметром $4 - 5$ см.

При проведении опытов оптические элементы размещаются на массивном оптическом столе и могут перемещаться вдоль оптической оси и в перпендикулярной ей плоскости. Расстояния измеряются линейкой. Фокусные расстояния линз указаны на их оправах.

## ЗАДАНИЕ

### A. Изучение характеристик голограммы точечного источника

В работе предлагается рассчитать расстояние от голограммы до точечного источника, который использовался при её создании:

а) по результатам измерения радиусов голографических колец, спроецированных на экран при помощи короткофокусной линзы;

б) по результатам измерения параметров проекционной установки, в которой голограмма используется как фокусирующая линза, а объектом служит предметная шкала.

Избегайте прямого попадания лазерного луча в глаз. Это крайне опасно для вашего зрения!

#### I. Настройка установки. Определение цены деления предметной шкалы

1. Включите лазер. Для определения цены деления предметной шкалы установите кассету с транспарантами вблизи лазера. Осветите лучом лазера шкалу (отверстие № 1 в кассете транспаранта) и получите на удалённом экране дифракиионную картину, созданную крестообразной шкалой.

Определив расстояние $\Delta x$ между дифракционными максимумами на экране и расстояние $L$ от шкалы до экрана, рассчитайте цену деления $D$ по известной формуле для дифракции Фраунгофера: $\lambda/D = \Delta x/L$.

2. Определите цену деления той же шкалы, используя линзу с фокусным расстоянием $F \approx 4$ см: получите в центре экрана увеличенное изображение предметной шкалы с чёткими делениями.

Измерьте расстояния от линзы до предметной шкалы $a$ и до экрана $b$ и рассчитайте увеличение системы. Определите расстояние $D'$ между изображениями штрихов и рассчитайте цену деления $D$ предметной шкалы. Сравните результаты, полученные разными методами. Какой из методов точнее в наших условиях?

#### II. Определение расстояния $d$ от голограммы до точечного источника

3. Получите на экране изображение голограммы точечного источника: для этого, перемещая кассету перпендикулярно лучу с помощью винта поперечной подачи (или плавно перемещая рейтер вдоль упора), осветите окно № 2 с голограммой. Перемещая голограмму по высоте, добейтесь того, чтобы три световых пятна на экране были расположены на одной высоте. Затем, совместив все три пятна,

получите на экране изображение голограммы (набор концентрических колец правильной формы). Смещая линзу вдоль луча на несколько миллиметров (при необходимости), получите наиболее контрастное изображение колец. Приложите к экрану лист бумаги, отметьте на нём радиусы нескольких тёмных колец и измерьте эти радиусы линейкой. Зная увеличение системы, рассчитайте размеры колец на голограмме, а затем определите расстояние $d$ от голограммы до точечного источника, который был использован при её создании.

4. Перемещая линзу вдоль луча, получите на экране изображение сначала мнимого $O_2$, а затем действительного $O_3$ точечного источника. Для каждого изображения измерьте необходимые геометрические размеры установки и, зная фокусное расстояние линзы, рассчитайте расстояния от точечных источников до голограммы.

5. Снова получите на экране изображение голограммы (совместите три световых пятна). Перемещая кассету перпендикулярно оптической оси, сместите голограмму ближе к границе светового пятта. Опять получите изображения действительного и мнимого источников и определите их расстояния до голограммы. Объясните, почему изображение источника смещено относительно оси системы.

### III. Изучение фокусирующих свойств голограммы

В этом опыте сама голограмма выполняет роль короткофокусной линзы. В качестве транспаранта, увеличенное изображение которого требуется получить с помощью голографической линзы, используется предметная шкала, закреплённая в отдельной оправе (цены деления этой шкалы и шкалы, закреплённой в кассете, одинаковы).

6. Сначала небольшим перемещением голограммы по вертикали добейтесь полного разделения пучков света на удалённом экране. По размерам световых пятен определите, какой из пучков соответствует действительному $O_3$, а какой - мнимому $O_2$ изображению точки (для этого можно приблизить переносной экран ближе к голограмме).

7. Установите переносной экран на расстоянии $\simeq 50$ см за голограммой (отсчёт всегда ведётся от источника света). Перед голограммой (вплотную к ней) поставьте предметную шкалу, закреплённую в отдельной оправе. Отодвигая рейтер со шкалой от голограммы, получите в одном из пятен резкое изображение делений крестообразной шкалы

Измерьте расстояние между штрихами $D'$ на экране и расстояние $b$ от экрана до голограммы. Используя эти данные, а также найденную ранее цену деления шкалы $D$, рассчитайте фокусное расстояние голографической линзы.

Сравните между собой найденные во всех опытах расстояния от точечного источника до голограммы.

## Б. Изучение характеристик голограммы объёмного предмета

В работе предлагается, рассматривая сквозь голограмму мнимое изображение предмета,

а) оценить угол падения опорной волны, который был выбран при создании голограммы;

б) убедиться, что изображение предмета восстанавливается по небольшой части голограммы;

в) оценить расстояние между линейкой и стержнем.

Кроме того, предлагается провести наблюдение действительного изображения предмета, чтобы

г) определить, чем отличаются изображения при повороте голограммы на 180° вокруг вертикальной оси;

д) проследить за изменением масштабов изображений при освещении голограммы сферической волной.

### IV. Юстировка системы

8. Соберите на оптическом столе расширитель пучка. Для этого определите положение лазерного пятна на удалённом экране; установите линзу с фокусным расстоянием 8-9 см на расстоянии $\approx 50$ см от экрана и, перемещая линзу в плоскости, перпендикулярной оптической оси, совместите центр светового пятна с точкой, где располагалось пятно в отсутствие линзы.

9. На расстоянии $\approx$ 10 см перед длиннофокусной линзой установите короткофокусную линзу с $f \approx 2$ мм. Перемещая короткофокусную линзу в плоскости, перпендикулярной лучу, вновь совместите центр светового пятна с его первоначальным положением.

10. Перемещением короткофокусной линзы вдоль луча добейтесь того, чтобы из собранного расширителя выходил параллельный пучок: при полном заполнении светом длиннофокусной линзы диаметр светового пятна на удалённом экране равен диаметру линзы, при неполном - можно контролировать размер пятна, перемещая вдоль луча переносной экран.

### V. Изучение мнимого изображения

11. Поместите голограмму в расширенный пучок лазера фотоэмульсией к лазеру и найдите мнимое изображение предмета. Для этого расположитесь за экраном с правой стороны. Медленно поворачивая голограмму вокруг вертикальной оси и перемещая глаз в горизонтальной плоскости на высоте луча, найдите изображение линейки в окне голограммы. Проследите за тем, как меняется ориентация мнимого изображения при перемещении глаза наблюдателя (убедитесь, что изображение объёмное).

12. По углу поворота голограммы оцените угол падения опорной волны, который был выбран при получении голограммы.

13. Постепенно закрывая голограмму листом бумаги, убедитесь в том, что изображение предмета восстанавливается даже по небольшой части голограммы. Объясните это явление.

14. Подумайте, как оценить расстояние от линейки до вертикального стержня, расположенного за ней, фиксируя по транспортиру угол, под которым наблюдается изображение. Оцените это расстояние.

15. Проведите наблюдение мнимого изображения предмета, используя в качестве лупы линзу с фокусным расстоянием 20 см.

**VI. Изучение действительного изображения**

16. Найдите действительное изображение предмета: голограмма должна при этом располагаться перпендикулярно падающему пучку света, глаз наблюдателя - с левой стороны от экрана на расстоянии от пластинки 50-80 см. Поскольку это изображение расположено между пластинкой и глазом, удобно рассматривать его через лупу.

Обратите внимание, что по мере удаления лупы от глаза вы сначала видите резкое изображение стержня и только потом - линейки. Объясните это явление.

17. Поверните голограмму на 180° вокруг вертикальной оси (фотоэмульсией от лазера). Найдите действительное и мнимое изображения. При каких углах падения восстанавливающей волны они возникают? Чем и почему отличаются эти изображения от наблюдаемых ранее?

18. Снова разверните голограмму эмульсией к лазеру. Перемещая короткофокусную линзу расширителя вдоль луча, определите, как изменяются масштабы действительного и мнимого изображений. Объясните эти изменения.

**ПРИЛОЖЕНИЕ**

Если голограмма записана на фотопластинке с толстым слоем эмульсии и при большом угле падения опорной волны, то она регистрирует объёмную картину интерференции предметной и опорной волн. Будем считать, что предметная волна с комплексной амплитудой $f_{\text{пр}}(\boldsymbol{r})$, где $\boldsymbol{r} = (x, y, z)$, распространяется в плоскопараллельном слое эмульсии толщины $h (0 \leqslant z \leqslant h)$. При записи голограммы опорная волна была плоской, её волновой вектор $\boldsymbol{k}_{\text{оп}}$ лежал в плоскости $x, z$ и составлял угол $\alpha$ с осью $z$. Запишем комплексную амплитуду этой волны в виде

$$f_{\text{оп}}(\boldsymbol{r}) = a_{\text{оп}}\, e^{-i(\boldsymbol{k}_{\text{оп}}\,\boldsymbol{r})} = a_{\text{оп}}\, e^{-ik(x\sin\alpha + z\cos\alpha)} \tag{1}$$

где $k$ - волновое число.

После экспонирования фотопластинки и её химической обработки в фотоэмульсии образуются кристаллы металлического серебра, локальная концентрация которых пропорциональна интенсивности интерференционной картины:

$$|f_{\text{пр}}(\boldsymbol{r}) + f_{\text{оп}}(\boldsymbol{r})|^2$$

Рассмотрим структуру эмульсии в случае, когда предметная волна является плоской волной с волновым вектором, перпендикулярным плоскости фотопластинки:

$$f_{\text{пр}}(\boldsymbol{r}) = a_{\text{пр}}\, e^{-ikz}$$

Интенсивность интерференционной картины

$$I(\boldsymbol{r}) = |a_{\text{пр}}|^2 + |a_{\text{оп}}|^2 + \underbrace{+2\,|a_{\text{пр}}|\,|a_{\text{пп}}|\cos(k[(1-\cos\alpha)z - x\sin\alpha] + \varphi)}\,,$$

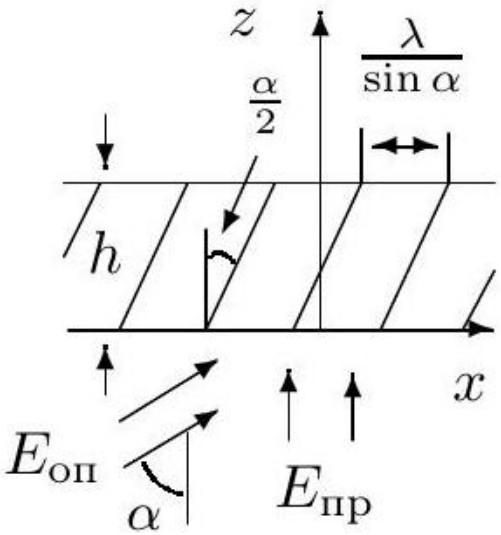
где $\varphi$ - постоянная разность фаз между предметной и опорной волнами. Отсюда видно, что в рассматриваемом случае линиями постоянной интенсивности являются прямые. В частности, интенсивность максимальна, если

$$z = \frac{x}{\operatorname{tg}(\alpha/2)} + \frac{2\pi n - \varphi}{k(1-\cos\alpha)}, \quad n = 0, \pm1, \pm2, \ldots$$

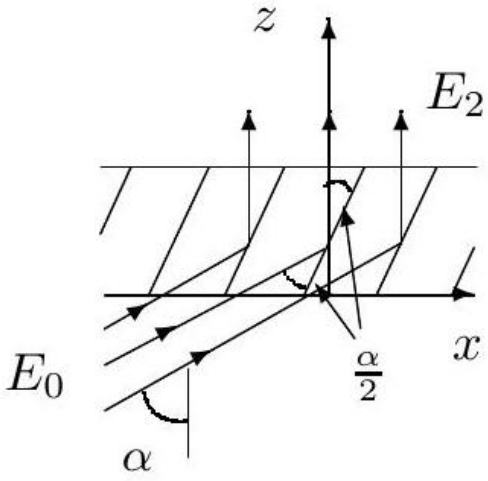
(здесь использовано $\mathrm{tg}\,\alpha/2 = (1-\cos\alpha)/\sin\alpha$). Кроме того, при постоянной координате $z$ интенсивность периодически меняется вдоль координаты $x$ с периодом

$$\Delta x = \frac{\lambda}{\sin\alpha}$$

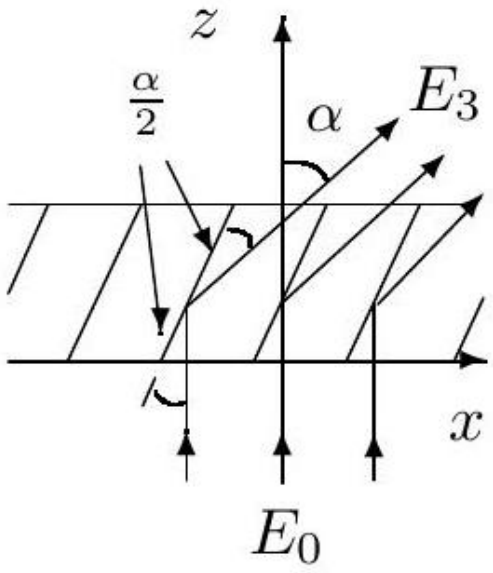
где $\lambda$ - длина волны света. На рис. 2а показано семейство прямых, вдоль которых интенсивность интерференционной картины постоянна. Эти прямые параллельны биссектрисе угла между волновыми векторами предметной и опорной волн.




а)




б)




в)

Рис. 2. Запись и восстановление фронта плоской волны

Пропорционально интенсивности $I(\boldsymbol{r})$ меняется локальная концентрация металлического серебра. Можно сказать, что в рассматриваемом случае эмульсия голограммы представляет собой периодическую последовательность наклонно расположенных зеркал, причём коэффициент пропускания и отражения зеркал меняется вдоль координат $x, z$. Эмульсию можно считать тонкой, если проекция $h\,\mathrm{tg}(\alpha/2)$ ($h$ - толщина эмульсии) линейного размера отдельного зеркала на ось $x$ много меньше периода следования зеркал $\Delta x$, т. е.

$$h\,\mathrm{tg}\,\frac{\alpha}{2} \ll \frac{\lambda}{\sin\alpha} \quad \text{или} \quad 2\frac{h}{\lambda}\sin^2\frac{\alpha}{2} \ll 1.$$

Толщина эмульсии на фотопластинках ЛОИ-2, используемых в работе,

$$h \simeq (7 \div 10)\lambda \simeq 5\text{mKM} \quad (\lambda \simeq 0,6\text{K})$$

При регистрации голограммы точечного источника опорная волна падает на фотопластинку нормально, а угол падения предметной волны на периферии голограммы (в районе десятого интерференционного кольца) $\alpha \simeq 10^{-2}$. Поэтому

$$\frac{2h}{\lambda}\sin^2\frac{\alpha}{2} \simeq 10^{-3}$$

и эту голограмму можно считать тонкой. Для голограммы предмета угол падения опорной волны $\alpha \simeq 40°$, и тот же параметр достигает значения $\simeq 2$. В этом случае эмульсию следует считать толстой.

Рассмотрим условия восстановления действительного и мнимого изображений для толстослойной голограммы. Пусть на голограмму падает плоская волна, волновой вектор $\boldsymbol{k}_0$ которой лежит в плоскости $x, z$ и составляет угол $\beta$ с осью $z$. Комплексная амплитуда этой волны

$$f_0(\boldsymbol{r}) = a_0 e^{-i\boldsymbol{k}_0\boldsymbol{r}} = a_0 e^{-ik(x\sin\beta + z\cos\beta)} \tag{3}$$

Падающая волна, проходя через эмульсию голограммы, будет частично рассеиваться на центрах восстановленного серебра. Можно считать, что небольшая область эмульсии вблизи точки $\boldsymbol{r} = (x, y, z)$ является источником вторичной (рассеянной) волны.

Амплитуда вторичной волны пропорциональна концентрации металлического серебра, которая в свою очередь пропорциональна интенсивности интерференционной картины предметной и опорной волн при записи голограммы, а фаза определяется фазой падающей волны. Исходя из этого, запишем комплексную амплитуду волны $f$, рассеянной в точке $\boldsymbol{r}$ в виде

$$f(\boldsymbol{r}) \sim |f_{\text{пр}} + f_{\text{оп}}|^2 f_0 = f_1(\boldsymbol{r}) + f_2(\boldsymbol{r}) + f_3(\boldsymbol{r})$$

где

$$\begin{aligned} f_1(\boldsymbol{r}) &= a_0 \left(|a_{\text{оп}}|^2 + |f_{\text{пр}}(\boldsymbol{r})|^2\right) e^{-i\boldsymbol{k}_0\boldsymbol{r}} \\ f_2(\boldsymbol{r}) &= a_0 a^*_{\text{оп}} f_{\text{пр}}(\boldsymbol{r}) e^{-i(\boldsymbol{k}_0 - \boldsymbol{k}_{\text{оп}})\boldsymbol{r}} \\ f_3(\boldsymbol{r}) &= a_0 a_{\text{оп}} f^*_{\text{пр}}(\boldsymbol{r}) e^{-i(\boldsymbol{k}_0 + \boldsymbol{k}_{\text{оп}})\boldsymbol{r}} \end{aligned} \tag{4}$$

Обычно $|a_{\text{оп}}|^2 \gg |f_{\text{пр}}(x,y)|^2$, поэтому $f_1$, как и в случае тонкой голограммы, описывает комплексную амплитуду прошедшей волны. Эта волна распространяется под тем же углом $\beta$, что и падающая. Слагаемое $f_2(\boldsymbol{r})$ пропорционально комплексной амплитуде предметной волны, причём соотношение между фазами предметной волны в разных точках сохраняется, если $\boldsymbol{k}_0 = \boldsymbol{k}_{\text{оп}}$. В этом случае $f_2(\boldsymbol{r})$ описывает мнимое изображение предмета. Таким образом, при наблюдении мнимого изображения восстанавливающая волна должна падать под тем же углом, что и опорная, т. е. $\alpha = \beta$. Отметим, что условие возникновения мнимого изображения в толстой и тонкой голограммах одно и то же. На рис. 2б показаны падающая и восстановленная плоские волны. Происходит как бы отражение падающей волны от системы наклонно расположенных зеркал.

При описании рассеянного поля удобно представить поле предметной волны в виде суперпозиции плоских волн:

$$f_{\text{пр}}(\boldsymbol{r}) = \sum_{\boldsymbol{k}_{\text{пр}}} a(\boldsymbol{k}_{\text{пр}}) e^{-i\boldsymbol{k}_{\text{пр}}\boldsymbol{r}} \tag{5}$$

где $a(\boldsymbol{k}_{\text{пр}})$ имеет смысл амплитуды волны с волновым вектором $\boldsymbol{k}_{\text{пр}}$ в дальней (фраунгоферовой) зоне. Используя определение (5), можно представить поле $f_2$ в виде

$$f_2(\boldsymbol{r}) = a_0 a^*_{\text{пп}} \sum_{\boldsymbol{k}_{\text{пр}}} a(\boldsymbol{k}_{\text{пр}}) e^{-i(\boldsymbol{k}_{\text{пр}} + \boldsymbol{k}_0 - \boldsymbol{k}_{\text{п}})\boldsymbol{r}}$$

При выполнении условия $\boldsymbol{k}_0 = \boldsymbol{k}_{\text{оп}}$ поле $f_2(\boldsymbol{r})$ с точностью до константы совпадает с полем предметной волны (5). В нашем случае предмет находился перед фотопластинкой, и рассеянное им излучение падало на фотопластинку под небольшими углами, т. е. волновые векторы $\boldsymbol{k}_{\text{пр}}$ расположены вблизи $\boldsymbol{k}_0 = (0,0,k)$. Поэтому мнимое изображение возникает непосредственно за голограммой на месте голографируемого предмета. Аналогично можно записать поле $f_3$:

$$f_3(\boldsymbol{r}) = a_0 a_{\text{оп}} \sum_{\boldsymbol{k}_{\text{пр}}} a^*(\boldsymbol{k}_{\text{пр}}) e^{-i(\boldsymbol{k}_{\text{п}} + \boldsymbol{k}_0 - \boldsymbol{k}_{\text{пр}})\boldsymbol{r}} \tag{6}$$

Поле $f_3$ в форме (6) представлено в виде суперпозиции плоских волн, волновой вектор которых определяется из условия

$$\boldsymbol{k}_3 = \boldsymbol{k}_{\text{оп}} + \boldsymbol{k}_0 - \boldsymbol{k}_{\text{пр}} \tag{7}$$

при естественном ограничении:

$$|\boldsymbol{k}_3| = |\boldsymbol{k}_{\text{оп}}| = |\boldsymbol{k}_0| = |\boldsymbol{k}_{\text{пр}}| = k. \tag{8}$$

В рассмотренном примере поле предметной волны представляет собой плоскую волну, фазовый фронт которой параллелен фотопластинке, т. е. $\boldsymbol{k}_{\text{пр}} \approx (0, 0, -k)$. Взаимная ориентация волновых векторов показана на рис. 3. Как ясно из рисунка, условия (7)-(8) выполняются, если и $\boldsymbol{k}_0 = (0, 0, k)$, т. е. восстанавливающая волна падает на голограмму нормально. При этом волновой вектор восстановленной волны $\boldsymbol{k}_3 = \boldsymbol{k}_{\text{оп}}$, и эта волна распространяется под углом $\alpha$ к оси $z$. На рис. 2в показана

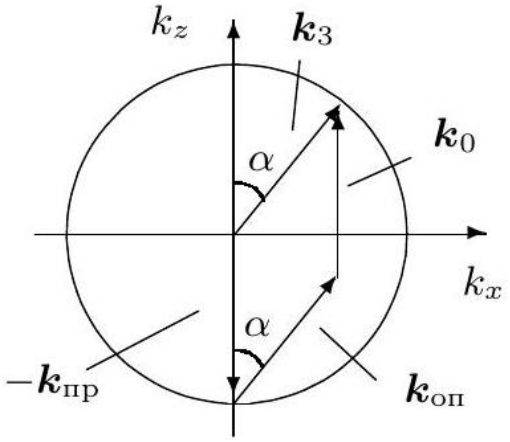



Рис. 3. Ориентация волновых векторов при восстановлении действительного изображения падающая и восстановленная волна для этого случая.

Если предмет рассеивает падающую на него волну под небольшими углами, а волновые векторы $\boldsymbol{k}_{\text{пр}}$ мало отличаются от $(0, 0, k)$, то поле $f_3(\boldsymbol{r})$ связано с действительным изображением предмета. Для наблюдения этого изображения нужно расположить голограмму перпендикулярно лазерному лучу и рассматривать её под углом $\alpha$, т. е. под тем же углом, под которым падала опорная волна. Изображение возникает между голограммой и глазом наблюдателя. Если набор волновых векторов $\boldsymbol{k}_0$ значительно отличается от $(0, 0, k)$, то удовлетворить условиям (7)-(8) одновременно для всех $\boldsymbol{k}_{\text{пр}}$ не представляется возможным, поэтому качество действительного изображения заметно хуже, чем качество мнимого изображения.

**Контрольные вопросы**

1. В чем сходство и различие голограммы точечного источника и зонной пластинки Френеля, фотографии колец Ньютона?

2. Что изменится в изображении голограммы точечного источника, наблюдаемого по схеме, изображённой на рис. 1, если устранить точечный источник $O_1'$, поместив в фокус линзы Л небольшой непрозрачный экран-фильтр?

3. Как изменится фокусное расстояние голограммы точечного источника, если она была получена при одной длине волны, а освещается светом с другой длиной волны?

4. Чем определяется минимальный размер голограммы, при котором можно восстановить изображение предмета?

5. Как изменятся масштабы действительного и мнимого изображений предмета, если голограмму освещать расходящейся или сходящейся волной?

6. Как зависит диапазон углов, под которым можно рассматривать мнимое изображение предмета, от разрешающей способности фотоэмульсии, от ширины спектра источника света?

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. §§8.11 – 8.15.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. IV, §§44, 54.

3. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. XI, 59, 60.

4. Гудмен Джн. Введение в фурье-оптику. М.: Мир, 1970. Гл. 8, 1, 2.

**Работа 4.3.6**

## 22.3.6 Diffraction of Light on Periodic Structures (Self-Reproduction)

Цель работы: Изучение явления саморепродукции и применение его к измерению параметров периодических структур.

В работе используются: лазер, кассета с сетками, мира, короткофокусная линза с микрометрическим винтом, экран, линейка.

При дифракции на предмете с периодической структурой наблюдается интересное явление: на некотором расстоянии от предмета вдоль направления распространения волны появляется изображение, которое потом периодически повторяется - репродуцируется (эффект Талбота, см. п. 2.3 введения к разделу).

Экспериментальная установка. Хорошим приближением к плоской волне в нашем эксперименте является излучение лазера. Луч лазера
падает перпендикулярно на периодический объект О, установленный в плоскости $P_0$ (рис. 1).

За плоскостью $P_0$ (в плоскостях $P_1 - P_N$) периодически по $z$ возникают изображения объекта, которые с помощью линзы Л можно поочерёдно проецировать на экран, установленный в плоскости Э. Если убрать линзу, то на экране наблюдается картина дифракции луча лазера на периодическом объекте.

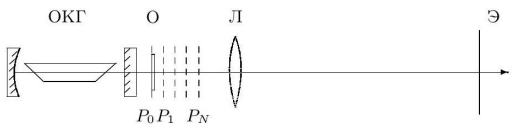


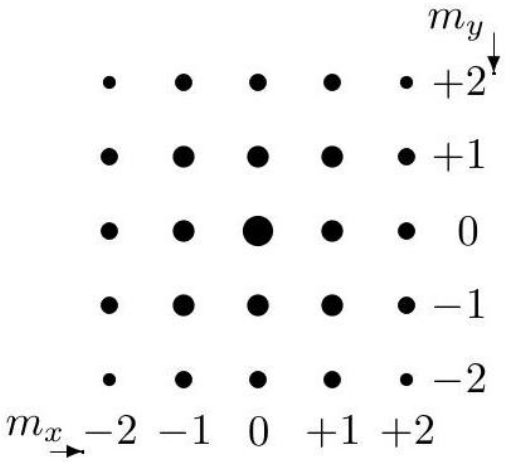

Рис. 1. Схема установки: ОКГ - оптический квантовый генератор (гелий-неоновый лазер), 0 - двумерная решётка, $P_N$ - плоскости, где наблюдаются репродуцированные изображения, Л - короткофокусная линза, Э - экран для наблюдения изображения объекта

Экран устанавливается достаточно далеко от объекта, так что продифрагировавшие лучи, соответствующие различным порядкам дифракции ($\sin\theta_n = n\lambda/d$), разделяются.

Измерив расстояние между дифракционными максимумами и расстояние от объекта до экрана, мы определим $\sin\theta_n$ и $d$.

В нашей работе в качестве периодических объектов применяется «мира» - набор различным образом ориентированных одномерных решёток разного периода, а также

двумерная решётка-сетка. Сетку можно рассматривать как две взаимно перпендикулярные решётки. Узкий пучок монохроматического света, пройдя через первую решётку с вертикальными штрихами, должен дать




совокупность максимумов, расположенных вдоль горизонтальной линии.

Световой пучок, соответствующий каж-

Рис. 2. Спектр решётки-сетки дому максимуму, проходя через вторую решётку, распадается на новую совокупность пучков, дающих максимумы вдоль вертикальной линии. В результате главные максимумы возникают тогда, когда одновременно выполняются условия

$$d \sin \theta_x = m_x \lambda, \quad d \sin \theta_y = m_y \lambda \tag{1}$$

где $m_x$ и $m_y$ - два целых числа, характеризующих порядки дифракционных максимумов, $\varphi_x$ и $\varphi_y$ - направления на главные дифракционные максимумы в горизонтальной и вертикальной плоскостях соответственно (рис. 2). Максимумы показаны кружками, размеры которых характеризуют интенсивность.

## ЗАДАНИЕ

В работе предлагается:

А) определить периоды сеток сначала по их спектру на удалённом экране; затем по увеличенному с помощью линзы изображению сеток и, наконец, по результатам измерения расстояний между репродуцированными изображениями сеток;

Б) повторить измерения для нескольких элементов миры.

### А. Исследование двумерных решёток

#### I. Определение периода решёток по их пространственному спектру

1. Включите в сеть блок питания лазера. Пронаблюдайте на экране распределение интенсивности в излучении лазера. Из-за переотражения в выходных стёклах лазера на его выходе может быть несколько лучей. Следите за лучом с максимальной яркостью.

Настройку системы следует вести, наблюдая за пятном света на листе бумаги. Яркость излучения зелёного лазера можно ослабить с помощью светофильтра.

**Внимание! Попадание прямого лазерного луча на сетчатку глаза крайне опасно!**

2. Закрепите кассету с двумерными решётками (сетками) вблизи выходного окна лазера так, чтобы в окошке под отверстием с сеткой был виден её номер (в каждой кассете 6 разных решёток). Вращением барабана (наружного кольца кассеты) решётки можно менять.

3. Для каждой сетки определите расстояние $x$ между соседними дифракционными максимумами на экране: измерьте расстояние $X$ между двумя достаточно удалёнными друг от друга максимумами и поделите на число промежутков $m$ между ними ($x = X/m = f()$).

4. Измерьте расстояние $L$ от кассеты до экрана. Длина волны лазера указана на установке.

**II. Определение периода решёток по изображению, увеличенному с помощью линзы**

5. Закрепите короткофокусную линзу на небольшом расстоянии от лазера. Временно удалите кассету с сетками из луча и центрировкой линзы
совместите световое пятно, сформированное линзой, с положением луча на экране в отсутствие линзы. Передвигая линзу с помощью микрометрического винта, сначала убедитесь, что световое пятно на экране неподвижно, затем переместите линзу как можно ближе к кассете.

6. Установите кассету с сетками между лазером и линзой и, перемещая её вдоль оси системы, получите на экране увеличенное изображение одной из сеток.

Изображение сетки периодически повторяется в пространстве, поэтому для того, чтобы среди множества репродуцированных изображений сетки можно было выделить её геометрическое изображение, на одну из сеток наложена тонкая проволочка, т. е. непериодический объект, изображение которого не репродуцируется. Найдите резкое изображение проволочки; убедитесь, что при смене сеток резкость изображений не нарушается.

7. Определите размеры $D$ клеток на экране. Для мелких сеток удобно прижать к экрану дополнительный листок бумаги и отметить на нём необходимые расстояния. Проведите измерения для всех сеток, для которых это возможно.

Обратите внимание, что по спектру можно определять размер мелкой сетки, который не может быть достаточно точно определён с помощью линзы, т. к. размер изображения на экране слишком мал.

8. Измерьте расстояния от линзы до сетки ($a$) и до экрана ($b$).

**III. Исследование эффекта саморепродукции с помощью сеток**

9. Получите на экране геометрическое изображение сетки. Перемещая линзу с помощью микровинта, определите по нониусной шкале координаты $z_N$ плоскостей саморепродукции, соответствующих чёткому изображению сетки на экране.

Повторите измерения для других сеток. Для крупных сеток полезно при измерениях строить графики $z_N = f(N)$, чтобы рассчитать периоды сеток $d_{\text{реп}} = f()$.

### Б. Исследование решёток миры

Ввиду сложности настройки на определённый элемент миры предлагается следующий порядок измерения периодов: для одного элемента сначала по саморепродукции, затем по увеличенному изображению и, наконец, по спектру.

10. Рассмотрите миру в отражённом свете. Убедитесь, что самые мелкие решётки расположены внизу. Найдите элемент миры с номером 25.

11. Установите миру на место кассеты (стекло миры обращено к линзе) таким образом, чтобы на экране было видно резкое увеличенное изображение цифры 25. Перемещая рейтер с мирой перпендикулярно лучу,
осветите одну из ближайших сеток (горизонтальную или вертикальную). Перемещением миры подведите пятно лазера к центру выбранной решётки и уточните настройку на геометрическое изображение с помощью микровинта.

12. Перемещая линзу с помощью микрометрического винта, определите по ониусной шкале координату плоскости, соответствующей изображению миры на экране по законам геометрической оптики, и координаты плоскостей саморепродукции.

13. Верните линзу в положение, соответствующее изображению по законам геометрической оптики. Определите параметры изображения миры на экране; измерьте расстояния от линзы до экрана и до миры.

14. Поворотом линзы на рейтере вокруг вертикальной оси выведите линзу из луча и проведите измерения спектра миры.

15. Повторите всю серию измерений для миры с номером 20.

### IV. Обработка результатов

16. По результатам измерений спектра рассчитайте период каждой решётки $d_{\text{сп}} = f($ №$)$, полагая $\sin\theta \approx \theta \approx x/L$.

17. По измерениям с линзой рассчитайте периоды всех сеток $d_{\text{Л}} = Da/b = f($ №$)$.

18. Постройте графики $z_N = f(N)$; по наклону прямых рассчитайте периоды сеток $d_{\text{реп}} = f($ №$)$.

19. Сведите результаты измерений $d$ разными методами в единую таблицу и оцените, какой из методов точнее.

20. Определите периоды решёток миры тремя способами и рассчитайте число штрихов на мм в выбранных элементах.

### Контрольные вопросы

1. Оцените, на каком расстоянии от экрана до решётки выполняется условие дифракции Фраунгофера.

2. Почему в эксперименте получается лишь несколько чётких репродуцированных изображений решёток?

3. Объясните, как получается выражение для направления на дифракционные максимумы.

4. Объясните, почему изображение проволочной сетки репродуцируется, а изображение одиночной проволочки не репродуцируется.

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. III, гл. 8, §§8.1, 8.2, 8.7.

## 22.4 Experiments about Spectral Instruments

### 22.4.1 Amplitude Diffraction Grating

Цель работы: Знакомство с работой и настройкой гониометра Г5, определение спектральных характеристик амплитудной решётки.

В работе используются: гониометр, дифракционная решётка, ртутная лампа.

**ЗАДАНИЕ**

В работе предлагается отъюстировать гониометр, исследовать спектр ртутной лампы, определить период и спектральные характеристики решётки.

**I. Настройка гониометра**

Проведите юстировку гониометра, руководствуясь правилами, изложенными в Приложении. Ознакомьтесь с устройством гониометра и назначением элементов настройки. Нумерация, написанная на элементах лабораторного экземпляра прибора, и нумерация, приведённая на рис. П4.1 Приложения, совпадают. Установите зрительную трубу на бесконечность. Из пункта «Установка поверхности столика перпендикулярно оси вращения прибора» выполните операции по предварительной настройке. Настройте коллиматор. Установите начало отсчёта углов, после чего винт 26 следует закрепить и не освобождать его до конца измерений.

**II. Установка решётки**

Основание оправы решётки и её штрихи могут быть не перпендикулярны друг другу, поэтому плоскость столика следует немного наклонить. Установите решётку на столик так, чтобы её плоскость была перпендикулярна оси зрительной трубы и параллельна одному из установочных винтов 8. В центре поля зрения расположена белая ахроматическая полоса (спектр нулевого порядка). Высота полос должна быть менее четверти поля зрения трубы. Поворачивая алидаду, найдите изображение линий максимального порядка, и винтом 8, параллельным плоскости решётки, верните центр изображения щели к горизонтальной оси отсчётного креста зрительной трубы.

**III. Исследование спектра ртутной лампы**

1. Подберите ширину входной щели коллиматора так, чтобы ширина линий жёлтого дублета была чуть больше промежутка между линиями двойного штриха окуляра зрительной трубы. Установите высоту щели, удобную для измерений (при короткой щели плохо виден двойной штрих, при слишком высокой - мешает кривизна изображения).

2. Измерьте угловые координаты спектральных линий ртути в $\pm 1$ порядке. Примерное расположение и относительная яркость основных линий приведена в Приложении (рис. П4.4, табл. 1).

3. Для оценки угловой дисперсии решётки измерьте угловые координаты линий жёлтого дублета для всех видимых порядков спектра, положительных и отрицательных.

4. Для оценки разрешающей способности спектрального прибора измерьте угловую ширину одной из линий жёлтого дублета по нулям интенсивности в нескольких порядках. Ширина входной щели коллиматора должна быть минимальной, позволяющей провести измерения.

### *IV*. Обработка результатов измерений

1. Для $\pm 1$ порядка постройте график зависимости $\sin \varphi_m$ от длины волны. Используя (4.1), определите шаг решётки.

2. Рассчитайте по линиям жёлтого дублета угловую дисперсию в спектрах разного порядка (формула (4.10)). Постройте график зависимости угловой дисперсии от порядка спектра, сравните эту зависимость с расчётной по формуле (4.11) для средней длины волны жёлтого дублета.

3. Оцените разрешимый спектральный интервал $\delta\lambda$. Согласно (4.17), для этого надо по имеющимся измерениям вычислить угловую полуширину жёлтой линии и знать угловую дисперсию. По формуле (4.15) оцените разрешающую способность для средней длины волны жёлтого дублета. Используя (4.18), оцените число эффективно работающих штрихов решётки и её эффективный размер.

4. Рассчитайте порядок спектра, при котором фиолетовая линия наложится на жёлтую.

### Контрольные вопросы

1. Дайте определение спектральных характеристик дифракционной решётки.

2. Сформулируйте критерий Релея.

3. Как меняются угловые характеристики спектров при изменении периода решётки, размера штриха, размера всей решётки?

4. Можно ли по спектру ртутной лампы определить в ваших опытах дисперсионную область?

5. Какие физические и технические причины определяют число эффективно работающих штрихов дифракционной решётки большого размера?

### Литература

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§46, 47.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 46, 50.

## 22.4.2 Phase Diffraction Grating

Цель работы: Знакомство с работой и настройкой гониометра Г5, определение спектральных характеристик фазовой решётки (эшелетта).

В работе используются: гониометр, эшелетт, ртутная лампа.

### ЗАДАНИЕ

В работе предлагается отъюстировать гониометр, исследовать спектр ртутной лампы для нескольких углов падения, определить технические и спектральные характеристики эшелетта.

### I. Настройка гониометра

Проведите настройку гониометра, руководствуясь правилами, изложенными в Приложении. Ознакомьтесь с устройством гониометра и назначением элементов настройки. Нумерация, написанная на элементах лабораторного экземпляра прибора, и нумерация, приведённая на рис. П4.1 приложения, совпадают. Установите зрительную трубу на бесконечность. Из пункта «Установка поверхности столика перпендикулярно оси вращения прибора» выполните операции по предварительной настройке. Настройте коллиматор. Установите начало отсчёта углов, после чего винт 26 следует закрепить и не освобождать его до конца измерений.

Во избежание необратимых повреждений эшелетта не прикасайтесь к его рабочей поверхности, не дышите на неё.

### II. Качественные наблюдения

Удерживая эшелетт в вытянутой руке, найдите отражённое изображение нити лампы накаливания, расположенной за вашей спиной. Вращая эшелетт, наблюдайте спектры различных положительных и отрицательных порядков. Определите рабочий порядок, в котором спектр наиболее интенсивен. Отметьте ориентацию направлений: источник света-эшелетт, эшелетт-направление на рабочий спектр. При проведении опытов это будут направления: коллиматор-эшелетт, эшелеттзрительная труба.

Отметьте, в каких порядках спектры начинают перекрываться. Оцените для этих порядков дисперсионную область и сравните её с (4.21) для средней длины волны 500 нм и полуширины 200 нм, что соответствует наблюдению глазом излучения лампы накаливания. Запишите паспортные данные эшелетта: рабочий порядок $m_p$, рабочую длину волны $\lambda_\text{p}$.

### III. Установка эшелетта

Основание оправы эшелетта и его ступеньки могут быть не перпендикулярны друг другу, поэтому плоскость столика следует немного наклонить.

1. Настройте зрительную трубу на наблюдение входной щели коллиматора, вертикальный размер изображения щели должен быть менее половины поля зрения, рекомендуемый начальный отсчёт угла 180°. Вам известны направления: ось коллиматора-эшелетт-рабочий порядок, поверните алидаду на 120° от начального положения, направление вращения, вправо или влево, должно позволить наблюдать рабочий порядок.

2. Поставьте эшелетт на столик рабочей поверхностью к коллиматору так, чтобы эшелетт был параллелен одному из винтов 8 и перпендикулярен другому. Вращая

только верхнюю часть столика, освободив винт 27 (винт 26 закреплён, чтобы не сбился начальный отсчёт угла), найдите ахроматическое (белое) изображение щели коллиматора, отражённое от эшелетта (спектр нулевого порядка). При этом угол падения света на плоскость эшелетта $\psi = (180^\circ - 120^\circ)/2 = 30^\circ$.

3. Винтом 8, перпендикулярным плоскости эшелетта, совместите центр изображения щели с горизонтальным штрихом отсчётного креста окуляра зрительной трубы. Отводя алидаду от коллиматора, найдите изображение линии в дальнем порядке и вторым винтом 8, параллельным плоскости эшелетта, устраните вертикальное расхождение. Вернитесь к ахроматическому изображению и уточните положение винтов 8 по приведённой процедуре. Допустимое вертикальное смещение линий менее трети радиуса поля зрения.

**IV. Исследование спектра ртутной лампы**

4. Подберите ширину входной щели коллиматора, при которой ширина линий жёлтого дублета чуть больше промежутка между линиями двойного штриха зрительной трубы. Установите высоту щели, удобную для измерений (при короткой щели плохо виден двойной штрих, при слишком высокой - мешает кривизна изображения).

5. Для угла падения $\psi = 30^\circ$ измерьте угловые координаты спектральных линий ртути в рабочем порядке. Примерное расположение и относительная яркость основных линий приведены в Приложении (рис. П4.4, табл. 1).

6. Для оценки разрешающей способности спектрального прибора измерьте угловую ширину одной из линий жёлтого дублета (по нулям интенсивности). Ширина щели коллиматора должна быть минимальной, позволяющей провести измерения.

7. Для установленного угла падения $\psi = 30^\circ$ измерьте угловые координаты жёлтого дублета во всех наблюдаемых порядках, положительных и отрицательных. Эти данные позволят оценить дисперсию в различных порядках спектра для фиксированного угла падения.

8. Измерьте угловые координаты линий жёлтого дублета в рабочем порядке для углов падения 45° и 60°. Для больших углов падения есть возможность найти изображение автоколлимационного креста, отражённого от эшелетта. В этом положении ось зрительной трубы перпендикулярна плоскости эшелетта, и можно ввести ещё одно удобное начало отсчёта углов. Например, угол падения света равен углу наблюдения ахроматической полосы, отсчитанному от нового начала.

Обратите внимание на появление спектров больших отрицательных порядков при увеличении угла падения света.

9. Эшелетт, ширина штриха которого сравнима с длиной волны, поляризует отражённый свет. Проведите следующий эксперимент. Используя отдельный эшелетт с плотностью штрихов 1200 штрихов/мм (полное число штрихов 180000, ширина отдельного штриха меньше 1 мкм), а в качестве источника излучения - свет настольной лампы (излучение имеет случайную поляризацию), определите с помощью поляроида преимущественное направление колебания электрического вектора в первом порядке спектра. Как связаны между собой это направление и ориентация штрихов? Объясните этот эффект.

**V. Обработка результатов**

1. Для угла падения 30° постройте график зависимости $\sin\varphi_m - \sin\psi$ от длины волны в рабочем порядке. Используя (4.3), определите шаг решётки.

2. Используя (4.5), рассчитайте угол скоса рабочей грани эшелетта.

3. Рассчитайте по формуле (4.12) угловую дисперсию, используя все измерения для жёлтого дублета.

Для угла падения 30° постройте график зависимости угловой дисперсии от порядка спектра. Отдельно выпишите значения угловой дисперсии в рабочем порядке для углов падения $30°, 45°, 60°$. Сравните
полученные экспериментальные данные с рассчитанными по формуле (4.12).

4. Оцените разрешимый спектральный интервал $\delta\lambda$ в рабочем порядке для угла падения 30°. Согласно (4.17), для этого надо вычислить угловую полуширину жёлтой линии и знать величину угловой дисперсии. По формуле (4.15) оцените разрешающую способность спектрального прибора, который включает эшелетт, гониометр, глаз экспериментатора. Используя (4.18), оцените число эффективно работающих штрихов эшелетта и его эффективный размер.

5. Из опытов с лампой накаливания определите дисперсионную область и сравните её с соответствующим порядком спектра.

**Контрольные вопросы**

1. Дайте определения спектральных характеристик эшелетта.

2. В чем заключается преимущество эшелетта по сравнению с амплитудной дифракционной решёткой?

3. Какой элемент спектрального прибора: эшелетт, гониометр, Ваш глаз (чувствительность, острота зрения) - оказал наибольшее влияние на точность оценки разрешающей способности?

4. Нарисуйте принципиальную оптическую схему установки для определения рабочего порядка эшелетта и его рабочей длины волны.

5. Нарисуйте качественную зависимость интенсивноти от угла излучения в дальней зоне для рабочей длины волны. Эшелетт работает в автоколлимационном режиме.

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 46-50.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§46, 47.

## 22.4.3 Study of a Prism Using a Goniometer

Цель работы: знакомство с работой и настройкой гониометра Г5, определение зависимости показателя преломления стекла призмы от длины волны, определение марки стекла и спектральных характеристик призмы.

В работе используются: гониометр, ртутная лампа, призма.

Показатель преломления материала призмы $n(\lambda)$ удобно определять по углу наименьшего отклонения $\delta(\lambda)$ (рис. 4.8). Минимальное отклонение луча, преломлённого призмой, от направления луча, падающего на призму, получается при симметричном ходе луча (в призме луч идёт параллельно основанию). Угол минимального отклонения $\delta$, преломляющий угол $\alpha$ (угол при вершине призмы) и показатель преломления связаны соотношением

$$n(\lambda) = \frac{\sin\frac{\alpha+\delta(\lambda)}{2}}{\sin\frac{\alpha}{2}}$$

Вывести это соотношение студентам предлагается самостоятельно.

## ЗАДАНИЕ

В работе предлагается отъюстировать гониометр, определить преломляющий угол призмы, оценить спектральные характеристики призМы.

### I. Настройка гониометра

Ознакомьтесь с устройством гониометра и назначением элементов настройки. Нумерация, написанная на элементах лабораторного экземпляра прибора, и нумерация, приведённая на рис. П4.1 Приложения, совпадают. Установите зрительную трубу на бесконечность. Из пункта «Установка поверхности столика перпендикулярно оси прибора» выполните операции по предварительной и точной настройке. Настройте коллиматор. Установите начало отсчёта углов, после чего винт 26 следует закрепить и до конца измерений не освобождать.

### II. Качественные наблюдения

1. Найдите резкую границу белого света и тени (подоконник, крышка освещённого лампой стола и т. д.). Рассмотрите эту границу через призму. Для горизонтальной границы преломляющее ребро призмы должно
быть горизонтальным. Объясните появление цвета границы тени для двух положений призмы (ребро призмы выше и ниже основания). Отметьте взаимную ориентацию направлений: источник света (граница)призма-спектр.

### III. Установка призмы и измерение преломляющего угла

Качество обработки граней призмы значительно хуже обработки пластины, использованной при точной настройке столика гониометра. Поэтому требуется подстройка столика под имеющуюся призму.

2. Поставьте призму так, чтобы её преломляющее ребро было вертикально, а одна из рабочих граней перпендикулярна одному из винтов 8.

3. Вращая либо алидаду при свободном винте 24, либо столик при свободном винте 27 (винт 26 после установки начала отсчёта угла освобождать нельзя), найдите отражение светящегося креста от рабочей грани и приведите его к центру винтом 8, перпендикулярным отражающей грани. Трубу наклонять не следует. Найдите отражение автоколлимационного креста от второй рабочей грани призмы и вторым винтом 8 скорректируйте его положение по вертикали.

Вертикальное смещение креста не должно превышать трети радиуса поля зрения.

4. Снимите отсчёт двух углов, при которых наблюдается изображение автоколлимационного креста, отражённого от рабочих граней призмы. Вначале алидада вращается от руки, затем точно винтом 25 при отжатом фиксаторе 24. Рассчитайте преломляющий угол призмы и ошибку его определения.

**IV. Исследование спектра ртутной лампы**

5. Освободите фиксаторы верхней части столика 27 и алидады 24. Вращая столик и алидаду, восстановите примерную ориентацию осей коллиматор-призма-зрительная труба, которая была при наблюдении границы света и тени. Найдите в спектре ртути длинноволновую жёлтую линию. Вращая столик рукой, а затем винтом тонкой подачи 29 при зажатом винте 27, установите столик в такое положение, при котором отклонение линии от оси коллиматора оказывается наименьшим. Измерьте угол наименьшего отклонения $\delta$. Не меняя положение столика, измерьте координаты обеих жёлтых линий. Примерное расположение основных линий спектра ртути, их относительная яркость и длина волны приведены в Приложении (рис. П4.4, табл. 1). Измерьте угол наименьшего отклонения для всех приведённых линий.

6. Для оценки разрешающей способности призмы измерьте угловую ширину одной из жёлтых линий жёлтого дублета (по нулям интенсивно сти). Ширина щели коллиматора должна быть минимальной, позволяющей провести измерения. Измерьте линейкой длину основания призмы.

**V. Обработка результатов**

7. Для измеренных углов наименьшего отклонения 6 -ти ярких линий спектра ртути вычислите значение показателя преломления и постройте график $n(\lambda)$.

8. Проведите на графике $n(\lambda)$ наилучшую прямую и определите среднее значение дисперсии показателя преломления $\frac{dn}{d\lambda}$. Определите марку стекла призмы. Для стекла марки ТФ1, ТФЗ, ТФ5 среднее значение дисперсии показателя преломления в спектральной области излучения ртути равно соответственно $-1,9 \cdot 10^3, -2,4 \cdot 10^3, -2,7 \cdot 10^3$ см$^{-1}$.

9. По измеренным угловым координатам линий жёлтого дублета в опыте с неподвижным столиком рассчитайте значение угловой дисперсии (4.14) и величину $\frac{a}{h}$, приравнивая для указанных двух линий $\Delta\varphi = \frac{a}{h}\Delta n$. Рассчитайте значение $\frac{a}{h}$ из соображения геометрической оптики, считая известным показатель преломления для жёлтой линии и полагая преломляющий угол призмы равным 60°.

10. Оцените разрешимый спектральный интервал $\delta\lambda$. Согласно (4.17), для этого надо по измеренным данным вычислить угловую полуширину жёлтой линии и знать угловую дисперсию призмы вблизи этой линии. По формуле (4.19) оцените разрешающую способность спектрального прибора, включающего призму, гониометр, глаз экспериментатора. Используя найденное усреднённое значение дисперсии показателя преломления, оцените размер основания призмы, сравните его с измеренным. Оцените максимальную разрешающую способность призмы по формуле (4.19).

**Контрольные вопросы**

1. Дайте определение спектральных характеристик призмы. Чему равна её дисперсионная область?

2. Выведите соотношение, связывающее угол наименьшего отклонения и значение показателя преломления.

3. Дайте качественное объяснение зависимости показателя преломления стекла от длины волны в оптическом диапазоне.

4. Оцените максимальный шаг амплитудной решётки, при котором её угловая дисперсия в 1-м порядке сравнима с угловой дисперсией вашей призмы. Для этого шага оцените размер решётки, при которой её разрешающая способность в 1-м порядке спектра сравнима с разрешающей способностью призмы.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §49.

## 22.4.4 Fabry–Perot Interferometer

Цель работы: измерение длины волны жёлтых линий ртути, жёлтого дублета натрия, определение спектральных характеристик интерферометра Фабри-Перо.

В работе используются: интерферометр Фабри-Перо, линзы, светофильтры, ртутная и натриевая лампы, катетометр КН-6.

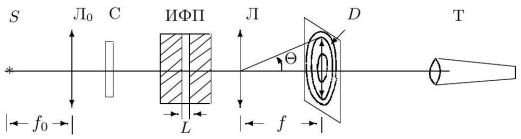



Рис. 1. Схема экспериментальной установки

Экспериментальная установка. Схема экспериментальной установки приведена на рис. 1. Свет от лампы $S$, пройдя через линзу $_0$ и светофильтр С, попадает на интерферометр Фабри-Перо (ИФП). Линза $_0$ служит для формирования пучка лучей (слегка сходящегося или слегка расходящегося). Интерференционные кольца наблюдаются в фокальной плоскости линзы Л. Картина рассматривается через зрительную трубу Т, сфокусированную на эту плоскость. Диаметры колец измеряются с помощью микроскопа катетометра. Зрительная труба Т, отсчётный микроскоп - элементы катетометра - прибора, предназначенного для измерения расстояний в вертикальной плоскости вдоль вертикальной оси. Подробное описание катетометра и инструкция к пользованию прилагаются к работе.

### ЗАДАНИЕ

В работе предлагается по результатам измерения диаметров интерференционных колец определить длины волн жёлтых линий ртути, жёлтого дублета натрия, расстояния между зеркалами интерферометра (по известной длине волны зелёной линии ртути), спектральные характеристики интерферометра.

### I. Юстировка системы

1. Включите лампу в сеть. Для ртутной лампы, кроме тумблера на блоке питания, следует нажать кнопку «ПОДЖИГ». Натриевую лампу следует включать только на время измерений из-за небольшого срока её службы.

2. На интерферометр должен падать слегка расходящийся пучок света, поэтому источник света $S$ располагается вблизи фокуса линзы $\Omega_0$. С помощью листа бумаги проследите за пучком света, падающим на интерферометр. Чем ближе интерферометр к линзе $\Omega_0$, тем большая часть пучка проходит через него.

3. Отодвиньте линзу Л и, приблизив глаз к интерферометру, рассмотрите систему колец невооружённым глазом. Если интерферометр отъюстирован правильно (зеркала параллельны), то при смещении глаза вправо-влево или вверх-вниз видимые размеры колец не изменяются. В том случае, когда при смещении глаза кольца сбегаются к центру или выплывают из него, обратитесь за помощью к лаборанту. Снова поставьте за интерферометром линзу Л. На фокальную плоскость этой линзы должна быть сфокусирована зрительная труба катетометра, на объективе которого указаны допустимые расстояния до плоскости наблюдения. Система колец должна быть резкой и равномерно освещена по вертикали (одинаковое число колец, расположенных выше и ниже их центра).

4. Пользуясь описанием катетометра, ознакомьтесь с устройством прибора, правилами его настройки и техникой измерений.

**II. Измерения**

5. В спектре ртутной лампы лабораторная установка позволяет наблюдать интерференционные кольца от зелёной линии, двух жёлтых и одной фиолетовой, в спектре натриевой лампы - жёлтый дублет. С помощью катетометра измерьте вертикальные координаты 5-6 диаметров для каждой спектральной линии. Рекомендуется, монотонно перемещая зрительную трубу катетометра снизу вверх, фиксировать координаты колец начиная с самого дальнего от центра. Пройдя центр, последовательно фиксируйте вторые координаты тех же колец. Если в центре системы колец наблюдается интерференционный максимум (относительно яркое пятно), то его при измерениях следует исключить. Удобно, пронумеровав предварительно кольца для каждой линии $i = 1, \ldots, 5, 6$

$)i = 1$ для кольца минимального диаметра), записывать соответствующие одному кольцу координаты друг под другом. Для спектра ртутной лампы измеряются диаметры зелёной и двух жёлтых линий (всего 30-36 отсчётов вертикальной координаты). Не забудьте отметить цвет
спектральной линии для спектра натриевой лампы - двух жёлтых линий (20-24 отсчётов).

6. Для оценки разрешаемой способности установки измерьте ширину (не полуширину) $\delta r$, ближайшего к центру кольца для зелёной и жёлтой линий в спектре ртути одной жёлтой линии в спектре натрия. Уровень половинной мощности при регистрации $\delta r$ оценивается на глаз.

Запишите фокусное расстояние $f$ линзы Л.

**III. Обработка результатов**

Диаметр интерференционного кольца $d$ и угол падения световой волны $\theta$ для малых $\theta$ связаны соотношением $\theta = \frac{d}{2f}$, где $f$ - фокусное расстояние линзы Л. Обе величины $\theta$ или $d$ можно использовать при обработке результатов измерений.

Для малых углов $\theta$ условие возникновения интерференционного кольца (4.6) приводится к виду

$$\theta_m^2 = 2 - \frac{\lambda}{L}m$$

При переходе от любого кольца к соседнему с большим диаметром порядок интерференции уменьшается на 1. Т.к. кольца пронумерованы $i = 1, \ldots, 5, 6$, возникает зависимость угла на максимум интерференции от его номера или диаметра кольца:

$$\frac{d^2(i)}{4f^2} = \theta^2(i) = \text{ const } + \frac{i\lambda}{L}$$

Если отложить по оси абсцисс номер кольца $i$, а по оси ординат величины $\theta^2(i)$ или $d^2(i)$, то по тангенсу угла наклона получившейся прямой можно определить базу интерферометра при известной длине волны или длину волны при известной базе. Для пары жёлтых линий можно вычислить среднюю длину волны $\bar{\lambda}$, используя квадраты угловых радиусов $\theta_1^2(i)$ и $\theta_2^2(i)$:

$$\frac{d_1^2(i) + d_2^2(i)}{8f^2} = \frac{\theta_1^2(i) + \theta_2^2(i)}{2} = \text{ const } + i\frac{\bar{\lambda}}{l}$$

и разность $\delta\lambda$ их длин волн

$$\frac{[d_1(i) - d_2(i)]\,[d_1(i) + d_2(i)]}{4f^2} = \theta_1^2(i) - \theta_2^2(i) = \text{ const } + i\frac{\delta\lambda}{L}$$

Отметим, что наблюдаемая пара жёлтых колец имеет один порядок интерференции $m$, если разность длин волн $\delta\lambda$ меньше дисперсионной
области интерферометра $\Delta\lambda$, как в случае жёлтого дублета натрия, и порядок спектра внутри пары отличается на 1 для жёлтых линий ртути.

1. Используя значение средней длины волны зелёной линии ртути $\lambda_3 = 5461$Å, определите базу интерферометра $L$.

2. Определите среднюю длину волны и разность длин волн для пары жёлтых линий ртути. Сравните полученные значения с приведёнными в приложении, табл. 1. Определите среднюю длину волны и разность длин волн для жёлтого дублета натрия.

3. Оцените максимальный порядок интерференции и дисперсионную область интерферометра (см. (4.22)).

4. Используя известное значение разности длин волн жёлтого дублета натрия $\delta\lambda = 6$Å, рассчитайте по результатам измерений спектра натрия угловую дисперсию интерферометра и проверьте полученную раннее зависимость (4.13).

5. Используя связь углового радиуса первого интерференционного кольца $\theta_1$, его ширины по уровню половинной мощности $\delta\theta_1$ и разрешающей способности $R = \frac{\lambda}{\delta\lambda} \approx \frac{1}{\theta_1\delta\theta_1}$, оцените аппаратную разрешающую способность установки в целом по зелёной линии ртути и двум жёлтым (ртути и натрия). Оцените число интерферирующих волн и коэффициент отражения поверхности интерферометра.

**Контрольные вопросы**

1. Дайте определения и приведите формулы расчёта спектральных характеристик интерферометра Фабри-Перо.

2. Если бы жёлтые линии ртути были не узкими, а перекрывали друг друга, т.е. имели ширину 21Å, то можно ли исследовать такой спектр вашим интерферометром?

3. Оцените минимальный диаметр зеркал интерферометра, при котором разрешается жёлтый дублет натрия.

4. Оцените длительность импульса излучения натриевой лампы, при которой интерферометр не разрешает жёлтый дублет. Для этого случая оцените путь, который проходит свет внутри интерферометра, сравните его с базой интерферометра.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. III. 36.

2. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. §8.9.

3. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. VII. 30.

## 22.4.5 Appendix about Spectral Instruments

### А. Устройство гониометра Г5

Внешний вид гониометра представлен на рис. П4.1. Коллиматор 3, столик 7 и алидада 17 со зрительной трубой 12 крепится на массивном основании 23. На столике 7 размещаются исследуемые объекты. Коллиматор закреплён неподвижно, а столик и алидада с трубой могут вращаться вокруг вертикальной оси.

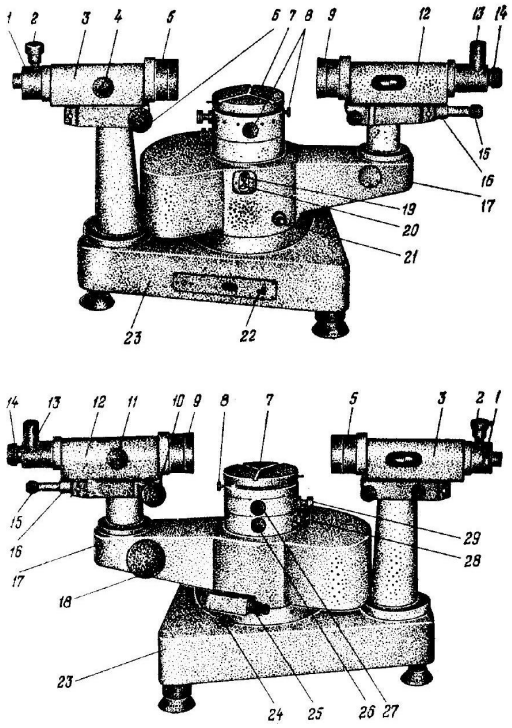



Рис. П4.1. Внешний вид гониометра Г5

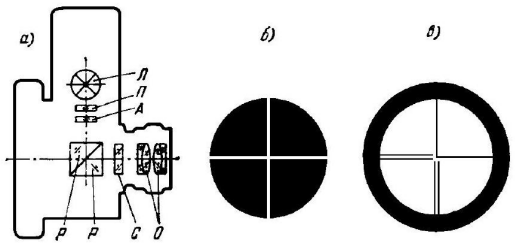



Рис. П4.2. Автоколлимационное устройство

Ширину коллиматорной щели можно менять от 0 до 2-х мм при помощи микрометрического винта 2, высоту - от 0 до 20 мм - при помощи диафрагмы с треугольным вырезом («ласточкин хвост»), надетой на щель. Винт 4 служит для настройки коллиматора на параллельный пучок. Зрительная труба 12 состоит из объектива 9 и окуляра 14 с автоколлимационным устройством 13. объективы коллиматора и зрительной трубы одинаковы. Фокусировка трубы производится винтом 11. Наклон коллиматора и зрительной трубы к горизонтально оси изменяется винтами 6 и 10 соответственно.

Схема окуляра зрительной трубы с автоколлимационным устройством приведена на рис. П4.2а.

Свет от лампы Л проходит через защитную стеклянную пластинку и попадает на автоколлимационную сетку А, содержащую две взаимно перпендикулярные щели (рис. П4.2б). Свет, прошедший через сетку А, попадает на две прямоугольные призмы Р и отражается от гипотенузной грани, на которую нанесён полупрозрачный слой с коэффициентом отражения 50%. Для юстировки гониометра на столик ставится предмет с плоской отражающей поверхностью. После отражения от неё параллельный пучок лучей возвращается назад в зрительную трубу и собирается в фокальной плоскости объектива. В этом случае светящийся крест можно увидеть через окуляр зрительной трубы. Кроме того, в окуляре имеется ещё одна сетка С, на которой изображён чёрный отсчётный крест (рис. П4.2в). Совмещённые изображения обоих крестов рассматриваются через окулярные линзы О. Резкость видимого изображения отсчётного креста регулируется вращением оправы окуляра 14.

Обе сетки окуляра, А и С (рис. П4.2а), расположены на строго одинаковых расстояниях от гипотенузных граней призмы Р, поэтому их одновременное наблюдение в окуляре возможно только при совпадении фокальных плоскостей объектива и окуляра (труба настроена на бесконечность).

Важнейшим узлом гониометра является устройство, служащее для отсчёта угла поворота зрительной трубы вокруг вертикальной оси, проходящей через центр столика. На этой оси крепится прозрачное кольцо (лимб), расположенное в корпусе прибора. На поверхности лимба нанесена шкала с делени ями. Лимб разделён на $3 \times 360 = 1080$ делений. Цена деления $20'$, оцифровка делений произведена через $1^\circ$. Шкалу лимба можно наблюдать через окуляр отсчётного устройства 16 при включённой подсветке (тумблер 22). Резкость изображения шкалы регулируется вращением оправы окуляра 15.

Оптическая система отсчётного устройства собрана так, что через окуляр можно наблюдать изображения штрихов двух диаметрально противоположных участков лимба, причём одно изображение прямое, а другое обратное (рис. П4.3). Кроме того, оптическая система позволяет перемещать эти изображения друг относительно друга, оставляя в покое как лимб, так и алидаду со зрительной трубой. Это перемещение штрихов измеряется при помощи оптического микрометра. Шкала микрометра рассчитана таким образом, что при перемещении её на 600 делений верхнее изображение штрихов лимба смещается относительно нижнего на 10. Следовательно, цена деления шкалы микрометра $1''$.

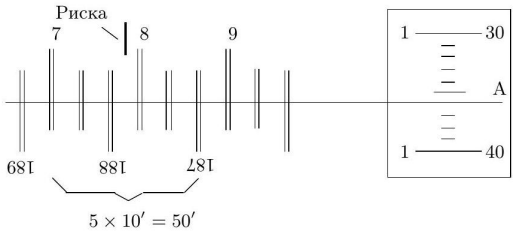



Рис. П4.3. Поле зрения отсчетного микроскопа ($7°51'36''$)

Поле зрения отсчётного микроскопа приведено на рис. П4.3. В левом окне наблюдаются изображения диаметрально противоположных участков лимба и вертикальный штрих для отсчёта градусов и десятков угловых минут, в правом - деления шкалы оптического микрометра и горизонтальная риска А для отсчёта минут и секунд.

### Б. Юстировка гониометра

Гониометр требует тщательной юстировки, которая заключается в настройке: а) зрительной трубы на бесконечность; б) поверхности столика и оптической оси трубы - перпендикулярно оси вращения прибора; в) коллиматора - на параллельный пучок лучей; г) оптической оси коллиматора перпендикулярно оси вращения прибора.

Любой вращающийся элемент системы должен свободно поворачиваться рукой, если освобождён соответствующий закрепляющий винт.

#### I. Установка зрительной трубы на бесконечность

1. Вращением оправы 14 настройте окуляр зрительной трубы на резкое изображение чёрного отсчётного креста (крест можно подсветить рассеянным светом настольной лампы).

2. Включите тумблер 22 - подсветка - и убедитесь, что светящийся крест автоколлимационной сетки виден, если посмотреть в трубу со стороны объектива.

Прислоните к оправе объектива плоскопараллельную пластину и, перестраивая объектив винтом 11, найдите резкое изображение светящегося креста, отражённого от стеклянной поверхности.

II. Установка поверхности столика перпендикулярно оси вращения прибора

Поле зрения зрительной трубы гониометра составляет менее 2°, поэтому этот этап настройки требует повышенного терпения и внимания.

3. Предварительная настройка. Установив зрительную трубу напротив коллиматора, зафиксируйте её положение винтом 24. Убедитесь, что рычажок 20 в верхнем положении (рычажок 19 нажат). Закрепите винт 26 и освободите винт 27. В таком положении верхняя часть столика может свободно вращаться от руки при неподвижной алидаде и лимбе. Поверните верхнюю часть столика так, чтобы один из винтов 8 был направлен вдоль оси трубы. Наклоняя столик двумя ориентированными взаимно перпендикулярно винтами 8, установите на глаз его плоскость горизонтально. При этом удобно следить за величиной щели между крышкой столика и торцом его основания-цилиндра.

4. Поставьте на столик уголковый отражатель (призму с углом при вершине 90°). Установите призму так, чтобы её гипотенузная грань была перпендикулярна оси зрительной трубы и оси одного из винтов 8. Вращая этот винт и слегка поворачивая рукой столик вправо и влево, добейтесь того, чтобы в поле зрения окуляра попал

светящийся крест, отражённый от призмы. Совместите центр отражённого креста с горизонтальным штрихом отсчётного креста (смещение по горизонтали роли не играет). Если отражённых крестов несколько, работайте с самым ярким.

5. Поднимите отражатель, поверните столик на 180° и поставьте отражатель на столик. Снова совместите центр автоколлимационного креста с горизонтальным штрихом отсчётного, на этот раз половину расстройки устраните винтом 8, а вторую половину - винтом 10 наклона трубы.

Снова поверните столик и призму на 180° и теми же действиями уточните настройку. Допустимо смещение креста на треть радиуса поля зрения. Теперь можно считать, что ось вращения столика лежит в вертикальной плоскости, перпендикулярной оси трубы.

Поверните столик на 90°, установите на него отражатель и вторым винтом 8 совместите изображение автоколлимационного креста с отсчётным. Теперь можно считать, что оси трубы и вращения столика находятся в одной вертикальной плоскости и они взаимно перпендикулярны.

6. Точная настройка. На этом этапе настройки вместо уголкового отражателя используется плоскопараллельная пластина. Плоскости этой пластины перпендикулярны её узкой полированной грани. На эту грань следует поставить пластину на столик так, чтобы её рабочие грани были примерно перпендикулярны осям трубы и одного из винтов 8. Вращая винт 8 и слегка поворачивая столик рукой, добейтесь того, чтобы в поле зрения окуляра попал отражённый светящийся крест. Совместите автоколлимационный и отсчётный кресты винтом 8 и винтом 10 наклона трубы, как в пункте 5. Не трогая пластину, поверните столик на 180° и уточните настройку винтами 8 и 10.

Поверните столик на 90°, поставьте на него пластину, вторым винтом 8 совместите изображения крестов. Проверьте настройку при повороте столика на 180°, допустимое расхождение изображений по вертикали не более трети радиуса поля зрения трубы.

**III. Настройка коллиматора**

7. Включите блок питания ртутной лампы и нажмите кнопку «ПОДЖЖИГ». Расширьте входную щель коллиматора и поставьте ртутную лампу вплотную к щели. Расположите лист бумаги за объективом коллиматора и перемещением источника света перпендикулярно оси коллиматора добейтесь максимальной освещённости центральной части светового пятна. Освободите алидаду (винт 24) и установите настроенную на бесконечность зрительную трубу напротив коллиматора. Уточните центровку изображения щели винтом 25 при зажатом винте 24. Наблюдая за изображением щели в зрительную трубу, вращением винта 4 (перемещением объектива коллиматора) настройтесь на резкое изображение щели. Ограничьте до нескольких миллиметров видимую высоту щели, перемещая вдоль неё диафрагму «ласточкин хвост».

Изменяя наклон коллиматора винтом 6, выведите резкое изображение щели на середину поля зрения окуляра.

Подберите рабочую ширину входной щели так, чтобы её видимая ширина составляла 1,5-2 расстояния между полосами двойного отсчётного штриха.

### В. Отсчёт углов

Глядя в окуляр 15 и вращая оправу окуляра, настройтесь на резкое изображение лимба (круговой шкалы) и вертикальной шкалы, расположенной справа от лимба. Вращением маховичка оптического микрометра 18 (без усилий!) совместите ближайший к отсчётной риске левый двойной оцифрованный штрих с ближайшим двойным штрихом нижней шкалы (если совместить штрихи без усилия не удаётся, вращайте маховичок 18 в другую сторону и совмещайте соседний нижний штрих); снимите отсчёт угла в градусах (7° на рис. П4.3). Для определения десятков минут просчитайте число интервалов между двойными штрихами от выбранного верхнего до нижнего, отличающегося на 180° (на рис. П4.3 от 7 до 187 между двойными штрихами 5 интервалов, каждый по 10′, т.е. 50′).

Число минут - это левая цифра вертикальной шкалы над отсчётной горизонтальной риской А (на рис. П4.3-1'). Число секунд - это отсчёт по правой части вертикальной шкалы (на рис. П4.3 положение горизонтальной риски А даёт 36″). Положение, показанное на рис. П4.3, соответствует отсчёту 7°51′36″. Отсчётное устройство гониометра обеспечивает точность измерений угла не хуже 5″, или $2,4 \cdot 10^{-5}$ рад (по паспорту Г5).

### IV. Установка отсчёта углов

1. Расположите трубу против коллиматора. Вращением микрометрического винта 25 при закреплённом винте 24 совместите двойной отсчётный штрих с изображением входной щели.

2. Для удобства отсчёта можно выставить начальный угол $\approx 180°$. Для этого вращением маховичка 18 установите горизонтальную отсчётную риску А на отсчёт 1′00″. Дополнительная минута рекомендуется потому, что при настройке точно на 0′00″ (на край шкалы) часто выходит из строя маховичок 18.

Поворачивая нижнюю часть столика сначала от руки при свободном винте 26, а затем микрометрическим винтом подачи 28 при зажатом винте 26 совместите верхний штрих $-180°$ и нижний 0°. При такой настройке углы, под которыми видны спектральные линии, будут отсчитываться от 0°1′00″.

Если настройка на 180° затруднена, запишите такое начало отсчёта, которое вы установили. Чтобы не сбить начало отсчёта, винт 26 не следует освобождать до конца измерений. Поворачивать столик можно, освобождая только винт 27 (вращается верхняя часть столика, но не шкала).

Полезно в процессе работы уточнять начало отсчёта.

### Г. Спектр ртутной лампы

Каждая линия спектра имеет свою ширину и тонкую структуру. Ниже приведены некоторые интегральные характеристики спектральных линий для лампы ДРШ-250.



12






Рис. П4.4. Спектр ртутной лампы ДРШ-250

Т а б л и ц а 1

Характеристики спектра ртутной лампы ДРШ-250

| № | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\lambda$ нм. | 579,1 | 577,0 | 546,1 | 491,6 | 435,8 | 404,7 |
| Цвет | желт. | желт. | зелен. | голуб. | синий | фиолет. |
| Яркость | 10 | 8 | 10 | 4 | 4 | 3 |

## 22.5 Experiments about Laser Radiation

### 22.5.1 Helium–Neon Laser

Цель работы: Изучение основных принципов работы гелий-неонового лазера, свойств лазерного излучения и измерение усиления лазерной трубки.

В работе используются: объект исследования - активный элемент от гелий-неонового лазера ЛГ-75 с блоком питания, дополнительный He − Ne-лазер для юстировки и измерений, модулятор излучения (обтюратор), фотодиоды, зеркала, поляроид, компьютер со звуковой картой.

Перед проведением работы необходимо познакомиться с устройством и общими принципами работы лазеров в теоретическом введении к разделу.

Существенное отличие гелий-неонового лазера от большинства других - малое усиление активного элемента, обычно интенсивность света увеличивается на $1-3\%$ на один проход. Чтобы измерить такое усиление, необходимо регистрировать интенсивность входного и выходного излучения с точностью доли процента или лучше. Это требует применения прецизионных приборов, которые обычно дороги. В данной работе точные измерения осуществляются с помощью простого и дешевого оборудования.

В лазере ЛГ-75 окошки лазерной трубки приклеены к торцам под углом Брюстера, что позволяет для одной из поляризаций исключить потери на отражение от этих окошек. В лазере генерируется излучение именно с этой поляризацией, для другой поляризации потери превышают усиление и лазерная генерация невозможна. Поэтому излучение лазера на исследуемой трубке линейно-поляризовано. В более современных лазерах зеркала резонатора приклеиваются непосредственно к торцам трубки и излучение таких лазеров неполяризовано.

Гелий-неоновый лазер. Рассмотрим механизм возникновения усиления в рабочей среде гелий-неонового лазера. Лазерная трубка заполняется смесью гелия и неона в соотношении от 5 : 1 до 10 : 1 с общим давлением порядка $10^2$ Па, при котором довольно легко возбудить постоянный электрический разряд. Рабочим лазерным веществом

является неон. Гелий используется для избирательного заселения верхнего рабочего уровня неона. Атомы гелия возбуждаются при столкновениях с разогнанными в электрическом поле разряда электронами. Передача энергии от возбуждённых атомов гелия к атомам неона осуществляется

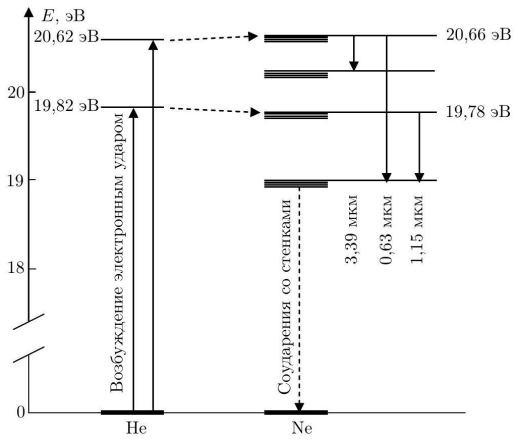



Рис. 1. Энергетическая схема работы гелий-неонового лазера

при столкновениях между ними. Известно, что наиболее эффективно передача энергии от атома к атому происходит в резонансном случае, то есть когда энергии уровней, между которыми происходит переход, близки. Упрощённая схема энергетических уровней атомов гелия и неона изображена на рис. 4. Из него видно, что энергии двух уровней атома гелия действительно близки к двум уровням неона, что приводит к эффективной передаче энергии от гелия к неону. На этом рисунке изображена лишь малая часть энергетических уровней неона, в действительности их гораздо больше, и все они в той или иной степени заселяются в электрическом разряде даже без добавки гелия, что может приводить к созданию инверсной населённости между некоторыми уровнями. И действительно, генерация лазерного излучения атомами неона получена в лабораторных условиях на более чем 200 переходах. Однако во всех промышленных лазерах на неоне для увеличения эффективности накачки используют селективное заселение верхних лазерных уровней
атомами гелия, поэтому такие лазеры называются гелий-неоновыми. На рисунке изображены три основных лазерных перехода с длинами волн 0,63 мкм (красное излучение), а также 1,15 и 3,39 мкм (невидимое инфракрасное). Кроме обычных лазеров с длиной волны 0,63 мкм выпускаются гелий-неоновые лазеры со сменными наборами зеркал, селективно отражающими излучение на одной из этих трёх длин волн, что позволяет выбирать длину волны генерируемого излучения. При этом для других длин волн из-за малого отражения пороговое условие генерации не выполняется.

Время жизни верхнего лазерного уровня для перехода 0,63 мкм составляет $10^{-8}$ с. Из принципа неопределённости $\Delta\nu \cdot \tau \geqslant 1$ можно получить оценку ширины линии $\Delta\nu \geqslant 10^8$ Гц. Реальная ширина спектра генерации обычного гелий-неонового лазера на порядок больше. Основной механизм уширения - эффект Доплера. При температуре 400 К полуширина спектра излучения газообразного неона равна $1,5 \cdot 10^9$ Гц. Полуширина линии лазерного излучения в $2 - 3$ раза меньше этой величины, поскольку генерация происходит только на тех частотах, где усиление не только больше единицы, но и превышает потери.

Следует отметить, что для поддержания инверсной населённости при работе непрерывного лазера необходимо не только заселение верхнего лазерного уровня, но и быстрое опустошение нижнего. В гелийнеоновом лазере это происходит при соударении атомов неона, находящихся на нижнем лазерном уровне, со стенками лазерной трубки,

при этом атомы передают энергию стенкам и сбрасываются ещё ниже, в основное состояние. Поэтому в современных лазерах трубки делаются с маленьким внутренним диаметром порядка 1-2 мм при длине 20-60 см. Дальнейшее уменьшение диаметра нецелесообразно из-за возрастания дифракционных потерь. Недостаточно быстрое опустошение нижнего лазерного уровня в гелий-неоновых лазерах ограничивает и предельный коэффициент усиления, достигаемый при некотором оптимальном разрядном токе. При дальнейшем увеличении тока нижний уровень не успевает опустошаться и эффективность генерации падает. Обычно достигается усиление всего $1 \div 3\%$ за один проход, т.е. $G = 1,01 \div 1,03$. При таком малом усилении генерация вынужденного излучения может быть получена, только если окна лазерной трубки либо очень хорошо просветлены, либо расположены под углом Брюстера к оси резонатора, при этом для одной из поляризаций потери на отражение от окошек исчезают. Отражение зеркал резонатора должно быть очень высоким, обычно $r_1 \geqslant 0,998, r_2 \approx 0,99$ (выходное зеркало). Такие отражения недостижимы с металлическими зеркалами (коэффициент отражения свеженапылённого алюминия $\approx 0,90$, серебра $\approx 0,98$). Поэтому используются
специальные зеркала, в которых на стеклянную подложку нанесены (обычно напылением) чередующиеся слои диэлектриков с сильно различающимися показателями преломления, например, $SiO_2$ с $n = 1,45$ и $TiO_2$ с $n = 2,5$. Толщины слоёв подобраны таким образом, чтобы все волны, отражённые от границ разделов слоёв, на выходе складывались в фазе, тогда при количестве слоёв $N \geqslant 10$ удаётся достичь отражения $r \geqslant 0,998 \div 0,999$. Сообщалось даже о получении коэффициента отражения более 0,9999 при числе слоёв несколько десятков.

**Предупреждение: напылённые зеркала, особенно старых лет выпуска, очень непрочные. Ни в коем случае не прикасайтесь к ним руками и другими предметами. Нежелательно даже дышать на них.**

Типичная конструкция гелий-неонового лазера изображена на рис. 2.




Рис. 2. Устройство гелий-неонового лазера

Обычно используется сферический или полусферический резонатор, предъявляющий более мягкие требования к точности юстировки зеркал и менее чувствительный к их угловым отклонениям по сравнению с плоским резонатором.

Экспериментальная установка. Схема экспериментальной установки приведена на рис. 3. На одном оптической рельсе расположены: головка промышленного He-Ne-лазера ЛГ-75 с исследуемой газоразрядной трубкой (11), заключённой в кожух (10), рейтер с полупрозрачным зеркалом (4), фотодиоды (5 и 6), а также 3 съёмных рейтера с выходным зеркалом (9), отрицательной линзой для наблюдения модовой структуры излучения исследуемого лазера или поляроидом для исследования поляризации выходного излучения лазера (8) и с белым экраном (7).

Юстировочный лазер (1) с белым экранчиком (2) и модулятор (3) закреплены на втором оптическом рельсе. Модулятор может быть повёрнут в разные положения: при измерении коэффициента усиления он модулирует пучок, идущий от юстировочного лазера, при измерении поляризации излучения исследуемого лазера он модулирует выходящее из

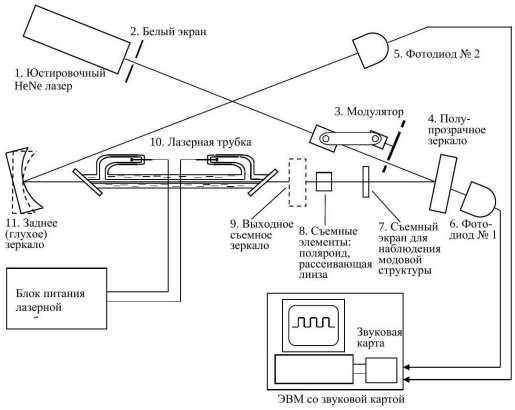



Рис. 3. Схема экспериментальной установки. Штриховыми линиями показано положение зеркал при получении лазерной генерации на исследуемой трубке

него излучение. В остальных случаях модулятор отводится в сторону, чтобы не перекрывать пучки.

Юстировочный лазер предназначен для настройки положения всех элементов установки и является источником зондирующего излучения для измерения усиления активной среды исследуемого лазера.

Зондирующий пучок сначала попадает на полупрозрачное зеркало (4). Часть излучения проходит сквозь зеркало и попадает на фотодиод № 1 (6), с которого снимается сигнал, пропорциональный интенсивности зондирующего пучка. Отражённая часть направляется в исследуемую трубку.

Такая методика позволяет исключить влияние нестабильности интенсивности зондирующего лазера во времени, которое составляет обычно несколько процентов. Сигналы с обоих фотодиодов подаются на звуковую карту компьютера и с помощью программы PhysLab

измеряются с высокой точностью. Однако, хотя сигналы с фотодиодов пропорциональны интенсивностям входного и выходного излучений, коэффициенты пропорциональности неизвестны. Они зависят не только от чувствительности фотодиодов и коэффициентов отражения полупрозрачного и глухого зеркал, но и от возможных потерь при прохождении через трубку, если края пучка «задевают» за стенки трубки вследствие неточной юстировки. Поэтому вычислить усиление трубки по отношению сигналов с фотодиодов невозможно. Для преодоления такой трудности используется следующий приём: вычисляется отношение сигналов с фотодиодов два раза. Один раз при выключенном питании трубки, когда усиления нет, второй раз при включённом питании, когда появляется усиление. Затем делится второе значение на первое и получается усиление. Таким образом, вычисляется «отношение отношений». Для увеличения точности и удобства измерений чувствительности фотодиодов подбираются таким образом, чтобы сигналы с них были приблизительно одной величины. Для этого пучок, попадающий на фотодиод № 1, ослабляется с помощью матовых пластинок, помещённых внутри кожуха фотодиода.

Сигналы с фотодиодов можно измерять достаточно точным цифровым амперметром или вольтметром постоянного тока. Однако при этом нужно вычитать из показаний темновой ток фотодиода, который сильно зависит от температуры и поэтому плавает во времени, что усложняет измерения. Самый простой путь устранения влияния темнового тока - промодулировать интенсивность излучения, например, вращающимся обтюратором и перейти к измерениям на переменном токе. При этом также устраняется влияние внешних засветок и так называемых фликер-шумов, амплитуда которых растёт по закону $1/f$, где $f$ частота.

Для получения генерации вынужденного излучения лазера на исследуемой трубке

используется съемное выходное зеркало (9), закреплённое в юстируемой оправе. Наблюдение модовой структуры излучения исследуемого лазера производится с помощью короткофокусной отрицательной линзы (8), позволяющей на коротком расстоянии увеличить размер выходного пучка, и съемного белого экрана (7). Изучение поляризации исследуемого лазера производится с помощью вращающегося поляроида, устанавливаемого вместо отрицательной линзы, модулятора и фотодиода № 2.

Проведение измерений. Измерение усиления трубки гелийнеонового лазера в данной работе производится с помощью компьютера со звуковой картой и программы PhysLab [1]. Эта программа превращает компьютер в измерительный центр, включающий в себя такие виртуальные приборы, как цифровой осциллограф с возможностью измерения пиковых значений и эффективных напряжений сигналов, анализатор спектра, синхронный детектор, позволяющий измерять сигналы, амплитуда которых в десятки и сотни раз меньше шумов, генератор и некоторые другие приборы. Основой всех этих виртуальных приборов является звуковая карта, в которой имеются аналогово-цифровые преобразователи (АЦП). Они могут периодически измерять амплитуду любых электрических сигналов по двум (иногда более) каналам и переводить амплитуду в цифровой код. Частота измерений может выбираться из ряда $11, 22, 44$ и иногда 48 кГц (в некоторых современных звуковых картах до 192 кГц). Измеренные значения могут накапливаться в памяти компьютера и затем с ними можно производить различные преобразования. Например, используя процедуру «быстрого преобразования Фурье» (FFT), можно вычислить спектр измеряемых сигналов, не производя собственно измерений спектра. В звуковой карте имеются также ЦАПы - цифро-аналоговые преобразователи, которые позволяют использовать звуковую карту как генератор сигналов различной формы. Если звуковая карта имеет разрядность 16 бит, теоретический динамический диапазон составляет $2^{16} \approx 6,6 \cdot 10^4$ (96 дБ) и можно ожидать относительной точности измерений порядка $2 \cdot 10^{-5}$. Реальные звуковые карты имеют значительные собственные шумы, которые «съедают» несколько младших разрядов. Типичный динамический диапазон дешевых звуковых карт по входу 70-80 дБ, у самых качественных 16-битных карт до 85-90 дБ [2]. Современные звуковые карты обеспечивают динамический диапазон до $\sim$ 120 дБ. В нашей установке может использоваться простая дискретная двухканальная звуковая карта либо встроенная в материнскую плату звуковая карта, которая, тем не менее, позволяет обеспечить относительную точность измерений $\sim 10^{-3}$.

В данной работе из всех виртуальных приборов используется только цифровой осциллограф. Для его запуска нужно дважды щёлкнуть мышкой по иконке PhysLab на экране монитора. Сначала появится сообщение: Calibration file is absent one default factor is used. Щёлкните по кнопке Ok, поскольку для данной работы нужны лишь относительные измерения, и абсолютная калибровка чувствительности не обязательна. Раскроется цифровой осциллограф (он по умолчанию запускается первым) и меню с перечнем других приборов.

[01] Разработчик PhysLab - С.П. Палто, Институт кристаллографии РАН.

[02] Напомним, что децибел (дБ) - это логарифмическая единица измерения отношения величин двух сигналов. По определению, для двух сигналов амплитуды $A$ и $B$ отношение их мощности в дБ равно $20\lg(A/B)$ или $10\lg\left(A^2/B^2\right)$.

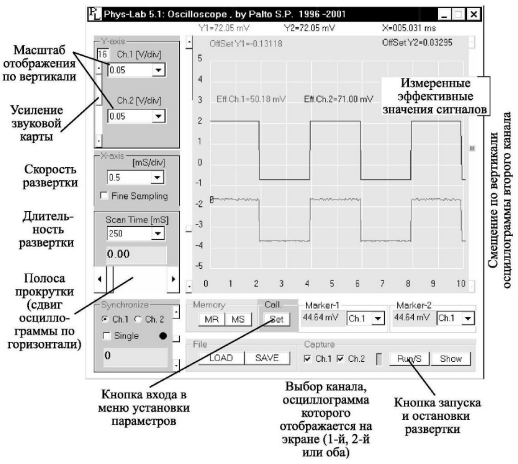



Рис. 4

Слева от экрана осциллографа расположены окошки, в которых можно выбрать скорость развёртки, длительность развёртки (она может быть много больше отрезка времени, умещающегося на экране, тогда осциллограмму можно передвигать по экрану с помощью бегунка слева внизу от экрана), масштаб отображения по вертикали (чувствительность) и регулировка усиления звуковой карты. В самом низу левой колонки расположены окошки выбора способа и уровня синхронизации запуска: по сигналу на первом или втором канале, по внутренним часам. В данной работе практически неважно, какой способ синхронизации выбран, немного удобнее синхронизоваться по сигналу на каком-либо канале.

Под экраном слева расположена кнопка Set (установка), по которой обязательно надо щёлкнуть, чтобы зайти в меню установок и проконтролировать их. Должен быть выбран линейный вход, частоту дискретизации звуковой карты лучше выбрать побольше)44 кГц), чтобы увеличить число измеряемых точек на длине выбранной развёртки и за счёт этого увеличить точность. Под экраном справа имеется кнопка Run/S (запуск/остановка). Щелчок по ней запускает развёртку осциллографа, повторный щелчок - останавливает, при этом на экране остаётся последняя оставшаяся в памяти осциллограмма. Слева от этой кнопки находится индикатор. Когда развёртка запущена, индикатор периодически мигает или светится красным цветом, показывая, что происходят измерения.

**Важно: все изменения временных параметров развёртки (скорость и длительность развёртки, частоту дискретизации) можно производить только при выключенном запуске, иначе программа может зависнуть, и её придётся перезапускать.**

Величину сигналов с фотодиодов в PhysLab можно измерять разными способами: например, определять размах осциллограммы с помощью маркеров (при перемещении маркера по экрану осциллографа всё время отображаются значения абсциссы и ординаты); можно выбрать режим вывода на экран пиковых значений сигнала относительно нулевой линии или эффективного значения напряжения в осциллограмме первого и второго каналов. В данной работе предлагается использовать последний способ, как наименее подверженный влиянию шумов. Для его реализации нужно щёлкнуть правой кнопкой мыши в любом месте экрана осциллографа, появится меню, из которого следует выбрать пункт View, затем EffCh1. Теперь эффективное значение напряжения в первом канале будет отображаться на экране осциллографа для каждой осциллограммы. Затем нужно повторить процедуру для второго канала. Можно таким же образом включить вывод на экран остальных величин: пикового и среднего значений в каждом канале, тогда будут одновременно выводиться 6 величин, из которых нам

нужны только эффективные значения напряжения. Они вычисляются компьютером по формуле

$$U_{\phi\phi} = \sqrt{\frac{1}{n}\sum U_i^2}$$

где $i$ - номер отсчёта, $n$ - полное число отсчётов. Может оказаться, что на экране высвечиваются ненулевые значения $U_{\text{эфф}}$, даже когда сигнал с фотодиода заведомо равен нулю, например, пучок света полностью перекрыт. Это вызвано сдвигом нуля в звуковых картах, особенно
старых. Легко показать, что связанная с этим сдвигом относительная ошибка квадратична по величине сдвига, она равна $k\,(U_{\text{сдв}}\,/U_{\text{эфф}}\,)^2$, где $k$ - коэффициент, зависящий от формы импульсов. Для последовательности прямоугольных импульсов со скважностью 2 (меандр) $k = 0,5$, при скважности 5 этот коэффициент увеличивается до 1,5. Поэтому, если отношение измеряемой величины при наличии и отсутствии сигнала превышает 30, наличие сдвига для сигналов с небольшой скважностью даёт относительную ошибку порядка $10^{-3}$ или менее, и ею можно пренебречь.

## ЗАДАНИЕ

1. Включите питание юстировочного лазера и убедитесь в наличии лазерного излучения. С помощью листочка белой бумаги ознакомьтесь с ходом лазерного пучка по установке.

2. Проведите юстировку оптической схемы установки, целью которой является хорошее прохождение пучка юстировочного лазера сквозь трубку. При этом выходное зеркало исследуемого лазера должно быть снято (удобнее его снимать вместе с рейтером), а модулятор отведен в сторону, чтобы не перегораживать излучение юстировочного лазера. Глухое зеркало также лучше снять вместе с рейтером, чтобы наблюдать прошедшее излучение на стене. Лазерный пучок должен проходить строго по центру трубки. Для этого сначала нужно добиться, чтобы пучок юстировочного лазера входил в трубку точно по центру (он хорошо виден на поверхности брюстеровского окошка трубки), затем обеспечить соосность пучка и лазерной трубки. Это делается качанием полупрозрачного зеркала №4 и, если этого недостаточно, качанием юстировочного лазера. Возможно, эту операцию придется повторить несколько раз. Если пучок какой-либо стороной касается стенки трубки, на экране на стене появляется кольцевой ореол. При правильной настройке пятно на экране должно быть равномерно освещенным, с ровными краями без ореолов и при малой расстройке хода пучка вверх-вниз и влево-вправо при качании полупрозрачного зеркала №4 практически не менять свою форму.

3. Измерьте усиление трубки.

Для этого нужно поставить глухое зеркало и развернуть его так, чтобы прошедший сквозь трубку пучок попадал на центр фотодиода №2. Фотодиод №1 за полупрозрачным зеркалом следует выставить таким образом, чтобы прошедший через зеркало пучок также попадал на центр.

Включите мотор модулятора и расположите диск модулятора таким образом, чтобы он прерывал пучок юстировочного лазера до попадания
на полупрозрачное зеркало. Выберите подходящие параметры осциллографа в PhysLab и пронаблюдайте сигналы с обоих фотодиодов. Они должны иметь одинаковую временную форму, не превышать 100 мВ и различаться по амплитуде не более чем в

два раза. Если это не выполняется, убедитесь в том, что выходное зеркало снято и проверьте юстировку. Добейтесь нормальных сигналов амплитудой $60 - 90$ мВ на экране осциллографа. Проведите по нескольку раз измерения с включённым и выключенным питанием трубки для трёх значений тока через трубку.

Важно: Если амплитуды сигналов не попадают в нужный интервал, например, хотя бы один превышает 100mB, отрегулируйте усиление звуковой карты с помощью вертикального ползунка, расположенного на левом краю экрана программы PhysLab. При сигнале, большем, чем 100mB, программа выдает только три значащие цифры величины сигнала, что не обеспечивает нужной точности.

Вычислите усиление при использованных значениях тока через трубку. Полезно вычислить усиление прямо на месте хотя бы для одного значения тока, чтобы убедиться в правильности измерений.

4. Добейтесь лазерной генерации на исследуемой трубке. Для этого разверните глухое зеркало так, чтобы пучок юстировочного лазера возвращался строго назад сквозь трубку. Поймайте отраженный пучок на экранчике №2 на юстировочном лазере. Добейтесь того, чтобы этот пучок попадал точно в центр экранчика. Затем поставьте выходное зеркало и также направьте отраженный от него пучок строго назад на центр экранчика. Должна появиться лазерная генерация. Если этого не произошло, повторите настройку более тщательно. Имейте в виду, что лазерные зеркала должны быть настроены параллельно друг другу с точностью $\approx 10^{-4}$ рад.

5. Исследуйте состояние поляризации излучения лазера на исследуемой трубке.

Закрепите в рейтере (8) поляроид. Поверните модулятор таким образом, чтобы пучок исследуемого лазера хорошо проходил сквозь отверстия модулятора и попадал на фотодиод № 2. Юстировочный лазер на время проведения этих измерений лучше выключить или перекрыть его излучение. Измерьте зависимость интенсивности излучения исследуемого лазера от угла поворота поляроида. Постройте зависимость относительной интенсивности от угла поворота поляроида, сравните с теоретической зависимостью для линейной поляризации.

6. Проведите наблюдение модовой структуры лазерного излучения. Поставьте на рельс вплотную к выходному зеркалу рейтер с короткофокусной линзой, а в рейтер (7) вставьте белый экран. Наблюдая за пят ном излучения лазера, с помощью малого поворота выходного зеркала попытайтесь получить одномодовый режим, трёхмодовый и многомодовый режимы.

### Контрольные вопросы

1. Почему не удаётся получить большего усиления в трубке $He - Ne$-лазера путём увеличения разрядного тока?

2. Зачем окошки в лазерной трубке приклеены под углом и почему наружные поверхности окошек смотрят вниз, а не вверх или вбок?

3. Зачем в лазерной трубке к рабочему веществу - неону добавляется гелий?

4. Зачем осуществляется модуляция пучка измеряемого излучения и можно ли обойтись без модуляции в данной установке?

## ПРИЛОЖКЕНИЕ

Устройство, принцип работы и схема включения фотодиода

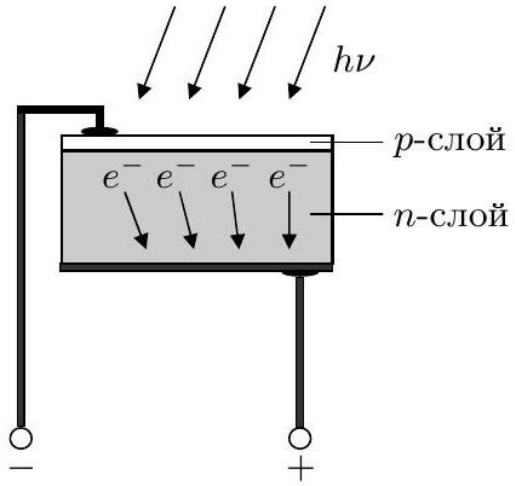



Рис. 5. Устройство фотодиода

Фотодиод представляет собой полупроводниковый диод на основе $p-n$-перехода между слоями полупроводников $p$ - и $n$ - типа. В отличие от обычных диодов (например, выпрямительных) $p$-слой делается очень тонким (десятки микрон). Сквозь этот слой и происходит освещение фотодиода светом, интенсивность которого нужно измерить. Поглощаясь в этом тонком слое, фотоны вызывают переходы электронов из валентной зоны наверх в зону проводимости, в результате чего появляются подвижные носители заряда. (Электроны, находящиеся в полностью заполненной энергетической зоне, не могут переносить заряд и обеспечивать прохождение тока, поскольку

не могут изменить своего состояния вследствие отсутствия свободных состояний). К $p-n$-переходу приложено постоянное обратное напряжение, так что без света ток через фотодиод практически не идёт. Если же под действием света появляются свободные носители заряда, электрическое поле «растаскивает» эти заряды к электродам и появляется фототок. При этом один поглощённый фотон перебрасывает наверх один электрон, поэтому энергетическую чувствительность фотодиода легко рассчитать. Например, на длине волны 620 нм энергия фотона равна $3,20 \cdot 10^{-19}$ Дж, в то время как заряд электрона равен $1,60 \cdot 10^{-19}$ Кл, поэтому чувствительность равна 0,5 Кл/Дж или 0,5 А/Вт. Чувствительность реальных фотодиодов ниже, потому что часть падающего излучения отражается от поверхности фотодиода и не участвует в создании носителей заряда. Замечательным свойством фотодиодов является их высокая линейность в широком диапазоне интенсивностей света. Известны примеры построения измерителей интенсивности света на фотодиодах с

диапазоном линейности 11 порядков. Для сравнения: фотоэлектронные умножители в зависимости от конструкции имеют диапазон линейности всего $2-3$ порядка, хотя и обладают большей чувствительностью. Поэтому с помощью фотодиодов легко проводить довольно точные измерения в широком диапазоне интенсивностей.

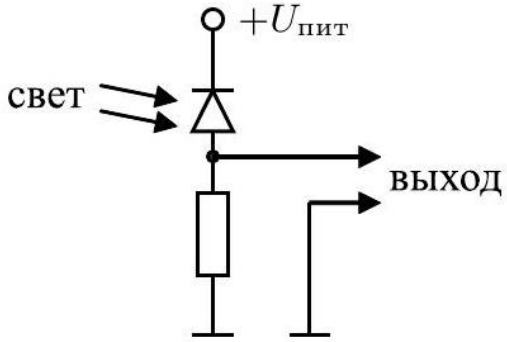



Рис. 6

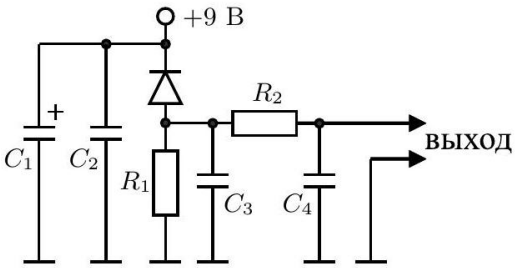



Рис. 7

Стандартная схема включения фотодиода показана на рис. 6.

В нашей установке для уменьшения влияния шумов и наводок в схему добавлены буферные конденсаторы в цепи питания, накопительный конденсатор $C_3$ и интегрирующая цепочка $r_2C_4$ на выходе (рис. 7).

$C_1$ - электролитический конденсатор большой ёмкости $C \approx 100$ мк Ф,

$C_2$ - быстрый керамический 0,1 мк Ф,

$C_3, C_4$ - керамические по 100 п Ф,

$R_1 - 8\kappa_{\text{К}}$,

$R_2 - 5,6$ кОм.

Использован фотодиод типа ФД 24 К с большой фоточувствительной площадкой диаметром 1 см.

## 22.5.2 Interference of Laser Radiation

Цель работы: исследование видности интерференционной картины излучения гелий-неонового лазера и определение длины когерентности излучения.

В работе используются: He-Ne-лазер, интерферометр Майкельсона с подвижным зеркалом, фотодиод с усилителем, осциллограф, поляроид, линейка.

Перед выполнением эксперимента следует ознакомиться с принципами работы лазеров в теоретическом введении к разделу.

Напомним коротко особенности конструкции и свойств излучения гелий-неонового лазера. Лазер состоит из двух зеркал, составляющих лазерный резонатор, и расположенной между ними газообразной усиливающей среды, состоящей из смеси гелия и неона. Типичное расстояние между зеркалами $-0,2 \div 1$ м. Смесь гелия и неона может находиться в отдельной трубке, закрытой с торцов стеклянными окошками, приклеенными под углом Брюстера с целью устранения потерь на отражение для одной из поляризаций. Лазер такой конструкции излучает свет с линейной поляризацией. Зеркала резонатора должны быть с высокой точностью настроены параллельно друг другу. В современных моделях лазеров, как правило, зеркала приклеиваются

непосредственно на торцы трубки, одновременно играя роль окошек и образуя резонатор. Этим достигается жёсткость резонатора и отпадает необходимость периодической подстройки зеркал. Такой лазер излучает неполяризованный свет (точнее, свет с хаотической поляризацией).

В лазере излучение распространяется по резонатору «туда и обратно». При этом максимальным усилением обладают волны, для которых набег фазы при полном обходе резонатора кратен $2\pi$. Это приводит к условию на разрешённые частоты и длины волн:

$$\frac{2\pi}{\lambda}2L = 2\pi m, \quad L = m\lambda, \quad \nu_m = \frac{mc}{2L}, \quad \Delta\nu_m = \nu_{m+1} - \nu_m = \frac{c}{2L} \tag{1}$$

где $L$ - длина резонатора, $m$ - целое число. Поэтому лазер генерирует отдельные типы колебаний, называемые модами, которые удовлетворяют условию (1). Отметим, что условие (1) в точности совпадает с условием максимума пропускания интерферометра Фабри-Перо, что не является удивительным, поскольку лазерный резонатор действительно представляет собой интерферометр Фабри-Перо.

Спектральная ширина отдельной моды определяется добротностью резонатора лазера и мощностью излучения (подробнее см. п. 2.3 введения к разделу). В гелий-неоновом лазере из-за малого усиления активной среды используются зеркала с высоким отражением, добротность резонатора большая и спектральная ширина моды может быть очень узкой, вплоть до единиц Гц. Реально из-за тепловых нестабильностей длины резонатора типичная ширина моды составляет $10^5$ Гц, что всё равно много меньше расстояния между модами ($10^8\Gamma_{\text{ц}}$ ). Количество генерируемых мод определяется шириной спектра усиления активной среды. Эта ширина складывается из естественной ширины линии излучения атомов неона и доплеровского уширения, вызванного тепловым движением атомов. Время когерентности и ширина спектра связаны соотношением неопределённости $\tau \cdot \Delta\nu \sim 1$. Для лазерного перехода атома неона с длиной волны 632,8 нм время затухания $\tau \approx 10^{-8}$ с, длина цуга в пространстве $L_{\text{цуга}} \approx 3$ м, ширина спектра $\Delta\nu \approx 10^8$ Гц. Доплеровское уширение на порядок больше, поэтому общая ширина спектра определяется именно доплеровским уширением. При температуре 400 К ширина по полувысоте спектра излучения газообразного неона равна $1,5 \cdot 10^9$ Гц. На такой ширине укладывается несколько мод при типичном расстоянии между модами $200 \div 300$ МГц, поэтому гелийнеоновый лазер с длиной резонатора $0,5 - 0,7$ м обычно одновременно излучает $3 - 7$ мод. Амплитуды и фазы этих мод флуктуируют во времени, но в среднем амплитуда мод вблизи максимума кривой усиления больше, чем на краях.

Вследствие тепловых нестабильностей длина резонатора все время меняется, в результате чего моды «переползают» с одного края контура усиления на другой, там исчезают, а на другом краю рождаются новые. Поэтому количество одновременно генерируемых мод и их положение нестабильно. При типичном коэффициенте теплового расширения твёрдых тел $10^{-5}$ достаточно изменения температуры на $0,05°$, чтобы длина резонатора изменилась на $\lambda/2$ и моды переместились на величину межмодового расстояния. Таким образом, температурная нестабильность резонатора приводит к медленным изменениям амплитуд колебаний в лазерных модах и числа самих мод. После примерно часа работы лазера характерное время перестройки резонатора составляет около 1 мин.

Время когерентности гелий-неонового лазера. С одной стороны, спектральная ширина каждой моды порядка $10^5$ Гц, поэтому из соотношения неопределённостей можно ожидать времени когерентности $10^{-5}$ с, соответственно длины когерентности $L_{\text{ког}} \approx tc = 3 \cdot 10^5$ cm $= 3$ км. С другой стороны, общая ширина спектра
порядка $10^9$ Гц и ожидаемое время когерентности $10^{-9}$ с, а длина когерентности 30 см.

Рассмотрим этот вопрос более детально.

Напомним, что временем когерентности называется такое время, на которое можно задержать один пучок относительно другого, чтобы ещё сохранялась способность к интерференции между ними при условии, что эти пучки получены из одного исходного пучка разделением на две части, например, полупрозрачным зеркалом. Длина когерентности это время когерентности, умноженное на скорость света, то есть это геометрическая разность хода пучков, при которой ещё сохраняется способность к интерференции. Излучение многомодового лазера - это не монохроматическая волна, это набор волн с разными частотами, удовлетворяющими условию (1). Между этими волнами происходят биения с частотой, равной разнице частот между модами, максимальная частота биений - это разность частот между крайними модами, то есть фактически полная ширина спектра излучения лазера. Характерное время, за которое существенно меняются амплитуда и другие параметры волны, - это период биений, равный обратной ширине спектра. Поэтому интерференция должна смазываться при задержке одного пучка относительно другого на время $t \approx 1/\Delta\nu_{\text{полн}} \approx 10^{-9}$ с, или на длину $\approx 30$ см.

С другой стороны, поскольку излучение в лазере последовательно распространяется по резонатору от зеркала к зеркалу, его характеристики (амплитуда и соотношение фаз между модами) должны в точности повторяться с периодом, равным времени полного обхода резонатора. Поэтому мы должны ожидать, что при геометрических задержках, кратных удвоенной длине резонатора, интерференция пучков должна восстанавливаться. Этот вывод, основанный на общих соображениях, можно получить более строго. Если один пучок волны с частотой $\nu_m$ задержать относительно другого на расстояние $\ell$, то между ними возникнет разность фаз $\Delta\varphi = kl = (2\pi\nu_m/c)\,\ell$ или с учётом (1) $\Delta\varphi = \pi m\ell/L$, где $k$ - волновое число, а $m$ - номер моды. То есть для разных мод с разными $m$ разность фаз будет разная, поэтому максимумы интерференционной картины будут располагаться в разных местах и картина смазывается. Если же задержка кратна удвоенной длине резонатора, то для всех мод разность фаз будет хотя и разной, но кратной $2\pi$.

Видность интерференционной картины. При сложении двух когерентных световых волн возникает интерференционная картина (см. раздел II). Если в плоскости наблюдения сходятся под малым углом $\alpha$ две плоских волны с длиной волны $\lambda_0$, то наблюдается интерференционная картина в виде последовательности тёмных и светлых полос с расстоянием между полосами (2.25):

$$\Delta x = \frac{\lambda_0}{\alpha} \tag{2}$$

Для оценки чёткости интерференционной картины в окрестности некоторой точки используют параметр видности (2.22):

$$\mathcal{V} = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{3}$$

где $I_{\max}$ и $I_{\min}$ - максимальная и минимальная интенсивности света интерференционной картины вблизи выбранной точки. Параметр $\mathcal{V}$ меняется в пределах от 0 (полное исчезновение интерференционной картины) до 1 (наиболее чёткая картина). Человеческий глаз может уверенно различать чередование светлых и тёмных интерференционных полос, если $\mathcal{V} > 0,1$.

Видность зависит от спектрального состава света, отношения амплитуд интерферирующих волн, разности хода между ними, поляризации интерферирующих пучков. Рассмотрим эту зависимость. Найдём выражение для интенсивности света в интерференционной картине. При этом можно не учитывать так называемые

межмодовые биения, приводящие к быстрым (с частотой $\Delta\nu \approx 200$ МГц) пульсациям интенсивности света в точке наблюдения. При визуальном наблюдении интерференционной картины или при использовании достаточно инерционного фотоприёмника эти пульсации усредняются.

Найдём видность интерференционной картины для одной моды лазерного излучения с частотой $\nu_m$. Пусть в плоскости наблюдения интерферируют под небольшим углом две волны с амплитудами $A_m$ и $B_m$. Если в точке наблюдения разность фаз между волнами равна $k_m\ell$, где $k_m = 2\pi/\lambda_m = 2\pi\nu_m/\mathrm{c}$ - волновое число, $\ell$ - разность хода, то интенсивность света в этой точке

$$I_m = A_m^2 + B_m^2 + 2A_mB_m\cos\left(k_m\ell\right) \tag{4}$$

При перемещении поперёк интерференционной картины разность фаз изменяется, и мы переходим от одного максимума к другому. При этом интенсивность света в максимуме интерференционной картины $I_{\max} = \left(A_m + B_m\right)^2$, а в минимуме $I_{\min} = \left(A_m - B_m\right)^2$. Поэтому видность

$$\mathcal{V}_1 = \frac{2\sqrt{\delta}}{1+\delta} \tag{5}$$

где введён параметр $\delta = B_m^2/A_m^2$. Видность $\mathcal{V}_1 = 1$ только тогда, когда интерферирующие волны имеют равную интенсивность. В установке
разделение пучка на два производится специальным делительным кубиком, при этом $\delta \approx 1$.

Найдём видность в случае, когда лазер генерирует одновременно несколько мод, но амплитуды разделяемых пучков одинаковы, т. е. $\delta = 1$ для всех мод. Для этого следует сложить мгновенные амплитуды всех интерферирующих волн, возвести в квадрат и усреднить по времени в течение $10^{-5}$ с (время отклика используемого в данной работе фотодиода), так как более быстрые изменения интенсивности фотодиод не воспринимает. Пусть генерируется одновременно $n$ мод с индексом $m$ от $m_0$ до $m_0 + n - 1$ и один из пучков проходит путь на величину $\ell$ больший, чем второй пучок, до места, где наблюдается интерференция. Интенсивность находится по формуле

$$I(\ell) = \overline{\left(\sum_{m=m_0}^{m_0+n-1} A_m\cos\left(\omega_m t + \varphi_m + k_m\ell\right) + A_m\cos\left(\omega_m t + \varphi_m\right)\right)^2} \tag{6}$$

где $A_m$ - амплитуда $m$-й моды, $\omega_m = 2\pi\nu_m = \pi cm/L$ - круговая частота, а $\varphi_m$ - фаза этой моды, черта означает усреднение по времени.

После возведения в квадрат этой суммы выражение будет содержать члены двух сортов: квадраты каждого члена суммы вида

$$A_m\cos^2\left(\omega_m t + \ldots\right)$$

и перекрёстные члены вида

$$2A_mA_i\cos\left(\omega_m t + \ldots\right)\cos\left(\omega_i t + \ldots\right)$$

Среднее значение $\overline{\cos^2\left(\omega_m t + \varphi\right)}$ равно $1/2$, перекрёстные члены, содержащие частоты разных мод с $m \neq i$, осциллируют с частотой межмодовых биений порядка сотен МГц и поэтому при усреднении по времени занулятся. Останутся только квадраты и перекрёстные члены с одинаковой частотой. Получим

$$I = \sum_{m=m_0}^{m_0+n-1} A_m^2 + \overline{A_m^2 \cos\left(\omega_m t + \varphi_m + k_m \ell\right) \cos\left(\omega_m t + \varphi_m\right)} \tag{7}$$

После преобразования произведения косинусов в сумму и усреднения по времени один из членов суммы также занулится и окончательно получим

$$I = \sum_{m_0}^{m_0+n-1} A_m^2 \left(1 + \cos k_m \ell\right) = \sum_{m_0}^{m_0+n-1} A_m^2 \left(1 + \cos \frac{\pi m}{L} \ell\right) \tag{8}$$

Из этого выражения видно, что разные моды не интерферируют друг с другом, а суммарный результат интерференции равен сумме интерференционных картин разных мод. Кроме того, фазы мод не влияют на картину интерференции, а соотношения амплитуд разных мод, напротив, имеют большое значение. Дальнейший анализ затрудняется тем, что амплитуды мод флуктуируют. Легко рассмотреть простейший случай, когда амплитуды всех мод одинаковы, $A_m = A = \text{const}$. Тогда, заменив косинусы комплексными экспонентами, $\cos x = \frac{e^{ix}+e^{-ix}}{2}$, и проведя суммирование получающейся геометрической прогрессии $\sum e^{imx} = e^{im_0 x} \frac{1-e^{inx}}{1-e^{xx}}$, находим

$$I = A^2 \left( n + \frac{\sin \frac{\pi \ell}{2L} n}{\sin \frac{\pi \ell}{2L}} \cos \left[ \frac{\pi \ell}{L} \left( m_0 + \frac{n-1}{2} \right) \right] \right) \tag{9}$$

В этой формуле последний множитель периодически меняет свой знак при небольших изменениях $\ell$, обеспечивая появление максимумов и минимумов при смещении поперёк интерференционной картины, а дробь с синусами определяет амплитуду изменения интенсивности. Поэтому видность получается равной

$$\mathcal{V}_2 = \left| \frac{1}{n} \frac{\sin \frac{\pi \ell}{2L} n}{\sin \frac{\pi \ell}{2L}} \right| \tag{10}$$

Мы ввели обозначение $\mathcal{V}_2$ для видности, обусловленной немонохроматичностью, в отличие от рассмотренной ранее $\mathcal{V}_1$, обусловленной разницей амплитуд интерферирующих пучков. Из формулы (10) видно, что при разности хода $\ell = 2L \cdot j$, где $j$ - целое число, видность $\mathcal{V}_2 = 1$, а число минимумов при изменении $\ell$ от 0 до $2L$ равно $n - 1$, где $n$ - число мод. На рис. 1 изображена зависимость видности от задержки $\ell$ для 3 мод (кривая а). По числу минимумов в принципе можно было бы определить число генерируемых лазером мод, но только если амплитуды мод равны, что в действительности не выполняется. В общем случае зависимость $\mathcal{V}_2$ от $\ell$ будет похожей на кривую (а) на рис. 1, но промежуточные максимумы и минимумы могут размыться или даже совсем исчезнуть. Например, для случая трёх мод, когда квадрат амплитуды центральной моды в два раза больше, чем у крайних мод, получается кривая (б) на рис. 1, у которой вместо двух минимумов остался один. Поэтому число мод лучше оценивать по ширине главного максимума в окрестности нулевой задержки. На рис. 1 приведено ещё несколько расчётных кривых для разного числа мод, интенсивность (квадрат амплитуды) которых плавно спадает от центра линии усиления к краям.

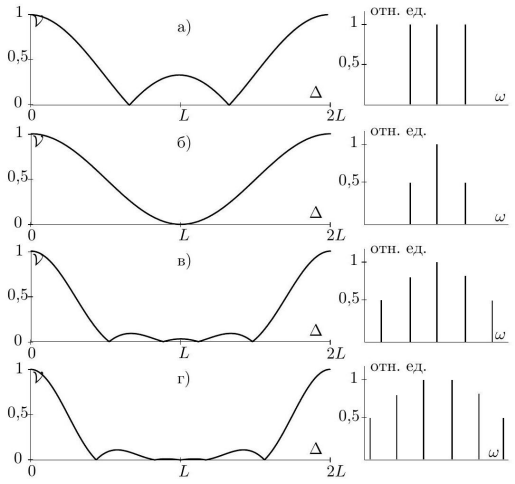



Рис. 1. Зависимость видности от задержки для разного количества генерируемых мод: а) и б) 3 моды, в) 5 мод, г) 6 мод. Справа приведены соотношения интенсивностей мод

При анализе видности мы предполагали, что световые волны поляризованы одинаково. Если же поляризация в волнах различна, то интерферируют компоненты только с одинаковой поляризацией, присутствие в одном из пучков излучения с поляризацией, перпендикулярной поляризации другого пучка, только даёт фон, уменьшая видность. В частности, если обе волны линейно поляризованы, их амплитуды не флуктуируют, а угол между плоскостями их поляризаций равен $\beta$, то в формуле (8) и в последующих выражениях для видности появится сомножитель $\mathcal{V}_3 = \cos\beta$. При равных амплитудах интерферирующих волн этот сомножитель имеет смысл видности, обусловленной разной поляризацией волн. Если световые волны поляризованы во взаимно
перпендикулярных плоскостях ($\beta = \pi/2$), то видность $\mathcal{V}_3$ обращается в ноль.

Несколько сложнее случай, когда источник света генерирует излучение с линейной поляризацией, но направление поляризации хаотически меняется в пределах от 0 до $\pi$. Если такое излучение разделить на два пучка и на пути каждого поставить по поляроиду с углом между направлениями разрешённой поляризации этих поляроидов $\beta$, то опять получим две волны с углом между плоскостями их поляризаций $\beta$, но амплитуды этих волн будут флуктуировать. Можно показать, что в этом случае $\mathcal{V}_3 = \cos^2\beta$.

Если имеют место все три фактора уменьшения видности: неравенство амплитуд, несовпадение поляризаций и разная оптическая задержка между интерферирующими пучками, то можно доказать, что результирующая видность является произведением

$$\mathcal{V} = \mathcal{V}_1 \cdot \mathcal{V}_2 \cdot \mathcal{V}_3 \tag{11}$$

Мы разобрали главные причины уменьшения видности, но есть и другие. Например, лазерные пучки, проходя через большое количество оптических элементов, могут портиться из-за дефектов этих элементов и дифракции на осевших пылинках, в результате чего интенсивность света становится неоднородной по поперечному сечению. При наложении таких пучков друг на друга происходит интерференция излучения разной амплитуды, даже при изначально равной интенсивности пучков. Поэтому в эксперименте практически невозможно достичь видности, строго равной 1.

Определение ширины спектра лазерного излучения и числа генерируемых мод. Геометрическая задержка $\ell$, при которой кривая видности заметно спадает, фактически является длиной когерентности. Разделив её на скорость света, получим время когерентности $\tau_{\text{ког}} = \ell/c$. Из принципа неопределённости $\tau_{\text{ког}}\,\Delta\nu_{\text{полн}} \sim 1$, где $\Delta\nu_{\text{полн}}$ - полная ширина спектра излучения лазера (с учётом всех имеющихся мод). Это

соотношение выполняется только по порядку величины, точное значение произведения $\tau_{\text{ког}}\,\Delta\nu$ зависит от формы спектра, а также от того уровня, по которому измеряется ширина. В нашем случае для определения ширины спектра рациональнее исходить из расчётных кривых, представленных на рис. 1. Их анализ показывает, что полная ширина

спектра связана с геометрической задержкой $\ell_{1/2}$, при которой видность падает вдвое, приблизительным соотношением

$$\Delta\nu_{\text{полн}} \approx 0,6\frac{\text{c}}{\ell_{1/2}} \tag{12}$$

а число мод равно

$$n \approx 1 + 1,2\frac{L}{\ell_{1/2}} \tag{13}$$

Экспериментальная установка. Для получения интерференционной картины используется интерферометр Майкельсона, смонтированный на вертикально стоящей массивной металлической плите. Схема установки приведена на рис. 2.

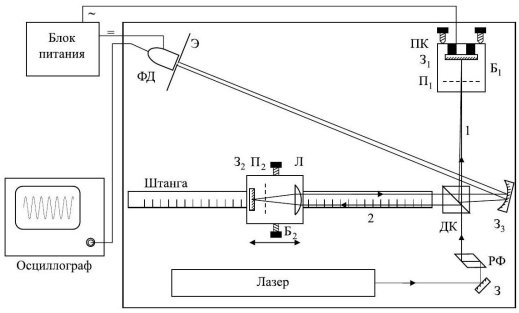



Рис. 2. Схема установки. З, $_1, 3_2, _3$ - зеркала. $\Pi_1$ и $\Pi_2$ - поляроиды. Б 1 $_1$ и Б Б2 - блоки № 1 и 2. ДК - делительный кубик, РФ - ромб Френеля.

ФД - фотодиод, Э - экран, ПК - пьезокерамика, Л - линза

Источником света служит гелий-неоновый лазер (средняя длина волны $\lambda_0 = 632,8$ нм). Пучок лазерного излучения отражается от зеркала 3 и проходит призму полного внутреннего отражения РФ (ромб Френеля), которая превращает линейную поляризацию излучения в круговую. Если в установке используется лазер, излучающий неполяризованный свет, то ромб Френеля не нужен, но он и не мешает выполнению работы. Далее лазерное излучение делится диагональной плоскостью делительного кубика ДК на два пучка.

Пучок 1 проходит поляроид $\Pi_1$, отражается под небольшим углом от зеркала $_1$, снова проходит поляроид $\Pi_1$ и, частично отражаясь от диагональной плоскости делительного кубика, выходит из интерферометра, попадает на зеркало $_3$ и далее на фотодиод ФД. Зеркало $3_1$

наклеено на пьезокерамику ПК, которая может осуществлять малые колебания зеркала вдоль направления распространения падающего пучка. Поляроид и зеркало с пьезокерамикой собраны в единый блок $_1$, который крепится к вертикально стоящей плите. В блоке $_1$ имеются юстировочные винты, которые позволяют регулировать угол наклона зеркала $_1$. В установке предусмотрена возможность вращения поляроида $\Pi_1$. Угол поворота отсчитывается по шкале, нанесённой на оправу поляроида.

Пучок 2 проходит линзу Л, поляроид $\Pi_2$, отражается от зеркала $_2$, снова проходит поляроид $\Pi_2$, линзу Л и делительный кубик, выходит из интерферометра, попадает на зеркало $_3$ и далее на фотодиод ФД. Таким образом, от зеркала $_3$ под небольшим углом друг к другу идут на фотодиод два пучка, прошедшие разные плечи интерферометра. Между ними происходит интерференция и образуются интерференционные полосы. Линза Л,

поляроид $П_2$ и зеркало $_2$ собраны в единый блок $_2$. Зеркало $_2$ установлено в фокальной плоскости линзы Л. Это сделано для того, чтобы падающий и выходящий из блока $_2$ пучки всегда были параллельны друг другу. Блок Б 2 может перемещаться вдоль пучка 2 по штанге, жёстко связанной с плитой интерферометра. Длина штанги 90 см. В установке предусмотрена возможность небольшого поперечного перемещения блока $Е_2$, что позволяет регулировать расстояние между падающим и выходящим из блока пучками. При измерениях блок $_2$ крепится к штанге при помощи двух винтов. Вдоль штанги нанесены деления через один сантиметр. При перемещении блока $_2$ вдоль штанги на величину $x_1$ геометрическая разность хода между пучками 1 и 2 изменяется на величину $\ell = 2x_1$.

Сферическое зеркало $3_3$ с небольшим фокусным расстоянием увеличивает картину интерференционных полос и позволяет наблюдать её на экране Э, расположенном в плоскости входного окна фотодиода.

Свет попадает на фотодиод ФД через узкую щель в центре экрана. Щель ориентируется параллельно интерференционным полосам. Ширина щели меньше расстояния между полосами. Сигнал фотодиода усиливается и подаётся на вход осциллографа. Для питания усилителя сигнала фотодиода и управления пьезокерамикой используется блок питания БП.

На пьезокерамику подаётся напряжение с частотой 50 Гц. При этом её длина изменяется с частотой 100 Гц. Величина удлинения зависит от приложенного напряжения и регулируется ручкой «Качание» на блоке питания. Обычно удлинение составляет несколько длин волн света. На эту величину перемещается вдоль пучка 1 зеркало $_1$. Интерференционная картина смещается на ширину полосы (одно колебание на экране
осциллографа), если зеркало $_1$ смещается на $\lambda_0/2 \sim 0,3$ мкм. При измерениях через входную щель фотодиода последовательно проходит несколько полос интерференционной картины, а на экране осциллографа наблюдаются колебания с изменяющимся периодом.

Настройка интерферометра. Пучок 1, падающий на блок $_1$, должен идти строго вертикально и проходить сквозь делительный кубик ДК. Расстояние от вертикальной плиты до центра пучка проверяют специальным шаблоном и при необходимости изменяют перемещением поворотного зеркала 3 (выполняется лаборантом).

Пучок 2, идущий к блоку $_2$, должен идти параллельно вертикальной плите и попадать на центр линзы Л. Перекрывают пучок 1 листом бумаги. Расположив блок $Б_2$ почти вплотную к делительному кубику, поворотным зеркалом подводят световое пятно к центру линзы (выполняется лаборантом). Переместив подвижный блок на дальний край штанги, совмещают пятно с центром линзы вращением кубика вокруг вертикальной оси и горизонтальной оси, перпендикулярной плите.

Перемещая блок $_2$, следует убедиться, что выходящий из него пучок проходит делительный кубик, а размер сечения пятна не увеличивается. Наблюдать пучок удобно на листе бумаги, размещённом между делительным кубиком и зеркалом $_3$. Допустимое смещение пучка для дальнего и ближнего положений блока на штанге $\approx 2 \div 3$ мм.

Установив блок Б $_2$ в положение, соответствующее нулевой разности хода и поворачивая его вокруг штанги, разводят световые пятна падающего и отражённого пучков на линзе Л на расстояние менее 2 мм. От этого расстояния зависит ширина интерференционных полос.

Сферическим зеркалом $_3$ направляют пучок от блока $_2$ на центр экрана, расположенного на входе фотодиода. Экран должен быть перпендикулярен направлению падающего пучка. Открыв пучок 1 и используя юстировочные винты на блоке $_1$, подводят к центру экрана пучок 1. Совместив оба пятна, вращением поляроида $П_1$ добиваются наибольшей чёткости интерференционной картины. Сняв экран,

устанавливают входную щель фотоприёмника параллельно интерференционным полосам и возвращают экран на место.

Измерение видности. Типичная осциллограмма сигнала фотодиода приведена на рис. 3. По осциллограмме можно найти следующие величины: фоновую засветку (линия 0 - перекрыты оба пучка 1 и 2); интенсивность света каждого из пучков (линии 1 или 2 - перекрыт пучок 2 или 1); максимума и минимума интенсивности интерференционной картины (открыты оба пучка). При этом параметр $\delta$, необходимый для
расчёта $\mathcal{V}_1$ в формуле (5), определяется отношением

$$\delta = \frac{h_1}{h_2} \tag{14}$$

Видность интерференционной картины рассчитывается по формуле

$$\mathcal{V} = \frac{h_4 - h_3}{h_4 + h_3} \tag{15}$$

Измерив величины $h_1, h_2, h_3$ и $h_4$, можно рассчитать $\mathcal{V}$ и $\mathcal{V}_1$, а затем определить видность при данной разности хода $\ell$ для угла между плоскостями поляризации пучков $\beta = 0\,(\mathcal{V}_3 = 1)$:

$$\mathcal{V}_2(\ell) = \frac{\mathcal{V}}{\mathcal{V}_1} \tag{16}$$

или при $\ell = 0, (\mathcal{V}_2 = 1)$ для известного угла $\beta$:

$$\mathcal{V}_3 = \frac{\mathcal{V}}{\mathcal{V}_1} \tag{17}$$

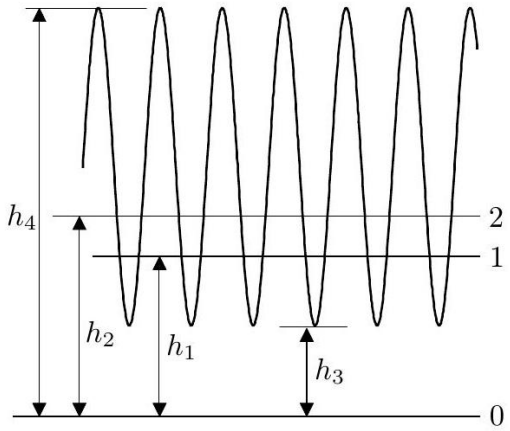



Рис. 3. Осциллограмма сигналов с фотодиода

## ЗАДАНИЕ

В работе предлагается исследовать: а) зависимость видности интерференционной картины от угла $\beta$ между плоскостями поляризации интерферирующих волн при нулевой разности хода; б) зависимость видности интерференционной картины от разности хода интерферирующих пучков для угла $\beta = 0$. По результатам измерений следует оценить спектральные характеристики лазерного излучения: ширину спектра генерации и число генерируемых мод.

**Внимание! Прямое попадание лазерного луча на сетчатку глаза крайне опасно для зрения!**

## I. Настройка системы

1. Включите лазер.
2. Убедитесь, что пучок лазера после поворотного зеркала З идёт вертикально.
3. Проверьте параллельность пучка 2 плите интерферометра, сравнив расстояния от центра пучка до плиты при двух положениях блока $_2$ сначала вблизи, а затем вдали от кубика. Установите блок Б $_2$ в положение, соответствующее нулевой разности хода, и настройте блок $_1$ так, чтобы оба отражённых пучка совместились на сферическом зеркале $З_3$.
4. Настройте сферическое зеркало так, чтобы отражённый от него пучок 2 попал на центр экрана, расположенного на входе фотодиода.
5. Получите на экране интерференционную картину, совместив оба световых пучка. Установите щель фотодиода параллельно интерференционным полосам.
6. Включите осциллограф и блок питания фотодиода и пьезокерамики. Получите на экране осциллографа устойчивую картину колебаний.

## II. Изучение влияния поляризации

7. Настройте поляроид $П_1$ в блоке Б $_1$ на максимальную видность (чёткость) интерференционной картины на экране фотодиода.
8. Поворотом поляроида $П_1$ установите минимальную чёткость интерференционной картины на экране ($\beta \approx 90°$). Введите дополнительный поляроид в пучок, идущий на экран, и объясните происходящие изменения.

## III. Измерение видности

9. Исследуйте зависимость видности интерференционной картины от угла $\beta$ поворота поляроида $П_1$ при нулевой разности хода ($\mathcal{V}_2 = 1$). Для этого измеряйте величины $h_1, h_2, h_3$ и $h_4$ на экране осциллографа. Измерения рекомендуется начинать с угла $\beta \approx 90°$.
10. Рассчитайте коэффициент $\mathcal{V}_3$, используя формулы (5), (14), (15) и (17). Постройте график $\mathcal{V}_3$ и сравните его с теоретическими зависимостями $\mathcal{V}_3 = \cos\beta$ и $\mathcal{V}_3 = \cos^2\beta$. Что можно сказать на основании этого сравнения о характере поляризации излучения используемого в установке лазера?
11. Исследуйте зависимость видности от разности хода между пучками. Для этого установите поляроид $П_1$ в положение, в котором интерференционная картина видна наиболее чётко ($\beta = 0°, \mathcal{V}_3 = 1$). Передвигая по штанге блок $_2$, найдите положение, где видность максимальна (нулевая задержка) и подстройте ещё раз угол поворота поляроида. Снимите зависимость величин $h_1, h_2, h_3$ и $h_4$ от координаты $x$ блока $_2$, начиная с минимального расстояния до максимально возможного. В области максимумов видности измерения следует проводить через 1-2 см, в промежутке между максимумами - через 5 см. При перемещении блока $_2$ можно подстраивать интерференционную картину, уменьшив предварительно амплитуду качания пъезокерамики до минимума. Это полезно сделать дважды: при среднем и дальнем положении подвижного блока на штанге.

В промежуточной области могут наблюдаться медленные колебания видности с характерным временем колебаний $\approx$ 1 мин. Эти колебания могут быть связаны с тепловым расширением резонатора, о котором мы говорили ранее. Если временные колебания видности заметны, следует проводить измерения в тот момент, когда размах колебаний $(h_4 - h_3)$ минимален (при этом число мод в доплеровском контуре максимально).

12. Рассчитайте коэффициент $\mathcal{V}_2$, используя формулы (5), (14) и (16). Постройте график зависимости видности $\mathcal{V}_2(x)$ от координаты блока $_2$. Определив по графику расстояние между максимумами, вычислите расстояние $L$ между зеркалами оптического резонатора лазера и межмодовое расстояние $\Delta\nu_m$ по формуле (1).

Определите задержку $\ell_{1/2}$ на половине высоты главного максимума и рассчитайте по формуле (12) диапазон частот $\Delta F$, в котором происходит генерация продольных мод. Оцените число генерируемых лазером продольных мод по формуле (13).

**Контрольные вопросы**

1. Объясните принцип действия ромба Френеля.
2. Чему равна видность интерференционной картины интерферирующих волн одинаковой интенсивности, если одна волна имеет плоскую поляризацию, а другая круговую?
3. Как изменяется кривая $\mathcal{V}_2(\ell)$ при уменьшении и при увеличении расстояния между зеркалами оптического резонатора лазера?
4. Из каких соображений следует выбирать ширину входной щели фотоприёмного устройства?
5. Как выглядит зависимость $\mathcal{V}(\ell)$ для одной моды; для двух мод?

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 22, 158, 227, 228.
2. Калитеевский Н.И. Волновая оптика. - М.: Высшая школа, 1978. 5.8, 7.3.
3. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 7.

## 22.5.3 Scanning Interferometer

Цель работы: знакомство с устройством и работой газового лазера непрерывного действия, со спектральными характеристиками лазерного излучения, а также с устройством и принципом действия сканирующего интерферометра Фабри-Перо.

В работе используются: He - Ne -лазер с блоком питания; сканирующий интерферометр Фабри-Перо; поляроид; пластинка $\lambda/4$; линза; фотодиод; электронный осциллограф.

Перед выполнением эксперимента необходимо ознакомиться с принципом работы лазеров в теоретическом введении к данному разделу и с теорией интерферометра Фабри-Перо (см. раздел IV).

В He-Ne-лазерах используются резонаторы, фактически представляющие собой интерферометр Фабри-Перо. Излучение распространяется вдоль оси интерферометра.

При этом генерируются моды (типы колебаний), для которых на длине резонатора укладывается целое число полуволн:

$$2L = m\lambda \quad (m - \text{ целое число }) \tag{1}$$

Из формулы (1) легко получить, что разность частот соседних мод (межмодовое расстояние) равна

$$\nu_{m+1} - \nu_m = \frac{c}{2L} \tag{2}$$

Для интерферометра с базой $L = 0,6$ м межмодовое расстояние равно 250 МГц. В то же время спектральная линия рабочего перехода неона имеет ширину порядка 1500 МГц, поэтому возможна одновременная генерация нескольких мод. Рис. 1 иллюстрирует увеличение числа мод генерации лазера с ростом усиления активной среды.

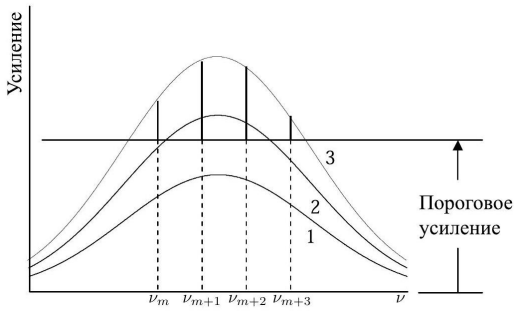



Рис. 1. Увеличение числа генерирующих мод при увеличении усиления

При небольшом усилении (кривая 1) генерации нет. В случае 2 генерация происходит только на 2 частотах $\nu_{m+1}$ и $\nu_{m+2}$, расположенных вблизи центра спектральной линии. Если усиление определяется кривой 3, генерация возникает на четырёх частотах от $\nu_m$ до $\nu_{m+3}$. Говорят, что в этом случае лазер одновременно работает на четырёх модах.

Для гелий-неонового лазера с достаточно длинной трубкой на переходе 632,8 нм многомодовая генерация является обычным режимом работы.

Сканирующий интерферометр. Для исследования межмодового состава излучения $He - Ne$-лазера в работе используется сканирующий интерферометр, представляющий собой высокодобротный интерферометр Фабри-Перо с периодически изменяемой базой. Его устройство схематически показано на рис. 2.

На жёстком массивном основании расположены две юстировочные головки $\Gamma_1$ и $\Gamma_2$, на которых укреплены зеркала $З_1$ и $З_2$. Зеркало $З_1$ установлено непосредственно на головке $\Gamma_1$, зеркало $З_2$ связано с головкой $\Gamma_2$ через пьезокерамический элемент П. Юстировочные головки снабжены винтами, не показанными на рис. 2, которые позволяют в

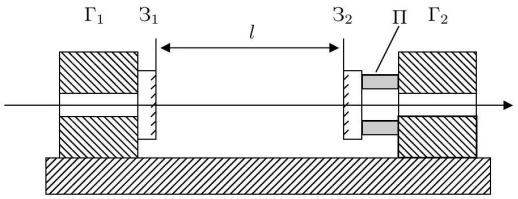



Рис. 2. Устройство сканирующего интерферометра

небольших пределах поворачивать зеркала относительно вертикальной и горизонтальной осей. С помощью головок $\Gamma_1$ и $\Gamma_2$ зеркала выставляются на параллельность.

Пьезокерамический элемент П позволяет периодически изменять базу интерферометра ($l \sim 10$ см) на величину порядка длины световой волны. Элемент имеет

форму полого цилиндра. Его внутренняя и наружная поверхности металлизированы и образуют цилиндрический конденсатор. Необходимое изменение длины цилиндра возникает при подаче напряжения в несколько сотен вольт.

Если вдоль оси интерферометра распространяется световое излучение с длиной волны $\lambda$, то при выполнении условия

$$2l = m\lambda \quad (m - \text{ целое число }) \tag{3}$$

аналогичного условию (1) для лазера, возникает резонанс. Внешнее излучение с длиной волны, удовлетворяющей условию (3), полностью проходит через интерферометр. Если на интерферометр падает излучение с различными длинами волн, то одновременно может возникнуть несколько резонансов. Собственные моды интерферометра отличаются по частоте на величину

$$\Delta\nu = \frac{c}{2L}$$

Величина $\Delta\nu$ называется дисперсионной областью спектрального прибора. В единицах $\lambda$ дисперсионная область сканирующего интерферометра равна

$$\Delta\lambda_{\text{c}} = \frac{\lambda}{m} = \frac{\lambda^2}{2l} \tag{4}$$

В нашей работе интерферометр Фабри-Перо используется как спектральный прибор высокой разрешающей силы. Разрешающая способность $R$ спектрального прибора определяется отношением

$$R = \frac{\lambda}{\delta\lambda} \tag{5}$$

где $\delta\lambda$ - минимальная разность длин волн, разрешаемая прибором вблизи длины волны $\lambda$. В случае интерферометра Фабри-Перо две линии считаются разрешимыми, если расстояние между максимумами их контуров равно или превышает ширину контура по уровню 0,5. Разрешающая способность интерферометра Фабри-Перо зависит от длины интерферометра $l$ и коэффициента отражения зеркал $r$:

$$R = \frac{2\pi l}{\lambda(1-r)} \tag{6}$$

(см. введение к разделу IV). В лазерной технике принято выражать разрешение интерферометра в единицах частоты:

$$\delta\nu = f\frac{\delta\lambda}{\lambda} = \frac{c(1-r)}{2\pi l} \tag{7}$$

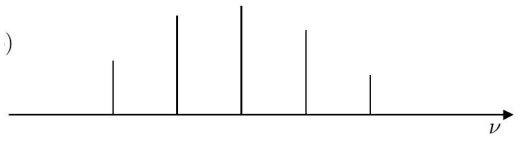
Как видно из (7), величина $(1-r)/\pi$ показывает, во сколько раз разрешаемая интерферометром разность частот меньше его дисперсионной области. Сканирующий интерферометр, применяемый в настоящей работе, имеет зеркала с коэффициентом отражения $r \approx 98,5\%$. Поэтому с его помощью можно разрешить две узкие спектральные линии, отличающиеся по частоте на величину порядка $0,005\Delta\nu$, т.е. (при $l = 10$ см) приблизительно на 7,5 МГц. Напомним, что межмодовое расстояние лазера при $L = 0,6$ м равно $\Delta\nu = 250$ МГц. Таким образом, наш сканирующий интерферометр вполне подходит для исследования модового состава лазерного излучения.

Рис. 3 даёт представление о соотношении между спектром генерации лазера и спектральной характеристикой сканирующего интерферометра (т.е. его спектром пропускания). Изменение расстояния между зеркалами сканирующего интерферометра

приводит к сдвигу нижней «гребёнки» по оси частот. При этом интерферометр последовательно настраивается на разные моды лазера.

Если одно из зеркал сканирующего интерферометра периодически перемещать вдоль оси, мощность прошедшего через интерферометр излучения периодически изменяется (рис. 4).

б)




а)

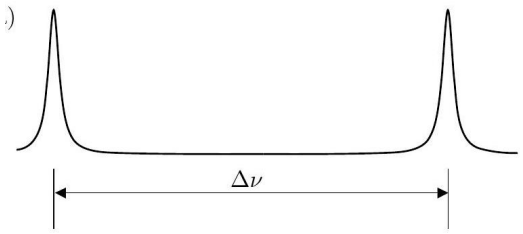



Рис. 3. Спектр генерации лазера (а) и спектр пропускания сканирующего интерферометра (б)

Динамика спектра генерации гелий-неонового лазера. На ширине контура усиления гелий неонового лазера обычно укладывается порядка десятка продольных мод. Поскольку генерация вынужденного излучения происходит обычно только на тех частотах, где усиление превышает потери, реально генерируется от 1 до $5 - 8$ мод в зависимости от длины резонатора. Излучение в разных модах возникает из усиленного спонтанного шума, поэтому амплитуда моды - случайная величина, которая медленно меняется случайным образом вследствие непрерывного «подмешивания» спонтанного излучения. Но амплитуды разных мод зависят друг от друга, поскольку излучение любой моды сбрасывает вниз возбуждённые атомы, уменьшая усиление для себя и для других мод. На этот процесс накладываются неизбежные флуктуации длины резонатора, связанные с изменением температуры и с механическими нестабильностями. При плавном изменении длины резонатора моды «ползут» по спектру в пределах контура усиления. Дойдя до края контура, мода исчезает, а на другой стороне контура усиления может родиться другая мода. Межмодовое расстояние при этом практически не меняется. Если частоты каких-либо двух мод оказались симметричны относительно центра контура усиления, то, как было отмечено в п. 2.5 введения к разделу, возникает неустойчивость и эти моды либо быстро сдвигаются по спектру, либо одна подавляет другую.

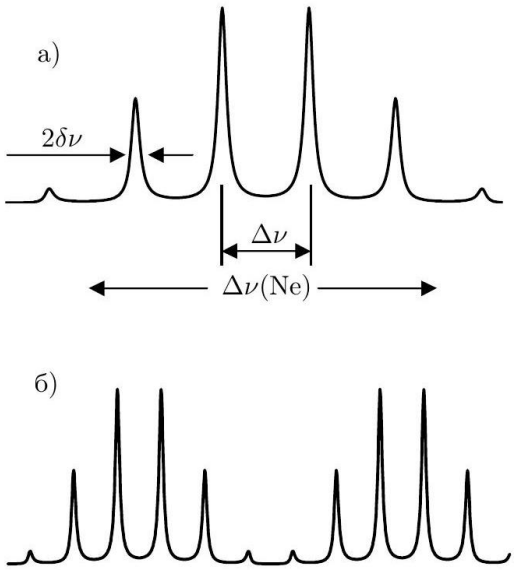



Рис. 4. Характерные осциллограммы: а) при небольшой амплитуде колебаний зеркала ($\leqslant \lambda/2$), б) при амплитуде колебаний, превышающей $\lambda/2$

Экспериментальная установка. Схема экспериментальной установки приведена на рис. 5.

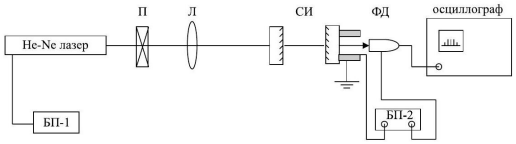



Рис. 5. Схема установки для исследования спектрального состава излучения лазера

Излучение He − Ne-лазера проходит через поляризационную развязку Р и линзу Л и поступает на вход сканирующего интерферометра (СИ). Поляризационная развязка предотвращает попадание в лазер излучения, отразившегося от элементов оптического тракта. Это излучение может существенно повлиять на работу лазера и даже привести к сры ву генерации. Развязка состоит из поляроида и пластинки $\lambda/4$, главные направления которой установлены под углом 45° по отношению к разрешённому направлению поляроида. После развязки П свет приобретает циркулярную поляризацию (например, по правому кругу). При отражении от передней поверхности линзы, от зеркала сканирующего интерферометра и т.п. свет распространяется в обратном направлении в виде левополяризованной волны. Такая волна, пройдя через пластинку $\lambda/4$, вновь приобретает линейную поляризацию. Однако направление колебаний в этой волне оказывается перпендикулярным направлению разрешённых колебаний поляроида, поэтому до лазера отражённая волна не доходит.

Линза Л служит для уменьшения расходимости пучка, поступающего на вход сканирующего интерферометра. Линза снабжена поперечными и продольными салазками для юстировки прибора на максимум сигнала.

Излучение, прошедшее сквозь сканирующий интерферометр, поступает на фотодиод ФД. Напряжение с фотодиода через усилитель подаётся на вертикальный вход электронного осциллографа ЭО.

Лазер питается от блока питания БП-1, фотодиод и усилитель от БП-2. Напряжение на пьезоэлемент сканирующего интерферометра подаётся с блока питания БП-2 и регулируется ручкой 1.

## ЗАДАНИЕ

В работе предлагается исследовать доплеровский контур спектральной линии излучения лазера: определить межмодовое расстояние и приборную ширину отдельной моды излучения лазера; оценить газокинетическую температуру в разряде; рассчитать дисперсионную область, разрешающую способность и коэффициент отражения зеркал сканирующего интерферометра.

**Внимание! Попадание прямого лазерного луча на сетчатку глаза крайне опасно для зрения!**

1. Включите в сеть осциллограф и блоки питания фотодиода и лазера.

2. Поставьте линзу посередине между лазером и сканирующим интерферометром. Временно выведите линзу из луча с помощью поперечных салазок.

3. Совместите прямой и отражённый пучки, осторожно вращая винты 1 и 2 на первом (по ходу луча) зеркале интерферометра: два пятна на пластинке $\lambda/4$ должны совпадать. Затем винтами 1 и 2 второго зеркала совместите пятна на первом зеркале.

4. Настройте поляризационную развязку. Для этого сначала вращением поляроида добейтесь максимальной яркости луча. Можно наблюдать за пятном на пластинке $\lambda/4$ или на листе бумаги.

Вращая пластинку $\lambda/4$, погасите отражённый луч: рассеянный свет на оправе лазерного окна не меняет яркости, если перекрывать луч между интерферометром и пластинкой $\lambda/4$. При этом яркие пятнышки (отражение от любой стеклянной поверхности) могут появляться и пропадать.

5. Слегка перемещая фотодиод вверх-вниз и в сторону, получите на экране осциллограмму типа показанной на рис. 4а. Добейтесь максимального сигнала.

6. Для уменьшения расходимости пучка (усиления сигнала) введите линзу в пучок. Чуть смещая линзу, добейтесь стабильности и симметрии картины. Это важнее получения максимума сигнала, т.к. при максимальном сигнале в лазер всё-таки идёт обратный луч, и лазер работает нестабильно.

Снова подстройте второе зеркало винтами 1 и 2, чтобы добиться совмещения пятен на первом зеркале. Убедитесь, что луч попадает на фотодиод, и добейтесь максимального сигнала на осциллографе.

7. Изменяя напряжение на пьезокристалле (амплитуду смещения зеркала СИ), проследите, как изменяется количество доплеровских контуров за один проход зеркала интерферометра. Подберите напряжение, при котором на экране укладывается $1-2$ контура.

8. Рассчитайте межмодовое расстояние резонатора в единицах $\nu$ и $\lambda$, используя формулу (2). Длина резонатора лазера $L$ указана на установке; $\lambda = 6328$Å.

Сосчитав число промежутков между модами на экране, оцените видимую ширину спектральной линии неона $\Delta\lambda$ (Ne) (т. е. ширину спектра генерации лазера).

Полагая, что ширина спектральной линии обусловлена эффектом Доплера (см. п. 1.3 введения к разделу) и что видимая ширина линии неона порядка полуширины

доплеровского контура ($\Delta\lambda(\mathrm{Ne}) \sim \Delta\lambda_D$), оцените среднюю скорость атомов неона $v_x$ и газокинетическую температуру $T$ в разряде:

$$\frac{\Delta\lambda_D}{\lambda} \approx \frac{v_x}{c}; \quad \frac{mv_x^2}{2} \approx \frac{kT}{2}. \tag{8}$$

Здесь $v_x$ - скорость молекул неона вдоль оси лазера; $m_{\mathrm{Ne}} = 20,2$ а.е.м.; 1 а.е.м. $= 1,66 \cdot 10^{-24}$ г.

9. Рассчитайте дисперсионную область $\Delta\lambda_{\text{си}}$ сканирующего интерферометра по формуле (4) для $l = 9$ см и сравните её с видимой шириной линии неона $\Delta\lambda(\mathrm{Ne})$.

10. Сравнив ширину отдельной моды на полувысоте с межмодовым расстоянием, оцените разрешение $\delta\lambda$ сканирующего интерферометра и разрешающую способность $R$ по формуле (5).

Оцените коэффициент отражения зеркал интерферометра $r$ по формуле (6).

11. Зарисуйте в тетради вид осциллограммы спектральной линии (качественно); укажите измеренные параметры.

12. Отключите осциллограф и оба блока питания.

**Контрольные вопросы**

1. Объясните основные процессы, приводящие к появлению инверсной заселённости уровней неона.

2. Почему плоскопараллельные пластинки, закрывающие с торцов рабочую трубку с гелий-неоновой смесью, устанавливают под углом Брюстера?

3. Что такое моды? Получите формулу (2) для межмодового расстояния.

4. В чём состоит условие возбуждения лазера?

5. Выведите формулу (6) для разрешающей способности интерферометра Фабри-Перо.

6. Как работает поляризационная развязка?

7. При какой длине резонатора He – Ne-лазера, независимо от уровня накачки, генерация возможна не более чем на одной моде (одномодовый лазер)?

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. XI, §225, 227, 228.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. XI, §122.

3. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 8, §8.9.

4. Квантовая электроника. Маленькая энциклопедия. - М.: Сов. энциклопедия, 1969. С. 110-115, 165-177, 325-329, 414-415.

## 22.6 Experiments about Interference and Tunneling of Radio Waves

### 22.6.1 Interference of electromagnetic waves in the millimeter range

Цель работы: изучение интерференции электромагнитных волн миллиметрового диапазона с применением двух оптических интерференционных схем, экспериментальное определение длины волны излучения и показателя преломления диэлектрика.

В работе используются: приёмно-передающая система радиоволн миллиметрового диапазона; металлические зеркала; микрометрический винт; проволочная решётка; пластина из диэлектрика.

Когерентные волны, необходимые для наблюдения интерференционной картины, в данной работе получаются путём расщепления волны из источника миллиметрового излучения по схемам, аналогичным представленным на рис. 6.1 оптическим схемам с плоскопараллельной пластинкой (а) и с полупрозрачной пластинкой в интерферометре Майкельсона (б).

Если в некоторой точке пространства происходит суперпозиция двух когерентных одинаково поляризованных волн с интенсивностями $I_1$ и $I_2$ и с разностью фаз $\varphi$, то интенсивность $I$ результирующего колебания определяется соотношением

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\varphi \tag{1}$$

Разность фаз $\varphi$ определяется разностью хода $\Delta$ между интерферирующими волнами:

$$\varphi = k\Delta \tag{2}$$

где $k = 2\pi/\lambda$ - волновое число. Интенсивность (1) максимальна при

$$\varphi = 2\pi m, \quad m = 0, 1, 2, \ldots \tag{3}$$

(интерференционный максимум) и минимальна при

$$\varphi = (2m + 1)\pi, \quad m = 0, 1, 2, \ldots \tag{4}$$

(интерференционный минимум). Число $m$ называется порядком интерференции или номером интерференционной полосы.

Экспериментальная установка. Источником радиоволн в работе является высокочастотный генератор Г4 − 115 на трёх отражательных клистронах, перекрывающих вместе полосу частот от 25,80 ГГц до

37,50 ГГц, разделённую на три поддиапазона, соответствующие отдельным клистронам. Генерирующий при выбранной настройке клистрон возбуждает в прямоугольном металлическом волноводе сечением $7,2 \times 3,4$ мм$^2$ электромагнитную волну, которая распространяется вдоль волновода и с помощью рупорной антенны излучается в пространство. Задача антенны заключается в том, чтобы сделать излучение более направленным.




Рис. 1. Приёмно-передающая система СВЧ-диапазона

Отражённое от препятствия электромагнитное излучение, попадая в рупорную антенну приёмника, распространяется по волноводу, в котором имеется детектор

высокочастотных колебаний, работающий в квадратичном режиме. Поэтому ток детектора пропорционален интенсивности $I$ волны, попадающей в приёмную антенну. Сигнал с выхода детектора усиливается и измеряется микровольтметром. Принципиальная схема приёмно-передающего тракта представлена на рис. 1.

Применяемый в настоящей работе передатчик излучает линейно поляризованную волну, электрический вектор $\boldsymbol{E}$ которой перпендикулярен широкой стенке волновода. Приёмник также может принимать только линейно поляризованную волну. Для установления связи в системе, изображённой на рис. 1, необходимо, чтобы широкие стенки волноводов передатчика и приёмника были параллельны друг другу.

Если одну из антенн повернуть относительно луча на некоторый угол $\alpha$, то интенсивность принимаемого сигнала будет изменяться по закону

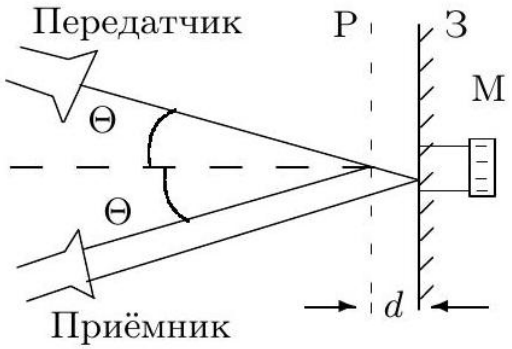
$$I = I_0 \cos^2 \alpha \tag{5}$$

Формула (5) выражает закон Малюса. Выполнение на опыте этого закона свидетельствует о том, что передатчик излучает, а приёмная система принимает линейно поляризованную волну.

Эксперимент с разворотом приёмной и передающей антенн аналогичен оптическому эксперименту с двумя поляроидами, расположенными на пути светового луча. В качестве отражающих зеркал в работе используются металлические листы; роль полупрозрачной пластинки выполняет решётка из медной проволоки. При установке решётки
необходимо проследить, чтобы направление проволок было параллельно вектору $\boldsymbol{E}$ падающей волны, так как электромагнитная волна, у которой вектор $\boldsymbol{E}$ перпендикулярен проводам, слабо отражается от решётки.

Интерференция радиоволн, отражённых от зеркала и решётки. Схема установки, используемой для этого опыта, приведена на рис. 2.

Металлическое зеркало 3 и проволочная решётка Р устанавливаются на некотором расстоянии $d$ друг от друга с помощью специальных фиксаторов. Приёмная и передающая антенны располагаются симметрично, так чтобы в приёмник попадала отражённая волна. Волна, излучаемая передающей антенной, частично отражается от решётки, а частично проходит через неё и отражается от зер-




кала. Зеркало может перемещаться при

Рис. 2. Схема с зеркалом и решёткой для наблюдения интерференции радиоволн помощи микрометрического винта М.

Между волнами, отражёнными от решётки и зеркала, возникает разность хода, равная

$$\Delta = 2d \cos \theta \tag{6}$$

где $\theta$ - углы падения и отражения волны, падающей на зеркало. При изменении разности хода (при изменении $d$) интенсивность волны в точке приёма, находящейся в области интерференции, изменяется в соответствии с формулой (2).

Интерферометр Майкельсона. В этом опыте используется установка, моделирующая оптический интерферометр Майкельсона (рис. 3). Зеркала $_1$ и $_2$ располагаются перпендикулярно осям передающей и приёмной антенн, которые в свою очередь должны быть взаимно перпендикулярны. Решётка Р располагается на пересечении осей под углом 45° к ним. Волна от передающей антенны расщепляется решёткой на две волны, распространяющиеся в направлении зеркал $_1$ и $_2$. После отражения от зеркал обе волны возвращаются к решётке. Каждая из этих волн после вторичного расщепления на решётке Р частично попадает в приёмную антенну.

Разность хода $\Delta$ возникает вследствие различия в расстояниях $l_1$ и $l_2$ между решёткой Р и зеркалами $_1$ и $З_2$:

$$\Delta = 2\,(l_1 - l_2) \tag{7}$$

При изменении длины одного из плеч интерферометра (при перемещении соответствующего зеркала) интенсивность в точке приёма изменяется в соответствии с формулой (1).

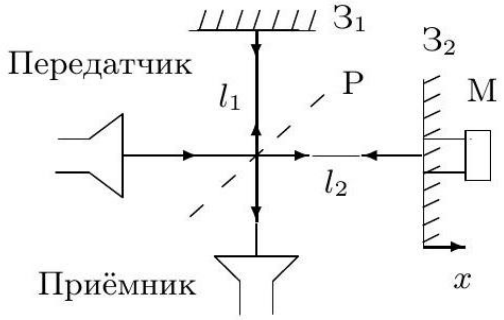



Рис. 3. Интерферометр Майкельсона на СВЧ-радиоволнах

Если на пути одного из лучей поставить пластинку толщиной $h$ с диэлектрической проницаемостью $\varepsilon$, то разность хода (7) изменится на величину $2h(n-1)$, где $n$ - показатель преломления вещества, из которого сделана пластинка. Это приводит к изменению интенсивности в точке приёма. Пусть в точке приёма до внесения пластинки наблюдался интерференционный максимум. Для того чтобы получить тот же максимум при наличии пластинки, нужно зеркало свободного плеча интерферометра (плеча, в котором нет пластинки) отодвинуть на расстояние $\delta x$, определяемое выражением

$$\delta x = h(n-1) \tag{8}$$

Зная $\delta x$, можно определить показатель преломления. Для этого, однако, мы должны быть уверены, что добавочная разность хода, внесённая пластинкой, не превышает длины волны $\lambda$, иначе возникает неопределённость в расчёте показателя преломления. Действительно, изменение разности хода на целое число длин волн смещает интерференционную картину на целое число полос. Такое смещение не может быть обнаружено в эксперименте, так как все интерференционные полосы идентичны. Применяемая в нашей работе диэлектрическая пластинка имеет достаточно малую толщину, так что вносимая ею дополнительная разность хода не превышает $\lambda$.

## ЗАДАНИЕ

В работе предлагается, настроив схему и проверив справедливость закона Малюса, исследовать интерференцию миллиметровых радиоволн. В схеме с зеркалом и решёткой - определить длину волны излучения по экспериментальной зависимости от смещения

зеркала интен сивности волны в зоне интерференции, а в схеме интерферометра Майкельсона - экспериментально проверить выполнение закона сложения интенсивностей при интерференции и определить показатель преломления материала диэлектрической пластинки.

### I. Подготовка приборов к работе и проверка закона Малюса

Мощность сигнала, получаемая от генератора Г4-115, невелика, поэтому излучение не представляет опасности для здоровья человека. Тем не менее заглядывать в открытый волновод при включённом генераторе не рекомендуется.

1. На дальнем от генератора конце стола закрепите с помощью фиксаторов металлическое зеркало с продольной подачей (если в оправу зеркала вставлена диэлектрическая пластина, её следует убрать). Рядом с рупором излучателя, вмонтированным в боковую стенку генератора, поставьте рупор приёмной антенны симметрично относительно разделительной линии так, чтобы оба рупора были направлены в сторону зеркала. При такой установке формула (5) для расчёта разности хода упрощается (на рис. $2\cos\theta \approx 1$). Проверьте, что длинные стороны рупоров горизонтальны.

2. Соедините коаксиальным кабелем выход приёмной антенны с входом 1 усилителя и подготовьте приборы к работе (см. «Техническое описание» на рабочем столе).

3. Проведите предварительную настройку установки. Для этого при любом заметном отклонении стрелки микровольтметра усилителя, перемещая отражающее зеркало при помощи подвижки и слегка поворачивая подвижку вручную вокруг вертикальной оси, настройтесь на максимум отражённого сигнала. Подстройте сигнал, передвигая приёмную антенну и слегка поворачивая её вокруг горизонтальной оси.

Методом последовательных приближений ещё раз отрегулируйте положение зеркала и антенн, чтобы настроиться на максимум сигнала. Следите, чтобы при выполнении всех упражнений отклонение стрелки выходного прибора не превышало $80 \div 90\%$ шкалы.

4. Пронаблюдайте за ослаблением сигнала, располагая перед зеркалом различные препятствия (стопку бумажных листов, диэлектрическую пластинку, металлический лист). Поставьте металлическую решётку перед приёмной антенной и проследите за изменением сигнала при повороте решётки на $90^\circ$ вокруг луча. Опишите наблюдаемые результаты и объясните их.

5. Снимите зависимость уровня сигнала $I$ от угла поворота $\alpha$ приёмной антенны относительно луча. Убедитесь, что излучаемая электромагнитная волна линейно поляризована.

Постройте график зависимости уровня сигнала $I$ от $\cos^2\alpha$ и убедитесь в справедливости закона Малюса.

### II. Интерференция волн, отражённых от зеркала и решётки

6. Закрепите на фиксаторах перед зеркалом металлическую решётку и поворотом решётки вокруг вертикальной оси настройте систему на максимум отражённого сигнала (рис. 2). Убедитесь, что при перемещении зеркала уровень сигнала в точке приёма изменяется. Перемещение зеркала осуществляется с помощью микрометрического винта. Один оборот винта - 100 делений - соответствует одному миллиметру.

7. Снимите зависимость интенсивности $I$ от координаты $x$ подвижного зеркала. Постройте график зависимости $I(x)$, определите по графику длину волны излучения $\lambda$ и сравните с аналогичной величиной, рассчитанной по частоте генератора.

### III. Интерферометр Майкельсона

8. Соберите схему согласно рис. 3. Для этого оба металлических зеркала и решётку закрепите на фиксаторах так, чтобы плечи $l_1$ и $l_2$ были примерно одинаковы. Настройте установку на максимум сигнала (см. п. 3).

9. Перемещая подвижное зеркало $3_2$, снимите зависимость координаты $x_m$ зеркала в точке интерференционного максимума от номера максимума $m$. По графику $x_m = f(m)$ определите длину волны излучения $\lambda$ и сравните с аналогичной величиной, найденной в предыдущем опыте.

10. Снимите зависимость интенсивности сигнала $I(x)$ подвижного зеркала $3_2$ в пределах одной длины волны.

Убирая поочерёдно зеркала $3_1$ и $3_2$, измерьте интенсивности $I_1$ и $I_2$ каждого из интерферирующих лучей.

Рассчитайте интенсивность сигнала, используя формулы (1) и (2) и принимая $\lambda = c/\nu$, а $\Delta = 2\delta x$, где $\delta x$ - смещение зеркала от начального положения (1-го максимума).

Постройте на одном графике и сравните экспериментальную и теоретическую зависимости $I = I(\Delta)$ в диапазоне изменения $\Delta$ от 0 до $\lambda$. Объясните результаты сравнения.

11. Настройте интерферометр на максимальную интенсивность. Поставьте в пазы подвижного зеркала $_2$ тонкую пластину из тефлона и, изменяя положение зеркала $3_2$, определите величину $\delta x$ - смещение интерференционного максимума от прежнего положения. Зная толщину пластины $h$, рассчитайте показатель преломления тефлона $n$ по формуле (8) и сравните результат с табличным значением.

### Контрольные вопросы

1. Выведите соотношение (1).

2. Как убедиться на опыте, что электромагнитная волна, излучаемая передающей антенной, линейно поляризована?

3. Как экспериментально определить направление колебаний вектора электрического поля в электромагнитной волне?

4. В чём заключается закон Малюса и как его экспериментально проверить?

5. При каких условиях можно наблюдать интерференцию ЭМ-волн?

6. Оцените длину и радиус когерентности для опыта с зеркалом и решёткой.

7. Укажите возможные причины расхождения измеренных в работе величин друг с другом и с известными значениями этих величин.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§26, 27, 33, 35.

2* Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 11-16, 29, 135, 137, 139.

3* Горелик Г.С. Колебания и волны. - М.: Физматгиз, 1959. Гл. VII-IX.

## 22.6.2 Tunneling of millimeter radio waves

Цель работы: экспериментальное исследование эффекта проникновения электромагнитных волн - туннелирования - через воздушный зазор между диэлектрическими призмами при полном внутреннем отражении на границе диэлектрик-воздух, а также моделирование интерферометра Майкельсона с использованием этого эффекта и измерение длины волны излучения и показателя преломления фторопласта для радиоволн миллиметрового диапазона.

В работе используются: генератор СВЧ-колебаний с рупорной антенной; приемная рупорная антенна и волновод; детектор; микроамперметр; металлические зеркала; две призмы и плоскопараллельная пластина из фторопласта; микрометрические винты.

Теоретические основы эффекта проникновения электромагнитных волн - туннелирования - через воздушный зазор между диэлектрическими средами при полном внутреннем отражении на границе диэлектрик-воздух изложены в вводной части раздела.

Экспериментальная установка. Туннелирование миллиметровых радиоволн через тонкий воздушный зазор переменной толщины изучается на установке, схема которой приведена на рис. 1. Источником радиоволн является высокочастотный генератор Г4 − 115 на трёх отражательных клистронах, перекрывающих полосу частот от 25,80 ГГц до 37,50 ГГц, разделённую на три поддиапазона. Генерирующий при выбранной настройке клистрон возбуждает в прямоугольном металлическом волноводе сечением $7,2 \times 3,4$ мм $^2$ электромагнитную волну, которая распространяется вдоль волновода и с помощью рупорной антенны $A_1$ излучается в пространство. Задача антенны заключается в том, чтобы сделать излучение более направленным. Электрический вектор волны, бегущей вдоль волновода и излучаемый антенной, перпендикулярен широкой стенке волновода.

На пути радиоволн устанавливаются две одинаковые прямые призмы $П_1$ и $П_2$ с почти прямоугольным (рис. 1) равнобедренным треугольником в основании. Уменьшение угла при вершине треугольника на 16° сделано для устранения обратных отражений. Призмы изготовлены из фторопласта (более точное название этого диэлектрика - фторопласт4), обладающего малыми потерями на высоких радиочастотах. Узкие грани призм ограничивают воздушную прослойку, ширина которой может изменяться с помощью микрометрических винтов $M_1$ и $M_2$.

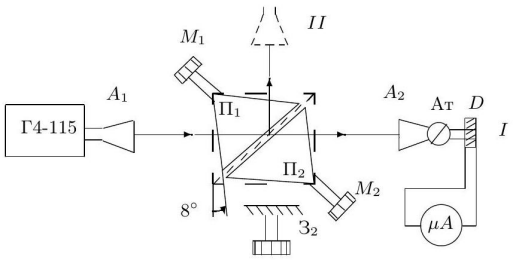



Рис. 1. Схема установки для исследования туннелирования миллиметровых радиоволн

Вторая рупорная антенна $A_2$ служит приёмником радиоволн. Попадая в антенну $A_2$, электромагнитная волна распространяется далее по волноводу, аналогичному волноводу генератора. Детектор $D$, расположенный в волноводе, подсоединяется к микроамперметру. Ток детектора пропорционален интенсивности принимаемого антенной электромагнитного излучения. Аттенюатор Ат позволяет ослаблять сигнал. В положении I антенна $A_2$ принимает сигнал, прошедший воздушный промежуток, в положении II - сигнал, отражённый от воздушного промежутка.

Небольшая реконструкция схемы рис. 1 позволяет смоделировать интерферометр Майкельсона (рис. 2). Воздушный зазор между призмами здесь используется в качестве делителя волны; зеркало $_1$ установлено неподвижно, зеркало $З_2$ может перемещаться с помощью микрометрического винта М.

Для измерения показателя преломления материала призм интерференционным методом перед неподвижным зеркалом устанавливается

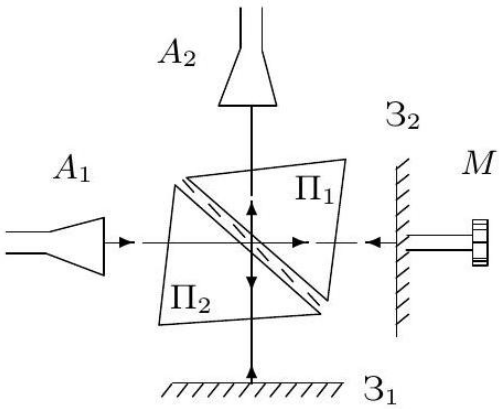



Рис. 2. Схема, моделирующая интерферометр Майкельсона пластина известной толщины $h$ из того же материала, что и призмы - фторопласта. В этом плече интерферометра возникает приращение длины «оптического» пути:

$$\Delta = 2h(n-1) \tag{1}$$

Это приращение можно скомпенсировать, передвинув подвижное зеркало на необходимое расстояние $\delta x$. Показатель преломления определяется из условия

$$\delta x = h(n-1) \tag{2}$$

Для толстых пластин, когда $\Lambda > \lambda$, необходимо учесть изменение порядка интерференции. Это можно сделать, зная приближённое значение показателя преломления фторопласта ($n \approx 1,5$).

Формулу, учитывающую изменение порядка интерференции, студентам предлагается вывести самостоятельно.

## ЗАДАНИЕ

В работе предлагается: провести исследование эффекта туннелирования радиоволн миллиметрового диапазона (снять зависимость интенсивностей прошедшей и отражённой волн от величины воздушного промежутка между призмами из фторопласта; убедиться, что энергия сохраняется, а интенсивность прошедшей волны убывает по экспоненциальному закону; рассчитать длину затухания и показатель преломления материала призм) и выполнить интерферометрические измерения в схеме Майкельсона, построенной с использованием этого эффекта (по зависимости сигнала от

смещения подвижного зеркала определить длину волны излучения; по дополнительному перемещению зеркала определить показатель преломления фторопласта пластины, добавленной на пути луча).

### I. Подготовка приборов к работе

Мощность сигнала, получаемая от генератора Г4-115, невелика, поэтому излучение не представляет опасности для здоровья человека. Тем не менее заглядывать в открытый волновод при включённом генераторе не рекомендуется.

1. Настройте генератор, руководствуясь Техническим описанием, расположенным на установке.

2. Установите столик с призмами (рис. 2) так, чтобы воздушный зазор был ориентирован под углом 45° к падающему лучу (на глаз). Для увеличения диапазона изменения воздушного промежутка установите винт левого микрометра $M_1$ в соответствующее положение. Вращением винта правого микрометра $M_2$ уберите воздушный промежуток.

3. Расположите приёмную антенну на одной прямой с передатчиком. Снимите металлическое зеркало, стоящее на пути луча.

Слегка поворачивая столик и приёмную антенну вокруг вертикальной оси, методом последовательных приближений добейтесь максимального отклика микроамперметра и закрепите оба рейтера.

4. Вращением ручек генератора)12 - установка частоты и 7 - установка генерации плавно) настройтесь на максимальную выходную мощность клистрона, наблюдая за показаниями микроамперметра.

Если ток слишком велик, уменьшите его, вращая ручку аттенюатора Ат, расположенную на приёмной антенне. В отсутствие зазора прошедший сигнал должен составлять $80 \div 90\%$ шкалы микроамперметра.

5. Вращением ручки 8 добейтесь загорания контрольной лампочки 10 и определите рабочую частоту клистрона по шкале 9. Рассчитайте соответствующую длину волны.

### II. Исследование туннелирования радиоволн

6. Снимите зависимость интенсивности прошедшей волны от величины зазора $\ell$ (используйте только правый микрометр и устраняйте люфт!). Если с увеличением зазора интенсивность падает монотонно, значит, призмы ориентированы правильно и можно продолжать измерения; если нет - попытайтесь небольшим поворотом столика избавиться от интерференции и добиться монотонного хода кривой.

7. Переставьте приёмник для измерения отражённого сигнала. Слегка поворачивая столик и приёмник, добейтесь, чтобы отклик микроамперметра на отражённый сигнал при максимальном зазоре был равен отклику на прошедший сигнал при нулевом зазоре.

Снимите зависимость интенсивности отражённой волны от величины зазора.

8. Для выполнения следующего упражнения установите такую величину зазора, при которой ток равен половине максимального. Переставляя приёмник, убедитесь, что $T \approx R \approx 0,5$.

9. Постройте на одном листе графики зависимости коэффициентов $T$ и $R$ от величины зазора $\ell$, отнормировав токи на величину $I_{\max}$. Проверьте, выполняется ли соотношение $T + R = 1$.

10. Постройте график $\ln T = f(z)$, где $z$ - показания микрометра. Проверьте, лежат ли полученные точки на одной прямой, как этого требует формула (6.33). По наклону прямой рассчитайте длину затухания $\Delta$, а затем по формуле (6.34) - величину $n \sin \varphi_1$ ($n$ - показатель преломления материала призм, $\varphi_1$ - угол падения волны на воздушный промежуток, $\lambda_2$ - длина ЭМ-волны в воздухе). Рассчитайте величину показателя преломления фторопласта $n$; при этом в условиях нашего опыта можно не учитывать, что входная плоскость призмы $\Pi_1$ наклонена на угол 8° по отношению к фронту падающей волны.

### III. Интерферометр Майкельсона

11. Соберите схему интерферометра Майкельсона (рис. 2), используя в качестве делителя воздушный зазор между призмами. Оптимальный раз мер зазора соответствует равенству $T \approx R \approx 0,5$. Установите на место неподвижное металлическое зеркало.

12. Снимите зависимость тока от координаты $x$ подвижного зеркала. По графику $I = f(x)$ определите экспериментальное значение длины волны ЭМ-излучения. Обычно максимумы размыты, поэтому определение длины волны будет более точным, если взять координаты, соответствующие максимальным производным $dI/dx$. Сравните экспериментальную длину волны с величиной, рассчитанной по рабочей частоте генератора.

13. Для измерения показателя преломления фторопласта интерференционным методом настройте интерферометр на максимальную интенсивность и поместите пластину известной толщины $d$ перед неподвижным зеркалом. Скомпенсируйте возникшее увеличение «оптической» длины пути, отодвинув от призм подвижное зеркало на необходимое расстояние $\delta x$. Рассчитайте показатель преломления фторопласта по формуле (2).

14. Сравните результаты измерения показателя преломления фторопласта $n$ интерференционным методом и методом туннелирования.

### Контрольные вопросы

1. Какой угол называется предельным углом полного внутреннего отражения? Чему он равен?

2. Что такое неоднородная волна? Как направлен вектор Умова-Пойнтинга в такой волне и почему?

3. В чем заключается явление туннелирования волн? Чем отличается это явление от обычного прохождения волн через тонкие прозрачные пластинки?

4. Каковы возможные причины расхождения измеренных в работе величин друг с другом и с табличными значениями этих величин?

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§35, 63 – 66.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 29, 135, 137, 139.

3. Калитеевский Н.И. Волновая оптика. - М.: Высшая школа, 1978. Гл. II.

## 22.7 Experiments about Polarization, Crystal Optics

### 22.7.1 Double Refraction

Цель работы: изучение зависимости показателя преломления необыкновенной волны от направления в двоякопреломляющем кристалле; определение главных показателей преломления в кристалле.

В работе используются: гелий-неоновый лазер, вращающийся столик с неподвижным лимбом, призма из исландского шпата, поляроид.

Перед выполнением работы необходимо ознакомиться с теорией двойного лучепреломления (см. п. 4 введения к разделу).

Величины $n_o = \sqrt{\varepsilon_\perp}$ и $n_e = \sqrt{\varepsilon_\parallel}$называют главными показателями преломления кристалла. Выразим, исходя из формулы (7.9), показатель преломления необыкновенной волны $n = \sqrt{\varepsilon}$ через главные показатели преломления $n_o, n_e$ и угол между оптической осью и волновым вектором $\theta$:

$$\frac{1}{n^2(\theta)} = \frac{\sin^2\theta}{n_e^2} + \frac{\cos^2\theta}{n_o^2} \tag{1}$$

Заметим, что при $\theta = \pi/2$ показатель преломления необыкновенной волны равен $n = n_e$, а при $\theta = 0$ он равен $n = n_o$. Для обыкновенной волны показатель преломления равен $n_o$ независимо от направления её распространения.

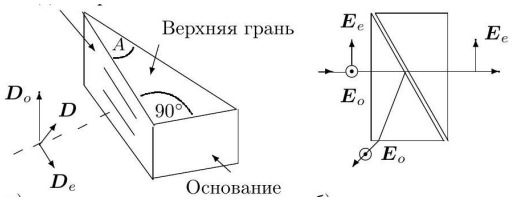
Если $n_o - n_e \ll n_o, n_e$ (для исландского шпата $n_o = 1,655, n_e = 1,485$ для $\lambda = 0,63$ мкм), формулу (1) можно упростить:

$$n(\theta) \approx n_e + (n_o - n_e)\cos^2\theta \tag{2}$$

**Двойное лучепреломление в призме из исландского шпата.**

Рассмотрим, как по преломлению лучей в кристаллической призме можно определить показатели преломления для обыкновенной и необыкновенной волны. В работе исследуется одна из двух призм, составляющих поляризатор (рис. 1).В исследуемой призме ось кристалла лежит в плоскости, параллельной верхней грани призмы, причём она параллельна входной грани призмы (длинному катету). При этом в обыкновенной волне вектор $\boldsymbol{D}_o$ перпендикулярен верхней грани призмы, а в необыкновенной волне вектор $\boldsymbol{D}_e$ параллелен верхней грани.

**Входная грань**




а)

б)

Рис. 1. а) Исследуемая призма из исландского шпата. Штриховкой указано направление оптической оси кристалла. б) Ход лучей в поляризационной призме

Волну, падающую на входную грань призмы, можно представить в виде суммы двух ортогональных линейно поляризованных волн. Преломление этих двух волн на грани призмы можно рассматривать независимо. Волна, в которой вектор $\boldsymbol{D}$ направлен вертикально (перпендикулярно верхней грани и оси кристалла), внутри кристалла будет распространяться как обыкновенная. Для этой волны выполняется закон Снеллиуса, а показатель преломления призмы для неё равен $n_o = \sqrt{\varepsilon_\perp}$. Волна, в которой вектор $\boldsymbol{D}$ направлен горизонтально, в кристалле будет распространяться как необыкновенная. Для этой волны также будет выполняться закон Снеллиуса [1], но с тем отличием, что показатель преломления призмы для неё будет зависеть от угла между осью кристалла и волновой нормалью.

Значение показателя преломления и угол, под которым преломилась волна в призме, можно найти, измерив угол падения на входную грань призмы $\varphi_1$ и угол $\varphi_2$ на выходе призмы (рис. 2). Запишем закон Снеллиуса для одной из волн применительно к первой и второй граням призмы:

$$\sin\varphi_1 = n\sin\beta_1$$
$$\sin\varphi_2 = n\sin\beta_2 = n\sin(A-\beta_1)$$

При этом мы выразили угол падения на вторую грань призмы $\beta_2$ через угол преломления на первой грани призмы $\beta_1$ и угол при вершине призмы $A$. Как видно из рис. 2, эти углы связаны простым соотношением:

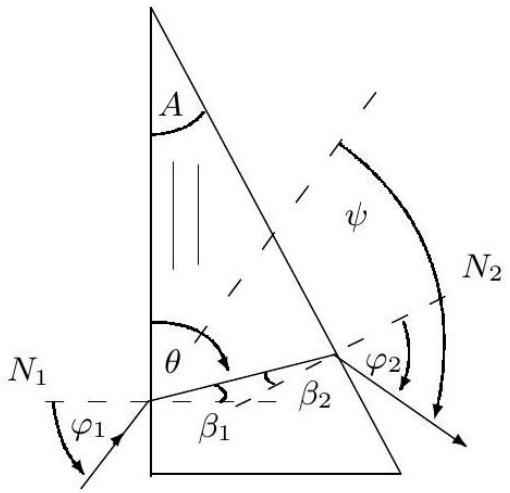



Рис. 2. Ход лучей в призме

$A = \beta_1 + \beta_2$. Учитывая, что угол преломления $\beta_1$ связан с углом $\theta$ между осью кристалла и волновой нормалью $\boldsymbol{N}$ соотношением $\theta + \beta_1 = \pi/2$, находим $n$ и $\theta$:

$$n = \frac{1}{\sin A}\sqrt{\sin^2\varphi_1 + \sin^2\varphi_2 + 2\sin\varphi_1\sin\varphi_2\cos A} \tag{3}$$
$$\cos\theta = \frac{\sin\varphi_1}{n}$$

[0]1 При этом следует помнить, что под направлением распространения волны в законе Снеллиуса подразумевается направление волнового вектора $k$, а не вектора Пойнтинга $S$.

Для обыкновенной волны $n$ не будет зависеть от угла $\theta$, а для необыкновенной волны зависимость $n$ от $\theta$ должна описываться выражением (1).

Показатель преломления призмы из изотропного материала удобно находить по углу наименьшего отклонения луча от первоначального направления. Угол отклонения луча призмой ($\psi$ на рис. 2) минимален для симметричного хода лучей, т.е. когда $\varphi_1 = \varphi_2$. Тогда показатель преломления можно рассчитать по формуле

$$n = \frac{\sin\left(\frac{\psi_m + A}{2}\right)}{\sin\left(\frac{A}{2}\right)} \tag{4}$$

где $\psi_m$ - угол наименьшего отклонения.

Если призма неизотропна, то этой формулой, строго говоря, можно воспользоваться только для обыкновенной волны, которая, как это было показано ранее, распространяется так же, как и в изотропной среде. Но если учесть, что угол при вершине призмы мал и при угле наименьшего отклонения преломлённый луч в призме распространяется под

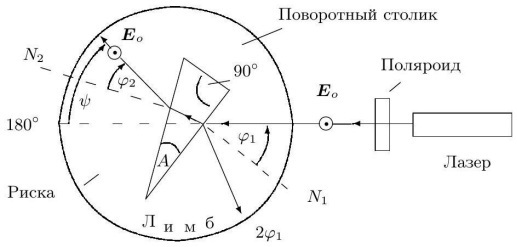



Рис. 3. Схема экспериментальной установки

углом к оси кристалла, близким к $\pi/2$, то в качестве оценки формулу (4) можно использовать для определения $n_e$.

Экспериментальная установка. Схема экспериментальной установки изображена на рис. 3. Источником излучения служит He − Ne-лазер ($\lambda = 0,63$ мкм). Излучение лазера поляризовано линейно за счёт наличия брюстеровских окошек в кювете лазера. Направление вектора $\boldsymbol{E}$ в луче можно изменять с помощью поляроида, установленного на выходе лазера. Исследуемая призма из исландского шпата закреплена в центре поворотного столика с неподвижным лимбом для отсчёта углов.

Преломляющий угол $A$ призмы (рис. 2) можно рассчитать, если известны угловые координаты нормалей $N_1$ и $N_2$ к преломляющим (рабочим) граням призмы, прилежащим преломляющему углу. Грань, противолежащая преломляющему углу, называется основанием призмы. Штриховкой указано направление оптической оси.

Обычно ход лучей в призме таков, что и падающий, и преломлённый лучи отклоняются от нормалей в сторону основания призмы, при этом углы $\varphi_1$ и $\varphi_2$ считаются положительными.

Угол падения $\varphi_1$ определяется по положению луча, отражённого от передней (входной) грани призмы (рис. 3). Из рис. 2 можно получить связь углов $\varphi_1$ и $\varphi_2$:

$$\varphi_2 = A + \psi - \varphi_1 \tag{5}$$

а угол $\psi$ - отклонение преломлённого луча от первоначального направления - определяется по разности отсчётов на лимбе между точками, куда попадает луч в отсутствие призмы, и точкой, куда попадает преломлённый луч.

При монотонном увеличении угла падения угол $\psi$ сначала уменьшается, а затем снова начинает увеличиваться. Минимальное отклонение соответствует симметричному ходу луча: внутри призмы луч идёт перпендикулярно биссектрисе угла $A$, а $\varphi_1 = \varphi_2$. Углы наименьшего отклонения $\psi_m$ различны для обыкновенного и необыкновенного лучей.

Угол $A$ подобран так, что призма может выполнять роль поляризатора: при нормальном падении луча на первую преломляющую грань из призмы выходит только один луч, а другой испытывает полное внутреннее отражение на второй грани. При повороте призмы на небольшой угол на экране появляются оба преломлённых луча. Можно подобрать такой угол падения, при котором исчезнет второй преломлённый луч. Область углов поворота призмы, в которой обеспечивается пространственное разделение лучей с взаимно ортогональной поляризацией, определяется относительной разницей главных показателей преломления $n_o$ и $n_e$.

## ЗАДАНИЕ

В работе предлагается измерить преломляющий угол призмы $A$ и определить главные показатели преломления $n_o$ и $n_e$ исландского шпата тремя способами: сняв зависимость $\psi = f(\varphi_1)$ для обыкновенной и необыкновенной волны; определив углы наименьшего отклонения $\psi_m$ для обыкновенной и необыкновенной волны; измерив для каждого из лучей угол падения $\varphi_1$ в условиях полного внутреннего отражения. Во время занятий необходимые расчёты могут быть выполнены на компьютере с помощью программы SIGMA PLOT.

1. Особенностью этой работы является то, что собственно измерения выполняются относительно быстро, но результат расчётов зависит от точности юстировки системы.

Чтобы отъюстировать установку, отцентрируйте экран по лучу лазера, т.е. убедитесь в том, что луч проходит под отметками «0» и «180», если его направить вдоль грани призмы (или над призмой). Настроив установку, закрепите оба винта на рейтере столика.

2. Определите угол $A$ при вершине призмы; для этого вначале нужно добиться, чтобы луч, отражённый от входной грани (длинного катета), шёл точно назад, заметить положение отсчётной риски на лимбе, а затем повторить эту операцию для второй рабочей грани (гипотенузы). По разнице этих двух отсчётов найдите угол $A$ и оцените точность измерения. Формулу для расчёта студентам предлагается вывести самостоятельно.

3. Определите разрешённое направление поляризатора: глядя через него на отражённый от горизонтальной поверхности (стола или подоконни ка) дневной свет, установите его в положение минимального пропускания. Так как отражённый свет преимущественно поляризован так, что вектор $\boldsymbol{E}$ направлен параллельно отражающей поверхности, у настроенного на минимум пропускания поляризатора разрешённое направление $\boldsymbol{E}$ вертикально.

4. Получите на лимбе изображения преломлённых лучей так, как показано на рис. 3 (падающий и преломлённый лучи отклоняются от нормалей к преломляющим граням в сторону основания призмы). Установите поляризатор в луче лазера перед призмой. Вращая поляризатор, определите, какой луч соответствует вертикально поляризованному свету, а какой - горизонтально поляризованному; определите, какой из лучей представляет обыкновенную волну, а какой - необыкновенную.

5. Вращая столик с призмой, снимите зависимость углов отклонения на выходе из призмы для обыкновенной и необыкновенной волн от угла падения луча на призму; удобно определять координату $2\varphi_1$ луча, отражённого от входной грани призмы - длинного катета, и координаты каждого из преломлённых лучей $(180^\circ + \psi)$ (или $180^\circ - \psi$).

Для проверки качества юстировки сначала проведите предварительную серию измерений, меняя угол падения $\varphi_1$ в диапазоне $10 - 70°$ через $10°$ ($2\varphi_1$ - через $20°$ до $140°$).

Для расчёта показателей преломления на компьютере с установленной программой SIGMA PLOT (для работы 406) подготовьте таблицу: $(\varphi_1), (\psi_o), (\psi_e)$. Обсчитайте результаты на компьютере.

6. Если рассчитанное значение $n_o$ зависит от угла падения, следует повторить юстировку, если нет - проведите основную серию измерений в том же диапазоне, меняя угол $\varphi_1$ через $5°$, а вблизи минимального угла $\psi_m$ - через $2,5°$.

Снова проведите расчёт на компьютере; покажите преподавателю, как выглядят графики, и запишите рассчитанные значения $\cos^2\theta, n_o(\theta)$ и $n_e(\theta)$ и окончательный результат $-n_o$ и $n_e$.

Распечатайте график на принтере. Если это невозможно, запишите рассчитанные на компьютере значения $\cos^2\theta, n_o(\theta), n_e(\theta)$ и постройте график самостоятельно.

7. Если работа с программой SIGMA PLOT невозможна, рассчитайте на калькуляторе 2-3 значения $n_o$ по формуле (3), а основную серию измерений обработайте дома: с помощью (5) и (3) вычислите $n_o$ и $n_e(\theta)$.

Постройте графики $n_o$ и $n_e(\theta)$ от $\cos^2\theta$ и определите главные показатели преломления $n_o$ и $n_e$. Сравните рассчитанные значения с табличными, приведёнными в описании работы, и оцените погрешности.

8. Из основной серии измерений определите средние значения углов наименьшего отклонения $\psi_m$; по формуле (4) рассчитайте показатели преломления $n_o$ и $n_e$ и оцените погрешности.

9. Для определения углов падения, соответствующих полному внутреннему отражению, сначала установите призму так, чтобы были видны оба преломлённых луча; затем, уменьшая угол падения, добейтесь для каждого из лучей выполнения условий полного отражения от второй грани призмы ($\varphi_2 = 90°$); определите соответствующие углы $\varphi_{1e}$ и $\varphi_{1o}$ с учётом знака.

Рассчитайте значения $n_o$ и $n_e$ по формуле (3) (с учётом знака угла падения) и сравните с результатами предыдущих измерений.

**Контрольные вопросы**

1. Как связаны между собой векторы $\boldsymbol{D}$ и $\boldsymbol{E}$ в анизотропной среде?

2. Как направлена оптическая ось в одноосном кристалле относительно главных осей эллипсоида диэлектрических проницаемостей?

3. Дайте определение главных показателей преломления.

4. Приведите пример, когда волна, распространяющаяся в кристалле, является обыкновенной и когда необыкновенной.

5. Получите из (1) приближённое выражение (2).

6. Как электромагнитная теория объясняет появление двух преломлённых волн при падении плоской волны на поверхность одноосного кристалла?

7. Как зависит показатель преломления обыкновенной и необыкновенной волны от угла между осью кристалла и волновой нормалью?

8. Как будет зависеть показатель преломления для обыкновенной и необыкновенной волны от угла преломления, если ось кристалла будет параллельна оси вращения призмы?

9. Как направлены векторы $\boldsymbol{E}$ в выходящих из призмы лучах?

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3. Гл. 10.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. 7, §§75, 76, 77.

3. Бутиков Е.И. Оптика. М.: Высшая школа, 1986. Гл. 4, 4.1-4.4.

4. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. 17, 107.

5. Ярив А., Юх П. Оптические волны в кристаллах. - М.: Мир, 1987. Гл. 4, §§4.1 – 4.7.

## 22.7.2 Pockels Effect

Цель работы: исследовать интерференцию рассеянного света, прошедшего кристалл; наблюдать изменение характера поляризации света при наложении на кристалл электрического поля.

В работе используются: гелий-неоновый лазер, поляризатор, кристалл ниобата лития, матовая пластинка, экран, источник высоковольтного переменного и постоянного напряжения, фотодиод, осциллограф, линейка.

Перед выполнением данной работы необходимо ознакомиться с теоретическим введением к разделу.

Эффектом Поккельса называется изменение показателя преломления света в кристалле под действием электрического поля, причём это изменение пропорционально напряжённости электрического поля. Вследствие эффекта Поккельса в кристалле либо появляется двойное лучепреломление, либо меняется его величина (если кристалл был двулучепреломляющим в отсутствие поля), либо, как в данной работе, одноосный кристалл становится двуосным.

Изменение показателя преломления кристаллов под действием внешнего электрического поля происходит за счёт анизотропных свойств кристаллов. Под действием постоянного электрического поля электроны смещаются в сторону того или иного иона (в случае кристалла ниобата лития $LiNbO_3$ - это ион Li или Nb), при этом меняется поляризуемость среды и связанный с ней показатель преломления. В первом приближении это изменение линейно относительно внешнего электрического поля. Эффект Поккельса может наблюдаться только в кристаллах, не обладающих центром симметрии. Вследствие линейности эффекта относительно внешнего поля $E_{эл}$ при изменении направления поля на противоположное должен меняться на противоположный и знак изменения показателя преломления $\Delta n$. Но в кристаллах с центром симметрии это невозможно, так как оба взаимно противоположных направления внешнего поля физически эквивалентны. Кристалл можно поместить между двумя скрещенными поляроидами таким образом, что в отсутствие внешнего электрического

поля пропускание света системой будет равно нулю. При подаче на кристалл внешнего поля появится наведённое двулучепреломление, которое изменит поляризацию прошедшего через кристалл света, и такая система начнёт пропускать свет. На этом принципе основаны многочисленные применения эффекта Поккельса в лазерной технике для оптических модуляторов, затворов и других устройств, управляющих лазерным излучением. Поскольку эффект Поккельса связан с изменением электронной поляризуемости под действием электрического поля, то он практически безынерционен - быстродействие устройств на его основе меньше $10^{-9}$ с.

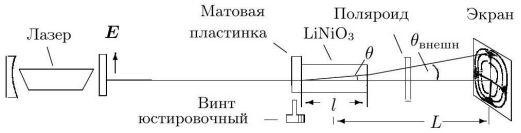



Рис. 1. Схема для наблюдения интерференционной картины

Рассмотрим сначала кристалл в отсутствие внешнего электрического поля. Кристалл ниобата лития является одноосным кристаллом, то есть кристаллом, оптические свойства которого обладают симметрией вращения относительно некоторого одного направления, называемого оптической осью z кристалла. Для световой волны, вектор электрического поля $\boldsymbol{E}$ которой перпендикулярен оси $z$, показатель преломления равен $n_o = \sqrt{\varepsilon_\perp}$, а для волны, вектор $\boldsymbol{E}$ которой располагается вдоль оси $z$, он равен $n_e = \sqrt{\varepsilon_{\parallel\parallel}}$, причём $n_e < n_o$, т. е. $LiNbO_3$ - «отрицательный кристалл».

В общем случае, когда луч света распространяется под углом $\theta$ к оптической оси $z$ (рис. 1), существуют два собственных значения показателя преломления $n_1$ и $n_2$: в обыкновенной волне (если световой вектор $\boldsymbol{E}$ перпендикулярен плоскости $(\boldsymbol{k}, \boldsymbol{e}_z)$, где $\boldsymbol{k}$ - волновой вектор луча, $\boldsymbol{e}_z$ - орт по оси $z$) показатель $n_1 = n_o$, а в необыкновенной (когда световой вектор $\boldsymbol{E}$ лежит в плоскости $(\boldsymbol{k}, \boldsymbol{e}_z)$) показатель преломления $n_2$ зависит от угла $\theta$: согласно формуле (7.9), можем записать

$$\frac{1}{n_2^2} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \tag{1}$$

Если перед кристаллом, помещённым между скрещенными поляроидами (рис. 1), расположить линзу или матовую пластинку, после которых лучи будут рассеиваться под различными углами, то на экране, расположенном за поляроидом, мы увидим тёмные концентрические окружности (коноскопическую картину) - результат интерференции обыкновенной и необыкновенной волн, точнее, проекцию их электрических полей на разрешённое направление выходного поляроида. В на шем эксперименте используется лазер, излучение которого поляризовано, поэтому входной поляроид можно не ставить.

Разность фаз между обыкновенной и необыкновенной волнами, приобретаемая при прохождении через кристалл длиной $\ell$, равна

$$\Delta\varphi = \frac{2\pi}{\lambda} \cdot \ell \cdot (n_1 - n_2)$$

Для обыкновенного луча $n_1 = n_o$ и не зависит от угла $\theta$ между направлением луча и осью $z$. Для необыкновенного луча $n_2$ зависит от угла $\theta$ и определяется уравнением (1). Считая, что $n_e$ и $n_o$ отличаются незначительно, для малых углов $(\sin\theta \approx \theta, \cos\theta \approx 1 - \theta^2/2)$ получаем

$$n_2 \approx n_o - (n_o - n_e)\,\theta^2$$

Таким образом,

$$\Delta\varphi = \frac{2\pi}{\lambda}\ell\,(n_o - n_e)\,\theta^2$$

Направлениями постоянной разности фаз служат конусы $\theta = \text{const}$, поэтому интерференционная картина представляет собой концентрические окружности. Интерференционные кольца перерезаны тёмным «мальтийским крестом», который выделяет области, где интерференция отсутствует. В этих направлениях распространяется только одна поляризованная волна (обыкновенная или необыкновенная). При повороте выходного поляроида (анализатора) на 90° картина меняется с позитива на негатив: везде, где были светлые места, появляются тёмные и наоборот.

Для случая, когда разрешённое направление анализатора перпендикулярно поляризации лазерного излучения (скрещенные поляризации), найдём радиус тёмного кольца с номером $m$. Для луча, идущего вдоль оси $z(m = 0)$, показатели преломления для двух волн совпадают, сдвиг фаз между ними равен нулю, поляризация излучения на выходе остаётся такой же, как на входе, и луч не проходит через анализатор. Картина не изменится при сдвиге фаз между обыкновенной и необыкновенной волной, кратном $2\pi$. Поэтому для $m$-го тёмного кольца $\Delta\varphi = 2\pi m$ или $\Delta\varphi = \frac{2\pi}{\lambda}\ell\,(n_o - n_e)\,\theta^2 = 2\pi m$. Если $L$ - расстояние от центра кристалла до экрана, то, учитывая закон преломления (закон Снеллиуса) на границе кристалла, при малых углах $\theta_{\text{внешн}} = n_o\theta$ (рис. 1) получаем выражение для радиуса кольца:

$$r_m^2 = \frac{\lambda}{\ell}\frac{(n_o L)^2}{(n_o - n_e)}m \tag{2}$$

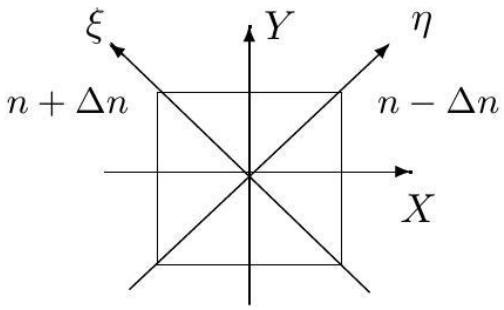



Рис. 2. Эффект Поккельса появление новых главных направлений при наложении электрического поля

Измеряя радиусы колец, можно найти разность $(n_o - n_e)$ - двулучепреломление кристалла.

Представим теперь, что мы поместили кристалл в постоянное электрическое поле $E_{\text{эл}}$ , направленное вдоль оси $x$, перпендикулярной оптической оси кристалла z. Луч света распространяется вдоль оси $z$, при этом для любой поляризации в отсутствие внешнего поля показатель преломления равен $n_o$. Свойства симметрии кристалла и его электрооптический тензор таковы, что в результате линейного электрооптического эффекта (эффекта Поккельса) в плоскости $(xy)$ возникают два главных направления $\xi$ и $\eta$ под углами 45° к осям $x$ и $y$ (рис. 2) с показателями преломления $(n_o - \Delta n)$ и $(n_o + \Delta n)$, то есть появляются «медленная» и «быстрая» оси, причём $\Delta n = A \cdot E_{\text{эл}}$ ($A$ - некая константа, зависящая только от типа кристалла).

Пусть свет на входе в кристалл поляризован вертикально, а на выходе стоит анализатор, пропускающий горизонтальную поляризацию. Разложим исходный световой вектор $E = E_0 e^{i(\omega t - kz)}$ по осям $\xi$ и $\eta$: $E_\xi = E_\eta = E_0/\sqrt{2}$. После прохождения кристалла между векторами $E_\xi$ и $E_\eta$ появится разность фаз:

$$\Delta\varphi = \frac{2\pi\ell}{\lambda} 2\Delta n = \frac{4\pi\ell}{\lambda} A E_{\text{эл}} = \frac{4\pi}{\lambda}\frac{\ell}{d} A U$$

где $U = E_{\text{эл}}\, d$ - напряжение на кристалле, $d$ - размер кристалла в поперечном направлении. Результирующее поле после анализатора это сумма проекций $E_\xi$ и $E_\eta$ на направление $x$, т. е.

$$E_{\text{вых}} = \frac{E_0}{2} e^{i(\omega t - k\ell)} \left(e^{i\Delta\varphi/2} - e^{-i\Delta\varphi/2}\right) = E_0 e^{i(\omega t - k\ell)} \sin\left(\frac{\Delta\varphi}{2}\right)$$

Интенсивность света пропорциональна квадрату модуля вектора электрического поля в волне:

$$I_{\text{вых}} \sim E E^* = E_0^2 \sin^2\left(\frac{\Delta\varphi}{2}\right)$$

поэтому

$$I_{\text{Вх}} = I_0 \sin^2\left(\frac{\Delta\varphi}{2}\right) = I_0 \sin^2\left(\frac{\pi}{2}\frac{U}{U_{\lambda/2}}\right) \tag{3}$$

Здесь

$$U_{\lambda/2} = \frac{\lambda}{4A}\frac{d}{\ell} \tag{4}$$

- так называемое полуволновое напряжение - имеет тот смысл, что при $U = U_{\lambda/2}$ сдвиг фаз между двумя волнами, соответствующими двум собственным поляризациям, $\Delta\varphi = \pi$ (разность хода равна $\lambda/2$), и интенсивность света на выходе анализатора, как следует из (3), достигает максимума.

Студенту предлагается показать, что при параллельных поляризациях лазера и анализатора

$$I_{\text{Вх}} = I_0 \cos^2\left(\frac{\pi}{2}\frac{U}{U_{\lambda/2}}\right) \tag{5}$$

Напряжение $U_{\lambda/2}$ называют также управляющим напряжением. Оно уменьшается, как это видно из (4), с уменьшением длины волны света $\lambda$ и с увеличением отношения $\lambda/d$ кристалла (это справедливо для поперечного электрооптического эффекта, который используется в нашем опыте). Характерная величина полуволнового напряжения в ниобате лития для видимого света составляет несколько сотен вольт.

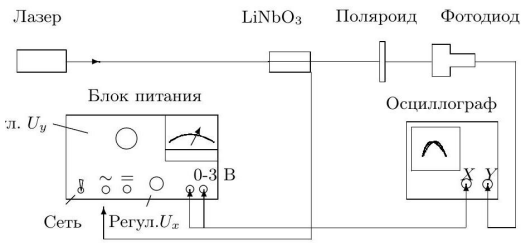



Рис. 3. Схема для изучения двойного лучепреломления в электрическом поле

Экспериментальная установка. Оптическая часть установки представлена на рис. 1. Свет гелий-неонового лазера, поляризованный в вертикальной плоскости, проходя сквозь матовую пластинку, рассеивается и падает на двоякопреломляющий кристалл под различными углами. Кристалл ниобата лития с размерами $3 \times 3 \times 26$ мм вырезан вдоль

оптической оси $z$. На экране, расположенном за скрещенным поляроидом, видна интерференционная картина.

Для $\lambda = 0,63$ мкм (длина волны гелий-неонового лазера) в ниобате лития $n_o = 2,29$.

Убрав рассеивающую пластинку и подавая на кристалл постоянное напряжение, можно величиной напряжения влиять на поляризацию луча, вышедшего из кристалла.

Заменив экран фотодиодом (рис. 3) и подав на кристалл переменное напряжение, можно исследовать поляризацию луча с помощью осциллографа.

## ЗАДАНИЕ

В работе предлагается: 1) измерив радиусы интерференционных колец, определить разность показателей преломления $n_o - n_e$; 2) подав на кристалл постоянное напряжение, получить свет, поляризованный по кругу; 3) определить полуволновое напряжение по фигурам Лиссажу на экране осциллографа.

### I. Юстировка системы

1. Соберите оптическую схему согласно рис. 1. Включите лазер и установите анализатор (без кристалла в схеме) так, чтобы лазерное излучение через него не проходило (скрещенные поляризации).

Чтобы убедиться, что лазерный луч поляризован вертикально, определите разрешённое направление анализатора: посмотрите сквозь поляроид на дневной свет от окна, отражённый от светлой поверхности (от стола, подоконника); вращая поляроид, найдите минимум освещённости и заметьте отсчёт угла на лимбе. Вблизи угла Брюстера в отражённом свете преобладает компонента светового вектора, параллельная плоскости стола («правило иголки»), следовательно, минимум отражённого света соответствует вертикальному разрешённому направлению поляроида.

2. Поставьте кристалл и установите перед ним вплотную к кювете матовую пластинку. Расстояние от кристалла до экрана определяет размер интерференционной картины и её контрастность: 70-80 см - оптимальное расстояние для тёмной комнаты, в светлое время дня - около 30 см.

3. Получите на экране интерференционную картину. Отклоняя кристалл с помощью юстировочного винта (рис. 1) и поворачивая рейтер с кюветой вокруг вертикальной оси, добейтесь совмещения центра коноскопической картины с положением луча на экране в отсутствие матовой пластинки.

Поверните анализатор на 90° и убедитесь, что коноскопическая картина изменилась на негативную. Верните анализатор в прежнее положение (горизонтальное разрешённое направление).

### II. Измерения

4. Измерьте радиусы тёмных колец $r(m)$ и расстояние $L$ от середины кристалла до экрана.

Постройте график $r^2 = f(m)$. По углу наклона прямой определите двулучепреломление $(n_o - n_e)$ ниобата лития, пользуясь формулой (2). Величины $n_o, \ell$ и $\lambda$ указаны на установке.

5. Ещё раз убедитесь, что направление лазерного луча совпадает с направлением на центр интерференционной картины и уберите матовую пластинку. Можно немного подъюстировать кристалл до получения минимальной интенсивности луча после прохождения анализатора.

Подключите разъём блока питания на постоянное напряжение (= (, установите регулятор напряжения на минимальное напряжение и включите блок питания в сеть.

С увеличением напряжения на кристалле яркость пятна на экране увеличивается и достигает максимума при $U = U_{\lambda/2}$. При $U = 2U_{\lambda/2} = U_\lambda$ яркость снова будет минимальной и т. д. Проделайте то же для параллельных поляризаций лазера и анализатора. Определите полуволновое напряжение ниобата лития.

6. Подайте на кристалл напряжение $U = \frac{1}{2}U_{\lambda/2} = U_{\lambda/4}$ (четвертьволновое напряжение). Поляризация на выходе кристалла должна быть круговой. Убедитесь в этом, вращая анализатор и наблюдая за яркостью пятна на экране.

7. Установите вместо экрана фотодиод (рис. 3) и подключите его к $y$-входу осциллографа. Убрав напряжение до нуля, переключите разъём с постоянного (=) на переменное напряжение ($\sim$). С трёхвольтового выхода блока питания подайте сигнал на вход $x$ осциллографа. Отклонение луча осциллографа по оси $x$, таким образом, будет пропорционально напряжению $U$ на кристалле, а по оси $y$ - интенсивности прошедшего через анализатор сигнала $I_{\text{вых}}$ .

8. Постепенно повышая напряжение на кристалле, наблюдайте на экране осциллографа фигуры Лиссажу, соответствующие зависимости $I_{\text{вых}}(U)$ для скрещенных поляризаций лазера и анализатора.

Слегка поворачивая кристалл, сделайте фигуру Лиссажу симметричной. Определите по фигурам Лиссажу полуволновое напряжение $U_{\lambda/2}$ как $\Delta U$, соответствующее переходу от максимума к минимуму сигнала на осциллограмме. Сравните значения полуволнового напряжения, полученные при постоянном и переменном напряжениях.

9. Зарисуйте фигуры Лиссажу для напряжений $U_{\lambda/2}, U_\lambda, U_{3\lambda/2}$ при скрещенных поляризациях лазера и анализатора. Проследите, как меняется картина при переходе к параллельным поляризациям.

**Контрольные вопросы**

1. Можно ли наблюдать эффект Поккельса в жидкости?

2. Почему при повороте анализатора на 90° интерференционная картина на экране меняется с позитива на негатив?

3. Что такое полуволновое напряжение $U_{\lambda/2}$? Как оно зависит от длины волны $\lambda$?

4. Почему при $U = \frac{1}{2}U_{\lambda/2}$ поляризация на выходе из кристалла круговая?

5. Оцените относительное изменение показателя преломления данного кристалла при подаче напряжения $U = U_{\lambda/2}$.

6. Выведите формулу (5).

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 148, 150-153.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§76, 91.

3. Борн М., Волъф З. Основы оптики. - М.: Наука, 1970. 14.3, 14.4.

4. Ярив А.О. Квантовая электроника. - М.: Советское радио, 1980. §§14.1 – 14.5.

## 22.7.3 Polarization

Цель работы: ознакомление с методами получения и анализа поляризованного света.

В работе используются: оптическая скамья с осветителем; зелёный светофильтр; два поляроида; чёрное зеркало; полированная эбонитовая пластинка; стопа стеклянных пластинок; слюдяные пластинки разной толщины; пластинки в 1/4 и 1/2 длины волны; пластинка в одну длину волны для зелёного света (пластинка чувствительного оттенка).

Перед выполнением работы необходимо ознакомиться с теоретическим введением к разделу.

Определение направления разрешённой плоскости колебаний поляроида. Определить направление разрешённых колебаний поляроида проще всего с помощью чёрного зеркала.

При падении света на отражающую поверхность под углом Брюстера, свет в отражённом луче почти полностью поляризован, а вектор $\boldsymbol{E}$ параллелен отражающей поверхности («правило иголки»). Луч света,
прошедший поляроид и отразившийся от чёрного зеркала, имеет минимальную интенсивность при выполнении двух условий: во-первых, свет падает на отражающую поверхность под углом Брюстера и, во-вторых, в падающем пучке вектор $\boldsymbol{E}$ лежит в плоскости падения.

Вращая поляроид вокруг направления луча и чёрное зеркало вокруг оси, перпендикулярной лучу, методом последовательных приближений можно добиться минимальной яркости луча, отражённого от зеркала, и таким образом определить разрешённое направление поляроида.

Измеряя угол поворота зеркала (угол Брюстера), нетрудно определить коэффициент преломления материала, из которого изготовлено зеркало. Описанный метод часто используется для измерения коэффициента преломления непрозрачных диэлектриков.

Получение эллиптически поляризованного света. Эллиптически поляризованный свет можно получить из линейно поляризованного с помощью двоякопреломляющих кристаллических пластинок.

Двоякопреломляющая пластинка имеет два взаимно перпендикулярных главных направления, совпадающих с осями эллипсоида диэлектрической проницаемости. Волны, поляризованные вдоль главных направлений, распространяются в пластинке с разными скоростями, не изменяя характера своей поляризации. Эти волны называются главными. Мы будем обозначать показатели преломления для главных волн через $n_x$ и $n_y$, где $x$ и $y$ - главные направления кристаллической пластинки (рис. 1).

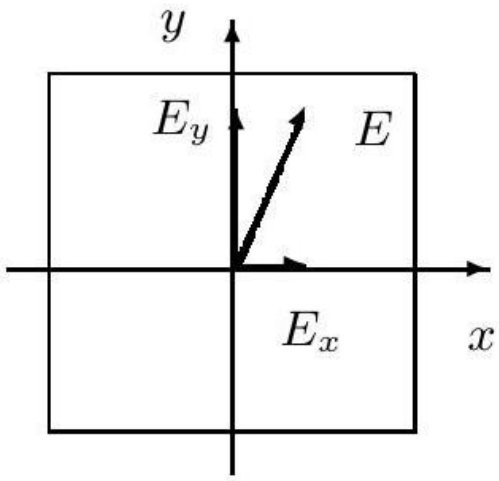



Рис. 1. Разложение линейно поляризованного света по главным направлениям двоякопреломляющей пластинки

Пусть на пластинку падает линейно поляризованная волна, электрический вектор которой ориентирован под некоторым углом $\alpha$ к оси $x$. Разложим вектор $\boldsymbol{E}$ на составляющие $E_x$ и $E_y$. На входе пластинки $E_x$ и $E_y$ находятся в фазе. На выходе из-за разности скоростей между ними появляется разность хода $d\,(n_x - n_y)$, при этом сдвиг фаз определяется соотношением (см. (7.12)):

$$\Delta\varphi = \frac{2\pi}{m} = kd\,(n_x - n_y) \tag{1}$$

где $k$ - волновое число (в пустоте), $d$ - толщина кристаллической пластинки. Как уже отмечалось, при сложении двух взаимно перпенди кулярных колебаний, обладающих некоторым сдвигом фаз, образуется колебание, поляризованное по эллипсу.

Рассмотрим практически важные частные случаи.

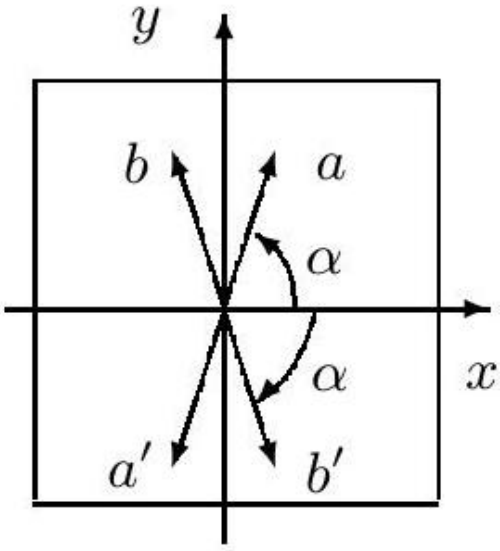



Рис. 2. Поворот направления колебаний с помощью пластинки в $\lambda/2$

а) Пластинка даёт сдвиг фаз $2\pi$ (пластинка в длину волны $\lambda$). В результате сложения волн на выходе пластинки образуется линейно поляризованная волна с тем же направлением колебаний, что и в падающей волне.

б) Пластинка даёт сдвиг фаз $\pi$ (пластинка в полдлины волны $\lambda/2$). На выходе пластинки снова образуется линейно поляризованная волна. Направление $bb'$ колебаний

этой волны повёрнуто относительно направления $aa'$ колебаний падающей волны (рис. 2). Как нетрудно сообразить, направление $bb'$ является зеркальным отображением направления $aa'$ относительно одного из главных направлений пластинки. Такую пластинку используют для поворота направления колебаний линейно поляризованного света.

в) Пластинка создаёт между колебаниями сдвиг фаз $\pi/2$ (пластинка в четверть длины волны). При сложении двух взаимно перпендикулярных колебаний, имеющих разность фаз $\pi/2$, образуется эллипс, главные оси которого совпадают с координатными осями $x$ и $y$. При равенстве амплитуд $E_x^{\max} = E_y^{\max}$ возникает круговая поляризация.

Следует отметить, что, говоря о пластинках $\lambda, \lambda/2, \lambda/4$ и т. д., всегда подразумевают какую-либо вполне определённую монохроматическую компоненту (например, пластинка $\lambda/2$ для зелёного света). Если на двоякопреломляющую пластинку падает не монохроматический свет, то на выходе из неё для разных спектральных компонент эллипсы поляризации будут различными.

Анализ эллиптически поляризованного света. Анализ эллиптически поляризованного света сводится к нахождению главных осей эллипса поляризации и к определению направления вращения электрического вектора.

Главные оси эллипса поляризации определяются с помощью анализатора по максимуму и минимуму интенсивности проходящего света. Направление вращения электрического вектора может быть найдено с помощью пластинки в четверть длины волны, для которой известно, какая из главных волн, $E_x$ или $E_y$, имеет большую скорость распространения (и соответственно меньшее значение показателя преломления).

Выберем для определённости координатные оси $x$ и $y$ на пластинке так, чтобы $n_x < n_y$. В этом случае главная волна $E_x$ имеет большую скорость распространения. Поместим такую пластинку на пути эллиптически поляризованного света и совместим главные направления пластинки $\lambda/4$ с главными осями эллипса поляризации. На выходе из этой пластинки сдвиг фаз между $E_x$ и $E_y$ вместо $\pi/2$ станет равным нулю или $\pi$. Свет окажется линейно поляризованным. Из двух возможных значений сдвига фаз, 0 или $\pi$, реализуется одно: то, которое соответствует имеющемуся в волне направлению вращения электрического вектора.

Рассмотрим, например, случай, когда электрический вектор в эллиптически поляризованной волне вращается против часовой стрелки, если смотреть навстречу лучу. В этом случае, очевидно, в волне, падающей на пластинку в $\lambda/4$, колебание $E_y$ отстаёт по фазе на $\pi/2$ от колебания $E_x$. При прохождении через пластинку разность фаз увеличивается до $\pi$. Таким образом на выходе из пластинки возникают линейно поляризованные волны со сдвигом фаз $\pi$. Сложение этих волн даёт плоскополяризованную волну, электрический вектор которой располагается во втором и четвёртом квадрантах координатной системы $x, y$.

Рассуждая аналогичным образом, найдём, что при вращении электрического вектора по часовой стрелке направление колебаний в линейно поляризованной волне, выходящей из пластинки, располагается в первом и третьем квадрантах. Определяя направление колебаний на выходе из пластинки с помощью поляроида, можно, таким образом, определить характер эллиптической поляризации (вращение против или по часовой стрелке).

Пластинка чувствительного оттенка. Выше предполагалось известным, какому из двух главных направлений пластинки в четверть длины волны соответствует большая скорость распространения света. Установить это можно различными способами, например с помощью пластинки чувствительного оттенка (так называют пластинку в $\lambda$ для зелёной спектральной компоненты, $\lambda = 560$ нм).

Пластинка имеет форму стрелы (рис. 3), вдоль оси которой расположено главное направление, соответствующее большей скорости распространения.

Если пластинка чувствительного оттенка помещена между скрещенными поляроидами и главные направления пластинки не параллельны направлениям разрешённых колебаний поляроидов, то при освещении белым светом пластинка кажется окрашенной в лилово-красный цвет. Это объясняется тем, что зелёная компонента линейно поляризованно го света при прохождении пластинки не меняет поляризации и задерживается вторым поляроидом. Для красной и фиолетовой компонент пластинка создаёт сдвиг фаз, несколько отличный от $2\pi$. На выходе из пластинки красная и фиолетовая компоненты оказываются поэтому эллиптически поляризованными и частично проходят через второй поляроид. Таким образом, в известном смысле наблюдаемый в указанном опыте цвет пластинки дополнителен к зелёному.

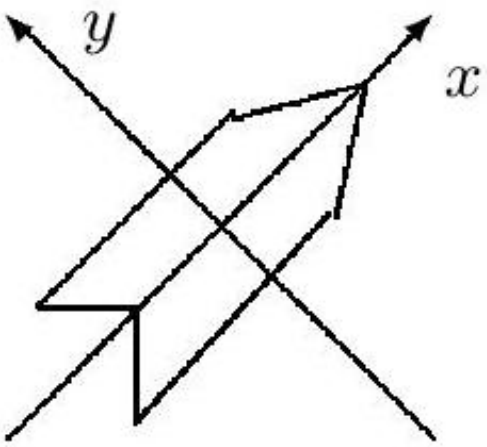



Рис. 3. Пластинка чувствительного оттенка

Если между скрещенными поляроидами поместить пластинку чувствительного оттенка ($\lambda$) и пластинку в $\lambda/4$ так, чтобы их главные направления совпадали, цвет пластинки изменится. Если у пластинки чувствительного оттенка и пластинки в $\lambda/4$ совпадут главные направления, соответствующие большей скорости распространения, то разность хода между $E_x$ и $E_y$ для зелёного света составит уже $5\lambda/4$. Это соответствует разности хода в $\lambda$ для света с большей длиной волны, т. е. для «более красного» света. При освещении этих пластинок (напомним, что они расположены между скрещенными поляроидами) белым светом теперь погасится не зелёная, а красная часть спектра, и проходящий свет будет казаться зеленовато-голубым. Если же главные направления, соответствующие большей скорости распространения, у пластинки чувствительного оттенка и у пластинки в $\lambda/4$ окажутся перпендикулярными, то проходящий свет приобретёт оранжево-желтую окраску (погасится фиолетово-голубая часть спектра).

Изменение цвета позволяет, таким образом, определить, какое из главных направлений пластинки в $\lambda/4$ соответствует большей скорости распространения.

Интерференция поляризованных лучей. Тонкие двоякопреломляющие пластинки, помещённые между поляроидами, кажутся окрашенными. Эта окраска может быть истолкована как результат интерференции поляризованных лучей. На рис. 4 представлена схема для случая скрещенных поляроидов.

Здесь $p_1p_1'$ - разрешённое направление колебаний поляризатора (первого поляроида); $x, y$ - координатная система, связанная с главными направлениями двоякопреломляющей пластинки; $p_2p_2'$ - разрешённое направление колебаний анализатора (второго поляроида). Волны $E_x$ и $E_y$ на выходе из пластинки когерентны, но не могут интерферировать, так как $\boldsymbol{E}_x \perp \boldsymbol{E}_y$. Волны $E_1$ и $E_2$ на выходе второго поляроида также являются когерентными

и к тому же поляризованы в одной плос кости. Эти волны интерферируют между собой. Результат интерференции определяется зависящим от длины волны сдвигом фаз между $E_1$ и $E_2$. В результате интерференции поляризованных лучей пластинка, освещаемая белым светом, кажется окрашенной.

Если поворачивать двоякопреломляющую пластинку, расположенную между скрещенными поляроидами, то соотношение амплитуд волн $E_1$ и $E_2$ и разность фаз между ними не изменяются. Это означает, что цвет пластинки при её поворотах не меняется, а меняется только интенсивность света. За один оборот пластинки интенсивность четыре раза обращается в нуль - это происходит при совпадении главных направлений $x$ и $y$ с разрешёнными направлениями коле-

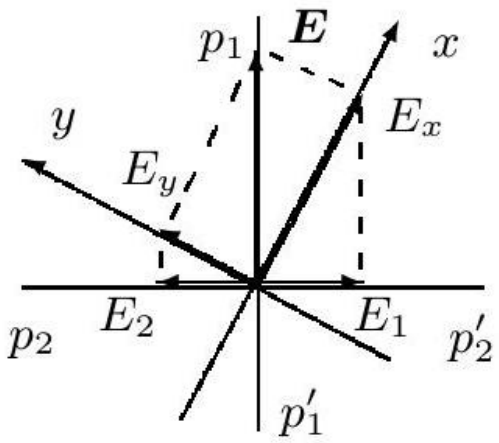



Рис. 4. К объяснению интерференции поляризованных лучей баний поляроидов.

Если же двоякопреломляющую пластинку оставить неподвижной, а второй поляроид повернуть так, чтобы разрешённые направления $p_1p_1'$ и $p_2p_2'$ совпали, то волны $E_1$ и $E_2$ приобретают дополнительный фазовый сдвиг на $\pi$ для всех спектральных компонент; при этом их амплитуды изменятся так, что цвет пластинки изменится на дополнительный.

Студентам предлагается самостоятельно это доказать.

## ЗАДАНИЕ

В работе предлагается с помощью чёрного зеркала определить разрешённые направления поляроидов; определить характер поляризации света, прошедшего стопу и отражённого от неё под углом Брюстера; оценить угол Брюстера для эбонита; выделить пластинки $\lambda/2$ и $\lambda/4$; определить направления большей и меньшей скоростей для пластинки $\lambda/4$; исследовать интерференцию поляризованных лучей.

1. Определите разрешённые направления поляроидов. Для этого

а) разместите на оптической скамье осветитель $S$, поляроид $P_1$ и чёрное зеркало (рис. 5);

б) поворачивая поляроид вокруг направления луча, а чёрное зеркало вокруг вертикальной оси, методом последовательных приближений добейтесь наименьшей яркости отражённого пятна. Определите разрешённое направление поляроида. Зарисуйте схему и опишите результат.

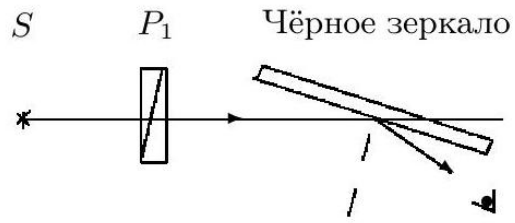



Рис. 5. Определение разрешённого направления поляроида

в) Поставьте вместо чёрного зеркала второй поляроид и определите его разрешённое направление, скрестив поляроиды.

2. Определите показатель преломления эбонита. Для этого

а) поставьте на скамью вместо чёрного зеркала (рис. 5) эбонитовую пластину и определите по лимбу угол Брюстера для эбонита, оцените неточность установки угла;

б) повторите измерения, добавив светофильтр Ф, и сравните результаты;

в) по углу Брюстера рассчитайте показатель преломления эбонита и сравните с табличным.

3. Исследуйте характер поляризации света в преломлённом и отражённом от стопы лучах.

Для этого поставьте вместо эбонитового зеркала (рис. 5) стопу стеклянных пластинок под углом Брюстера.

Осветите стопу неполяризо-

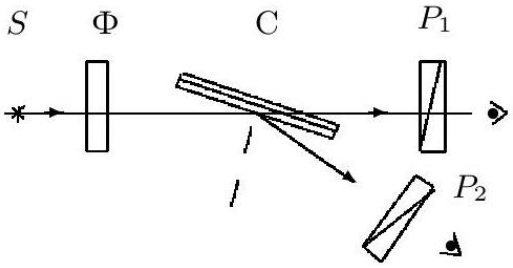



Рис. 6. Исследование стопы ванным светом и, рассматривая через поляроиды (рис. 6) отражённый от стопы и преломлённый лучи, определите в них ориентацию вектора $\boldsymbol{E}$. Опишите результат.

4. Определите главные направления двоякопреломляющих пластин:

а) поставьте кристаллическую пластинку между скрещенными поляроидами (рис. 7);

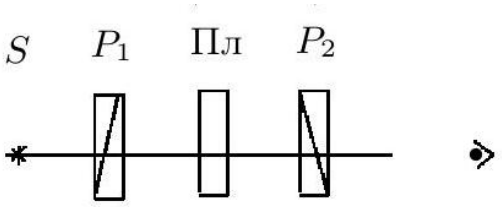



Рис. 7. Определение главных направлений в пластинках

б) вращая пластинку вокруг направления луча и наблюдая за интенсивностью света, проходящего сквозь второй поляроид, определите, при каком условии главные направления пластинки совпадают с разрешёнными направлениями поляроидов;

в) повторите опыт для второй пластинки.

5. Для выделения пластин $\lambda/2$ и $\lambda/4$

а) добавьте к схеме, изображённой на рис. 7, зелёный фильтр;

б) установите разрешённое направление первого поляроида горизонтально, а главные направления исследуемой пластинки - под углом 45° к горизонтали;

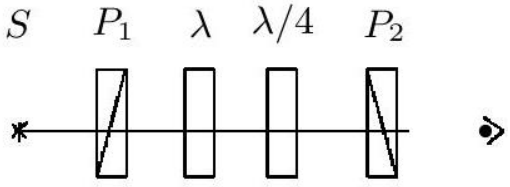
в) с помощью второго поляроида установите, какую поляризацию имеет свет, прошедший пластинку: круговую или линейную с переходом в другой квадрант. Опишите результат.

6. Определите «быструю» и «медленную» оси в пластинке $\lambda/4$:

а) поставьте между скрещенными поляроидами пластинку чувствительного оттенка, имеющую вид стрелки, и убедитесь, что эта пластинка не меняет поляризацию зелёного света;

б) уберите зелёный фильтр и убедитесь, что стрелка имеет пурпурный цвет; объясните это явление;

в) добавьте к схеме пластинку $\lambda/4$




(рис. 8), главные направления которой совпадают с главными направлениями пластины $\lambda$ и ориентированы

Рис. 8. Определение направлений под углом 45° к разрешённым направ-

большей и меньшей скорости лениям скрещенных поляроидов.

При повороте рейтера со стрелкой на 180° вокруг вертикальной оси цвет стрелки меняется от зелёно-голубого до оранжево-жёлтого. В каком случае «быстрые» оси обеих пластин совпадают? Объясните явление и опишите результат.

г) Для последнего упражнения установите направление «быстрой» оси пластины $\lambda/4$ горизонтально.

7. Исследуйте интерференцию поляризованных лучей. Для этого расположите между скрещенными поляроидами мозаичную слюдяную пластинку. Она собрана из 4 - х узких полосок слюды, лежащих по сторонам квадрата (две полоски «толщиной» $\lambda/4$ и по одной - $\lambda/2$ и $3\lambda/4$). В центральном квадратике слюды нет. Главные направления всех пластинок ориентированы параллельно сторонам квадрата.

а) Вращая пластинку, наблюдайте за изменениями (цвета или интенсивности) в отдельном квадратике.

б) Не трогая пластинки, вращайте второй поляроид. Чем отличается эффект? Опишите и объясните результат.

8. Определите направление вращения светового вектора в эллиптически поляризованной волне.

а) Нарисуйте эллипс поляризации для вектора $\boldsymbol{E}$, вышедшего из пластинки $\lambda/4$, и укажите, какая из его осей соответствует большей скорости. Рядом нарисуйте две вышедших из пластинки синусоиды: $x(t)$ и $y(t)$ со сдвигом фаз в четверть периода. Проанализируйте графики и определите направление вращения электрического вектора в эллиптически поляризованной волне.

б) Поставьте зелёный фильтр, а за ним между скрещенными поляроидами - пластинку произвольной толщины (это может быть пластинка $\lambda/4$ с соседней установки).

в) Получите эллиптически-поляризованный свет. Для этого установите разрешённое направление первого поляроида под углом $10 - 20°$К горизонтали так, чтобы вектор $\boldsymbol{E}$ падающего на пластинку света был расположен в первом квадранте. Установите разрешённое направление второго поляроида вертикально и, вращая пластинку, найдите минимальную интенсивность света, прошедшего второй поляроид. Так будет получен эллипс поляризации с вертикально ориентированной малой осью.

г) Для определения направления вращения светового вектора установите между поляроидами дополнительную пластинку $\lambda/4$ с горизонтально ориентированной «быстрой» осью. В этом случае свет на выходе из второй пластинки будет линейно поляризован. Если пластинки поодиночке дают эллипсы, вращающиеся в разные стороны, то поставленные друг за другом, они скомпенсируют разность фаз, и вектор $\boldsymbol{E}$ на выходе останется в первом и третьем квадрантах. Если же световой вектор перешёл в смежные квадранты, значит, эллипсы вращаются в одну сторону. А как вращается вектор $\boldsymbol{E}$ после пластинки $\lambda/4$, было определено в п. 8 а.

**Контрольные вопросы**

1. Покажите, что при выполнении условия Брюстера отражённый и преломлённый лучи взаимно перпендикулярны.

2. Как отличить свет с правой и с левой круговой поляризацией?

3. Неполяризованный свет проходит через двоякопреломляющую пластинку $\lambda/4$. Что можно сказать о поляризации света на выходе из пластинки?

4. Как отличить естественный свет от света, поляризованного по кругу, и от смеси естественного света со светом, поляризованным по кругу?

5. Объясните изменения интенсивности и цвета, наблюдаемые в опытах по интерференции поляризованных лучей.

6. Почему свет от вечернего неба поляризован?

**Литература**

1. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. XVI-XVIII, XXIII, XXVI, §§142–146.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. VII, 75-79.

3. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. II, гл. 11; ч. III, гл. 10.

4. Горелик Г.С. Колебания и волны. - М.: Физматгиз, 1959. Гл. X.

5* Поль Р.В. Оптика и атомная физика. - М.: Наука, 1966. Гл. X.

6. Дитчберн Р. Физическая оптика. - М.: Наука, 1965. Гл. XII.

## 22.7.4 Rotation of the Plane of Polarization

Цель работы: исследовать вращение плоскости поляризации света в оптически активных растворах и в магнитном поле.

В работе используются: сахариметр; магнитный поляриметр; автотрансформатор (ЛАТР); амперметр; милливеберметр; стеклянные трубки с раствором сахара и дистиллированной водой; образцы из стекла.

Оптически активные вещества. При прохождении плоскополяризованного света сквозь некоторые вещества обнаруживается, что плоскость, в которой колеблется электрический вектор, постепенно поворачивается. Такие вещества получили название оптически активных. К их числу принадлежат как твёрдые (например, кварц), так и жидкие тела (например, скипидар, раствор сахара).

Направление вращения плоскости поляризации принято устанавливать для наблюдателя, смотрящего навстречу световому пучку. Если плоскость поляризации поворачивается по часовой стрелке, то вещество называют правовращающим, или положительным, а в противном случае - левовращающим, или отрицательным.

Основы теории вращения плоскости поляризации были заложены Френелем (1817 г.). Он показал, что оптическая активность является особым типом двойного лучепреломления. В оптически активных веществах скорость распространения света различна для лучей, поляризованных по правому и левому кругу. Для правовращающих веществ

$$v_{\text{П}} > v, \quad n_{\text{П}} < n$$

а для левовращающих веществ

$$v_{\text{П}} < v, \quad n_{\text{П}} > n$$

где $v_{\text{п}}$ и $v_{\text{л}}$ - скорости распространения света с правой и левой циркулярной поляризацией, а $n_{\text{п}}$ и $n_{\text{л}}$ - соответствующие показатели преломления.

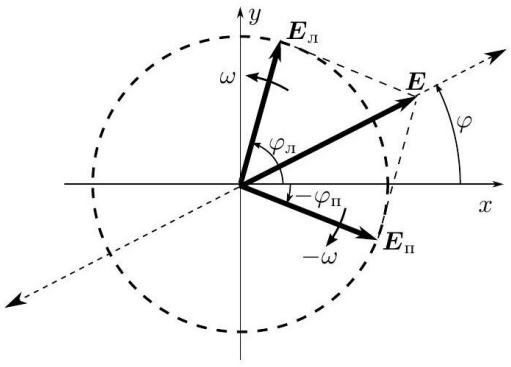



Рис. 1. Две циркулярно поляризованные волны (правая и левая) и их сумма - линейно поляризованная волна

Рассмотрим две монохроматические плоские волны одинаковой амплитуды, имеющие круговую поляризацию, и распространяющиеся в активной среде вдоль оси $z$. Пусть одна волна является левополяризованной, другая - правополяризованной. Векторы электрического поля этих волн лежат в плоскости $xy$ (см. рис. 1), а их угол поворота относительно оси $x$ совпадает с фазой волны в момент времени $t$ в данной точке оси $z$: в левополяризованной волне вектор $\boldsymbol{E}_{\text{л}}$ вращается против часовой стрелки (если смотреть навстречу волне), и угол поворота равен $\varphi_{\text{л}} = \omega t - k_{\text{л}} z$; и наоборот, в правополяризованной волне вектор $\boldsymbol{E}_{\text{п}}$ вращается по часовой стрелке, а угол поворота

равен $\varphi_{\text{п}} = -(\omega t - k_{\text{п}}\, z)$ ($k_{\text{л}}$ и $k_{\text{п}}$ - волновые числа лево- и правополяризованной волн в среде).

Поскольку $|\boldsymbol{E}_{\text{п}}| = |\boldsymbol{E}_{\text{л}}|$, суммарный вектор $\boldsymbol{E} = \boldsymbol{E}_{\text{п}} + \boldsymbol{E}_{\text{л}}$ направлен по биссектрисе, то есть под углом $\varphi = \frac{1}{2}(\varphi_{\text{п}} + \varphi_{\text{л}}) = \frac{1}{2}(k_{\text{п}} - k)\, z$ к оси $x$. Иными словами, фаза суммарной волны равна

$$\varphi = \frac{1}{2}(k_{\text{п}} - k_{\text{Л}})\, z \tag{1}$$

и не зависит от времени. Таким образом, при фиксированном $z$ суммарный вектор $\boldsymbol{E}$ не вращается - он, как нетрудно видеть, колеблется по гармоническому закону с циклической частотой $\omega$. Значит, сумма двух циркулярно поляризованных волн есть не что иное, как волна, поляризованная линейно. При этом формула (1) даёт зависимость угла поворота $\varphi$ плоскости её поляризации от координаты $z$.

Угол поворота плоскости поляризации в результате прохождения активной пластинки толщиной $\ell$ можно выразить через длину волны
света в пустоте $\lambda_0$ и показатели преломления среды $n_{\text{п}}$ и $n_{\text{л}}$. Из формулы (1) найдём

$$\varphi = \pi\,(n_{\text{п}} - n_{\text{Л}})\,\frac{\ell}{\lambda_0} \tag{2}$$

Если для данной среды $v_{\text{п}} > v_{\text{л}}$ и соответственно $n_{\text{п}} < n_{\text{л}}$, то угол $\varphi$ поворота плоскости поляризации отрицателен, т. е. вращение плоскости поляризации происходит вправо (по часовой стрелке), а для веществ, для которых $n_{\text{п}} > n_{\text{л}}$ ($v_{\text{п}} < v_{\text{л}}$), поворот происходит влево (против часовой стрелки).

Молекулярная оптика объясняет различие $n_{\text{п}}$ и $n_{\text{л}}$ асимметричным строением молекул (для активных жидкостей) или асимметричным строением кристаллов (в случае активных кристаллических тел). Такие молекулы (кристаллы) не имеют ни центра, ни плоскости симметрии. Все оптически активные вещества существуют в виде двух модификаций, обладающих соответственно правым и левым вращением. Эти две модификации являются зеркальными изображениями друг друга.

Угол $\varphi$ поворота плоскости поляризации прямо пропорционален толщине $\ell$ слоя раствора и концентрации активного вещества $C$:

$$\varphi = \alpha \ell C \tag{3}$$

Коэффициент пропорциональности $\alpha$ зависит от природы вещества и носит название постоянной вращения. Постоянная вращения очень слабо зависит от температуры и приблизительно обратно пропорциональна квадрату длины волны света.

Принято выражать $\alpha$ в градусах, $\ell$ в дециметрах и $C$ - в г /cm$^3$. В этом случае постоянная $\alpha$ называется удельным вращением и обозначается через $[\alpha]$.

Магнитное вращение плоскости поляризации. В 1846 г. Фарадей обнаружил, что плоскость поляризации света, проходящего сквозь вещества, не обладающие оптической активностью, испытывает вращение, если поместить эти вещества в магнитное поле, направленное вдоль светового луча. Угол $\varphi$ поворота плоскости поляризации пропорционален длине пути $\ell$ света в веществе и индукции магнитного поля $B$:

$$\varphi = \beta \ell B \tag{4}$$

Коэффициент пропорциональности $\beta$ носит название постоянной Верде. В таблицах значение $\beta$ обычно приводится в угловых минутах на сантиметр и гаусс.

Знак вращения определяется для наблюдателя, смотрящего по направлению магнитного поля. У большинства веществ плоскость поляризации поворачивается

вправо, - такие вещества называются положительными. В противном случае вещество называют отрицательным.

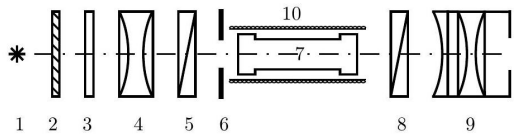



Рис. 2. Схема экспериментальной установки

Экспериментальная установка. Вращение плоскости поляризации света в предлагаемой работе изучается с помощью модифицированного технического прибора - сахариметра, оптическая схема которого изображена на рис. 2. Монохроматический свет от натриевой лампы 1 падает на матовое стекло 2, проходит через дополнительный светофильтр 3, который убирает все спектральные линии, кроме двойной жёлтой, проходит конденсорную систему 4, создающую параллельный пучок и поляризатор 5 (поляроидная плёнка). За поляризатором находится круглая диафрагма 6, наполовину перекрытая кварцевой пластинкой, являющейся пластинкой $\lambda/2$ для жёлтой линии натрия, поворачивающей плоскость поляризации света, прошедшего через неё, примерно на 50 градусов. Затем свет проходит через исследуемый образец (трубку с раствором) 7, которая находится внутри соленоида 10, по которому может течь постоянный ток, затем через анализатор 8 (поляроидная плёнка) и попадает в зрительную трубу, через которую ведётся визуальное наблюдение.

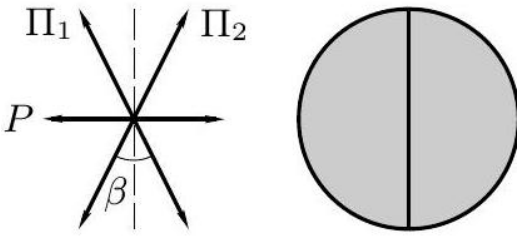



Рис. 3

При наблюдении в зрительную трубу видны два соприкасающихся поля зрения, разделённых чёткой границей. Если направление разрешённых колебаний анализатора расположено симметрично относительно плоскостей поляризации обеих половин пучка, то оба поля зрения имеют одинаковую освещённость.

Если при прохождении света через исследуемое вещество его плоскость поляризации повернётся, то фотометрическое равенство нарушится - одно поле окажется ярче другого. Равенство освещённостей
можно восстановить, поворачивая анализатор. Угол поворота анализатора определяется по лимбу с нониусом.

## ЗАДАНИЕ

Внимание! Угол поворота плоскости поляризации в установке невелик, поэтому важно проводить измерения с максимальной возможной точностью. Для этого необходимо правильно подобрать условия измерения и добиться хорошей адаптации глаза к ним. Следует проводить измерения в затемнённой комнате, чтобы чувствительность глаза к изменению освещённости поля зрения (контрастная чувствительность) была максимальной. Следует избегать засветки глаз яркими источниками света - это снижает контрастную чувствительность. Также необходимо выбрать режим измерения, не приводящий к чрезмерному утомлению глаз, что тоже снижает чувствительность и приводит к ухудшению воспроизводимости измерений. Время измерения ограничено также тем, что при протекании значительного тока в

соленоиде выделяется тепло, что может привести к его перегреву и выходу из строя. Поэтому при измерении постоянной Верде измерения необходимо проводить быстро, а ток должен протекать через соленоид только во время измерения.

**I. Исследование вращения плоскости поляризации в оптически активных растворах**

1. Включите натриевую лампу и дайте ей прогреться в течение $3-5$ минут.

2. Не помещая образец в соленоид, настройте зрительную трубу так чтобы увидеть резкое изображение входной диафрагмы сахариметра. Поворачивая анализатор, убедитесь, что поле зрения разделено на две половины, освещённость которых по-разному меняется при повороте анализатора, убедитесь, что изображение границы раздела полей зрения резкое. Добейтесь равенства освещённостей обеих половин поля зрения и снимите отсчёт по лимбу с нониусом с помощью отсчётной лупы. Проведите опыт несколько раз и убедитесь, что результаты измерения воспроизводятся с точностью не хуже примерно 0,1 градуса. (Точность будет наибольшей, если вы добьётесь исчезновения границы при минимальной освещённости обоих полей зрения.)

3. Поместите в соленоид исследуемый образец и измерьте угол поворота плоскости поляризации не менее 5-и раз. Убедитесь в хорошей воспроизводимости результатов измерений. За измеренное значение примите среднюю величину.

4. Повторите измерения для нескольких растворов глюкозы и фруктозы с известной концентрацией и определите концентрацию неизвестных растворов.

**II. Исследование магнитного вращения плоскости поляризации**

5. Поместите исследуемое вещество в поляриметр. Получите резкое изображение границы раздела полей зрения. Определите нулевую точку прибора, добившись исчезновения границы раздела полей зрения. Проведите не менее пяти независимых измерений и усредните результаты.

6. Отградуируйте соленоид. Меняя ток в катушке, снимите зависимость показаний милливеберметра от величины тока. Постройте график зависимости $B(I)$. На тех установках, где милливеберметр отсутствует, указан коэффициент пропорциональности между $I$ и $B$.

7. Измерьте угол $\varphi$ поворота плоскости поляризации при 5-6 значениях тока в катушке. Постройте график $\varphi(B)$. Измерьте длину трубки. По графику, зная длину трубки, определите постоянную Верде. Сравните результат с табличным.

**Контрольные вопросы**

1. В чём заключается отличие оптической активности от двойного лучепреломления?

2. Как объясняет теория Френеля поворот плоскости поляризации в оптически активных веществах?

3. Что называют удельным вращением?

4. Как работает поляризационная призма Николя?

5. Опишите принцип устройства и работы поляриметра.

6. При исследовании магнитного вращения плоскости поляризации иногда применяют устройства с многократным прохождением света в образце в направлении по и против магнитного поля. Объясните, почему в этом случае эффект поворота плоскости поляризации накапливается.

**Литература**

1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§94, 95.

2. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. XXX.

# 23 Fundamentals of Experimental Optics

(мб в теоретическую часть потом перенесу)

## 23.1 Optical Fibers (???)

Обсудим подробно оптоволокна
(мб будет еще отдельная запись, но скорее всего нет, потому что тут тоже много про них будет)

### 23.1.1 Overview of Applications of Optical Fibers

**Мотивация заниматься волоконной оптикой (??)**

(пропишу потом, прикладной все-таки раздел)

**Интернет в основом реализован оптоволокном**

То есть оно используется, вот.

**используются в лазерах**

Сварка, маркировка делается лазерами, созданной какой-то фотонической компанией.

### 23.1.2 Elements of Optical Fibers

**оптический усилитель**

они применяются где-то

**модуляторы**

хз

### 23.1.3 Typical Phenomena in Optical Fibers (??)

(потом это все отдельно посмотрю!)

**Об изменении поляризации при прохождении по волокну**

(в оптику, Лесовик (?) говорит, эффект большой)

**Об утечке моды из оптоволокна**

**Об изменении параметров света за счет сплющивания световода**

если геометрия меняется, может новые эффекты появятся?

**Уравнение волны в оптоволокне**

как она выглядит?

**О затухании световой волны в волноводе отсутствии усилителей**

насколько там она затухнет?
скорее это в оптику пропишу

## 23.2 Waveguides

(подробнее это скорее всего в электродинамике)

### 23.2.1 Propagation of Light in a Waveguide

(write basic theory about it, maybe it is written in electrodynamics)

### 23.2.2 Estimations of Typical Effects in Waveguides

(maybe I'll create a special note about them, there are a lot of different general questions...)

# 24 Improvement and Creation of Optical Equipment

## 24.1 Improvement of Optical Equipment

### 24.1.1 How to make a better cameras?

(find people who know it and ask, a section which will give a lot of money)

### 24.1.2 How to make a better glasses?

(find people who know it and ask, a section which will give a lot of money)

## 24.2 Creation of Optical Equipment

(I'll write later about some of them)

### 24.2.1 How to Create a Lamp (???)

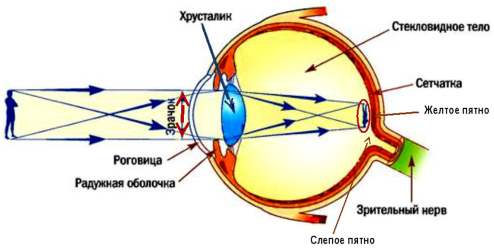
(find people who know it and ask, a section which will give a lot of money)

# 25 Vision and Eyes

## 25.1 How Eyes Work

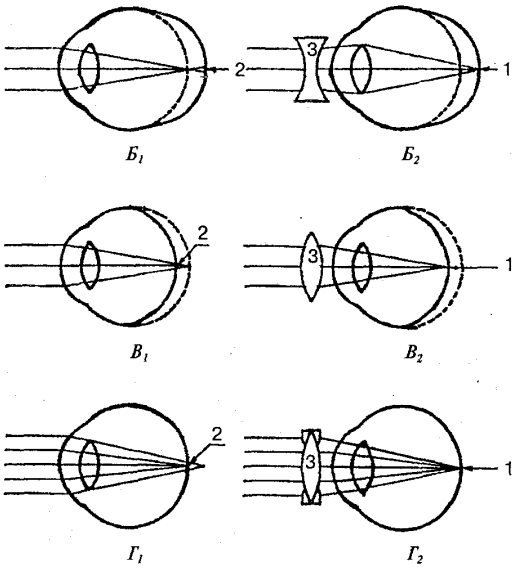
### 25.1.1 Basic Models

**Устройство человеческого глаза**




**О том, как другие создания видят мир**

**Элементарно об отклонениях в зрении**




**Можно ли видеть ушами?**

(хз на самом деле, честно, мб и да?)

**О коррекции зрения лазером (??)**

(?? почему такое вообще возможно.)

**Другие вопросы о зрении**

**Какие цвета документов лучше, темная тема или светлая?**

(Пока остановился на том, что лучше документы на светлой компилировать, потому что так привычнее. Вообще, мб потом на темную и перейду, пока тут просто рассуждения. Надо бы потом подобрать цвета для записей, какие вообще лучше (!??!??!))

(тут рассждения, потом подумаю про них, просто есть такой тоже момент в оптике)

Темная сторона не просто популярна у рыцарей-джедаев, страдающих манией величия, но и среди обычных пользователей Интернета.

Некоторым людям нравится атмосфера, в то время как другие используют темные темы, особенно потому, что они вроде как должны уменьшить нагрузку на глаза и сэкономить заряд батареи.

Это имеет интуитивный смысл: более яркие огни кажутся немного более грубыми для глаз, а белый – более высокоэнергетическим цветом, поэтому он, вероятно, использует больше батареи, верно?

Однако, как обычно, реальность немного сложнее.

Несмотря на то, что более темные цвета потребляют меньше энергии для генерации, не все типы экранов настроены для этого.

Насколько ваше зрение улучшится с темной темой, во многом зависит от ваших глаз и ситуации с окружающим освещением на вашем экране.

Темные темы экономят энергию? Это относительно просто: ваше устройство использует OLED или AMOLED?

Если это так, то темные темы могут спасти вас!

Если вместо этого в вашем устройстве используется ЖК-дисплей или экран другого типа, изменение цвета не сильно повлияет на время автономной работы.

Экраны AMOLED работают, пропуская электрические токи через органические соединения, чтобы осветить отдельные пиксели.

Однако, если пиксель черный, он просто останется выключенным, не получая никакой энергии для освещения.

Правда, черный (hex 000000) – единственный цвет, который отключит пиксель.

Любой другой цвет потребляет больше энергии, при этом белый цвет является самым дорогим для отображения с точки зрения потребления электроэнергии.

AMOLED обычно более энергоэффективен, чем ЖК-дисплей, но не в том вопросе, что касается белого фона.

Даже Google, чьи телефоны Pixel используют AMOLED, начал продвигать темные темы как способ экономии заряда аккумулятора на устройствах Android.

Они продемонстрировали снижение расхода батареи до 63% за счет темных тем на дисплеях AMOLED и добавили эту функцию в более новые версии Android.

С другой стороны, устройства с любой формой ЖК-экрана вообще не имеют влияния времени автономной работы на цвет.

Яркость, да; цвет, нет ЖК-дисплеи имеют несколько слоев, которые вносят вклад в конечное изображение, но частью уравнения является подсветка, состоящая из группы светодиодов.

Если ваш экран включен, каждый из этих светодиодов включается, потребляя одинаковое количество энергии независимо от цвета, на который переключается каждый отдельный пиксель.

Сами пиксели не потребляют никакой энергии на ЖК-экране.

Наука здесь достаточно ясна, что приятно.

AMOLED / OLED выигрывает от темных тем, а LCD – нет.

Тёмные темы лучше для глаз? Здесь все становится немного сложнее: потому что темные темы иногда лучше для ваших глаз, но не всегда.

С точки зрения читабельности вердикт ясен: черный текст на белом фоне – лучший вариант.

Это сводится к свойствам цвета, света и радужной оболочки человека.

Белый по существу отражает каждую длину волны в цветовом спектре, что означает, что нашим ирисам не нужно широко раскрываться, чтобы поглотить достаточно света,

оставляя их в их естественной форме.

Поскольку линза не деформируется более широкой диафрагмой, мы можем видеть вещи более резко, особенно высококонтрастный цвет, такой как черный, который фактически поглощает длины волн, а не отражает их.

Белый текст на черном фоне не работает так же здорово, так как наша радужная оболочка должна открываться больше, чтобы получить свет, деформируя линзу и заставляя ее выглядеть так, как будто белые буквы кровоточат на черном фоне (так называемый эффект «ореола»), Это особенно верно для 50

Разобрались с читабельностью, но как насчет усталости глаз?

Долгое время начинать с белого экрана, конечно, хуже.

Плюсы:

Темные темы могут снизить нагрузку на глаза в условиях низкой освещенности (в ночное время или в тусклом рабочем пространстве). Высокая контрастность текста и фона снижает нагрузку на глаза. Приложения, для которых требуется подсветка синтаксиса (например, программирование), легче увидеть с помощью темных тем.

Минусы:

Темные темы могут создать большую нагрузку на ваши глаза в условиях высокой освещенности, поскольку они размывают текст. 100% контраст (белый текст на черном фоне), как правило, труднее читать и может вызвать повышенную нагрузку на глаза. Приложениям, требующим чтения длинных отрывков текста, обычно будет труднее читать в белом тексте на темной теме.

Что это на самом деле означает?

В основном, темные темы наиболее полезны, когда вы находитесь в условиях низкой освещенности или вам не нужно читать большие куски текста.

Им также необходимо поддерживать относительно высокий уровень контрастности для удобства чтения, но переход с чистого белого на чистый черный, вероятно, не является хорошим выбором.

Если вы много читаете на своем экране, вам, вероятно, следует придерживаться темы темного освещения (возможно, светло-серого фона с черным текстом) и пытаться контролировать яркость экрана и ситуацию с окружающим освещением, а не менять цвета.

### 25.1.2 Proof that Light is Transferred to Electric Signal

(??? ask specialists, how to proof it???)

# 26 Other Devices

## 26.1 Screens and Displays (??)

(?? как они работают?? укажу в дополнения потом.)

## 26.2 Optoelectronics

(по куробатову всё)

## 26.3 Very Exotic Devices

(по куробатову всё)

### 26.3.1 Single photon counters

вот это интересно.

#### конструкция

потом разберусь.

#### применение в разных точках Вселенной

Володя там фигню какую-то говорил, соберу тут теорию для ответа ему.

### 26.3.2 single-pixel cameras

(see T. Tao, I have no idea what is it?)

## 26.4 Optical coherence tomography today: speed, contrast, and multimodality by Drexler, Liu, Kumar, Kamali, Unterhuber, Leitgeb

**Abstract**

In the last 25 years, optical coherence tomography (OCT) has advanced to be one of the most innovative and most successful translational optical imaging techniques, achieving substantial economic impact as well as clinical acceptance. This is largely owing to the resolution improvements by a factor of 10 to the submicron regime and to the imaging speed increase by more than half a million times to more than 5 million A-scans per second, with the latter one accomplished by the state-of-the-art swept source laser technologies that are reviewed in this article. In addition, parallelization of OCT detection, such as line-field and full-field OCT, has shortened the acquisition time even further by establishing quasi-akinetic scanning. Besides the technical improvements, several functional and contrast-enhancing OCT applications have been investigated, among which the label-free angiography shows great potential for future studies. Finally, various multimodal imaging modalities with OCT incorporated are reviewed, in that these multimodal implementations can synergistically compensate for the fundamental limitations of OCT when it is used alone. © The Authors. Published by SPIE under a Creative Commons Attribution 3.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: 10.1117/1.JBO.19.7.071412]

### 1 Introduction

Optical coherence tomography (OCT) is one of the most innovative and rapidly emerging optical imaging modalities in the last decades [1–4] because it is capable of perfectly noninvasively exploiting the wealth of morphologic and functional tissue information in the first few millimeters of organs. Since the beginning of OCT in the late 1980s and beginning of the 1990s, [5–7] more than 50 OCT companies have been created, more than a hundred research groups are involved in OCT, over a thousand OCT patents have been granted, and more than ten thousand research articles have been published-mostly in ophthalmology, followed by technology-related journals and cardiovascular publications (http://www.octnews.org/). In ophthalmic diagnosis especially, OCT is the fastest adopted imaging technology in the history of ophthalmology. While 108 million x-rays, 30 million sin-gle-photon emission computed tomographys (CT), positron emission tomographys, and CTs, and 26 million magnetic resonance (MR) imaging

examinations have been performed in 2010, $\sim$ 32 million ophthalmic OCT scans have already been conducted (http://www.octnews.org/). In >110 years of x-ray imaging development, the bionizing radiation dose has been reduced by 1500 times; imaging speed is 257,000 faster; its contrast significantly increased and the images are much less blurry. [8] In <20 years of OCT development, on the contrary, the axial resolution has been improved by > 10 times; the imaging speed is increased by more than half a million times; image contrast is greatly enhanced and many functional extensions of OCT have been developed.

The original idea of OCT was to enable noninvasive optical biopsy, i.e., the in situ imaging of tissue microstructure with a

resolution approaching that of histology without the need for tissue excision and processing. [9] Thus, we have the following prerequisites: (1) visualization of subcellular morphology (i.e., ultrahigh isotropic resolution), (2) three-dimensional imaging (resulting in the need for ultrahigh imaging speed), (3) sufficient (tissue) contrast, (4) molecular, biochemical sensitivity (i.e., sensitive to absorption of diagnostically relevant endogenous or exogenous chromophores), (5) localized (quantitative) functional tissue information, and (6) sufficient (diagnostically relevant) tissue penetration. Axial resolution improvement (prerequisite 1) has been the key technological milestone in the history of OCT in its first decade, [10–14] which has been achieved by using ultra-broadband sources. OCT imaging speed improvements (prerequisite 2 ) evolved in its second decade, which were accomplished by Fourier domain (FD)/spectral domain (SD) [15–18] and swept source (SS) OCT. [19–23] While in FD/SD OCT speed is mainly determined by the readout time of the camera in a spectrometer, the wavelength-tuning speed of swept sources is the decisive factor in SS OCT. In the past 5 to 10 years, different swept source technologies have emerged to significantly improve imaging speed in (commercial) OCT systemsespecially the ones using wavelengths $> 1\mu m$. In the present paper, the state-of-the-art (ultra)high-speed SS technologies are reviewed and compared (cf. Sec. 2.) and an outlook of even further improving OCT imaging speed via parallelization (line- and full-field OCT) of OCT detection is presented (cf. Sec. 3). OCT contrast enhancement (prerequisite 3) has been accomplished by introducing polarization-sensitive OCT, [24–28] phasesensitive OCT, [29–33] optical coherence elastography, [34–39] spectroscopic low coherence interferometry, [40–43] elastic scattering spectroscopy, [31,44–46] and nonlinear interferometric vibrational imaging (NIVI), as well as employing endogenous or exogenous contrast. [47–52] In this paper, a recently developed contrast
improvement for OCT, named label-free optical angiography, is reviewed (cf. Sec. 4). Some of the above-mentioned contrast enhancement OCT extensions are targeted to enable OCT molecular and biochemical sensitivity (prerequisite 4) as well as localized (quantitative) functional tissue information (prerequisite 5), with the most successful one being Doppler OCT. Though having been improved in many aspects, OCT so far is still not sensitive enough to provide absorption contrast or biochemical and molecular information. Therefore, multimodal optical imaging modalities with OCT incorporated have become a hot research topic. In this paper, the following combinations are reviewed: OCT with multiphoton tomography (MPT, for subcellular resolution); OCT with nonlinear microscopy (for label-free molecular tissue information, cf. Sec. 5.1); and OCT with photoacoustic imaging (for enhanced absorption sensitivity and penetration depth-prerequisite 6, cf. Sec. 5.2). Previous reviews of OCT focused on advanced screening in the fields of primary care, [53] microscopy, [54] rapid tissue screening, [55] phantoms, [56] high speed, [57] and translational research, [58] while recent application-specific OCT reviews covered ophthalmology, [59–61] cardiology, [62–66] dermatology, [67–70] novel applications in pulmonary medicine, [71] cancer, [72] and optical coherence elastography, [73] as well as OCT post- and signal processing. [74,75] This review focuses on high-speed technology (SS light technology and parallelization of OCT detection), label-free angiography, and multimodal OCT.

[0]*Address all correspondence to: Wolfgang Drexler, E-mail: Wolfgang.Drexler@ meduniwien.ac.at

## 6 Summary and Outlook

Although ultrahigh (axial and isotropic) resolution and ultrahigh speed for OCT have been demonstrated on a research level, there is still some time to wait for their commercializationespecially for the latter one. Particularly for the wavelength region above $1\mu$ m, it seems that SS technology has significant potential for enabling OCT imaging speeds beyond 400kHz− up to 1 MHz . Therefore, seven different state-of-the-art SS technologies have been reviewed in this article. A recently developed akinetic, all-electronic swept source promises the highest flexibility and unprecedented performance among its peers along with the features of being cost-effective and compact. The optically pumped VCSEL-based swept source offers by far the longest coherence length and fastest sweep range of commercial sources. To make it commercially attractive, though, electrically pumped VCSELs with identical specifications as the optically pumped ones are a necessity. FDML technology is unique for ultrahigh speed and high optical output power OCT research applications-its widespread commercialization is still unclear at this point in time. There is, at the moment, only one company offering MEMS-based swept sources in five wavelength regions with very competitive specifications. Parallelization of OCT detection based on LF and FF

OCT-preferably swept source based-or digital holoscopy might increase OCT scanning speed even further in an efficient, quasi-akinetic (scanning free) manner. Camera technology (line and area), especially in the wavelength region above $1\mu$ m, has to improve (especially in terms of speed and noise by maintaining sufficient pixels) and become more cost-effective to commercialize these technologies, though.

Although numerous functional and contrast-enhanced extensions of OCT have been developed, they have had limited clinical impact so far as compared to pure morphological OCT systems. This is related to the fact that functional tissue information is inherently more challenging to quantify and reproduce. One recently emerging, contrast-enhancing techniquealso referred to as phase variance, speckle correlation based label-free optical coherence angiographyhas great potential for providing additional specific diagnostic information. Already highly established in ophthalmology, this technique is being extensively investigated, at the moment, in other clinical applications, such as in dermatology, neurology, and cancer research, in general.

Multiscale, multiresolution imaging promises to be the future of biomedical and clinical diagnosis, enabling multiscale 3-D morphologic and functional imaging at different resolutions in order to accomplish morphofunctional performance, i.e., the fusion of anatomic and functional imaging at the cellular or even molecular resolution level. Optical imaging, on the other hand, has the potential to enable low-cost, noninvasive, accurate, rapid alternatives in order to address global medical needs, addressing noninvasively to minimally invasive optical imaging from molecular to (sub)cellular to individual organs enabling point-of-care diagnosis. Having already gained broad acceptance in ophthalmic diagnosis, OCT is being comprehensively investigated in other clinical fields, too, such as in cardiology, dermatology, and gastrointestinal as well as oncologic imaging, in general. However, despite the various efforts to improve OCT's performances, its scattering-based contrast mechanism has limited OCT in certain respects, e.g., visualization of subcellular tissue morphology and extraction of biochemical and, in general, molecular tissue information due to its lack of sensitivity to absorption. Therefore, incorporating OCT with other imaging modalities into a hybrid platform has become a hot topic in optical imaging with the ultimate challenging goal to provide easy-to-use, cost-effective optical technologies with higher sensitivity and specificity and to enable a step-change in diagnostics for a range of diseases. Among them, hybrid OCT/MPT, OCT/nonlinear microscopy/spectroscopy, and OCT/photoacoustic imaging are under vigorous investigation, with each of them bearing unique complementary features and potential applications. Unfortunately-especially for nonlinear optical microscopy and photoacoustic imaging-light sources are still based on expensive, large footprint, not extremely reliable (as compared to OCT light sources) technol-

ogy that can mainly be used and maintained by high-level specialists. Hence, core photonics technology for optical imaging still has to be developed focusing on a reduction in size, reliability, ease-of-operation, and cost-effectiveness as well as the capability to supply more than one imaging modality. Cost reduction would facilitate widespread application throughout the healthcare system, thereby enabling earlier diagnosis, which, in turn, means earlier treatment in the disease process and an improved prognosis.

Other application-oriented hurdles that have to be overcome with multimodal OCT imaging include the inherent limitation of (nonlinear optical) microscopic techniques suffering from small FOVs providing only tunnel vision-in terms of clinical diagnosis. In addition to time-intensive postprocessing-based mosaicking and stitching of smaller FOVs resulting in a sufficiently large one, prescreening with macroscopic techniques (OCT) over a large FOV with subsequent zooming in at the subcellular/biochemical (with, e.g., MPM) or even label-free molecular-level nonlinear optical microscopy (Raman, CARS, or stimulated Raman scattering) on the suspicious areas seems to be a clinically viable way. The clinical impact of nonlinear optical microscopy is still under scientific investigation, although dermatologic and, recently, endoscopic diagnosis, lipid detection for cancer diagnosis, pharmacokinetic applications, as well as improved detection of tumor borders show significant clinical promise.

Hybrid photoacoustic imaging and OCT are an exquisite example of combining two modalities with complementary, synergistic visualization performances: OCT being based on scattering depicts tissue morphology and photoacoustic imaging being sensitive to absorption reveals endogenous or exogenous chromophore-based contrast. In addition, the latter one offers the ability to quantify chromophore concentrations, e.g., oxy-
and deoxy-hemoglobin and, therefore, oxygenation or melanin concentration. Hence, oncologic applications investigating angiogenetic factors as well as a better understanding of tumorgenesis seem to be promising applications.

Combining OCT with complementary optical imaging modalities offers higher sensitivity with modalities enabling high specificity (best of both worlds) and, therefore, fulfilling diagnostic requirements. Furthermore, critical parameters (imaging speed, penetration depth, resolution, as well as biochemical/molecular sensitivity) are significantly improved, which is essential for the successful introduction of a new diagnostic standard.

## 26.4.1 2 Swept Source Lasers for OCT-Light Source Technology Enhancing Imaging Speed

Despite the boom of new SS technologies in recent years, it is noteworthy that none of them is at the moment even close in maturity and reliability to that of the superluminescent laser diodes (SLDs) that have been already employed in (commercial) OCT systems. Nevertheless, swept sources seem to be the OCT technology of the future since they have the potential to offer significantly higher scanning speeds, extended depth range with significantly reduced sensitivity roll-off, reduced fringe washout from sample motion or rapid transverse scanning, and improved light detection efficiency due to dual balanced detection. SS OCT avoids the need for spectrometers with line scan cameras, but requires a high-speed, narrow line width swept source. Spectrometers have limited spectral resolution due to the grating resolving power, beam spot size, and finite pixel dimensions of the line scan cameras. Narrow line width swept source enables spectral resolution in SS OCT, which can be much higher than that of SD OCT. For OCT applications using wavelengths $> 1\mu m$, SS OCT seems to be the most promising technology of choice because indium gallium arsenide (InGaAs) line cameras are limited in speed and number of pixels, and are more expensive than silicon line cameras. SD OCT suffers from significantly more SNR roll-off with scanning depth than that achievable by using narrow line width swept sources. The seven most important swept laser technologies (from six

companies; cf. Table 1) are reviewed and compared in terms of their potential for (commercial) OCT systems. Only those technologies that have been commercialized or are on the verge of commercialization have been reviewed. Novel emerging scientific approaches for swept sources are not considered in this paper.

### 2.1 Axsun Technologies Inc., Massachusetts ("a Volcano Company")

The laser engine consists of a swept laser module, control electronics, $k$-clock, balanced receiver, and data acquisition board, which samples on k-clock transitions (information in this section is from www.axsun.com). More specifically, it contains a reflective microelectromechanical system (MEMS) tunable FabryPerot filter, a broadband gain chip, and a fiber reflector that forms the other end of the laser cavity and serves as the output coupler. Filter tuning is accomplished by changing the drive voltage on the MEMS filter. Fiber extension brings the equivalent air length of the cavity to 104 mm such that there are a handful of laser cavity modes underneath the filter at all times. The fact that this external cavity laser operates with a cluster of modes, rather than a single mode, leads to a coherence length that is an order of magnitude or more smaller than single-mode semiconductor lasers. Multimode operation can also increase relative intensity noise (RIN) though. The $100-$kHz laser operates with two pulses traveling in the cavity at once separated by half the cavity roundtrip time. With longer path mismatches, pulses can interfere with their neighbors, leading to the coherence revival phenomenon. This external cavity laser uses a digital $k$-clock, which is derived from a fiber-based Mach-Zehnder interferometer. Main envisaged innovations include doubling the scanning speed to 200 kHz and increasing coherence length by adjusting the filter bandwidth.

### 2.2 Hamamatsu Photonics K.K. (NTT-Advance Technology Corp.), Tokyo, Japan

A high-speed KTa1-xNbxO3 (KTN) light deflector is used in an external-cavity laser (Littman-Metcalf cavity configuration) that has no moving parts (mechanical free high-speed operation) (information in this section is from www.ntt-at.com). It deflects light using an electro-optic effect. KTN has a very large electrooptic effect ( $20\times$ higher than lithium niobate), which changes a refractive index by an applied voltage and bends the path of a light beam in a direction. The KTN crystal is precharged by applying $\pm 500$ V dc for 10 s and then scans the laser wavelength by applying a $\pm 400$ V sinusoidal voltage to the crystal. However, KTN simultaneously exhibits the characteristics of a cylindrical convex lens because of the trapped electrons. This convex lens is compensated with a cylindrical concave lens because the lens power inside the cavity degrades the instantaneous line width of the laser. At the moment, the scanning wavelength range is $\sim 80$ nm with 20 mW output power and up to 200 kHz (sinusoidal) sweep speed with a $\pm 400$ V deflector driving voltage at 1310 nm . The duty ratio changes from 50:50 to 70:30-at the moment, 200 kHz drive signal with one-way imaging-enabling a 50% duty cycle. Main envisaged innovations include 100 nm sweep width, 20 mW optical output power at 1310 nm with 200 kHz sweep rate and 8 mm coherence length-equivalent to 4 mm scanning depth ( -6 dB signal roll-off), standard deviation of timing jitters between adjacent A-lines: $< 200$ps, which corresponds to a phase difference of $< 0.05$rad at a path difference of 1.5 mm of a Michelson interferometer.

### 2.3 Insight Photonics Solutions Inc., Lafayette, Colorado

Sweep flexibility of the akinetic laser is purely software-driven. The all-semiconductor laser's distributed Bragg reflector-

Table 1 Comparison of state-of-the-art swept source technology.

| | |
|---|---|
| Device/system | Sweep speed |
| Axsun optical coherence tomography (OCT) engine | 100 kHz |
| NTT-AT KTN based, mechanical free swept source | 200 kHz |
| Insight akinetic swept source | 4 to 400 kHz |
| Santec MEMSbased swept source | 50kHz/100kH |
| Thorlabs MEMSVCSEL swept source | 50 to 580 kHz |
| Exalos external cavity laser with resonant MEMSbased onedimensional scanning mirror | 40 to 150 kHz |
| FDML laser | 3200 kHz at 1 |

Note: KTN, KTa1-xNbxO3; FDML, Fourier domain mode locked; MEMS, microelectromechanical systems.

like structure increases the finesse of the cavity by a factor of approximately nine times, resulting in a narrow line width (information in this section is from www.sweptlaser.com). The small dimensions of the laser cavity and the fact that the entire cavity is on a single rigid structure substantially reduce cavity length variation. Superior line width and spectral performance are natural outcomes of the very short waveguide design. Experimental data show that moving from any wavelength to an adjacent wavelength with the all-semiconductor laser takes $\sim 2.5$ ns. Because the duty cycle is software-controlled, it is adjustable and, hence, can be set to almost any unidirectional

value from 5 to 95%. The all-semiconductor laser is inherently linear and does not require an external optical $k$-clock. The laser self-generates an internal electronic $k$-clock. The laser forces the wavelength to be correct at each of the evenly spaced clock transitions. The all-semiconductor laser does not need the extra cost of parts and integration of an optical $k$-clock, and eliminates the challenges frequently associated with the nonuniform triggering that can occur with external optical $k$-clocks. The result is direct triggering of the data acquisition. No postacquisition resampling is needed, avoiding ghost images and reducing computation time. The akinetic laser has a typical RMS linearity of 0.0012% at

200,000 sweeps per second ( $< \pm 1\text{pm}; \pm 0.2\text{GHz}$ ) and a linearity span that is typically $< \pm 0.002\% (< \pm 2\text{pm})$. Wavelength repeatability of 0.5 pm (standard deviation) and $\pm 2.5$pm span (peak amplitude) have been confirmed for the all-semiconductor laser. In air, a wavelength repeatability of 0.5 pm produces a phase error of 0.5 mrad at 0.2 mm ( 5 mrad at 2 mm ). Coherence revival occurs if a laser is simultaneously oscillating at multiple longitudinal modes of the cavity. Because the akinetic swept laser has a single longitudinal mode, the issues with coherence revival in mechanically tuned lasers are largely avoided. Output power in the all-semiconductor akinetic laser is directly controlled on the chip and can be selected from software. Over the life of a laser, the center wavelength will only vary by $\sim$ 80pm based on life-test measurements taken on similar laser devices. The nonactive coherence length of the all-semiconductor laser is in the range of several tens of meters. This is substantially higher coherence than the $\sim$ 200 mm of currently available all-semiconductor lasers. Main envisaged innovations are user-adjusted coherence length and a software platform allowing detailed control over the laser's behavior. In lab experiments, all-semiconductor lasers have been demonstrated at 1 million sweeps per second. In the long run, it is believed that the akinetic technology can be pushed to roughly 2 million sweeps per second. With an akinetic laser, other power profiles may be tried by the user, under program control, for application-specific optimization without any other changes to the system. The same technology (with a modified substrate) will be employed for 1060 nm . The electronic drive circuitry is identical for 1310, 1550, and 1060 nm .

It is also within the bounds of the technology to work with applications at 850 and > 1640 nm. The all-semiconductor technology will allow the production of lasers with wavelength coverage of ∼ 140 nm and wider in the future. The wafer-scale laser manufacturing technology used for the all-semiconductor laser also grants cost-effectiveness to this product. Use of high-volume telecom supplies for the internal electronics may further reduce the costs for this type of laser.

### 2.4 Santec Corporation, Komaki, Japan

Santec is one of the first companies to launch a swept source especially optimized for OCT imaging: an external cavity laser (ECL) with a polygon mirror and a diffraction grating, resulting in high-performance sources with a desirable linear sweep behavior in $k$ (frequency domain), but suffering from insufficient scaling toward higher speed due to the bulkiness of the polygon mirror, which, in combination with the high price, prevents widespread cost-effective use (information in this section is from www.santec.com). ECLs with a polygon mirror are available at 1300 nm(20kHz/ > 170 nm/ > 20 mW; 50kHz/140 nm/15 to 30 mW; 1060 and 1310 nm/60 mm double pass coherence length/ 1 to 3kHz/6 to 8 mW; 0.15% linearity/ 1310 nm/20kHz/8 mm coherence length > 20 mW ) and at 1050 nm ( 30 to 50kHz/70 to 120 nm/ > 12 mW/6 mm double pass coherence length). These sources will remain research/academic light sources due to aforementioned reasons. Santec also offers MEMS-based swept sources-at least at 1310 nm so far: 100kHz/ > 20 mm double pass coherence length (−6 dB)/105 nm sweep range/ > 50 W optical output power, which is also available as an OEM product. This 1310 nm MEMS-based swept source is very similar to the Axsun 1310 nm swept source and is, therefore, considered a proper back-up by some OCT companies. Main envisaged innovations include an MEMS-based swept source at 1050 nm .

### 2.5 Thorlabs Inc., Newton, New Jersey

These lasers are the first (in 2012) commercially available MEMS vertical-cavity surface-emitting laser (MEMS-VCSEL) swept sources at 1050 nm ( 100 nm sweep range at -10 dB ) and 1310 nm ( 150 nm at -10 dB sweep range) (information in this section is from www.thorlabs.us). The current tuning range of 150 nm at 1310 nm is the largest reported for any MEMSVCSEL and is comparable to the 160 nm reported for the FD mode-locked (FDML) laser. First demonstrations of coherence lengths > 100 mm and axial scan rates up to 1.2 MHz have been accomplished. Ultrahigh speed, high-resolution imaging (up to 580 kHz ), high-speed, long-depth range imaging ( 100 kHz ), and ultra long-range imaging ( 50 kHz ) have been demonstrated. A further advantage of the MEMS-VCSEL is wavelength flexibility. Adaptation of the same technology to other wavelengths from 450 to 2300 nm appears to be feasible. The VCSEL is optically pumped at 980 nm (for 1310 nm ) or 850 nm (for 1050 nm ) through a top dielectric mirror, generating tunable 1310 nm emission, which emerges from the top mirror and is fiber-coupled and amplified by a low-noise semiconductor optical amplifier. Efforts to date have focused on optically, rather than electrically, pumped devices-this makes the source more complex and expensive. Though the ultimate low-cost device will be electrically pumped, optical pumping provides a number of performance advantages over electrical pumping, including larger tuning range and better spectral purity. However, MEMS-VCSEL, like other MEMS-based tunable lasers, can be linearized through drive waveform preshaping. Linearized drive waveforms at 200 kHz axial scan rates have also been demonstrated. The duty cycle is > 90% with symmetric forward and backward sweeps. Unidirectional frequency sweep of this source results in a 400 kHz A -scan rate with a duty cycle of 50% at 1050 nm . The system phase stability, defined as the standard deviation of the phase differences between sequential A-scans, measured from a mirror in the patient interface at a signal-to-noise ratio of 57.5 dB (ratio of peak to mean noise) and a depth of 0.3 mm was 1.5 mrad , which approaches the theoretical

limit. Main envisaged innovations include MEMS actuator designs to increase mechanical resonance frequency to support higher-frequency drive, development of electrical instead of optical pumping for VCSELs for lower cost and miniaturization (since it would eliminate the pump, wavelength division multiplexer coupler, and isolator), and higher bandwidth and optical power at 1050 nm with improved booster amplifier. A further increase of the tuning range to 200 nm may be possible by increasing the number of quantum wells in the gain region, using a wider bandwidth top suspended mirror, and by further increasing the free spectral range. In addition, it should also be possible to multiplex two or more VCSELs with offset bandwidths in order to obtain increased sweep ranges and improve axial resolutions.

### 2.6 Exalos AG, Schlieren, Switzerland (Exalos Inc., Pennsylvania)

These lasers are ECLs with a resonant MEMS-based onedimensional (1-D) scanning mirror and a diffraction grating, resulting in fast tunable sources with demonstrated sweep frequencies up to 200 kHz (information in this section is from www.exalos.com). Due to the use of high-Q resonant MEMS scanners, those sources perform a quasi-sinusoidal sweep in $k$, which requires linearization and signal postprocessing.

However, the sweep behavior is highly deterministic and provides ultrastable long-term phase stability. The laser architecture itself is similar to polygon mirrors, flexible in wavelength, and allows for realizing swept sources from the visible to the nearinfrared (NIR). Due to the miniaturized filter architecture, this approach allows for compact, cost-effective swept sources. At the moment, the company is the only provider that offers swept sources between 400 and 1700 nm with sweep rates from 2 to 200 kHz in a miniaturized optical butterfly package that also allows the integration of optical reference ( $k$-clock) interferometers or other optical reference filters for spectral calibration. 10dB sweep ranges as wide as 80 nm at 840 nm, 120 nm at 1060 nm, 100 nm at 1220 nm, 150 nm at 1300 nm , and 200 nm at 1550 nm , respectively, have been achieved. Clean imaging performance results in no secondary coherence peaks as well as sharp and narrow point spread function peaks without side lobes throughout the imaging. Main envisaged innovations include next-generation ultrahigh-speed 1-D MEMS scanners with mechanical resonance frequencies in the range of 75 to 150 kHz , resulting in 150 to 300 kHz swept sources. Ultrabroadband semiconductor optical gain chips at 840,1050 , and 1310 nm enable ultrawide optical sweep ranges and hence unprecedented axial resolution.

### 2.7 Fourier Domain Mode-Locked Laser

So far, the main impact of FDML lasers has been the demonstration of OCT systems with dramatically higher imaging speed. [21,22,76–78] The first versions have pushed the speed from several 10 kHz line rate, which have been standard for the first FD OCT systems, to several 100 kHz , and later on, FDML lasers have helped to break the barrier of a 1 MHz line rate with swept sources. Besides the higher imaging speed, FDML lasers have been proven useful for many different applications, where good phase stability, long coherence, low laser noise, or similar conditions are required. Despite these many initial applications, only a few applied or clinical studies using FDML have been published. This is partly caused by the difficulty to build proper OCT systems that can handle the high imaging speed and the fast generated huge data sets by these multi-megahertz (MHz) OCT systems. It is interesting to see how the high-speed FDML results have triggered vibrant research efforts to realize nonFDML sources, which can achieve similar performance. It can be expected that the availability of more than one swept laser technology for MHz OCT will spur research on applications to find out where MHz OCT imaging speeds are required. FDML lasers can achieve ultrahigh sweep rates of up to 5.2 MHz by buffering or multiplexing the sweeps. FDML works optimally at 1.3 and 1.5$\mu$ m wavelengths where optical fiber dispersion and loss are negligible. However, dispersion

can be compensated for using fiber Bragg gratings to improve performance at 1 and 1.3$\mu$ m wavelengths. The fact that the laser light is seeded from the last round-trip and that FDML is a real stationary laser-operating regime reduces the RIN of the laser. The good saturation of the laser gain medium enables very high output powers of 100 mW and more.

Necessary swept source technological specifications for future envisaged (commercial) OCT systems in the coming years would be at least 200 to 400 kHz (with preferred unidirectional sweep), with a 100 to 150 nm sweep range (ideally not at -10 dB but as flattop in linear scale and with good optical flatness), optical output power $>$ 10 mW, coherence length of 100 to 150 mm , duty cycle ( $>$ 90% ), high phase stability,
linear-in-frequency sweep (picometer range; tenths of gigahertz), and low RIN noise ( $\sim -150$ dB/Hz ). At the moment, the perfect swept source for (commercial) OCT in terms of optimum specs, size, price, robustness, and future potential does not exist.

The akinetic, all-semiconductor Insight swept source is extremely easy to use, very flexible, compact, suitable for a miniaturized OCT system (OCT on a chip), and has the potential to be very economical with unprecedented performance. However, it has only been partially explored from an OCT imaging point of view. The VCSEL MEMS-based swept source is no doubt the most developed long-coherence high-speed swept source at the moment. Despite its momentary impressive performance, the short micron scale length of the VCSEL cavity will probably be a risk of generating artifacts due to slight misalignments or long-term instabilities. Furthermore, electrical pumping would be preferred to reduce complexity, costs, and size. Exalos is currently the only provider for swept sources at all five wavelength regions, also including 850 nm . Most of the swept source technology companies entered the market only quite recently. Therefore, an absolutely fair judgment regarding lifetime, stability, and other specifications is challenging. In general, nonmechanical/akinetic swept sources should be a better choice due to the proneness to failure of the mechanical scanning mechanism. The NNT-AT KTN-based swept source is nonmechanical, but seems to be significantly harder to operate and does not promise flexibility and performance features like the akinetic all-semiconductor, akinetic Insight swept source.

## 26.4.2 3 Parallelization of Optical Coherence Tomography-Pushing Imaging Speed Limits

During the last two decades, OCT has seen dramatic improvements in resolution, sensitivity, and speed. For in vivo imaging, resolution and speed are strongly linked, since motion blurring degrades not only resolution, but also sensitivity. The faster the acquisition, the shorter is the measurement time, which is, in particular, important for human in vivo imaging. The speed performance improvement is impressive: first implementations of OCT systems achieved a few depth scans per second. Nowadays, based on rapidly tuning SS technology, even several millions of such A-scans have been demonstrated. This allows, on one hand, dense sampling of large tissue patches, allowing for a comprehensive insight into tissue morphology, and, on the other hand, new functional imaging modalities profited from an enhanced flexibility. Nevertheless, even ultrafast systems ultimately strike the physical boundaries given by the detection process. Especially for in vivo imaging, the applied optical power is limited by laser safety regulations. Increasing, therefore, the speed, i.e., reducing the detection time, without being able to adjust the applied power, reduces the sensitivity, and degrades contrast and image quality. It is, therefore, questionable if systems much faster than 100 to 400 kHz will be used for clinical OCT applications.

A natural way out of this dilemma is to parallelize the detection. The theoretical advantage of full-field parallel OCT (FF OCT) regarding its intrinsic better sensitivity and potentially higher speed has been recognized already with early time domain OCT (TD OCT). Sensitivity in the shot noise limit scales proportional with the number of detected photons backscattered

from the sample. It increases, therefore, with sensor recording time and with optical power. Flying-spot OCT covers
each sampling point for a shorter time as the speed is increased. Parallel recording does not suffer from this limitation. In fact, if the time needed for a flying-spot system to sample a tomogram of, let us say, $N$ lateral points is $T$, then each scan has a recording time of only $T/N$. In parallel OCT, in contrast, the full time $T$ is available for recording at all parallel points. This compensates for the spread of power across the parallel detection channels. The advantage in sensitivity for a parallel system comes now from the possibility to apply more power while still complying with laser safety regulations, since a larger patch of the sample is illuminated. Although feasibility of parallelization of OCT for in vivo imaging has been demonstrated, there are some serious challenges that need to be resolved in the future to make it suitable for practical clinical applications. In the following, we will review developments in FF and line-field (LF) OCT and discuss current limitations and possible solutions for establishing parallel OCT and exploiting its potential advantages.

### 3.1 Full-Field Time Domain Optical Coherence Microscopy/Tomography

Conventional TD OCT uses a low-coherence light source to achieve micrometer axial resolution. The light is focused on the sample using a low numerical aperture (NA) microscope objective (MO) to keep the confocal gate long enough to cover the full imaging range in the sample. This avoids the need to move the sample in the $Z$-direction. Instead, the sample at a given lateral point is scanned in depth by modulating the optical path length of the reference arm at a high speed. By scanning the sample in the lateral direction, in vivo B-scans ( $X-Z$ cross-section) with high axial resolution can be obtained. [79] But, the use of a low NA sacrifices the lateral resolution, which is typically on the order of $\sim 10$ to $30\mu$ m. Beaurepaire et al. [80] demonstrated that by simultaneously irradiating the FF of view on the sample with a low-coherence light source and detecting the backscattered light using a high-NA MO and a two-dimensional (2-D) charge-coupled device (CCD) array in a Michelson interferometric setup, a 2-D en face image ( $X-Y$ cross-section) with high lateral resolution can be obtained. A lateral resolution of two was achieved using an MO of NA = 0.25. The axial resolution was $8\mu$ m, limited by the low-coherence light-emitting diode light source with a center wavelength at $\lambda_0 = 840$ nm and bandwidth $\sim \Delta\lambda = 31$ nm (output power = 20 mW ). As the confocal gating was narrow due to the MO with a high NA, the sample was scanned in depth to several hundred microns by moving in the sample in the $Z$-direction to create a three-dimensional (3-D) volume image. The magnitude of the interferometric signal from each pixel of the CCD array was extracted in parallel using a lock-in detection technique. [80,81] A combination of linearly polarized illumination and modulation of detected interferometric data using a photo-elastic birefringence modulator and an analyzer ( 45 deg ) were used to create phase-shifted data. [80] A linear combination of four phase-shifted data (phase shifted by $0, 90, 180$, and 270 deg) is used to produce a signal at a given pixel, proportional to $A\sin(2\pi\delta/\lambda)$ and $A\cos(2\pi\delta/\lambda)$, where $\lambda$ is the center wavelength of the illumination, $A$ is the amplitude of the coherent backscattered light, and $\delta$ is the optical path length difference between the light from the reference mirror and backscatters in the corresponding sample voxel. [80] The CCD camera (Dalsa, CA-D1, pixels) offered a frame rate of 200 Hz , but the frame grabber (IC-PCI, Imaging Technology Inc.) limited the frame rate to 50 Hz . Sixty-four quadruplet images were
acquired to produce an en face image [ $500\times500\mu m^2$ field of view (FOV)] in 5 s with a reported sensitivity of $\sim 100$ dB ( $300\mu$ W power on sample). Images of plant cells, such as onion cells, were shown up to the depth of $\sim 200$ to $350\mu$ m. Dubois et al. [82] further improved the resolution of the system by modifying the system into a Michelson interferometer with a Linnik type configuration with MOs of the same NA in both the reference and the sample arms. The same light source, signal modulation, and detection system were used as in Ref. 80. A higher lateral

resolution of 0.7$\mu$ m and axial resolution of 2.8$\mu$ m were reported using the MO with a high NA of 0.5 . A depth penetration of only up to 150$\mu$ m was achieved. The lateral resolution degradation with depth was attributed to the increase in multiple scattering of light with depth in the sample. The measured sensitivity of the system was reported to be $\sim$ 82 dB. En face images of onion cells with fields of view of $160 \times 160\mu m^2$ and $370 \times 370\mu m^2$ were acquired in 0.5 and 1 s , respectively. However, the speed was not enough to image a moving sample, and thus, the imaging was restricted to ex vivo samples. Grieve et al. [83] demonstrated 3-D ocular tissue imaging with cellular-level resolution using an ultrahigh-resolution FF TD OCT setup. The system was based on a Michelson interferometer with a Linnik type configuration. A tungsten halogen lamp was used as a light source with a center wavelength of $\lambda_o = 770$ nm and a bandwidth of $\Delta\lambda = 350$ nm. The low coherence of the light source provided a very high axial resolution of 0.7$\mu$ m. A high lateral resolution of 0.9$\mu$ m was achieved using an MO with NA = 0.5. A piezoelectric transducer (PZT) in the reference arm was used to modulate the optical path length in the reference arm to produce the phase-shifted images. An en face image ( $500 \times 500$ pixels) was acquired in 1 s . However, again due to the limited imaging speed, only ex vivo ocular tissue samples (cornea, lens, retina, choroid, and sclera) of rat, mouse, and pig were imaged. It was reported that a sample displacement of $< 1\mu m$ was necessary during the time scale of image acquisition. [83] Oh et al. demonstrated better depth penetration in tissue by using a Xenon arc lamp as a light source and an InGaAs camera (SU320 MSW-RS170, 12 bit, 60 Hz ) in an ultrahigh-resolution FF OCT setup. [84] InGaAs cameras have better spectral response in the wavelength range of the Xenon lamp source ( 0.9 to 0.4$\mu$ m wavelength range) as compared to Si-cameras. Moreover, at longer wavelengths, the effect of multiple scattering is reduced. Depth penetration of $\sim 800\mu m$ was demonstrated by this setup in human thyroid tissue. The reported axial and lateral resolutions achieved were 1.9 and 2$\mu$ m, respectively. [85] An en face image was recorded in 2 s with a sensitivity of 86 dB . Fig. 1 shows the TD FF OCT system based on linnik configuration used by Boccarra et al.

FF TD OCT with an on-pixel signal demodulation and amplification using a smart pixel detector array was demonstrated by Laubscher et al. [87] Although already achieving video rate volumetric imaging, its small sensitivity limited its use for imaging biological tissue.

FF TD OCT has also been extended to endoscopic imaging. It is based on the coupling of two distinct interferometers: one external to the probe and another placed at the distal end of the probe. The external interferometer is used to modulate the source spectrum. The interference between the light reflected from the reference arm and the light back-scattered at different depth in the sample obtained by the interferometer at the distal end is superimposed with the interference signal from the external interferometer on a 2-D camera. The path length difference

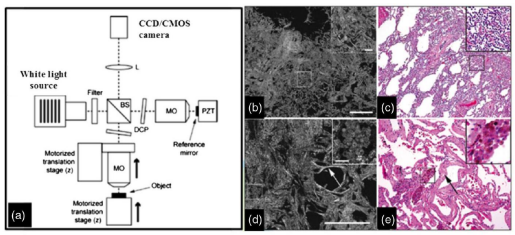



Fig. 1 (a) Schematic of the FF TD OCT system used by Boccara et.al. Figure adapted from Ref. 85. (b-e) Figures taken from Ref. 86 showing comparative full-field optical coherence tomography (FFOCT) and H&E images of non-neoplastic lung with false positive diagnosis. (b, c) Images of non-neoplastic tissue with collapse of normal lung architecture. Boxed areas and insets show dense connective tissue where it is difficult to rule out presence of tumor. ( $\mathrm{d}, \mathrm{e}$ ) Images of non-neoplastic tissue showing clusters of smoker's macrophages (boxed areas and insets) and thickened alveolar septa (arrows).

in the external interferometer determines the depth in the sample. [88–90] Oh et al. used a Michelson interferometer as an external interferometer, a Linnik interferometer at the distal end of the optical fiber probe, and a fiber bundle for 2-D image collection. A spatially and temporally incoherent Xenon arc lamp light source was used. [88] Cellular-level imaging at a sensitivity of $\sim$ 77 dB was demonstrated. However, the imaging speed was too slow (en face imaging in 8 s ) for any practical clinical application. Also, the use of a Linnik configuration and bulk optics at the distal end of the probe prevented any in situ imaging. Ford et al. improved the design by miniaturizing the endoscopic probe by using a Fizeau (common-path) interferometer at the distal end. [89] However, the sensitivity achieved was too low for OCT imaging ( $\sim$ 60 dB ). Latrive and Boccara were able to demonstrate in vivo and in situ cellular-level imaging using a rigid endoscopic probe. [90] They also employed a common path interferometer at the distal end and used a Xenon arc lamp as the light source. The axial and lateral resolutions achieved were 1.8 and 3.5$\mu$ m, respectively. An en face image was obtained in 1 s with a sensitivity of 80 dB . Ex vivo images of human breast tissue up to a depth of 40$\mu$ m and in vivo human skin imaging up to a depth of 20$\mu$ m were demonstrated. [90]

### 3.2 Line-Field OCT

The concept of axial-lateral parallel TD OCT was demonstrated by Zeylikovich et al. in 1998, which provides a X-Z cross-section image instead of an en face image. [11] The setup typically consists of a free space Michelson interferometer with a broadband light source illuminating a line on the sample using a cylindrical lens in the sample arm. The reference arm has a reflective diffraction grating set in the Littrow configuration with respect to the incoming collimated reference beam. The interference of the image of the grating and the backscattered light from the sample is captured by a 2-D camera. The reference arm is modulated using a PZT to produce phase-shifted signals. The diffraction grating introduces a continuous optical path length delay along the grating dispersion direction. This optical delay allows scanning the sample in depth at each point of the line
illuminating the sample, thus encoding ( $X - Z$ ) cross-section information. Watanabe et al. demonstrated such a system for high-speed in vivo imaging of a human finger. [92] A sample volume of $5.8(x) \times 2.8(y) \times 2.0(z)\mathrm{mm}^3$ (corresponding to $512 \times 250 \times 512$ pixels) was imaged at six volumes per second with a horizontal mechanical scan. An ultrahigh-speed complementary metal oxide semiconductor (CMOS) camera (Photron, 10 -bit resolution, at 3000 fps ) was used. However, the detection sensitivity was only 76 dB (after $2 \times 2$-pixel binning). [92]

It has been shown that FD OCT has a sensitivity advantage of 20 to 30 dB over conventional

TD OCT. [18,93] Sensitivity well above 80 dB can be achieved in the case of an FD OCT even with low light levels and high-speed detection. This also allows for better depth penetration in the sample. This sensitivity advantage of FD OCT has led researchers to develop LF FD OCT, where an $X$ - $Z$ cross-section image of the sample can be acquired at once rather than an $X-Y$ cross-section (en face plane) as in FF TD OCT. The setup typically consists of a Michelson interferometer with an anamorphic illumination system to produce a line illumination on the sample and the reference mirror. The broad bandwidth light reflected from the sample and the reference arm is combined and dispersed using a diffraction grating and detected by a 2-D CCD/CMOS camera. Hence, a spectrally resolved interference signal of each point on the line illuminating the sample can be obtained in parallel. A simple 1-D FFT (after $\lambda \rightarrow k$ interpolation, $k$ is wavenumber) along the spectral dimension of the data yields an $X-Z$ cross-sectional image. [94–99] Zuluaga and Kortum were among the first to demonstrate a working LF FD OCT system using a technical sample. [94] Grajciar et al. developed the first LF FD OCT system (shown in Fig. 2) for in vivo measurement of the human eye. [95] In vivo images of the anterior chamber, iris, and cornea of the human eye were recorded using an SLD light source ( $\lambda_o = 811$ nm, FWHM $\Delta\lambda = 17$ nm ). The detected sensitivity was 89 dB with 2 mW power on the sample. The reported axial and lateral resolutions were 17 and 100, respectively. A full tomogram of size 8 mm($X$) $\times$ 3.8 mm($Z$) was recorded in 1 ms . The degradation of lateral resolution due to cross-talk was reported for the SLD source. A spatially and temporally incoherent halogen

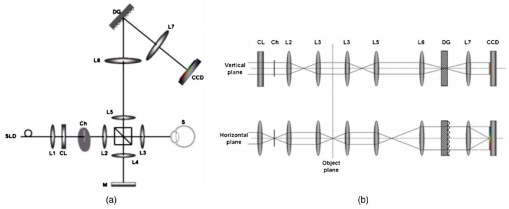



Fig. 2 (a) Schematic of the line field Fourier domain OCT system used by Grajciar et al. (b) Diagrams of optical paths for the two orthogonal planes of the anamorphic optical system. Figures are taken from Ref. 95.

lamp source ( $\lambda_o = 800$ nm, $\Delta\lambda = 100$ nm ) was also tried to reduce the cross-talk. However, the output power was too low for in vivo imaging. Nakamura et al. reported an improved LF FD OCT system for 3-D human retinal imaging. [96] In vivo retinal images over a lateral area of $2.1 \times 2.3$ mm$^2$ with 256 cross-sectional ( $X-Z, 256$ A-lines /s ) were recorded in 1.3 s . The effective imaging speed of $51,500$ A-lines /s achieved was limited by the frame rate (201 fps) of the camera (ATMOS1M60, Atmel Corp., USA, 12 bit). The lateral resolution was 16.4$\mu$ m along the parallel direction and 9.7$\mu$ m along the scanning direction. The axial resolution achieved was 7.4$\mu$ m using an SLD light source ( $\lambda_o = 840$ nm, FWHM $\Delta\lambda = 50$ nm ). It was reported that the ocular aberrations degraded the image quality in the LF OCT system in comparison to the flying-spot FD OCT systems. [96] A sensitivity of 89.4 dB was achieved with 9.7 mW power on the sample. A lateral sensitivity drop of -30 dB was reported at the edges of the measurement, which was attributed to factors such as the spatial coherence cross-talk due to the aberrations, the tolerance, and the extent of the spot size. [96] 3-D LF FD OCT for in vivo dermatological investigation has also been reported by Yasuno et al. [97] Zhang et al. demonstrated a combined adaptive optics (AO) and LF FD OCT system for retinal imaging. [98] A flood illumination system was included to aid focusing at different layers in the retina for AO aberration correction. An 843-nm SLD light source ( $\Delta\lambda = 49.4$ nm ) was used for LF FD OCT, whereas a $788-$nm SLD was used for AO aberration correction. B-scans ( $100\mu m \times 500\mu m$

) were acquired at 500 Hz with dynamic aberration compensation at $\sim$ 14 Hz. The reported axial and lateral resolutions were 5.7 and 3$\mu$ m, respectively. A maximum detected sensitivity of up to 94 dB was reported. The exposure at the cornea for $843-$ nm SLD was 1 mW with 1/3deg line illumination at the retina, whereas for the 788nm SLD, the exposure level at cornea was 0.73 mW with a 1 deg patch illumination at retina. These exposure levels were well below the maximum permissible exposure (MPE) limit according to the American National Standards Institutes (ANSI) standards. [98] It was shown that the detected sensitivity depends on ocular aberrations. A significant improvement of $\sim$ 11.4 dB was achieved by correcting aberrations such as
astigmatism, third order, and higher orders with a focus on the cones. A degradation of 13.1 dB was reported when a defocus of $\sim$ 200$\mu$ m was introduced. Images of subcellular structure in the cone photoreceptors were recorded. The problems such as cross-talk and drop in lateral sensitivity as encountered in other LF FD OCT systems were also reported. A full-range (complex conjugate cancelled) LF FD OCT achieved by dispersion control was reported by Witte et al. [99] The algorithm was adapted from a dispersion encoded full-range method introduced for FD OCT. [100] The robust, fast, and 2-D OCT compatible dispersion encoded full-range method based on an efficient peak-finding algorithm was used to suppress the FD OCT intrinsic mirror image by more than -40 dB and extend the axial measurement range from $z = -1.6$ mm to $+1.6$ mm . Full-range single-shot in vivo images of fingertip and the anterior chamber of mouse eye in vitro were shown. Alternatively to sampling the full signal spectrum for each point along the line using an imaging spectrometer, it is possible to sample the spectrum in time with a line array sensor employing a frequency-sweeping laser. Thus, the problem of spectral cross-talk seen in the spectrometer-based LF FD OCT system is avoided and imaging optics is simplified. Such a system has been reported by Lee et al. [101] The axial resolution achieved with this system was 8.3$\mu$ m. A tomogram ( $512\times1024$ pixels) was acquired at a speed of 45 fps with a sensitivity of $\sim$ 88 dB. [101]

### 3.3 Full-Field Swept Source OCT

The development of high-speed sweeping (several hundred kilohertz) laser sources has led researchers to investigate their application in an FF OCT setup employing an ultrahigh-speed 2-D camera to increase the imaging speed even more than those of the point scanning SS OCT systems. [102] The FF SS OCT setup usually consists of a free space Michelson interferometer and involves area illumination with a sweeping wavelength laser source on the sample and the reference mirror. The detection of the generated interference signal is performed by a highspeed 2-D camera. The 2-D interference signal is sequentially recorded for each wavelength, and 1-D FFT (after $\lambda \rightarrow k$
interpolation) is performed along the spectral dimension to get a 3-D volume image with depth information of the sample. Since FF SS OCT is also an FD imaging modality, it enables higher sensitivity as compared to TD OCT systems. [17,18] The interference signal with a varying wavelength can also be recorded in the Fresnel or the Fourier plane instead of the image plane, and digital holographic reconstruction can then be performed on the recorded data to yield OCT equivalent images. [103–105] In the year 2000, Kim demonstrated 3-D imaging of a millimeter-sized biological sample using wavelength scanning wide-field digital interference holography. [103] A wavelength sweeping ring dye laser ( $\lambda_o = 601.7$ nm ) was used with a narrow sweeping range of 3.08 nm . As a result, a low axial resolution of only 100$\mu$ m could be achieved. Povazay et al. demonstrated an improved FF SS OCT system using a broadband Ti-sapphire laser ( $\lambda_o = 800$ nm, $\Delta\lambda = 100$ nm ) with an acousto-optic frequency tunable element (line width $< 0.4$ nm ). [106] The axial and lateral resolutions achieved were 3 and 4$\mu$ m, respectively. The detected sensitivity was $\sim$ 83 dB with an average power per A-scan of $\sim$ 20$\mu$ W. An acoustic mode-mixer was used to generate different speckle patterns formed due to the different transversal

mode of the multimode fiber used for illumination during the exposure time to average out the speckle noise. Due to the speed limitations, only ex vivo images of a biological sample (fruit fly) were demonstrated. [106] Later, in 2010, Bonin et al. demonstrated in vivo retinal imaging using an FF SS OCT system equipped with an ultrahigh-speed CMOS camera (Y4, Redlake/IDT, Tallahassee, Florida). [102] A-scan rate of 1.5 million A-lines/s was demonstrated for a sample volume of size $640(\mathrm{X}) \times 24(\mathrm{Y}) \times 512(\mathrm{Z})$ pixels. An SS laser form Superlum Ireland (Broadsweeper BS840-01, $\lambda_o = 850$ nm, $\Delta\lambda = 50$ nm ) with an output power of 5 mW was used. The detected sensitivity was only $\sim 72$ dB with an equivalent power on the sample limited to $92\mu$ W/ pixel. It was noted that at least a 100 Hz sweep rate was required for in vivo imaging in order to avoid any significant sample motion blurring. [102]

Wide-field imaging does not introduce a confocal depth gating. This important feature enables collection of out-of-focus light beyond the Rayleigh range. If the amplitude and phase of the out-of-focus signal is known, then numerical holographic reconstruction can be applied to get the focused image. Thus, a uniform lateral resolution can be obtained over the whole imaging range. Hillmann et al proposed Holoscopy, i.e., a combination of wide-field digital lensless holography and SS OCT to allow focusing beyond the Rayleigh range. [105] The optical setup consists of a lens-less free space Michelson interferometer with a convex mirror in the reference arm to produce a spherical reference wave. The holographic data were acquired with the same sweeping laser as in Ref. 102. For 3-D image reconstruction, the recorded holographic image at each wavenumber was multiplied with the conjugate of the reference wave and propagated to a specific distance using the angular spectrum approach. Finally, the 1-D FFT ( $k \rightarrow z$ ) was calculated to yield an image with depth information. This process was repeated for different reconstruction distances to focus on different depths in the sample. The focused region in each reconstruction was filtered and they were stitched together to get an image with depth-invariant lateral resolution. The axial resolution achieved was $14.7\mu$ m, whereas a lateral resolution of up to $7\mu$ m was achieved. Improvement of over a 30 Rayleigh range was demonstrated using an iron-oxide phantom sample. The detected sensitivity was $\sim 72$ dB. In vivo imaging of a human fingertip was also
shown using an ultrafast camera (FASTCAM SA5, Photron USA Inc., frame rate = 7000fps ). [105] Hillmann et al. further proposed an efficient holoscopy reconstruction method to obtain focused 3-D volume images in a single step. [104] It involves interpolating the recorded 3-D holographic data on a nonequidistance grid in a 3-D spatial frequency space, similar to the inverse scattering reconstruction called inverse synthetic aperture microscopy (ISAM) proposed for FF SS OCT, [107] and then calculating the 3-D FFT to get an OCT equivalent image of the sample. However, unlike ISAM, this method does not calculate the regularized pseudo inverse of the forward-propagating kernel. Instead, the recorded object field is numerically propagated back into the sample volume. A decrease by a factor of 15 in the processing time was demonstrated for an imaging NA of 0.14.[104]

ISAM and holoscopy reconstruction provide 3-D volume images focused over the whole imaging range. But these reconstruction techniques work well only when the sample has a uniform refractive index, the imaging NA is low, and defocus is the main problem. [104,107] However, at a higher NA ( $> 0.2$ ), the effect of higher-order aberrations, such as spherical, astigmatism, coma, etc., becomes dominant. As a result, proper image reconstruction becomes difficult. A computational AO technique has been suggested to reduce higher-order aberration in the case of a phase stable point scanning FD OCT system. [107] But this technique is based on an optimization method that is iterative in nature and increases the computational cost. Numerical complexity increases if the sample is highly inhomogeneous, causing variable aberration with depth. [108] Recently, Kumar et al. proposed a digital aberration correction method, which is, in fact, a digital equivalent of a Hartmann sensor, to deal with variable higher-order aberrations in FF SS OCT. [109] In this method, a 2-D FFT is performed on the aberrated image field

corresponding to a layer in the sample to get to the spatial Fourier plane, and then the Fourier data are segmented into small subapertures. 2-D inverse fast Fourier transform (IFFT) s on these segmented Fourier data yield images that are shifted with respect to each other due to aberrations. Cross-correlation of the intensity of these images with respect to a reference image gives shift data from which local slope information of the wavefront error can be calculated. Knowing the slope information, the wavefront error can be calculated using a suitable basis function, such as Zernike's or Taylor monomials. Phase correction can then be applied digitally on the whole Fourier data and an aberration-free image is finally obtained in a single step by calculating the 2-D IFFT. Furthermore, it was shown that this method does not require any a priori knowledge of the system parameters. However, this method assumes that the data are uniformly distributed in the spatial frequency plane, i.e., the sample gives diffuse reflection on illumination, and also the spatial coherence is maintained over the FOV. This condition is usually fulfilled when spatially coherent light source is used to image tissue. Aberration corrections on a technical and a biological sample (grape) were demonstrated. [109] Figure 3 shows the SS FF OCT setup used by Kumar et. al and the results obtained by applying digital defocus correction on a grape sample.

FF TD OCT systems achieve cellular-level en face images over the depth of few hundred microns into the biological samples with histological-level image quality without the need to stain and histologically prepare the sample. [82–85,88,89] The coherence gating provided by the broadband light source in combination with the confocal gating due to the use of MOs with high NA causes rejection of multiple-scattered

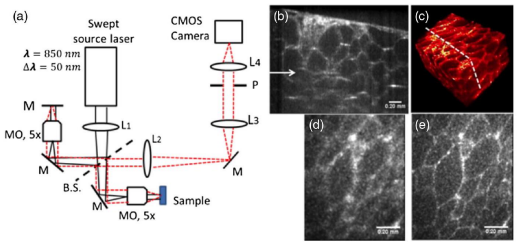



Fig. 3 (a) Schematic of the FF swept source (SS) OCT system used by Kumar et al. (b) A tomogram of the grape sample. (c) Three-dimensional (3-D) image volume with dotted line showing location of the tomogram shown in (b). (d) Enface image obtained at the depth of \$424.8 \mu \mathrm{˜m}\$ in the grape sample indicated by arrow in (b). (d) is the digitally focused image of (e) obtained using the algorithm described in Ref 109. Figures are taken from Ref. 109.

light and yields high-resolution images. FF TD OCT systems performing rapid thick tissue histopathology have been successfully commercialized. [86] The axial-lateral parallel TD OCT system using an ultrahigh-speed camera has been shown to produce in vivo $X - Z$ cross-section images. [92] However, due to the sensitivity limitation of the TD OCT configuration, and increased cross-talk due to multiple scattering with depth, the useful imaging range is limited to few hundred microns in tissue samples. [18,93] The LF FD OCT systems have been shown to produce in vivo images with higher sensitivity ( $> 80$ dB ) over a longer imaging range. [95–99] However, they suffer from problems such as cross-talk (speckle noise) due to multiple-scattered light affecting the lateral resolution, spectral cross-talk along the dispersion direction, and a drop in sensitivity at the edges of the images due to the Gaussian illumination. These problems can be mitigated by using a spatially incoherent light source [95] with uniform illumination along the line and by improving the imaging optics. The FF SS OCT system has also been shown to produce in vivo retinal images using an ultrahigh-speed camera. [102] However, the sensitivity was limited to $\sim$ 72 dB because of the low output power of the sweeping laser source used.

Also, the coherent noise caused by backreflections from the lenses in the optical system, which is difficult to avoid in the FF OCT setup, causes reduction of the dynamic range of the detector. [102] A lensfree holoscopy system combining wide-field digital holography with FD OCT has been suggested to mitigate this problem. Wide-field imaging allows for the collection of out-of-focus light and FD imaging provides easy access to the phase information of the detected signal. This enables numerical refocusing of out-of-focus signals beyond the Rayleigh range to produce 3-D image volume with uniform lateral resolution. [104,105] A digital adaptive optics technique based on subaperture correlation has also been implemented in FF SS OCT to correct for higherorder aberrations that can occur in high-NA imaging of a nonuniform sample. [109]

The use of an SS laser in the FF OCT setup has simplified the whole imaging system as no scanning of the sample is needed. The LF SS OCT requires scanning only in one transverse direction. LF and FF OCT setups employ a line and an area
illumination onto the sample, respectively, and, hence, allow for a higher MPE limit as compared to any point scanning system where the illumination beam is focused into the sample. The use of higher power theoretically implies higher sensitivity even at higher speed when the exposure time is very short. Although the feasibility of parallelization of OCT for in vivo imaging has been demonstrated, there are some serious challenges that need to be resolved in future to make it suitable for practical clinical applications. Due to parallelization, there is an inherent problem of cross-talk or speckle noise caused by multiple-scattered light. This makes FF SS OCT unsuitable for imaging tissue where multiple scattering is high. However, FF SS OCT may work for ophthalmic imaging where tissues are relatively transparent and multiple scattering is less pronounced. LF FD OCT or LF SS OCT is expected to suffer less from spatial cross-talk as compared to FF SS OCT as it rejects multiple-scattered light along the scanning direction. But, at present, there is no study that directly compares and quantifies the multiple-scattering effects in FF and the LF OCT setup. The use of multimode fibers with mode-mixing techniques has been suggested for destroying the lateral spatial coherence to reduce cross-talk. [106,110] However, the destruction of lateral spatial coherence may destroy phase correlation over the FOV, which is necessary for holographic reconstruction of focused 3-D volume data or any digital aberration correction method. The use of programmable array illumination may be a viable option to separate signals from adjacent parallel channels in time and space to reduce crosstalk while preserving the phase correlation across the lateral FOV. [111] But the feasibility of such a system for in vivo imaging has to be tested in the future. There are other technological limitations such as the nonavailability of high-speed and highpower broadband commercial sweeping laser sources at 800 nm suitable for ophthalmic imaging using FF or LF SS OCT. Currently, the strongest bottleneck seems to be the commercially available ultrafast cameras that are economical in cost with low read noise when run at high speed. [102] The reduced cross-talk advantage (along the scanning direction) of the LF OCT system over the FF OCT needs to be quantified in the future, and its application to skin tissue imaging needs to be
investigated. However, for tissue imaging, an 1100 nm SS laser source would be needed, as this wavelength range enables reduced multiple scattering and deeper penetration into the tissue. Also, the commercially available fast CCD/CMOS cameras use silicon-based detectors, which have low sensitivity around 1100 nm . The InGaAs cameras have higher sensitivity around this spectral range, but, currently, they offer lower frame rates as compared to most CCD/CMOS cameras. It is hoped that with technological developments, these issues can be resolved and parallelization in OCT may turn out to be a viable option for imaging in the future.

### 26.4.3 4 Label-Free Optical Angiography

A shortcoming of OCT is its missing tissue specificity as its contrast mechanism is backscattering. Functional extensions of OCT have the potential to partially mitigate this drawback. They are, therefore, extensively investigated and received additional impetus by the recent technical advances in high-speed sensors and laser technology. Above all, Doppler OCT has matured during the last years to become a reliable method for quantifying blood flow in tissue, as well as for depth-resolved label-free micro-angiography providing unprecedented details of tissue capillary networks. In the following, the approaches for optical angiography are presented, and their advantages, applications, and still open questions are discussed.

The gold standard for performing angiography in the eye is fluorescence angiography (FA). In other organs, CT angiography or MR angiography are mostly used, as they give access to deep vascular structures within the body. Their disadvantage, which is shared with FA, is the need for administering contrast agents that might lead to adverse side effects for the patients. This hinders the possibility for the screening of large populations, which would be advantageous for detecting vascularrelated diseases early on, with a huge socioeconomic benefit. Optical angiography will not be able to change the imaging paradigm for deep vascular structures; however, for superficial vascular and microvascular plexuses, including retinal and choroidal vessels, the situation is very different. The eye itself can be seen as a window to the brain, as neural retinal tissue and its perfusion is directly accessible by optical means. Optical angiography based on OCT provides concise 3-D vascular maps that allow distinguishing between different retinal perfusion layers, and also the important choroidal vasculature. Especially, the development of new light source technology operating in the NIR region at $\sim 1\mu m$ wavelength helped to improve contrast for deeper choroidal structures, as longer wavelengths experience less scattering. Apart from ophthalmology, there is also an increasing need and vivid interest to access vasculature and perfusion in dermatology. Recent work demonstrated the capabilities to contrast skin microvasculature in healthy skin as well as in different pathological alterations, including skin cancer. Most pathologies start in the superficial skin layers in the first hundreds of micrometers that are easily accessible by OCT. There are characteristic vascular patterns that might, in the future, be used as biomarkers of disease. For deeper organs, such as the gastro-intestinal tract, or for head and neck diseases, specific delivery probes are needed.

Given the exciting prospects of OCT angiography, it has been the subject of active and proliferative research during the last decade. Early developments based on TD OCT include speckle and phase variance and power Doppler imaging. [112,113] The disadvantages of TD OCT were, however, its limited sensitivity and
speed for 3-D angiography. This changed with the introduction of FD modalities, [18,114] which paved the way for the exciting developments in optical angiography that we see today. Methods based on FD OCT soon adapted the phase approaches from TD OCT. [115,116] However, the differences of the data acquisition process and optical setup in FD versus TD OCT opened a large playfield for many new approaches to optical angiography. For example, spectrometer-based FD OCT suffers from fringe washout and sensitivity loss when imaging moving structures. [117] Resonant Doppler OCT turns this apparent downside into an advantage when contrasting moving structures, by dynamically tuning the reference arm delay, matching speeds in both arms. [118] Other approaches use spatial frequency filtering techniques, by making use of the fact that moving structures appear shifted in spatial freqeuncy. [119,120] Alternatively, one can operate on the Doppler histogram, [121] or use support vector machines operating on phase difference data. [122] Advanced postprocessing techniques to filter out vessel structures from static bulk can also operate directly on OCT intensity data. This has been successful for retinal vascular structures as well as for choroidal vasculature. [123–125] Employing dual-beam setups that probe the sample in parallel on laterally displaced positions enhances the sensitivity for contrasting vessels with a large range of velocities. Both beams scan the same spot after an adjustable time interval

given by the lateral displacement and the scanning speed. [126–128] The possibility of high-speed data acquisition with FD OCT certainly marked a breakthrough for optical angiography. High speed translated directly to smaller motion artifacts, which is important when visualizing small capillary structures, as well as for keeping signal correlation by dense sampling. [129,130] The latter is of particular importance for high-quality vessel contrasting against static bulk tissue in many angiography approaches. Furthermore, high speed enables specific Doppler analysis methods, [131] enhanced velocity bandwidth, [132,133] as well as the possibility of using flexible scanning patterns to further increase the velocity sensitivity. [134,135]

In general, flow contrasting requires comparing two or more scans preferably taken at the same location. The velocity sensitivity of the contrasting technique scales with the time difference between the scans. Even though angiography aims not necessarily at quantifying the flow, it is important to observe a flow signature by looking either at phase differences, intensity, or speckle changes, or at changes of the full complex OCT signal. In order to observe signatures from small capillary flow, relatively long time intervals of typically a few milliseconds between acquisitions are required. They are equivalent to only few hundred scans per second. Obviously, at such an A-scan rate, the acquisition is too slow for in vivo volumetric imaging and would also be heavily affected by motion artifacts. On the other hand, with high-speed OCT, it is possible to achieve Bscan or tomogram rates of several hundred scans per second, which is exactly the timing needed for high-sensitive flow contrasting, including the smallest retinal capillaries. Thus, instead of calculating signal differences between A-scans, one considers successive B-scans or parts of B-scans.

The most impressive results with optical angiography have been achieved when calculating the amount of signal decorrelation between two or more scans. [136] As mentioned before, it is possible to observe decorrelation or variance of phase, [137,138] speckle, [139,140] or the full complex signal. [141] An important factor, apart from the time interval, is the oversampling ratio, which is the number of A-scans per spot size of
the scanning beam. Clearly, if the oversampling is too low, signal decorrelation will occur for all structures even without bulk motion, again degrading the blood flow contrast. This imposes important restrictions concerning the covered angular extent on the retina, as typical imaging times for a full volume should be kept to a few seconds. This is, of course, different if a tracking system is employed, allowing for stitching of several recorded volumes of smaller areas. In this case, the B-scan rate might even be slower, as long as the proper correlation between two acquisitions for the static bulk tissue is maintained.

The strategy for obtaining DOCT micro-angiographies with high motion sensitivity is displayed in Fig. 4. It is based on calculating signal differences between B-scans that are taken at the same vertical position. B-scan rates are typically a few hundred hertz. The signal changes are analyzed pairwise between adjacent B-scans in postprocessing. Averaging over several difference images taken from the same lateral position helps to obtain a motion contrast volume $A(x, y, z)$ with improved vascular contrast and motion sensitivity. [142]

$$A(x,y,z) = \frac{1}{M(y)} \sum_{i=0}^{N-1} \left\{ \left[ \sum_{x,z} D\left(x, y_i, z\right) \right] < T \right\} \cdot D(x,y,z),$$

where $D\left(x, y_i, z\right)$ stands for tomograms that represent, in general, differences between either the intensity, the phase, or the full complex OCT signal of successive tomograms at the vertical position $y$. The tomogram or fast scanning direction is denoted by index $x$. It is, however, important to exclude the difference tomograms with excessive bulk motion artifacts due to involuntary patient motion. This is accounted for by using a threshold $T$ for the average signal decorrelation across a full tomogram.

The number $M$ of remaining average difference tomograms at each position $y, D\left(x, y_i, z\right)$, can be formally written as $M(y) = \sum_{i=0}^{N-1} \{[D\left(x, y_i, z\right)] < T\}$, with $N$ the number of Bscans measured at the same position $y$. The logic operation in the brackets [see also Eq. (1)] yields 1 or 0 for TRUE or FALSE, respectively. Easy to implement in particular for SS

OCT is to analyze changes in the intensity signal. Phase sensitive analysis is strongly affected by trigger jitter, which can, in principle, be solved by using a hardware trigger provided by the source or a reference interferometer signal. However, if the source does not provide a $k$-trigger signal for synchronizing the acquisition, intensity-based imaging has clear advantages by being less complex. In order to contrast moving structures, one calculates differences between logarithmically scaled intensity tomograms taken at the same position $y$, i.e., $D(x, y_i, z) = \log[I(x, y_{i+1}, z)] - \log[I(x, y_i, z)]$ Static tissue will ideally cancel out, whereas dynamically changing structures, which cause fluctuations in intensity or speckle, remain [Fig. 4(b)].

Most tissue has flat vascular network structures, which are usually best appreciated from en face views. They are obtained by integrating the difference signatures over selected depth ranges as shown in Fig. 4(d) for the inner retina capillary structures. The micro-angiography displays integrated capillary structures of the retinal nerve fiber layer down to the outer nuclear layer. The resulting fine vessel network in the parafoveal region allows for precise assessment of vascular density as well as of diagnostically important anatomical features, such as the size of the foveal avascular zone.

A closer look at Fig. 4(b) reveals, however, intrinsic artifacts of highly sensitive motion contrast techniques being signal tails or shadows below vessel cross-sections. They are due to phase

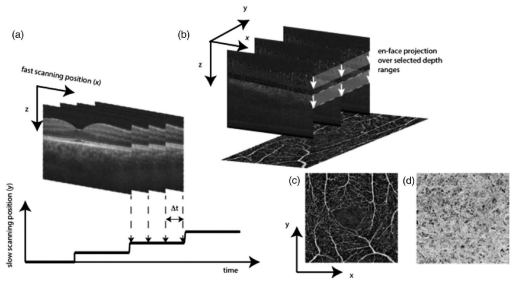



Fig. 4 (a) Recording scheme for Doppler OCT (DOCT) angiography. (b) Calculated 3-D DOCT angiography. (c) Taking the en-face projection by, e.g., plotting the maximum intensity results in comprehensive en-face angiography maps. (d) Dense choroidal vasculature obtained by integration over dashed depth region in (b).

and speckle decorrelation of light that needs to transverse the strongly scattering blood volume. Multiple scattering, together with the dynamic inhomogeneous refractive index distribution within the vessel lumen, causes phase as well as speckle decorrelation even for light backscattered from static structures below the vessel. Masking or weighting with the original intensity information can reduce this artifact, yielding clearer 3-D rendered angiography maps. An improvement of this artifact has further been accomplished by using Bessel beams for functional tissue imaging. [143] SS OCT usually operates in the longerwavelength region above a $1\mu$ m central wavelength. Those wavelengths experience less scattering than the ones at 800 nm , allowing, therefore, for better contrast in deeper tissue layers. [144,145] In ophthalmology, it is of increasing interest to achieve better contrast for choroidal structures that lie below the strongly scattering pigment epithelium [Fig. 4(d)].

Figure 5 demonstrates the performance of an intensity-based OCT angiography of a healthy human retina. The image has been obtained with a high-speed FDML SS OCT system, operating at a $1.06\mu$ m central wavelength and achieving a 1.7 MHz A-scan rate. [142] Such speed allows for close to 50 deg large FOV angiography maps of the retina in a fully noninvasive label-free fashion.

The possibility to assess the microvascular structure down to the level of capillaries opens exciting new perspectives in preclinical research, in particular for cancer research. Due to the easy label-free contrasting scheme, neovascular growth can be studied in situ over time. Figure 6(b) gives an impression of the level of detail achieved with phase variance OCT imaging. The angiography shows features such as satellite metastases that are not visible in the structural OCT image [Fig. 6(a)]. Figures 6(c) to 6(e) demonstrate the sensitivity gain by including an increasing number of frames for the contrasting algorithm. [146]

The method performs equally well even in strongly scattering tissue like human skin. Figure 7 demonstrates the performance of contrasting micro-angiographic details for pathologic skin conditions. A case of basal cell carcinoma, being a nonmelanoma skin cancer, is shown. The gold standard in dermatology is the dermascope, which is basically an epiluminescence microscope to image skin lesions. It yields no in-depth information of the
skin structure, as OCT does. Therefore, OCT is gradually penetrating the dermatology market as well, through dedicated commercially available systems. Still, structural information alone is unspecific, and additional biomarkers of disease are extensively investigated. Most diseases come along with alterations of the microvascular pattern. Thus, methods that are capable of visualizing the patterns in depth are of high diagnostic value. Recent results of OCT angiography indicated that alterations of the vascular pattern are characteristic for different pathologic conditions, such as inflammation, or skin cancer. [147] The expectations to use functional OCT for enhanced diagnosis in dermatology are, therefore, high.

Apart from pure angiography, OCT has also the potential to quantify perfusion in vessels. [112,148,149] For flat vessel beds, signal variance calculations show a linear dependence on the flow. [150] Since most angiographic structures in skin, retina, or the choroicapillaries are parallel to the tissue surface, those methods are currently investigated in detail. Quantitative assessment of the velocity through signal correlation analysis needs, however, fast scan rates in order to avoid complete decorrelation of flow signals. So far those methods have been successfully applied only to well-fixated samples. [151] The reason is the number of correlated scans necessary to perform flow analysis with sufficient dynamics. Bulk motion as well as measurement time might, therefore, be critical for human in vivo applications.

For retinal angiography, OCT is in competition with fluorescein or indoyanine green (ICG) angiography. The latter needs intravenous administration of contrast agents, which might lead to adverse side effects for the patient. OCT angiography, on the other hand, does not need dyes and can, therefore, be frequently applied. However, as a motion contrast technique, it does not detect dye leakage. This limits its application for diseases where vascular permeability is altered. Still, for choriocapillary vasculature, the missing sensitivity to leakage is, in fact, of advantage, as the choriocapillaris are fenestrated. [152] This leaves the contrasting of choriocapillaris with fluorescence angiography more challenging due to dye leakage. Most studies of the choriocapillaris have, therefore, been based on postmortem histology or on observation of the temporal perfusion patterns.

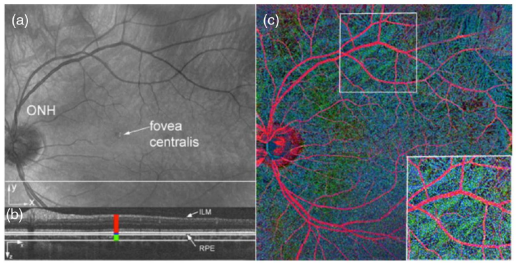



Fig. 5 Wide field optical angiography over 48 deg. (a) Fundus projection of OCT intensity data. (b) Tomogram taken at indicated line in (a). The colors represent depth ranges for calculating angiography projections. (c) Color coded OCT wide field angiography. White box region is shown zoomed in inlay (from Ref. 142).

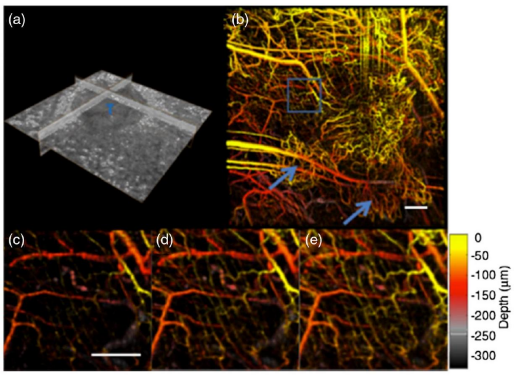



Fig. 6 (a) OCT intensity image of a tumor. (b) Microvascular projection image calculated over eight Bscans at each vertical position. Blue arrows point to satellite metastases that are not visible in (a). The color indicates relative depth of the vessels. (c) to (e) Magnified region [box in (b)] showing the improvement on increasing the number of B-scans ( $N = 2, 4, 8$, respectively). Scale bar denotes \$250 \mu \mathrm{˜m}\$ (from Ref. 146).

Knowledge about perfusion pathways is a first step toward understanding the metabolic need. Important additional information is certainly the oxygen consumption. This might be estimated via knowledge of oxygen saturation in the arterial and venous vascular tree. The most common approach for systemic oxygen saturation is pulse oximetry, which, however, is not spatially resolved. Oximetry based on differential spectroscopy, on the other hand, is a spatially resolved method that has been successfully applied to retinal vessels. [153,154] Another approach combines Doppler OCT with hyperspectral imaging for assessing blood oxygenation in tumor vasculature. [155] A relatively young biomedical imaging technique that uses absorption as a contrast mechanism is photoacoustic tomography (PAT) or microscopy (PAM). [156] Multispectral PAM allows for calculating
blood oxygen saturation even in strongly scattering neural tissue. [157] It is, therefore, natural to combine OCT that yields high-resolution anatomic contrast and PAT/PAM that provide absorption or metabolic information. Hybrid imaging approaches with OCT are discussed in Sec. 5.

OCT angiography is a very promising functional extension of OCT that holds great potential for providing additional specific biomarkers of disease. In ophthalmology, the noninvasive and label-free alternative to FA might soon be adapted for commercial OCT platforms. The most important advantage, since no trained specialist is needed for dye administration or for operating the instrument, is the capability of frequent monitoring and large population screening.

This is expected to reduce the number of invasive angiographies performed, to

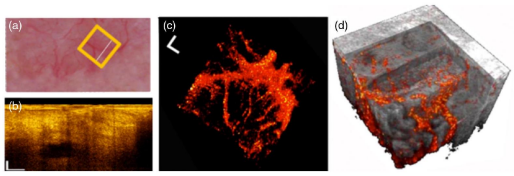



Fig. 7 (a) Dermascope image of basal cell carcinoma; yellow box denotes field of view of OCT. White line indicates OCT scan position for tomogram displayed in (b). (c) OCT angiography projection. (d) Overlay of OCT angiography (red) to OCT intensity data (gray) (from Ref. 147).

help enhancing early diagnosis of major ophthalmic diseases, as well as to support the understanding of the pathophysiologic mechanisms behind them. In other medical fields, such as dermatology, neurology, cancer research, cardiology, or for endoscopic organ screening, its application is extensively investigated. Further improvement for the clinical applications may also be expected from improved calculation speed for real-time perfusion contrasting using graphics processing units. [158] Without doubt, OCT angiography can be currently seen as the most successful functional extension of OCT that meets immediate diagnostic needs.

## 26.4.4 5 Multimodal Optical Coherence Tomography

### 5.1 Hybrid OCT with Fluorescence and Nonlinear Microscopy

There is a strong need to assess tissue structure and molecular information in vivo and in real time with high sensitivity and specificity to circumvent biopsy for improved cancer diagnostics and therapy. OCT provides high-resolution, cross-sectional large-area imaging of tissue microstructures in situ and in real time with several millimeters of depth penetration, but one major drawback of OCT is the lack of chemical selectivity, resulting in difficulty in differentiating between pathological and normal tissue with similar morphological or optical properties. [159] To study biological tissue at a cellular level, higher resolution is needed. Confocal microscopy and multiphoton microscopy (MPM) achieve submicrometer resolution with penetration depth limited to several hundreds of micrometers due to severe scattering in biological samples. Compared to confocal microscopy, MPM uses NIR ultrafast lasers to evoke the nonlinear process; therefore, it does not require a pinhole for excluding out-of-focus fluorescence. Due to the use of longer wavelengths, higher penetration depths can be achieved. Ti:sapphire lasers are capable of simultaneously and efficiently addressing the needs for MPM and OCT, respectively, and permit the construction of a combined label-based and label-free imaging platform. Such a platform offers numerous benefits, such as the inherent 3-D sectioning capability, the deeper penetration depth into the sample compared to visible light, and the use of lower average power, which prevents photo damage and photo toxicity, making it an extremely interesting modality for live cell and in vivo imaging. The most common nonlinear optical modalities are two-photon excitation fluorescence (TPF), [160] second harmonic generation (SHG), [161] third harmonic generation (THG), [162] and coherent Raman scattering. [163] Each technology provides unique and complementary benefits. High axial and lateral resolution comparable to confocal microscopy without a pinhole can be achieved. Imaging can be performed with either two galvanometric mirrors to scan the beam across the sample or a scanning unit for point-by-point scanning. Filters and dichroic beam splitters are essential to avoid detection of the fundamental laser beam and Rayleigh scattered light. The nonlinear signal can be detected with a single-element detector [photomultiplier tube (PMT)]. TPF and SHG signals are generated simultaneously

and can be collected with two photomultipliers. Distinguishing between normal and abnormal tissue types based on morphological and molecular composition will expand the range of non-invasive clinical diagnostic techniques. Noninvasively detecting both chemical and structural changes in healthy as well as diseased tissue can help to diagnose diseases at an early stage, long before structural changes occur. [161–163] Linear
and nonlinear chemically selective imaging methods, such as Raman scattering, coherent anti-Stokes Raman scattering (CARS), or MPM, can be combined with OCT to enable molecular sensitivity in OCT and to simultaneously provide structural and functional information of the tissue. This can greatly boost biomedical applications, such as detecting early diseases and therapeutic interventions. Any single modality typically furnishes only an incomplete picture of the tissue optical response; hence, an approach that integrates complementary optical imaging modalities is needed for a more comprehensive nondestructive and minimally invasive tissue characterization.

### 5.1.1 Multimodal OCT and Raman

In Raman spectroscopy (RS), monochromatic illumination of a sample can lead to a small fraction of weakly scattered photons, indicating a frequency shift that is directly related to the vibrational or rotational states of the probed molecular bonds. RS is a label-free imaging technique because the signal is purely derived from the vibrational or rotational states probed in the material. [164] A simple sketch of combined OCT/RS is shown in Fig. 8.

The ability to use the structural information of OCT with subsequent RS from two independent modalities to identify early dental caries was demonstrated by Ko et al. [165] Their research showed that the Raman peak sensitivity of enamel is dependent on the Raman sampling arrangement. Further, the obtained Raman spectrum from sound and carious enamel varied with the Raman sampling depth, resulting in difficulties in using RS alone to identify lesions. The additional data obtained from OCT were important for identifying critical related structural changes and depths for the RS, suggesting the usefulness of combining structural and chemical information.

Patil et al. [166,167] demonstrated a combined NIR OCT and Raman system with separate detection channels providing morphological and biochemical information for normal and malignant tissue characterization, including skin cancer. OCT/RS was performed on cancerous skin [basal cell carcinoma (BCC)] and normal skin. OCT alone was capable of indicating hyper-reflective and hypo-reflective features in both; hence, critical reflectivity features cannot be classified as malignant by OCT alone. However, the large morphological scan was used to guide the RS to a region of interest for obtaining biochemical information. A clear differentiation between malignant and normal tissue was achieved by the obtained Raman spectrum, which contained notable differences in specific wavenumber regions. The presence of BCC induced a stronger scattering intensity in the spectral bands associated with proteins ( 1330 and 1440 $cm^{-1}$ ) and nucleic acids ( 1090 $cm^{-1}$ ). The increase in the nuclear-to-cytoplasm ratio is often indicated with tumors in histopathology. In conclusion, the OCT/RS system was able to combine imaging

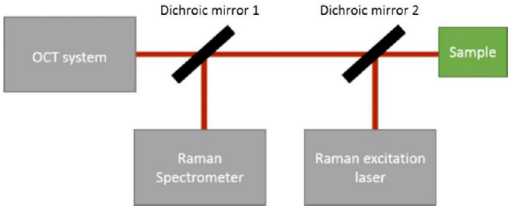



Fig. 8 Simplified depiction of a combined OCT and Raman system.

and spectroscopy for classification of similar morphological structures found in malignant and healthy tissues. Patil et al. [166] also reported an integrated OCT/RS system using a deepdepleted, back-illuminated, thermo-electrically cooled CCD array as a single detector for both Raman and OCT, respectively.

Simultaneous OCT/RS with ex vivo human and porcine retinal samples was demonstrated by Evans et al. [168] The OCT/RS measurements were performed under in vivo imaging conditions. To retrieve the Raman signal with an acquisition time of 200 ms, 4 mW was applied on the sample. In contrast, Patil et al. [166] used 40 mW and 5 s , respectively, to obtain the Raman signal. The reduction of the source power and acquisition time implicated an impaired RS signal-to-noise ratio (SNR). To correct for the drop-off in SNR, several Raman spectra from adjacent OCT A-scans were averaged. The simultaneous OCT/RS was able to distinguish the different structures in a polystyrenepolymer phantom, but it was not able to visualize the main retinal layers of the ex vivo human and porcine retinal samples due to the low SNR.

Ashok et al. [169] performed a multimodal approach using RS and OCT successively from independent modalities to discriminate malignant colon from normal colon tissue. Their work was based on quantitatively combining the morphological information (OCT) and the biochemical information (RS) to enhance sensitivity and specificity. A texture analysis was performed on OCT data to quantify OCT images to further combine the texture parameters with the chemical information obtained by RS. Principal component analysis was used for both Raman and OCT, respectively, for classification. Table 2 shows the evaluated classification accuracy for each modality, demonstrating the potential of achieving accurate discrimination of malignant or normal tissue by combining both data sets.

All the work mentioned above demonstrated the advantage of combining RS with OCT. However, in each study, the obtained RS was volume integrated from different depths, resulting in an average RS outcome, which can result in difficulties in achieving a more detailed biochemical analysis for a given tissue. Khan et al. [170] demonstrated a combined OCT/RS system capable of layered tissue analysis. Depth-resolved RS was achieved due to a confocal Raman configuration rejecting out-of-focus Raman scattered light. The performance of the OCT/ RS system was validated with a paraffin-Acetaminophen phantom and promising results were also obtained on a mucosal tissue sample from goat. Biochemical maps from the epithelium and stroma were retrieved.

### 5.1.2 Multimodal OCT and CARS

CARS is a four-wave mixing (FWM) process where three incident fields (pump photon $\omega_{\mathrm{p}}$, Stokes photon $\omega_{\mathrm{s}}$, and probe photon) interact with the sample to generate a fourth blueshifted

Table 2 Comparison of sensitivity and specificity of the classifier with data acquired from each modality. ${ }^{169}$

| | Sensitivity (%) | Specificity (%) | Accuracy (%) |
|---|---|---|---|
| Raman | 89 | 77 | 82 |
| OCT | 78 | 74 | 75 |
| Combined | 94 | 94 | 94 |

anti-Stokes field with frequency $\omega_{\mathrm{as}} = \omega_{\mathrm{p}} - \omega_{\mathrm{s}} + \omega_{\mathrm{p}}$.[171] When the frequency difference of the pump and Stokes photon matches a molecular vibrational resonance $\Omega = \omega_{\mathrm{p}} - \omega_{\mathrm{s}}$, the CARS signal is enhanced, making CARS a coherent labelfree chemical specific imaging modality. [172]

In molecularly sensitive OCT, CARS was used for nonlinear optical contrast enhancement. [52,173] The contrast enhancement was obtained with a method called NIVI. [48,49] The interferom-

etry setup for molecularly sensitive OCT is based on a conventional OCT setup. However, NIVI differs from conventional linear interferometry. A second wavelength-shifted beam is needed to generate the CARS signal. The experimental setup is shown in Fig. 9.

In Fig. 9 (left), a regenerative amplifier is used to create a pump pulse for the sample arm and to pump the SHG optical parametric amplifier (OPA). The OPA creates two pulses, a Stokes pulse for the sample arm and a coherent anti-Stokes pulse for the reference arm to cause interference with the CARS signal retrieved in the sample arm. Molecularly sensitive OCT imaging of a biological sample was performed from a thin lipid-

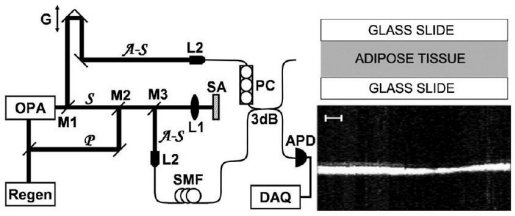



Fig. 9 Experimental setup. Left: Regen, regenerative amplifier; G, gal-vanometer-controlled optical delay; M1 to M3, long-pass dichroic mirrors with different cutoffs; P, pump; S, Stokes; A-S, anti-Stokes; L1, 30mm focal-length lens; SA, sample; L2, 10× microscope objective; SMF, single-mode fiber; PC, polarization controller; APD, avalanche photodiode; DAQ, data acquisition. Right: Coherent anti-Stokes Raman scattering image from lipid resonance at $2845 \mathrm{~cm}^{-1}$ of a thin lipid-dense layer of beef tissue sandwiched between two glass slides. [52]

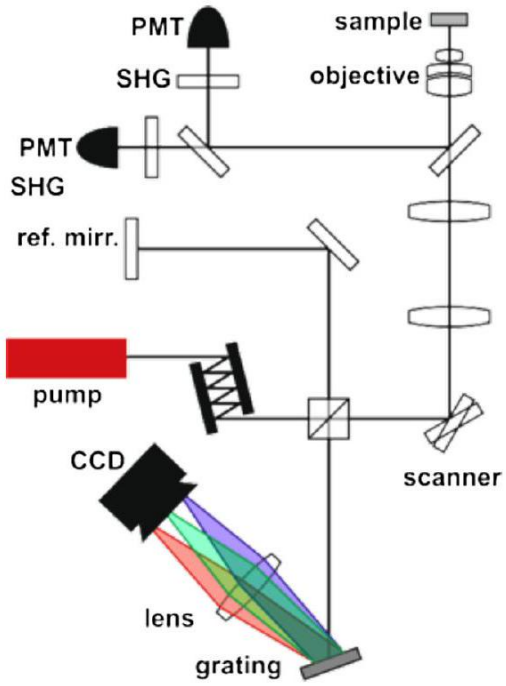



Fig. 10 Typical combined spectral domain and nonlinear optical microscopy setup.

dense layer of beef tissue sandwiched between two glass slides. The CARS signal was retrieved from a lipid resonance at 2845 $\mathrm{cm}^{-1}$, as shown in Fig. 9 (right).

### 5.1.3 Multimodal OCT and MPM

There is a great interest in combining MPM with optical coherence microscopy (OCM). A typical schematic of a combined platform is shown in Fig. 10. In OCT and OCM, imaging contrast is due to the difference in refractive index of different tissues. For many applications, changes in the sample linear scattering properties are small and difficult to measure. MPM provides biochemical and metabolism information which is related to nonlinear effects based on the simultaneous interaction of two or more photons within a molecule related to higher order susceptibilities. Hence, multiphoton excited
fluorescence and/or harmonic generation provide molecular contrast and specificity with submicron resolution. The theoretical estimation of the lateral resolution can be derived from the Rayleigh criterion and the axial resolution is referred to as the depth of focus of the illuminating beam. For a high-NA objective, the axial resolution can be estimated as approximately three to four times its lateral resolution. The integration of OCT and MPT in a combined setup allows for depthresolved high-resolution structural and functional imaging and is demonstrated in many biomedical and clinical applications.

The concept of TPF was first described by GöppertMayer [174] in 1931 and first observed after the invention of the laser to achieve the high photon flux density. [175] It is a three-stage process of photon emission by a molecule or fluorophore. The probability of a two-photon absorption occurring

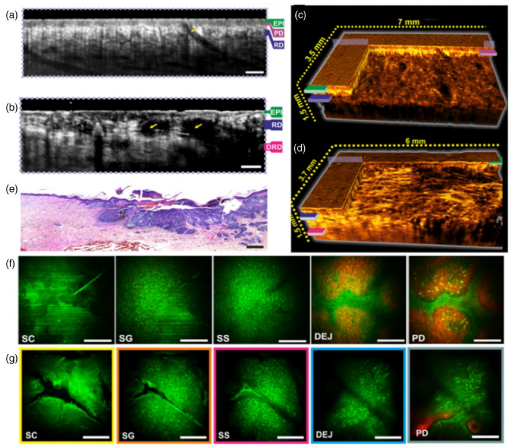



Fig. 11 In vivo clinical MPT/OCT in basal cell carcinoma in the dorsal forearm. (a) and (c) Two-dimensional (2-D) and 3-D OCT of normal skin. (b) and (d) 2-D and 3-D OCT of basal cell carcinoma. (e) Histology of excised basal cell carcinoma after imaging. (f) MPT of normal skin at different depths. (g) MPT of basal cell carcinoma. MPT was acquired in the center of 3-D OCT. SC, stratum corneum; SG, stratum granulosum; SS, stratum spinosum; DEJ, dermalepidermal junction, and PD, papillary dermis.

in a fluorophore is extremely low. Hence, a high photon flux density is required. NIR femtosecond lasers have sufficient peak intensity to enable a high probability for two or three photon excitations raising the fluorescent molecule into an excited state. TPF can be detected from exogenous contrast agents like ICG or green-fluorescent protein (GFP). Label-free two-photon imaging is possible by selecting targeting intrinsic cellular molecules such as NADH and flavins. This approach is usually called two-photon autofluorescence imaging. Features extracted from NAD(P)H and keratin autofluorescence in TPF images have been used to differentiate between normal, precancerous, and cancerous squamous epithelia. The potential of molecular imaging with multiphoton excitation fluorescence was first demonstrated in 1990[160] by the scanning of chromosomes in dividing cells (LLC-PK-1; American Type Culture Collection) stained by

cellular DNA labeled with a UV excitable fluorescent stain (Hoechst 33258) and a mode-locked dye laser. Although TPF has enhanced our knowledge on how cells behave within intact tissue environments, it still has limitations inherent to fluorescence microscopy. It can only visualize fluorescing features with the FOV of a few hundred microns. Label-free MPM is also possible via three-photon processes and can be used to image tryptophan and serotonin. [176]

Unlike TPF, SHG is a parametric process which does not involve photon absorption and emission. [161] It only occurs in noncentrosymmetric molecules when the electric field of the excitation light is strong enough. Due to the intense laser pulse, a nonlinear polarization is generated. SHG is the lowestorder nonlinear optical process in which second-order nonlinear optical susceptibility is responsible for the generation of light at the second-harmonic frequency. The quadratic power dependence of SHG on the refractive index provides better optical contrast for imaging tissue structures than conventional linear reflectance microscopy. The first biological SHG imaging experiment dates back to 1986, showing the orientation of collagen fibers in a rat tail tendon. [177] Imaging can be performed with a single widely tunable NIR femtosecond laser. The successful combination of SHG with OCT was first reported in 2004 with an ultrafast Ti:sapphire laser with 110 fs pulse duration. [178] Two collagen layers were sandwiched among glass slides. Since then a vast number of implementations by other groups have been reported. [49,179] Collagen is the most abundant extracellular structural protein of the vertebrate body, making up $\sim 6\%$ of muscle tendons. [180] Collagen, in the form of elongated fibrils, is mainly found in bone, cartilage, skin, blood vessels, tendon, and cornea. Due to its longevity, it constitutes a prominent target of structural modifications usually associated with various physiological processes, such as diabetes, renal diseases, aging, and cancer. [181]

THG occurs even in centrosymmetric structures and can visualize interface heterogeneities as well as myelin sheaths in studies of the nervous system in animal models. [182] Since SHG and THG do not excite coupling of electronic levels, photobleaching can be suppressed, enabling observation of structures over extended periods. Squier et al. [162] demonstrated THG microscopy with dynamic living specimens for the first time in 1998 with a 100 fs excitation pulse at $1.2\mu$ m and 250 kHz repetition rate resulting in a signal at 400 nm arising from the cytoblasmic streaming and statolith movement of the chara plant rhizoids. Both are parameters for the vitality of the plant.

A number of groups have demonstrated a multimodal OCT and nonlinear optical imaging platform with Ti:sapphire lasers
to obtain structural and biochemical information from tissue. Beaurepaire et al. [183] were the first who demonstrated the potential of TD OCM and TPF microscopy with an ultrafast Ti:sapphire laser (Spectra Physics, Santa Clara, California) with a pulse duration of 80 fs and 82 MHz repetition rate. The possibility of coupling functional and structural imaging based on the different contrast mechanisms derived from fluorescence and backscattered light was demonstrated in vivo with a drosophila embryo transfected to express GFP in its cell nuclei. Furthermore, the combination of OCT and fluorescence spectroscopy has been reported for diagnosis of carcinoma in esophageal and cervical tissue by Kuranov et al. [184] Pan et al. [185] were the first who reported an experimental study enhancing efficiency and sensitivity of early bladder cancer diagnosis with an aminolevulinic acid hydrochlorid (ALA) fluores-cence-guided endoscopic OCT for ex vivo rat bladder imaging in the rat model. In vivo bladder imaging was performed as well. [186] Jiang et al. [178] used a femtosecond laser pulse with 110 fs pulse duration and 76 MHz repetition rate to generate secondharmonic signals from collagen harvested from rat tail tendon sandwiched among glass slides and combined it for the first time with OCT. This technique added contrast and enhanced the resolution of OCT due to the strong dependence of SHG on molecular and tissue structures. In the same year, Yeh et al. [187] combined OCT with a superluminescent diode centered at 1310 nm and MPM with an ultrafast Ti:sapphire laser (150 fs, 800 nm ) in two separate systems and compared it with histopathology to monitor laser

thermal injury and subsequent wound healing in organotypic, skin-equivalent tissue models. The in vitro skin-equivalent rat tissue model was composed of dermis with type I collagen and fibroblast cells and epidermis of differentiated keratinocytes. High-resolution MPM imaging using SHG revealed alterations in collagen microstructure organization with subsequent matrix reconstruction. Fibroblast cell migration in response to injury was monitored by MPM using TPF. Vinegoni et al. [188] reported an integrated microscope that combined SD OCM, TPF, and SHG signals to allow for simultaneous acquisition of structural and functional information based on a single Ti:sapphire laser with a $60-$ nm bandwidth and $80-$ MHz pulse repetition rate (Femtolasers, Vienna, Austria). An average sample power between 1 and 10 mW was focused with a $20 \times 0.95$MO onto fibroblast cells transfected with GFP-labeled vinculin and stained with a nuclear dye (Hoechst 33342). Also, dynamic cell-scaffold interactions under more physiological 3-D conditions and in vitro smooth muscle tissue from a transgenic GFP mouse were studied.

A noninvasive dual-modality imaging system was demonstrated that incorporated OCT and fluorescence lifetime imaging microscopy (FLIM) capable of simultaneously characterizing the 3-D tissue morphology and its biochemical composition into a single optical system for the first time. [189] A spectrom-eter-based OCT based on a SLD centered at 830 nm provided high-resolution morphological volumetric tissue images with an axial and lateral resolution of 7.3 and $13.4\mu$ m, respectively was combined with a pulsed Q-switched tripled Nd:YAG laser for excitation. The multispectral FLIM based on a direct pulse-recording approach (upon 355 nm laser excitation) provided 2-D superficial maps of the tissue autofluorescence intensity and lifetime at three customizable emission bands with a $100-\mu$m lateral resolution. Both systems shared the same excitation/illumination optical path to facilitate the coregistration of OCT volumes and FLIM images. The
developed OCT/FLIM system was capable of a maximum A-line rate of 59 kHz for OCT and a pixel rate of up to 30 kHz for FLIM and was validated with standard fluorophore solutions and subsequently applied to characterize two biological tissue types: postmortem human coronary atherosclerotic plaques, and in vivo normal and cancerous hamster cheek pouch epithelial tissue. Sets of biomarkers in a Syrian hamster cheek pouch model of epithelial cancer were identified. OCT provided morphological information for cancer, including thickening of the epithelial layer and loss of layered structure, and FLIM provided chemical information, including increased nicotinamide adenine dinucleotide and reduced collagen emission, to discriminate benign lesions. Additional lifetime characteristics provided a better quantification of tissue biochemistry, as lifetime measurements can resolve fluorophores with overlapping spectra, such as collagen and NADH. The fluorescence signal was detected with a microchannel plate-PMT and simultaneous multispectral time-resolved fluorescence detection was achieved by separating the fluorescence emission in three bands to selectively quantify the fluorescence emission of collagen ( $390\pm20$ nm ), NADH ( $452 \pm 22.5$ nm ), and flavin adenine dinucleotide (FAD, $550 \pm 20$ nm ).

Multimodal OCT setups were also used to investigate the eye. Dobre et al. [190] developed a dual-channel OCT-ICG fluorescence system based on an ophthalmic OCT confocal imaging system with the confocal channel tuned to the fluorescence wavelength range of ICG. A SLD was implemented to visualize en face OCT slices and corresponding ICG angiograms of the ocular fundus simultaneously, demonstrating the potential for a more complete view of the pathologic conditions in ocular diseases. [191] A combined nonlinear optical microscopy and OCM imaging system based on an ultrabroad bandwidth Ti:sapphire laser (sub 10 fs ) was used to measure the mechanical and microstructural response of the central cornea from New Zealand white rabbits postmortem. [192] Autofluorescence and SHG imaging were performed to image cornea anatomy in normal and diseased states and to characterize the collagen microstructural response to changes in intraocular pressure. A trimodal microscopy with MPM, OCM, and OCT for multiscale and multicontrast imaging was demonstrated in leaves and fish cornea samples (cf.

Fig. 12) by a sliding for rapid scanning over a large area and a high-resolution zoom-in for cellular-level structures on regions of interest. [193] Cross-sectional OCT over millimeter FOV and en face high-resolution MPM with submicrometer resolution from the same sample position could be obtained by switching the focusing objective. All layers of a fish cornea were imaged at a frame rate of $\sim$ 100 frames /s with OCT and cells and collagen fibers were visualized at a frame rate of 0.4 frames /s. Dai et al. [194] demonstrated in a proof of principle study with an ultrafast Ti:sapphire laser simultaneous OCT and autofluorescence microscopy to address biomolecules like lipofuscin. Ex vivo ocular tissues and a rat retina in vivo were successfully imaged. A single 12 fs broadband (100 nm) source (Femtolasers) was used to combine threechannel MPM and OCT in a single platform by Tang et al. [195] The ultrafast Ti:sapphire laser was dispersion compensated to simultaneously provide short pulses necessary for efficient MPM excitation and the broad bandwidth required for high-resolution OCT. MPM and OCT channels were coregistered with a lateral resolution of $\sim 0.5\mu$ m and an axial resolution of $\sim 1.5\mu$ m, showing the potential of providing simultaneous functional and structural information from cells and an extracellular
matrix with a 3-D organotypic epithelial tissue model. SHG revealed information about the cellular and extracellular collagen matrix, the autofluorescence signal added information from the cell body of fibroblasts and OCT visualized the internal structure.

Yazdanfar et al. [196] demonstrated a trimodal microscope for multifunctional imaging of endogenous contrast by combining the functionality of OCM, TPF, and SHG in fluorescent microspheres and ex vivo skin tissue. A combined instrument of two imaging modalities to acquire information on the cardiac function in larval Drosophila melanogaster was demonstrated. [197] The imaging platform was able to sequentially provide cross-sectional OCT and C-scan laser scanning fluorescence microscopy images with a SLD and a continuous-wave solidstate laser (FDC488, JDSU, Milpitas, California) operating at 488 nm , where the major excitation peak of GFP is located. With this dual-imaging system, the heart could be easily located and visualized within the specimen and the change of the heart shape in a cardiac cycle could be monitored.

Various laser sources were integrated into multimodal MPM and OCT to optimize the performance for such platforms. Since the skin and eye are easily accessible with OCT, many studies were related to the skin or eye. A dual-spectrum laser source based on fiber continuum generation in combination with a widely tunable Ti:sapphire laser was used for in vivo OCT and TPM high-resolution tissue imaging to target a broader range of fluorescent markers. [198] Jeong et al. [199] proposed a combined TPF and OCT using individually optimized sources. TPM was performed with a Ti:sapphire laser and provided molecular and cellular information. OCT was performed with a swept source centered at 1300 nm and provided simultaneous largearea screening of simple microsphere specimens to demonstrate the spatial coregistration in the small intestine and ear tissues of mouse models ex vivo. König et al. [200] reported the first clinical study with a variety of skin diseases and disorders, such as skin cancer, psoriasis, hemangioma, connective tissue diseases, pigmented lesions, and autoimmune bullous skin diseases. Two different OCT systems were combined with MPT and dermoscopy to provide 2-D color images of the skin surface. A stack of cross-sectional OCT wide-field images with a typical FOV of $5 \times 2$ mm$^2$ gave fast information on the depth and volume of the lesion. MPT provided $0.36 \times 0.36$ mm$^2$ horizontal/diagonal optical sections within seconds of a particular region of interest with superior submicron resolution down to tissue depths of $200\mu$ m.[201] The clinical diagnostic potential of this multimodal MPT/OCT system in dermatology had been demonstrated in a further study by acquiring sequential 3-D MPT and OCT images of various dermatological disorders [202] (cf. Fig. 11). Recently, a combined MPT/OCT system capable of simultaneously acquiring high-resolution OCM and MPM was demonstrated in a single platform to enable simultaneous acquisition and coregistration for in vivo skin imaging and dynamic imaging of

blood flow in capillaries. [203] A widely tunable ultrafast Ti:sapphire laser was used to excite the endogenous proteins NADH and FAD at 760 nm . A part of the Tisapphire laser beam was launched into a photonic crystal fiber for high-resolution OCT for structural imaging. [203] Furthermore, in vivo 3-D multimodal optical imaging was used to search for evidence of epidermal regeneration by bone marrow (BM)-derived cells during various skin repair processes. [204] In vivo imaging was demonstrated for visualizing GFP BM-derived cells in the skin, as well as for detecting collagen in the dermis layer. OCT structural and

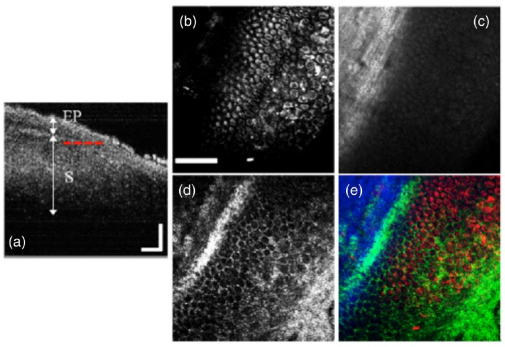



Fig. 12 OCT, multiphoton microscopy (MPM), and optical coherence microscopy (OCM) images from fish cornea sample. (a) Large field of view (FOV) cross-sectional OCT image where the epithelium (EP) and stroma ( S ) could be differentiated. Dashed line shows the region where the MPM/OCM is performed. (b) to (d) High-resolution two photon excitation fluorescence (TPF), second harmonic generation (SHG), and OCM images. (e) Overlay of the TPF, SHG, and OCM images with color codes of red, blue, and green, respectively. Scale bars in (a) and (b) represent 100 and \$50 \mu \mathrm{˜m}\$, respectively. (b) to (e) share the same FOV. [193]

microvascular imaging was demonstrated for assessing the state of skin following wounding and grafting. This study demonstrated the potential of in vivo multimodal imaging to contribute to a better understanding of the general mechanisms by which skin regeneration is influenced by BM-derived cells. Since multimodal imaging without an endoscope or fiber delivery probe is mainly limited to cell, skin, and ocular imaging, thorough efforts have been undertaken to integrate an endoscope or build a fiber delivery probe. Pan et al. [185] suggested that an ALA fluorescence-guided endoscopic OCT could enhance the efficiency and sensitivity of early bladder cancer diagnosis and demonstrated in animal studies that the specificity of fluorescence detection of transitional cell carcinoma was significantly enhanced by fluorescence-guided OCT. Tumlinson et al. [205] have demonstrated a combined miniaturized OCT and laser-induced fluorescence (LIF) spectroscopy imaging catheter for in vivo mouse colon imaging to monitor the disease progression in mouse colon longitudinally and to identify colorectal adenomatous mouse colon. [206–208] OCT images showed significant changes in mucosal layer thickness and loss of visibility of tissue boundary lines between healthy and diseased tissue. Similarly, the LIF spectra of healthy and diseased tissues were distinguishable by the emissions at 405 and 680 nm . The autofluorescence signal measured by LIF provided biochemical information that was correlated to a dietary metabolite. Coregistered OCT and MPM enabled simultaneous imaging of tissue morphology and molecular information of mouse intestinal tissues at high resolution ( $10\mu$ m ) over a 2 to 3 mm FOV with the potential for small animal imaging and clinical imaging. A fiberbased multimodal imaging system that combined MPM with FD OCT simultaneously was suggested. [209] It consisted of a fiber laser centered at $1.04\mu$ m, fiber couplers, and a double-clad fiber and
was tested with a thin slice of fixed rabbit heart. OCM and MPM shared the same excitation

light path in the core of a double-clad fiber. The OCM signal was collected by the core without significant losses and the MPM by the clad of the double-clad fiber with 20% collection efficiency. Xi et al. [210] reported a compact all-fiberoptic scanning multimodal endomicroscope capable of simultaneous OCT and TPF imaging. Both imaging modalities shared the same miniature fiber-optic scanning endomicroscope consisting of a double-clad fiber for the $1310-\mathrm{nm}$ swept source for en face OCT and a $1550-\mathrm{nm}$ light source (passive mode-locked) for TPF. The endomicroscopy system was used to image in vitro cell cultures and ex vivo tissue with ICG as fluorescence dye (cf. Fig. 13).

Multimodal and multiscale imaging has found widespread biomedical applications and will further advance with improved optical probes. It can be used to increase the understanding of the molecular and functional changes associated with disease progression. TPF, SHG, THG, and CARS can be seamlessly integrated into a single, unified microscopy platform and OCT can easily be added to obtain complementary structural and functional information within tissue samples and cells. The progress of multimodal optical imaging platforms has been facilitated by advances in ultrafast laser technology, high-performance specialized optical filters, and high-sensitivity detectors. Proper coregistration is needed for large-area prescreening with zoom-in function for chemical specific information.

### 5.2 Dual-Modality Optical Coherence Tomography and Photoacoustic Imaging

Having been clinically readily investigated and applied in ophthalmology [211] and intravascular imaging [212] for over a decade,

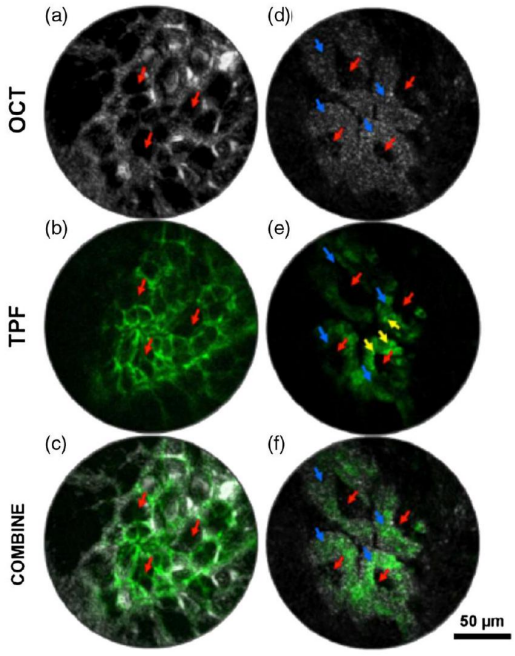



Fig. 13 (a) OCT, (b) TPF, and (c) superposed images of mouse adipose tissue with local indoyanine green (ICG) administration. Each red arrow in (a) and (b) indicates one of the adipocytes visualized under both imaging modalities. (d) OCT, (e) TPF, and (f) superposed images of mouse small intestine tissue with local ICG administration. Blue arrows (right downward) shown in (d) and (e) indicate villus structures and red arrows (left downward) indicate lacteals. The stronger fluorescence dots indicated by yellow arrows at center shown in (e) may be either enterocytes or lymphocytes. Both sets of images show great correlation between two imaging modalities. ${ }^{210}$

OCT has gained great success, and has become even more successful with the ever increasing imaging speed and finer resolution introduced by new optical sources, which enabled OCT to achieve MHz A-line rate. [213] However, due to the highly scattering nature of biological tissues and the contrast mechanism of OCT, which relies on the difference of tissue scattering properties, the penetration depth is limited to be within a few millimeters in most cases, even with the $1300-$ nm sources whose beam gets less scattered than the $800-$ nm sources. [4] For interconnected soft tissues whose scattering properties are similar, OCT's contrast is also hampered. In order to overcome these disadvantages, efforts have been made to combine OCT with other imaging modalities, such as MPM (Ref. 183) and confocal fluorescence microscopy (CFM), [214] but both MPM and CFM use fluorescence as the contrast mechanism and cannot give OCT any extra imaging depth. Some advanced OCT modalities, such as Doppler OCT[215] and polarization-sensitive OCT (Ref. 216), can generate fluid flow information and bring extra differentiation capability by exploiting the birefringence of biological samples, but the inherent contrast stemming from OCT and the imaging depth limit set by the scattering tissues are yet to be breached.

Compared to the well-developed and mature optical imaging modality OCT, photoacoustic imaging is a relatively novel imaging modality that has attracted much research interest in the past decade. Photoacoustic imaging uses the photoacoustic effect-a certain chromophore absorbs a short laser pulse's energy; this absorbed energy is partially converted to heat; then under thermoelastic expansion, the heat is further converted to a local pressure rise and, hence, an acoustic wave-the photoacoustic wave; by detecting the laser pulse-induced photoacoustic waves, one can reconstruct in 3-D the chromophore distribution in the illuminated volume. [217] Different from OCT that detects backscattered light, photoacoustic imaging detects acoustic waves. Since the photoacoustic waves are inherently broadband, detection of different bandwidths of the photoacoustic wave can generate a variety of different imaging depths ranging from several hundred micrometers by submicron-resolution PAM (Ref. 218) to the centimeter level in whole-body illumination PAT with a resolution of $\sim 100\mu m$.[219] A table comparing OCT and photoacoustic imaging is given below in Table 3 for a brief comparison of some important features of these two modalities.

From Table 3 we can see that OCT and photoacoustic imaging systems have distinctive but complementary contrast mechanisms. While most PAM imaging systems have imaging depth comparable with that of OCT, PAT systems can image beyond the optical diffusion regime. In optical-resolution PAM (ORPAM), [220] the lateral resolution is determined by the focal spot size of the excitation laser beam, therefore, it is comparable to that of OCT. The complementary information of these two modalities makes it interesting to compare the results for the

Table 3 Comparison of OCT and photoacoustic imaging.

| | OCT | Photoacoustic in |
|---|---|---|
| Contrast mechanism | Difference in tissues' optical scattering properties | Difference in tiss |
| Imaging depth | $\sim 2$ mm | $\sim 0.42$ to $> 10$ r |
| Lateral resolution | Diffraction limited, generally within a few tens of micrometers | from $\sim 0.5$ to $>$ |
| Axial resolution | Determined by laser source, generally within $10\mu$ m | Determined by t |
| Imaging speed | Video rate | Generally slow d |

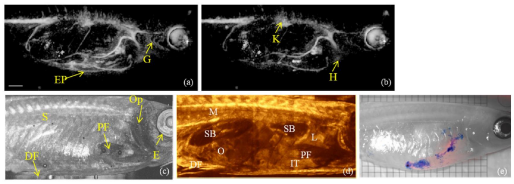



Fig. 14 (a) and (b) are photoacoustic tomography maximum intensity projection (MIP) images of a zebrafish expressing E2-Crimson in its exocrine pancreas imaged at 611 nm and 640 nm , respectively; (c) and (d) are OCT MIP and B-scan images of the same fish, respectively; (e) is an epifluorescence image showing the E2-Crimson labeled exocrine pancreas (red) overlaid with the E2-Crimson distribution calculated from photoacoustic images. G: gills; EP: exocrine pancreas; K: kidney; H: heart; S: spine; DF: dorsal fin; PF: pectoral fin; OP: opercular; E: eye; M: muscle; SB: swim bladder; O: oocyte; L: liver; IT: intestinal tracts. ${ }^{222,223}$

same sample imaged by these two methods. Multiple studies have employed these two techniques for the same object using separate OCT and photoacoustic imaging systems, such as the study of chemically induced bovine cartilage osteoarthritis where OCT provides the general structural background and ORPAM shows the locations of clefts and fissures in the cartilage due to osteoarthritis, [221] or the zebrafish exocrine pancreas imaging where an FD OCT system gives the morphological background of the adult zebrafish, while PAT resolves the fluorescent protein (FP) labeled exocrine pancreas and the blood vessels [222] (cf. Fig. 14), or the clinical test of port wine stain (PWS) depth by using OCT to measure the epidermal thickness and using photoacoustic imaging to test the epidermal-to-PWS distance. [224]

All these reports justify the benefits of combining these two imaging modalities together. This section reviews the various implementations of dual-modality OCT/photoacoustic imaging system designs, their characteristics, and their applications in biomedical studies, as well as the technical challenges of the dual-modality systems and their potential applications.

### 5.2.1 Implementations of dual-modality OCT/photoacoustic imaging system

Due to the various submodalities of OCT and photoacoustic imaging, different implementations of OCT/photoacoustic imaging systems have been tested and published. In this section, three major categories are reviewed, namely OCT/photoacoustic endoscopy, OCT/PAM, and OCT/PAT, each having unique features and applications.

Combined OCT/photoacoustic endoscopy. Driven by the increasing need for internal body imaging, OCT endoscopes were introduced in the 1990s and demonstrated their feasibility in inner cavity imaging. [225] This specific type of OCT has since undergone a series of improvements in terms of miniaturization, imaging speed, and safety. Compared to OCT endoscopy, photoacoustic endoscopy has suffered a prolonged development phase due to the technical challenges of the comparatively thick multimode fiber's flexibility [226] as well as the detection of photoacoustic waves by miniaturized probes. Real in vivo photoacoustic endoscopy imaging was accomplished only recently. [227] Despite these challenges, a few designs of combined OCT/photoacoustic endoscopes have been tested, as are shown in Figs. 15 and 16. One major difference between the two designs is how the excitation light for photoacoustics and the OCT sample beam are delivered to the sample. In Fig. 15, we can see that the excitations for photoacoustic imaging and the OCT sample arm are delivered by separate fibers while an ultrasound transducer takes additional space in the probe design.

This design makes the outer diameter of the endoscope 5 mm . Compared to the design shown in Fig. 15, Fig. 16 uses a combined OCT/photoacoustics probe, of which OCT and photoacoustic imaging share the same light delivery fiber to the sample. The use of a gradient

index (GRIN) lens and a prism pair enables the ultrasound transducer to be mounted right on top of the prism pair and reduces the probe diameter to 2.3 mm .

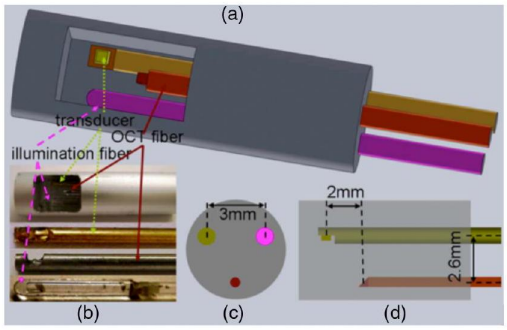



Fig. 15 Design of a combined miniaturized OCT/photoacoustic endoscope head. (a) Sketch of the general design. (b) Photograph of the OCT fiber, photoacoustic illumination fiber, and the transducer, respectively. (c) Relative placement of the probe. (d) Side view of the probe head. ${ }^{228}$

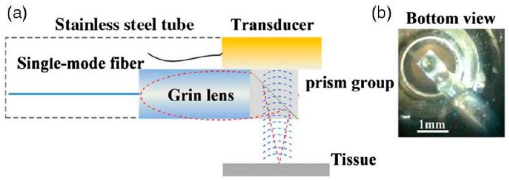



Fig. 16 An OCT/photoacoustic probe design employing a single fiber. (a) Schematic of the design (red dashed line: photoacoustic excitation beam; blue dashed line: photoacoustic wave). (b) Photograph of the bottom view of the probe. ${ }^{233}$

With a few to several hundred micrometer lateral resolution of the photoacoustic part achieved from the design in Fig. 15 and an SNR of 25 dB for the photoacoustic part in Fig. 16, along with a mechanically scanned stepper motor for 2-D and 3-D imaging, these two dual-modality OCT/photoacoustics miniaturized probes are tested by imaging ovarian tissue and mouse ear, respectively. However, the imaging speed is quite limited by the use of a stepper motor. Unlike the circumferential cavity images quite often seen from OCT endoscopic studies, these two studies gave only traditional B-mode images. The imaging speed, sensitivity, and, hence, application are quite limited for these preliminary prototypes.

Combined $OCT/PAM$. Several different implementations of OCT/PAM dual-modality imaging system have been introduced since 2009, including an SD OCT system combined with a transmission mode OR-PAM system, [229] an SD OCT system combined with a laser scanning optical resolution PAM (LSORPAM) system, [230] and a dual-modality system in which OCT and PAM use the same NIR supercontinuum laser source [231] or the same visible laser source. [232] Two representative designs are given in Figs. 17(a) and 17(b).

The resolutions of the system in Fig. 17(a) for OCT are $20\mu$ m laterally and $6\mu$ m axially, while those of the LSORPAM are 2.8 and $23\mu$ m.

Using a laser source with a much wider bandwidth and proper focusing, the OCT resolutions in Fig. 17(b) reach $5\mu$ m laterally and $5.9\mu$ m axially. A 75 MHz central frequency ultrasound transducer in the second generation (G2) OR-PAM system shown in Fig. 17(b) was able to achieve a lateral resolution of $5\mu$ m, the same as that of the OCT system combined with it, and an axial resolution of $14\mu$ m. The use of a pair of scanning mirrors to direct the PAM

excitation beam, which is given in the schematic of Fig. 17(a), enables this system to perform fast scanning over a sample. This is compared to the design in Fig. 17(b) that uses linear stage for scanning a better choice in regard of imaging speed. However, the G2 OR-PAM design in Fig. 17(b) uses a rhomboid prism and an acoustic lens to maximize the sensitivity, [220] and the FOV can also be much larger than that of Fig. 17(a).

The OCT/PAM systems, represented by the designs shown in Fig. 17, have been applied in in vivo studies, such as the metabolic rate of oxygen in mouse ear, [234] the neovascularization in 3D porous scaffolds, [235] epilepsy mapping with high temporal and spatial resolutions and dual optical contrasts, [236] and retinal imaging in rats. [237] Results of the OCT/PAM imaging of the neovascularization in mouse ear is given in Fig. 18 as an example.

The OCT/PAM systems can well demonstrate the advantages of the dual-modality imaging system by the comparable resolutions of OCT and OR-PAM, the similar imaging depth of 1 to 2 mm , and the complementary optical imaging contrast. With the help of a spectroscopic PAM system, which is capable of extracting the concentration of a certain absorber [238] and the fluid flow rate that can be calculated by Doppler OCT, [234] OCT/ PAM systems can be used not only in morphological and angiogenesis studies, but also in functional and dynamic biological studies.

Combined OCT/PAT. As is reviewed in the previous section, OCT/PAM generally cannot exploit the depth advantage of photoacoustic imaging. Also, LSOR-PAM has the inherent problem of a very limited FOV within only several square millimeters, and G2 OR-PAM is limited in imaging speed by the

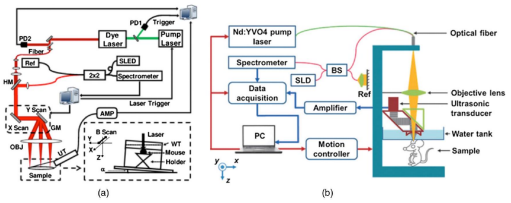



Fig. 17 Representative designs of the OCT/PAM system. (a) combines a spectral domain (SD) OCT system with a laser scanning optical resolution PAM system. The sample is placed in a tilted water tank for DOCT measurement. ${ }^{234}$ (b) employs a second generation optical resolution PAM (OR-PAM) system with an SD OCT system. [235] PD, photodiode; SLED, superluminescent emitting diode; WT, water tank; UT, ultrasound transducer; OBJ, objective; GM, galvanometer scanning mirrors; HM, hot mirror; Ref, OCT reference arm; SLD, superluminescent diode; BS, beam splitter.

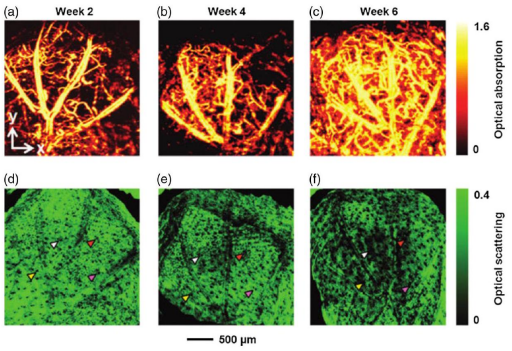



Fig. 18 Maximum amplitude projection (MAP) images of OR-PAM [(a) to (c)] and OCT [(d) to (f)] showing chronic vasculature and scaffold/tissue construct. The scaffolds can be seen in the OCT images clearly while the blood vessels stand out of the background in the PAM images. The arrowheads indicate scaffold degradation caused pore size increase. ${ }^{235}$

slow mechanical scanning. Although fast B-mode scanning is made possible by using a voice coil for OR-PAM (Ref. 239) and is applicable in a combined OCT/PAM system, it only enables fast scanning in a certain frame. To fully use the beyond optical mean free path feature of photoacoustic imaging, and to enable a large FOV with coregistered OCT/photoacoustics dual-modality imaging, an OCT/PAT system was introduced and tested in vivo. [240]

Since this report, the system has been further developed and an SS OCT system with higher bandwidth has been incorporated with the optical detection whole-body illumination PAT system. The newest design of the system is given in Fig. 19. The working mechanism of the optical detection PAT system can be found in Ref. 222. In brief, a multimode fiber (green line in Fig. 19) delivers the excitation pulse to illuminate the whole sample sitting on top of the Fabry-Perot interferometer (FPI) sensor, where distilled water is used as the acoustic couplant. The interrogation beam from the telecom wavelength continuous wave laser probes the FPI sensor in a point-by-point manner and detects the acoustically modulated reflectivity change. This detected modulation linearly correlates to the intensity of the photoacoustic waves, and by using 3-D PAT reconstruction algorithms, [241] the absorber distribution can be mapped in 3-D. For the SS OCT system, a dual-balanced detector (DBD1) is used to compensate for the jitter of the swept source, while DBD2 can extract the interferogram for further postprocessing and image reconstruction. Both systems are controlled by homemade LabVIEW control applications.

This OCT/PAT system has an FOV that is limited only by the size of the FPI sensor, which, in this case, is 3 cm $\times$ 5 cm. The sensitivity of the SS OCT part of the system achieves 95 dB . While the imaging depth of SS OCT does not go beyond a few millimeters, the PAT part gives an effective penetration depth of $>$ 5 mm in soft tissue. Though sub- 100$\mu$ m resolution is achievable both axially and laterally for the PAT part, due to the low repetition rate of the optical parametric oscillator (OPO) used in our system ( 50 Hz ), a scan step size of 100$\mu$ m is generally applied to reduce the imaging time. Since the imaging time is two orders of magnitude different between the two modalities, the OCT/PAT system runs in a sequential order. The results of in vivo human palm imaging using an OCT/ PAT system are given in Fig. 20. From the human palm imaging results, we can see that the vasculature is very well resolved by PAT, including detailed branching information. At the same time, OCT shows the stratum information of skin and the layers in the skin can be distinguished from the OCT B-mode image. The en face projection images and the 3-D volume renderings of the fused OCT/PAT images perfectly demonstrate the dual optical contrast and complementary information feature of OCT/ PAT. They also illustrate the large scanning area easily achievable by using the optical detection mechanism for photoacoustic imaging.

### 5.2.2 Challenges of OCT/photoacoustic imaging

Having drawn research interests over the past few years, combining OCT and photoacoustic imaging still has some challenges to overcome. The very first one is the imaging speed mismatch. In order to perform real-time functional studies of dynamic biological processes, and also to reduce motion artifact, imaging speed is very critical. OCT, with faster swept sources coming to the market and newer wavelengths being introduced, has no problem when combined with photoacoustic imaging to reach video rate over a large scanning area. However, mechanical scanning of the probe head or the fast scanning available only in single slices for the photoacoustic imaging drags down the imaging speed of the combined dual-modality system. Although LSOR-PAM claims fast scanning speed, the effective FOV is very limited. The obliquely positioned

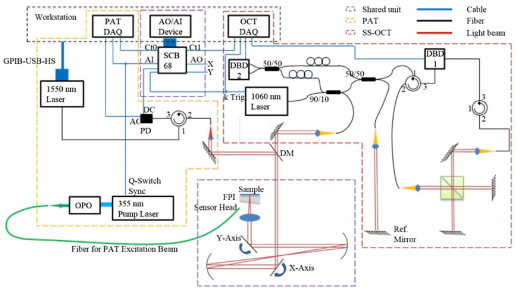



Fig. 19 Schematic of the OCT/PAT system. An optical detection PAT system using an Nd:YAG pumped OPO as the whole-body illuminating source is combined with an SS OCT system featuring a $1060-\mathrm{nm}$ swept source with 100 nm bandwidth and 100 kHz sweep rate. Data acquisition of the two systems is performed with individual data acquisition devices (DAQ) and is synchronized with the scanning of the 2D galvanometer scanning mirrors with the AO/AI devices. Both the OCT part and the PAT part share the same scanning unit with the sample sitting on top of the Fabry-Perot polymer film sensor for coregistered imaging. AO, analog output; AI, analog input; OPO, optical parametric oscillator; FPI, Fabry-Perot interferometer; Ref, reference arm; DBD, dual-balanced detector; PD, photodiode; Ct, counter; DM, dichroic mirror. [223]

unfocused ultrasound transducer further reduces the sensitivity of the photoacoustic part when combined with OCT. Though 2D fast scanning, using a focusing transducer in its focal spot in the reflection mode OR-PAM has been reported [242] and the small focal spot can be mosaicked to form a larger FOV. [243] It has to use a comparatively low-bandwidth ultrasound transducer for its relatively larger focal spot size, which in turn reduces the achievable axial resolution dramatically that is acoustically determined. Also, the nonuniform piezoelectric response of the low-bandwidth transducers to the photoacoustic sources in their focal spots makes it difficult to normalize the signals acquired in the FOV and, hence, is problematic in spectroscopic applications, where it is necessary to extract chromophore concentrations in photoacoustic imaging. The OCT/PAT system suffers the same imaging speed issue due to the low repetition rate of OPO with only 50 Hz currently. Therefore, OCT/photoacoustic imaging systems with the features of real time and large FOV can be a future direction. The second challenge could be the combination of different submodalities of OCT with photoacoustic imaging. Currently, TD OCT, SD OCT, and SS OCT have been tested to combine with different photoacoustic systems. One study already showed the benefit of using Doppler OCT with PAM for the metabolic rate of oxygen measurement. [234] But in their system, the water tank has to be tilted for the flow rate measurement. Newer three-beam Doppler OCT, [244] if it can be combined with PAM, will make the measurement of flow rate easier without a priori knowledge.

Polarization-sensitive OCT combined with PAM or PAT is also of special interest due to the extra contrast that can be retrieved by polarization-sensitive OCT. If the detection of the optical detection whole-body illumination PAT system can be parallelized, meaning a line's or a field's photoacoustic waves can be detected simultaneously, then combining line-field OCT or full-field OCT or OCM with PAT will revolutionize the imaging speed and enable simultaneous functional OCT/photoacoustic imaging.

### 5.2.3 Outlook of OCT/photoacoustic imaging

As a hybrid imaging modality, OCT/photoacoustic imaging brings the benefits of OCT and photoacoustic imaging together. Although facing a few technical challenges, the application of OCT/photoacoustic imaging, from endoscopy to whole-body tomography, still has a lot to explore. Not only would it be potentially powerful in all the biomedical, preclinical, and clinical fields where OCT and photoacoustic imaging have been investigated, knowing that OCT has already been reported for biometric identification studies, [245] OCT/photoacoustic imaging may also be used in biometric applications. In summary, in light of the new laser sources, new detection methods, and new applications, OCT/photoacoustic dual-modality imaging is highly expected to get more and more research and application interests in the years to come.

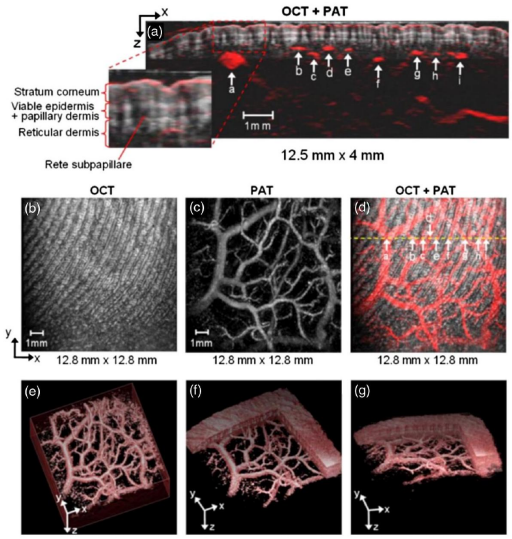



Fig. 20 Human palm imaging using OCT/PAT. (a) B-mode image where OCT differentiates the layered structures of the skin while PAT shows the blood vessels beneath. (b) En face OCT MAP image in which the ridges of the skin can be clearly seen. (c) PAT MAP image giving the distributions of the blood vessels over a depth range of 5 mm . (d) fuses (b) and (c). (e) to (g) 3-D representations of the imaged palm. ${ }^{240}$

## 26.4.5 Appendix Basic Terminologies and Key Performance Parameters in Optical Coherence Tomography

Optical coherence tomography (OCT) systems are usually based on the Michelson interferometer. The light from the source is split by a beam splitter into the sample and the reference arm. The interference of the backreflected light from the sample and the reference mirror is then captured by a detector. OCT systems can be broadly classified as follows (Fig. 21).

### A1 Time Domain OCT

In this case, a broadband light source is used. Due to the low coherence of the light source, the interference signal is obtained only when the optical path length of the sample and reference arm is matched to within the narrow coherence length. The reference arm is then scanned to match the optical path length of the reflections from within the sample. The recorded interference signal at different depths or relative time delays between reference and sample is then demodulated to generate a reflectivity depth profile or A-scan.

### A2 Fourier Domain OCT

In this case, the spectrum of the interference signal is detected. Fourier domain (FD) OCT can be futher classifed as spectrometer based and swept source (SS) based. In the case of a spectrometer-based system, a broadband light source is used and the interference signal is split into different optical frequencies using a diffraction grating and detected by a one-dimensional (1-D) CCD array as shown in Fig. 21, whereas in case of a swept source-based system, the interference signal is detected by a photo diode as the spectrum of the light source is scanned with a narrow instantaneous line width in time. The detected intensity spectrum is then Fourier transformed to obtain the depth-resolved reflectivity profile (A-scan) of the sample as shown in Fig. 21.

The key technological parameters associated with the OCT are described briefly below.

### A2.1 Axial/Longitudinal Resolution

In OCT, the axial resolution depends on the coherence length of the light source. The coherence length is the spatial width of the field autocorrelation measured by the interferometer. Since the light travels twice the distance to a given sample arm interface, OCT axial resolution is given by the round trip coherence length, which is half the standard coherence length. The envelope of the field correlation is equivalent to the Fourier transform of the power spectrum. The axial resolution is, therefore, inversely proportional to the spectral bandwidth of the light source as $\Delta z = 2\ln(2)\lambda_o^2/(\pi\Delta\lambda)$, where $\Delta\lambda$ is the FWHM bandwidth
and $\lambda_o$ is the center wavelength of the light source (assuming a Gaussian spectrum, refer to Fig. 21).

### A2.2 Lateral/Transverse Resolution

Lateral resolution in OCT is independent of the bandwidth of the light source and, hence, independent of the axial resolution. It is limited by geometical optics constraints of the sample arm design, given by Abbe's criteria of the lateral resolution $\Delta x = 0.61\lambda_o/\mathrm{NA}$, where NA is the numerical aperture of the objective lens.

### A2.3 Depth of Focus

It is dependent on the NA of the imaging optics. It is twice the Rayleigh range $z_R$, given by $2z_R = 2\lambda_o n/\mathrm{NA}^2$, where $n$ is the refractive index of the sample. Hence, increasing the NA increases the resolution (see above) but decreases the depth of focus quadratically. Thus, in OCT, a compromise has to be found between the lateral resolution and the depth of focus.

### A2.4 Imaging Depth

In the case of FD OCT, it depends on the center wavelength $\lambda_o$ and the spectral resolution $\Delta\lambda_{\text{full}}/N$, given by $z_{\max} = \lambda_o^2 N/(4n\Delta\lambda_{\text{full}})$, where $\Delta\lambda_{\text{full}}$ is the total bandwidth of the light source and $N$ is the number of spectral data points. In the case of TD OCT, the imaging depth is given by the reference arm travel range.

### A2.5 Detection Sensitivity

OCT, being an interferometric detection technique, boosts the weak backscattered optical field by effectively multipling it with a strong reference field. In the shot noise limit, the power signal-to-noise ratio (SNR) performance of the OCT
system is given by $\text{SNR} = 10\log(\eta P_{\text{s}}/E_{\text{p}}B)$, where $P_{\text{s}}$ is the power at the detector backscattered from a single interface in the sample arm, which is proportional to the A-scan peak height, $B$ is the bandwidth of electronic detection, $\eta$ is the photodetector efficiency, and $E_{\text{p}} = h\nu$ is the energy of the photon associated with the center frequency $\nu$.$P_{\text{s}}$ is assumed to be much lower than the power from the reference arm. Sensitivity in OCT is defined as the inverse minimal sample reflectivity, which leads to an SNR of 1, i.e., $\Sigma = 1/R_{S,\,\min} \mid \text{SNR} = 1$. In the case of SS OCT, assuming a single sample reflector, the total power associated with the peak value of the interferometric portion of the A-scan after Fourier transform is, by Parvesal's theorem, half the sum of the power associated with each spectral point. In order to obtain an A-scan peak, $N$ instantaneous sample exposures are done by an illuminating swept source if $N$ spectral points are recorded. In the case of time domain (TD) OCT, the power associated with the peak value of an A -scan is due to a single exposure. The instantaneous exposure power (under the constraint of ANSI exposure) in the case of SS OCT can be same as in TD OCT. Thus, the power $P_{\text{ss}}$ associated with the peak value of an A-scan in the case of SS OCT is about $N$ times the power $P_{\text{td}}$ associated with the A-scan peak of TD OCT. [17] Usually with bandwidth $B$ being same for both TD and SS OCT (17), and $N \geq 1024$, a senstitivtiy advantage of $\sim$ 20 to 30 dB is obtained for SS OCT over TD OCT. In the case of a spectrometer-based system, the detection of the spectral interference signal is done with a 1-D CCD array as shown in Fig. 21. The A-scan in this case is acquired by a single exposure as opposed to the case of TD OCT, where reference arm scanning is required. However, the total power is distributed among $N$ channels of the spectrometer. Thus, the power associated with the peak after Fourier transform is $P_{\text{sd}} = P_{\text{td}} = P_{\text{ss}}/N$. However, the bandwidth $B$ in this case is reduced by a factor $1/N$(17), and therefore, the effective SNR is same as in the case of SS OCT. The bandwidth can be related to the

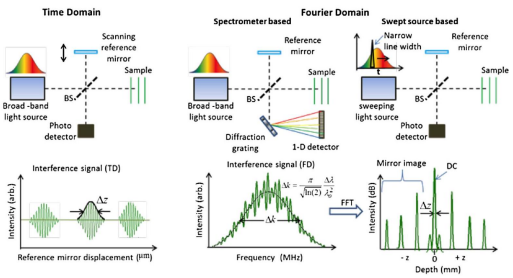



Fig. 21 Schematic of different OCT modalities. OCT systems can be classified into time domain (TD) and Fourier domain (FD) systems. FD OCT systems can be divided into spectrometer based and swept source based systems.

exposure or A -scan time $\tau$ as $\text{B} = 1/(2\tau)$. Hence, we see that FD OCT systems have 20 to

30 dB of sensitivity adavantage over a TD OCT system.

Obviously, there is a tradeoff between the electronic bandwidth or the data acquisition rate and the SNR. Faster data acquisition speed results in degradation of SNR. Furthermore for the shot noise limited performance, an optimum optical power output from the light source is required for sample illumination. However, the power of the illumination is restricted by the maximum permissible ANSI exposure limit. Thus, there is a limitation to the maximum power that can be applied during the short exposure time of a high-speed OCT system. Hence, the OCT system design should be optimized according to the optical power limitations and data acquistion speed requirement.

#### A2.6 Depth Penetration

OCT imaging at different wavelengths can be used to enhance tissue contrast and penetration, as well as to measure absorbing or scattering properties of various pigments and structures. Tissue scattering and absorption properties are strongly dependent on tissue morphology, such as cell and nuclei size, shape, and density, as well as the physiological state of certain chromophores. Hence, OCT penetration depth is significantly affected by light scattering within biological tissue, which is scaled as $1/\lambda_o^k$, where the coefficient $k$ is dependent on the size, shape, and relative refractive index of the scattering particles (14). The difference in tissue scattering, which usually dominates over chromophore absorption, provides structural contrast for OCT. Most biological tissues absorb in the visible and near-IR wavelength range because of the presence of hemoglobin and melanin. In the range between 0.8 and 1.8$\mu$ m, scattering and water absorbption are the predominant mechanisms limiting image penetration depth. Since scattering depends strongly on wavelength and decreases for longer wavelengths, significantly better image penetration depth can be achieved with light centered at 1.35$\mu$ m than at 0.85$\mu$ m. It has been shown that the optimum wavelengths for imaging in opaque biological tissues are in the range of 1.3 to 1.5$\mu$ m. In this range, imaging depths of 1 to 2 mm can be achieved (14). An exception is human retinal imaging, where water absorption becomes dominant due to the ocular media. There, the optimum wavelength is either between 750 and 850 nm , or in the region of 1.06$\mu$ m. Water absorption becomes dominant for wavelengths $> 1.8\mu$ m. A strong water absorption peak at $\sim 1.43\mu$ m can also reduce the spectral width of a broad bandwidth light source centered at 1.3$\mu$ m as a function of depth in biological tissue, since most soft tissues have 50 to 90% water content. Hence, by choosing the center wavelength and bandwidth of the OCT light source from within the range of 500 to 1500 nm , axial resolution and contrast as well as penetration may be adjusted to the imaging procedure and medical application.

# 27 Hypothetical Devices

## 27.1 Invisibility

(??? write how to make it to some extent???)

# Part IX
# Other Topics and Applications

## 28 Other Effects and Methods

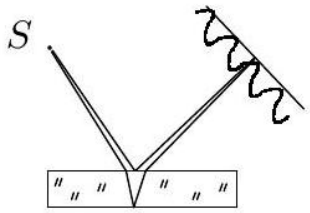
### 28.1 Interference and Tunneling of Radio Waves

#### 28.1.1 1. Introduction

В предыдущих разделах были представлены лабораторные работы, в которых изучались физические явления в видимом диапазоне электромагнитных (ЭМ) волн. К этому диапазону относится узкий интервал длин волн $\Delta\lambda \approx (4 \div 8) \cdot 10^{-5}$ см, в котором свет воспринимается глазом человека. Под светом, однако, понимают не только видимый свет, но и примыкающие к нему относительно широкие области ЭМизлучения: инфракрасную и ультрафиолетовую. Весь электромагнитный спектр принято делить на радиоволны, инфракрасное, видимое, улътрафиолетовое, рентгеновское и гамма-излучения. Подчеркнём, что эти участки спектра различаются не по своей физической природе, а по способам генерации и приёма излучения. Границы между участками условны и сами участки частично перекрываются. Различными, но, естественно, пересекающимися оказываются и области практического применения ЭМ-излучения.

В данном разделе предложены лабораторные работы, в которых такие общие для волн различной физической природы явления, как интерференция, отражение и преломление на границе раздела сред, наблюдаются в радиоволновом участке спектра электромагнитных волн. Переход к значительно более длинным, чем в видимом диапазоне, волнам упрощает реализацию заимствованных из оптики интерференционных схем, а также даёт возможность наблюдать важный эффект туннелирования волн при полном внутреннем отражении на границе раздела двух сред.

Для возникновения стационарной картины интерференции волн любой физической природы должно выполняться условие когерентности. На практике во всех интерференционных схемах когерентные волны получаются путём расщепления (деления) одной волны. Две из большого
числа интерференционных схем, применяемых в оптике, показаны на рис. 6.1. На рис. 6.1а изображена схема, в которой когерентные волны получаются при отражении исходной волны от двух поверхностей плоскопараллельной пластинки. На рис. 6.1б изображён ход лучей в интерферометре Майкельсона, в котором расщепление лучей достигается при помощи полупрозрачной пластинки.




а)

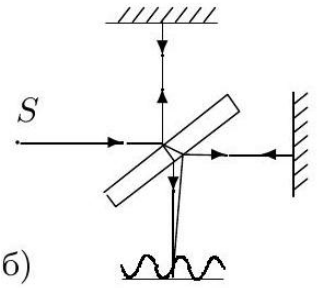



Рис. 6.1. Примеры интерференционных схем, применяемых в оптике

Получаемые при делении «дочерние» волны сохраняют частоту и поляризацию, однако для их когерентности должны выполняться ещё ограничения на ширину спектра излучения и угловой размер источника. Всякая волна занимает конечный спектральный интервал $\Delta\nu$, связанный с длительностью цуга (временем когерентности) $\tau$ соотношением неопределённости

$$\tau \cdot \Delta\nu \sim 1 \tag{6.1}$$

В оптическом диапазоне даже у хороших источников света, использующих излучение узкой спектральной линии, величина $\tau$ оказывается порядка $10^{-10}$ с, поэтому интерференция может наблюдаться только при разностях хода порядка сантиметра и менее. В большинстве практических случаев условия наблюдения интерференции в оптике оказываются ещё значительно более жёсткими. Так, например, при использовании источников белого света (лампа накаливания, дуга и т. п.) интерференция может наблюдаться только при разности хода в несколько микрометров. Именно поэтому интерференцию в оптике, как правило, легко наблюдать на тонких плёнках и трудно на толстых (несколько миллиметров) стеклянных пластинках. По той же причине в оптическом интерферометре Майкельсона для компенсации большой разности хода приходится использовать специальную пластинку - компенсатор (эта пластинка на рис. 6.1б не показана). Единственное исключение представляет излучение лазера, у которого время когерентности $\tau$ может быть на несколько порядков больше, чем у самых хороших нелазерных источников.

При переходе от оптики к радиоволнам миллиметрового или сантиметрового диапазонов длина электромагнитной волны $\lambda$ увеличивается примерно в $10^4 \div 10^5$ раз. Поэтому существенно изменяются масштабы явления, экспериментальные методы и условия наблюдения интерференции. В частности, в указанных диапазонах длин волн легко достижимо время когерентности $\tau$ порядка $10^{-5} \div 10^{-3}$ с, определяемое в данном случае стабильностью источника радиоизлучения, так что длина цуга $c\tau$ может достигать сотен километров. К тому же ослабляется ограничение на угловой размер источника $\gamma < \lambda/\rho$, где $\rho$ - поперечный размер области различимости интерференционной картины (радиус когерентности). Таким образом, в диапазонах миллиметровых и сантиметровых радиоволн существенно облегчаются условия наблюдения интерференции.

Миллиметровый диапазон радиоволн)0,1 см $< \lambda \leqslant$ 1 см), которому в частотном спектре соответствует полоса крайне высоких частот 30 ГГц $\leqslant \nu <$ 300 ГГц (КВЧ), широко используется в радиоастрономии, радиоголографии, радиорелейной связи, метеорологии, медицине и т. д. В свою очередь сантиметровый диапазон)1 см $< \lambda \leqslant$ 10 см), которому в частотном спектре отвечает полоса сверхвысоких частот 3 ГГц $\leqslant \nu <$ 30 ГГщ (СВЧ), используется в радиолокации, радиоголографии, спутниковом телевидении, навигации и т. п. Отметим, что выделенные здесь участки спектра радиоволн соответствуют международному регламенту радиосвязи. В отечественной литературе к

СВЧрадиоволнам часто относят полосу частот от 0,3 ГГц до 300 ГГц, отвечающую диапазону длин волн от 1 m до 1 мм, который при этом называют микроволновым. В некоторых случаях для упрощения мы будем поступать так же, надеясь, что эта неточность не отразится на понимании вопроса.

## 28.1.2 2. Generation of Microwave Radio Waves

В качестве источников ЭМ-излучения малой мощности (от 5 мВт до 5 Вт) в диапазонах дециметровых, сантиметровых и миллиметровых волн обычно используются генераторы на отражательных клистронах. Отражательный клистрон, схематически изображённый на рис. 6.2, представляет собой электровакуумный прибор, имеющий один тороидальный резонатор 1, дважды пронизываемый электронным потоком 2. Возвращение электронов осуществляется с помощью отражателя 3, находящегося под отрицательным постоянным потенциалом $U_{\text{отр}}$ по отношению к катоду.

Процесс генерации протекает следующим образом. Электроны, эмитируемые катодом 4, ускоряются постоянным анодным напряжением $U_{\text{a}}$

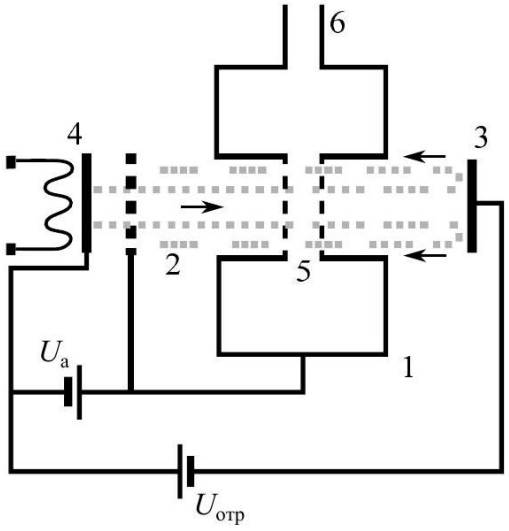



Рис. 6.2. Схема отражательного клистрона

и попадают в узкий сеточный зазор резонатора 5, в котором имеется продольное высокочастотное электрическое поле. Это поле периодически ускоряет и замедляет электроны, модулируя электронный поток по скорости. Двигаясь далее в пространстве дрейфа до отражателя и обратно, электроны постепенно образуют сгустки за счёт того, что быстрые электроны догоняют медленные. Геометрия прибора и распределение потенциалов на электродах подобраны так, чтобы при обратном пролёте сгустки электронов проходили через резонатор в те моменты, когда в нём имеется тормозящее высокочастотное поле. При этом часть кинетической энергии электронов преобразуется в энергию поля резонатора, компенсируя затраты на полезное излучение.

Частота генерации отражательного клистрона допускает практически безынерционную электронную перестройку в пределах $\approx 1\%$ от своего среднего значения за счёт изменения напряжения на отражателе, а также механическую перестройку в пределах $\approx 20\%$ путём изменения конфигурации резонатора.

Энергия СВЧ-колебаний из клистрона поступает в радиоволновод 6, по которому распространяется в виде бегущей ЭМ-волны к нагрузке, в роли которой при излучении в открытое пространство выступает антенна. В рассматриваемом диапазоне длин волн особенно эффективна антенна в форме рупора, который улучшает согласование антенны с открытым пространством, благодаря чему увеличивается излучаемая

мощность. Кроме того, рупор делает излучение более направленным за счёт увеличения площади излучающего отверстия по сравнению с волноводом.

Чем шире раскрыв рупорной антенны, тем уже направленным будет излучение.

### 28.1.3 3. Propagation of Em Waves in Radio Waveguides

Остановимся немного подробнее на распространении ЭМ-волн в радиоволноводах - важных элементах высокочастотной радиотехники, служащих для транспортировки электромагнитной энергии в интервале длин волн от 1 м до 1 мм. Волноводы (далее будем использовать этот термин) делаются из металлических труб и могут иметь любой профиль, но из технологических соображений их сечения делаются круглыми или прямоугольными. В миллиметровом диапазоне волноводы могут быть сделаны и из диэлектрика, чаще всего - в виде стержней.

Чтобы найти структуру ЭМ-поля в прямоугольном металлическом волноводе, надо решить уравнения Максвелла с соответствующими граничными условиями. Решение этой задачи приведено во многих учебниках. Мы построим ЭМ-поле в волноводе, воспользовавшись методом Бриллюэна, складывая падающую и отражённые от стенок плоские волны.

Плоские электромагнитные волны, распространяющиеся в однородной изотропной среде, обычно представляют выражениями

$$\boldsymbol{E}(\boldsymbol{r},t)=\boldsymbol{E}_0e^{i(\boldsymbol{kr}-\omega t)}=\boldsymbol{A}(\boldsymbol{r})e^{-i\omega t} \tag{6.2}$$

Здесь $\boldsymbol{E}_0$ - амплитуда напряжённости электрического поля $\boldsymbol{E}$, $\boldsymbol{r}(x,y,z)$ - радиус-вектор точки наблюдения, $t$ - время, $\omega=2\pi\nu$ - круговая частота, $\boldsymbol{k}\,(k_x,k_y,k_z)$ - волновой вектор. Направление вектора $\boldsymbol{k}$ совпадает с направлением распространения волны, а модуль этого вектора равен

$$k=\frac{2\pi}{\lambda}=\frac{\omega}{v} \tag{6.3}$$

где $v$ - фазовая скорость волны, которая в пустом пространстве совпадает со скоростью света $c$. Векторная величина

$$\boldsymbol{A}(\boldsymbol{r})=\boldsymbol{E}_0e^{i\boldsymbol{kr}}=\boldsymbol{E}_0e^{i(k_xx+k_yy+k_zz)} \tag{6.4}$$

называется комплексной амплитудой волны. Комплексная запись (6.2), (6.4) позволяет использовать удобную для выкладок экспоненциальную функцию. При этом надо помнить, что физический смысл имеет только действительная часть комплексного выражения.

Рассмотрим отражение плоской ЭМ-волны от идеально проводящей, бесконечно протяжённой плоской поверхности $x=0$ (рис. 6.3). Пусть вектор напряжённости электрического поля падающей волны $\boldsymbol{E}$ параллелен этой плоскости. В наших обозначениях вектор $\boldsymbol{E}$ направлен по оси $y$ (на нас). Оба вектора напряжённости поля $\boldsymbol{E}$ и $\boldsymbol{H}$ лежат в плоскости одинаковой фазы волны, им перпендикулярен волновой вектор $\boldsymbol{k}$, составляющий с векторами поля правовинтовую тройку.

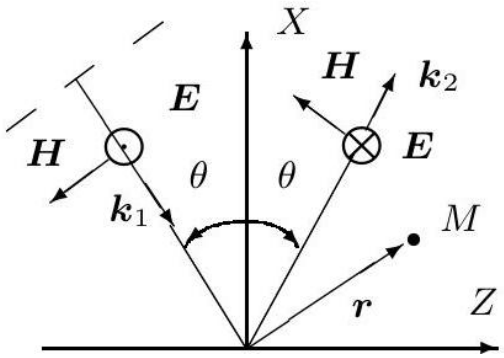



Рис. 6.3. Отражение плоской волны от проводящей плоскости

В некоторую произвольную точку $M$ над проводящей плоскостью приходят две волны: падающая волна с напряжённостью $\boldsymbol{E}_{\text{пад}}$ и отражённая от плоскости - с напряжённостью $\boldsymbol{E}_{\text{отр}}$. Из граничного условия на проводящей поверхности: $\boldsymbol{E}_\tau(t) = 0$, где индексом $\tau$ отмечена тангенциальная составляющая напряжённости электрического поля, вытекает, что в отражённой и падающей волнах совпадают углы падения и отражения, частоты, волновые числа и модули амплитуд. Таким образом,

$$\boldsymbol{E}_{\text{пад}} = \boldsymbol{E}_0 e^{i(\boldsymbol{k}_1\boldsymbol{r}-\omega t)}, \quad \boldsymbol{E}_{\text{отр}} = -\boldsymbol{E}_0 e^{i(\boldsymbol{k}_2\boldsymbol{r}-\omega t)},$$

где $k_1 = k_2 = k = \omega/c$. Знак «минус» в отражённой волне связан со сдвигом фаз на 180°, возникающим при отражении волны от проводящей поверхности. Суммарное электрическое поле в точке $M$ при этом равно

$$\boldsymbol{E} = \boldsymbol{E}_0 \left(e^{i\boldsymbol{k}_1\boldsymbol{r}} - e^{i\boldsymbol{k}_2\boldsymbol{r}}\right) e^{-i\omega t} \tag{6.5}$$

Результат вычисления скалярных произведений $\boldsymbol{k}_{1,2}\boldsymbol{r}$ в (6.5) удобно представить формулой

$$\boldsymbol{k}_{1,2}\boldsymbol{r} = k(\mp x\cos\theta + z\sin\theta) \tag{6.6}$$

Подставив (6.6) в (6.5), найдём

$$\boldsymbol{E} = -2i\boldsymbol{E}_0 \sin(kx\cos\theta) e^{i(kz\sin\theta-\omega t)} \tag{6.7}$$

Реальная часть комплексного выражения (6.7) равна

$$\boldsymbol{E}_{\text{Re}} = 2\boldsymbol{E}_0 \sin(kx\cos\theta)\sin(kz\sin\theta - \omega t) \tag{6.8}$$

Видно, что формулы (6.7), (6.8) описывают волну с амплитудой $2\boldsymbol{E}_0\sin(kx\cos\theta)$, продолъънъм волновым числом

$$k_z = k\sin\theta \tag{6.9}$$

бегущую в направлении $z$ с фазовой скоростью

$$v_\Phi = \frac{\omega}{k_z} = \frac{c}{\sin\theta} \tag{6.10}$$

Отметим две важные особенности этой волны: 1) её фазовая скорость больше скорости света; 2) при фиксированном угле $\theta$ амплитуда поля гармонически зависит от $x$ и не меняется со временем. Иначе говоря, в результате интерференции падающей и отражённой волн в пространстве над проводящей поверхностью в направлении оси $x$ образуется система стоячих волн с поперечным волновым числом

$$k_\perp = k_x = k\cos\theta \tag{6.11}$$

Электрическое поле стоячей волны равно нулю в точках

$$x_n = \frac{m\pi}{k\cos\theta}; \quad m = 0, 1, 2, \ldots \tag{6.12}$$

Таким образом, поверхности нулевого электрического поля представляют собой плоскости, параллельные отражающей поверхности (для которой $x = 0, m = 0$). Расположим в первой из них, то есть на расстоянии $x_1 = \pi/k\cos\theta = a$, вторую проводящую поверхность. Эта поверхность не исказит полученного распределения поля, так как на ней автоматически удовлетворяется граничное условие $\boldsymbol{E}_\tau(t) = 0$. Точно такие же плоскости можно поставить, например, при $y = 0$ и произвольном $y = b$. Эти плоскости нормальны электрическим силовым линиям, и на них выполняется граничное условие $\boldsymbol{E}_\tau(t) = 0$.

Итак, мы показали, что в волноводе прямоугольного сечения может распространяться ЭМ-волна, которую в пределах волновода можно рассматривать как результат суперпозиции двух плоских волн. Каждая плоская волна является чисто поперечной, так что электрическое и магнитное поля перпендикулярны направлению их распространения. В суммарной волне электрическое поле имеет только составляющую $E_y$ и, следовательно, перпендикулярно оси волновода, а магнитное поле имеет составляющие $H_x$ и $H_z$.

Электромагнитное поле в волноводе не является чисто поперечным, а имеет продольные составляющие.

В рассмотренном случае отлична от нуля продольная составляющая магнитного поля, и поэтому такую волну называют магнитной ($H$ волна). Мы могли бы взять другую поляризацию исходной падающей волны (вектор $\boldsymbol{H}$ направлен по оси $y$), и тогда возникла бы электрическая волна с $E_z \neq 0$ ($E$-волна).

Посмотрим на соотношение (6.12) с другой стороны. Если даны две параллельные проводящие плоскости, расположенные на расстоянии $a$ друг от друга, то между ними могут распространяться волны с углом падения $\theta$ из дискретного набора, определяемого равенствами

$$\cos\theta_m = \frac{m\pi}{ka} = \frac{m\lambda}{2a} = \frac{m\pi c}{a\omega} \tag{6.13}$$

в которых учтены соотношения (6.3) для свободного пространства. Так как

$$\cos\theta_m \leqslant 1 \tag{6.14}$$

то для каждого $m$ существуют наименьшие критические волновое число и частота, а также наибольшая критическая длина волны, соответствующие $m = 1$, при которых волна ещё может проходить через волновод:

$$k_{\text{Кр}} = \frac{\pi}{a}, \quad \omega_{\text{Кр}} = \frac{c}{a}, \quad \lambda_{\text{Кр}} = 2a \tag{6.15}$$

Выражения (6.9) и (6.10) для продольного волнового числа и фазовой скорости ЭМ-волны, распространяющейся в волноводе, теперь можно записать так:

$$k_z = k\sin\theta = \frac{\omega}{c}\sqrt{1 - (\omega_{\text{Кр}}/\omega)^2} \tag{6.16}$$

$$v_\Phi = \frac{\omega}{k_z} = \frac{c}{\sin\theta} = \frac{c}{\sqrt{1 - \cos^2\theta}} = \frac{c}{\sqrt{1 - (\omega_{\text{Кр}}/\omega)^2}} \tag{6.17}$$

Из выражений (6.16), (6.17) получаются дисперсионное соотношение $\omega = \omega(k_z)$ для бегущей волны в волноводе и формулы для групповой скорости $v_г = d\omega/dk_z$ этой волны:

$$\omega = \sqrt{\omega_{\text{Кр}}^2 + (ck_z)^2} \tag{6.18}$$

$$v_\Gamma = c\sqrt{1 - \left(\frac{\omega_{\text{Кр}}}{\omega}\right)^2} = \frac{c^2}{v_\Phi} \tag{6.19}$$

Как видно из (6.17), (6.19), фазовая скорость - скорость перемещения поверхности постоянной фазы - в волноводе всегда больше скорости света в пустоте, а групповая - скорость распространения сигнала - всегда меньше. Важно отметить, что фазовая скорость зависит от частоты. В таких случаях говорят, что среда (в данном случае волновод) обладает дисперсией.

Полученные выше результаты относятся к волне в прямоугольном волноводе с синусоидальным распределением поля по координате $x$ и однородным распределением по $y$, как это следовало из формул (6.7),

(6.8). Ясно, однако, что никакой выделенности оси $x$ нет, и поэтому точно так же может образоваться синусоидальное распределение поля вдоль оси $y$. Для каждого вида $E$ - и $H$-волн получается бесчисленное множество решений, каждое из которых имеет свои критические частоту и длину волны. В случае прямоугольного волновода с поперечными размерами $a$ и $b$ все возможные поперечные и продольные волновые числа определяются по формулам

$$k_\perp^2 = k_x^2 + k_y^2 = \left(m\frac{\pi}{a}\right)^2 + \left(n\frac{\pi}{b}\right)^2, \quad m, n - \text{ целые числа} \tag{6.20}$$

$$k_z = \pm\sqrt{k^2 - k_\perp^2} \tag{6.21}$$

где знаки «» показывают возможные направления распространения волны. Величина $m$ представляет собой полное число полупериодов изменения той или иной составляющей поля вдоль линии, параллельной широкой стенке волновода ($a$), а $n$ - то же для узкой стенки ($b$). Эти же символы употребляются и в обозначениях волн: соответственно $E_{mn}$ или $H_{mn}$.

Из формулы (6.21) видно, что если в свободном пространстве у волны $k^2 < k_\perp^2$, то в волноводе для неё $k_z = \pm i\sqrt{k_\perp^2 - k^2}$, и волна затухает, не распространяясь. Это затухание не связано с каким либо потерями. Так как для $E$-волн $m, n \neq 0$, а для $H$-волн $m^2 + n^2 \neq 0$ (в этом студентам предлагается убедиться самостоятельно!), то из формулы (6.20) мы получаем, что минимальным $k_{\text{кр}} = \pi/a$ и, соответственно, минимальной критической частотой $\omega_{\text{кр}} = \pi c/a$ и максимальной критической длиной $\lambda_{\text{кр}} = 2a$ в прямоугольном волноводе обладает волна $H_{10}$. Поэтому эту волну называют основной и обычно используют для передачи электромагнитной энергии по прямоугольным волноводам. Очевидно, что для волновода заданного сечения существует полоса частот, ограниченная снизу критической частотой волны $H_{10}$ ($\lambda_{\text{кр}} = 2a$), а сверху - критической частотой следующей распространяющейся волны (например, $H_{10}$ с $\lambda_{\text{кр}} = 2b$ или $H_{20}$ с $\lambda_{\text{кр}} = a$). В этой полосе частот электромагнитная энергия переносится только одним типом волн, что существенно облегчает её дальнейшее использование.

## 28.1.4 4. Tunneling of Microwave Radio Waves

Как отмечалось в предыдущем пункте, плоскую электромагнитную волну в однородной изотропной среде обычно представляют выражениями (6.2). В

максвелловской теории электромагнитного поля они возникают как решение волнового уравнения при действительных значениях $k_x, k_y$ и $k_z$. Однако волновое уравнение формально допускает решения
вида (6.2) и с мнимыми значениями $k_x, k_y$ или $k_z$. Такие волны также имеют реальный физический смысл и называются неоднородными в отличие от однородных плоских волн с действительными компонентами волнового вектора (см., например, [1]). В качестве примера приведём волну вида (6.2) с мнимым значением $k_z = \pm i\varkappa$:

$$\boldsymbol{E} = \boldsymbol{a} e^{\mp \varkappa z} e^{i(k_x x + k_y y)} e^{-i\omega t} \tag{6.22}$$

Выражение (6.22) описывает бегущую волну, амплитуда которой экспоненциально затухает (или нарастает) по оси $z$ на характерной длине $\varkappa^{-1}$. Волна распространяется вдоль плоскости $xy$, направление распространения определяется компонентами $k_x, k_y$.

Неоднородные волны появляются во многих физических задачах, например, при дифракции света на решётке. Неоднородные волны возникают и вблизи границы раздела двух сред при полном внутреннем отражении света.

Рассмотрим отражение и преломление электромагнитной волны на границе раздела двух однородных диэлектрических сред. Формулы, определяющие интенсивность, направление распространения и поляризацию отражённых и преломлённых волн, могут быть получены из граничных условий для векторов $\boldsymbol{E}, \boldsymbol{D}, \boldsymbol{H}$ и $\boldsymbol{B}$:

$$E_{1\tau} = E_{2\tau}, \quad D_{1n} = D_{2n}, \quad H_{1\tau} = H_{2\tau}, \quad B_{1n} = B_{2n} \tag{6.23}$$

Здесь индексами «  »тмечены тангенциальные, а « $n$ » - нормальные составляющие векторов, цифра 1 относится к первой, а 2 - ко второй среде. Поскольку в электромагнитной волне электрическое и магнитное поля связаны между собой, из четырёх соотношений (6.23) независимыми остаются только два. Обычно используют условия для тангенциальных компонент полей.

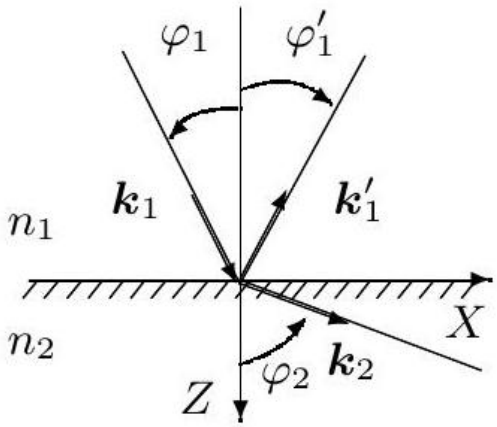



Рис. 6.4. Отражение и преломление волны на границе раздела двух сред

Выберем координатную систему так, как это изображено на рис. 6.4. Ось $z$ совпадает с нормалью к поверхности раздела сред. Ось $x$ лежит в плоскости падения светового луча.

Пусть $\boldsymbol{E}_1, \boldsymbol{E}_1'$ и $\boldsymbol{E}_2$ - электрические поля в падающей, отражённой и преломлённой волнах соответственно:

$$\begin{aligned} \boldsymbol{E}_1 &= \boldsymbol{a}_1 e^{ik_1(x\sin\varphi_1 + z\sin\varphi_1)} e^{-i\omega t} \\ \boldsymbol{E}_1' &= \boldsymbol{a}_1' e^{ik_1'\left(x\sin\varphi_1' + z\sin\varphi_1'\right)} e^{-i\omega_1' t} \\ \boldsymbol{E}_2 &= \boldsymbol{a}_2 e^{i(k_{2x}x + k_{2z}z)} e^{-i\omega_2 t} \end{aligned} \tag{6.24}$$

Здесь $\varphi_1$ - угол падения, $\varphi_1'$ - угол отражения и $\varphi_2$ - угол преломления (рис. 6.4). Компоненты $k_x, k_z$ в падающей и отражённой волнах выражены через $k_1, k_1'$ и синусы углов падения и отражения, а $k_{2x}$, $k_{2z}$ - не преобразованы, чтобы можно было описывать как однородные, так и неоднородные волны во второй среде. Предполагается, что все три вектора $\boldsymbol{k}_1, \boldsymbol{k}_1'$ и $\boldsymbol{k}_2$ лежат в плоскости $xz$.

На границе раздела (при $z = 0$) должны выполняться граничные условия (6.23). Первое из них даёт: $\boldsymbol{E}_{1\tau} + \boldsymbol{E}'_{1\tau} = \boldsymbol{E}_{2\tau}$, или, с учётом формул (6.24),

$$\boldsymbol{a}_{1\tau} e^{ik_1 x \sin\varphi_1} e^{-i\omega t} + \boldsymbol{a}'_{1\tau} e^{ik_1' x \sin\varphi_1'} e^{-i\omega_1' t} = \boldsymbol{a}_{2\tau} e^{ik_{2x} x} e^{-i\omega_2 t}$$

Это равенство должно выполняться при любых значениях $x$ и $t$. Поэтому

$$\omega_1 = \omega_1' = \omega_2 \tag{6.25}$$

$$k_1 \sin\varphi_1 = k_1' \sin\varphi_1' = k_{2x} \tag{6.26}$$

Равенство (6.25) показывает, что частоты отражённой и преломлённой волн равны частоте падающей волны. Падающая и отражённая волны распространяются в одной и той же среде, а значит, $k_1 = k_1'$. Из первого равенства (6.26) при этом следует:

$$\sin\varphi_1 = \sin\varphi_1'$$

то есть угол падения равен углу отражения.

Если предположить, что волна во второй среде однородна, то $k_{2x} = k_2 \sin\varphi_2$ и, следовательно, на основании (6.3), (6.25) и (6.26) можно записать:

$$\frac{\sin\varphi_2}{\sin\varphi_1} = \frac{k_1}{k_2} = \frac{v_2}{v_1} = \frac{n_1}{n_2} = n \tag{6.27}$$

Здесь $n_1, n_2$ - показатели преломления первой и второй сред соответственно, а $n = n_1/n_2$ - относительный показатель преломления этих сред. Мы получили, таким образом, обычные законы отражения и преломления, известные из оптики.

Из безусловного неравенства $\sin\varphi_2 \leqslant 1$ вытекает, что формула (6.27) теряет смысл при падении ЭМ-волны на границу раздела со стороны «оптически» более плотной среды ($n_1 > n_2$), когда угол падения $\varphi_1$ превышает критическое значение $\varphi_{\text{пр}}$ , называемое предельным углом полного внутреннего отражения:

$$\sin\varphi_{\text{пр}} = \frac{k_2}{k_1} = \frac{n_2}{n_1} = \frac{1}{n} \tag{6.28}$$

Это означает, что при $\varphi_1$ наше предположение об однородности волны во второй среде в случае полного внутреннего отражения оказывается несправедливым.

Попытаемся теперь удовлетворить граничным условиям и вытекающему из них соотношению (6.26), предположив, что волна во второй среде является неоднородной. При $\varphi_1 > \varphi_{\text{пр}}$ получим

$$k_1 \sin\varphi_1 > k_1 \sin\varphi_{\text{пр}} = k_1 \frac{k_2}{k_1} = k_2$$

Сравнивая это соотношение с (6.25), находим

$$k_{2x} > k_2 \tag{6.29}$$

Но

$$k_2^2 = k_{2x}^2 + k_{2z}^2 \tag{6.30}$$

где мы учли, что в нашем случае $k_{1y} = k_{2y} = 0$. Разрешая уравнение (6.30) относительно $k_{2x}$ с учётом неравенства (6.29), получаем

$$k_{2z} = \pm\sqrt{k_2^2 - k_{2x}^2} = \pm i\sqrt{k_{2x}^2 - k_2^2} = \pm i\sqrt{k_1^2 \sin^2\varphi_1 - k_2^2} \tag{6.31}$$

Величина $k_{2z}$ оказывается мнимой. Следовательно, волна во второй среде является неоднородной и описывается выражением вида (6.22), где $k_y = k_{2y} = 0, k_x = k_{2x} = k_1 \sin\varphi_1$, а величина $\varkappa$ даётся формулой

$$\varkappa = \sqrt{k_1^2 \sin^2\varphi_1 - k_2^2} \tag{6.32}$$

Таким образом, при $\varphi_1 > \varphi_{\text{пр}}$ электромагнитное поле во второй среде (например, при переходе световой волны из стекла в воздух) экспоненциально затухает (или нарастает) с удалением от поверхности раздела. На основании закона сохранения энергии в выражении вида (6.22) перед положительной величиной $\varkappa$ следует брать знак «-», соответствующий затухающей волне.

Экспоненциальную функцию, описывающую затухание волны с удалением от поверхности раздела, удобно записать в виде $\exp(-z/2\Lambda)$, где $\Lambda = 1/2\varkappa$. Тогда интенсивность волны, пропорциональная квадрату амплитуды, изменяется с расстоянием по закону

$$I \propto e^{-z/\Lambda} \tag{6.33}$$

Длина затухания $\Lambda$ с учётом (6.32) и (6.28) может быть представлена выражениями

$$\Lambda = \frac{1}{2\sqrt{k_1^2 \sin^2\varphi_1 - k_2^2}} = \frac{1}{2k_2\sqrt{n^2 \sin^2\varphi_1 - 1}} = \frac{\lambda_2}{4\pi\sqrt{n^2 \sin^2\varphi_1 - 1}} \tag{6.34}$$

То обстоятельство, что при полном внутреннем отражении колебания присутствуют не только в первой, но и во второй среде, является необходимым следствием граничных условий (6.23). Формулы (6.33) и (6.34) позволяют количественно исследовать затухание электромагнитных колебаний во второй среде.

Из курса электродинамики известно, что последовательное применение граничных условий (6.23) позволяет найти соотношения между амплитудами (и интенсивностями) отражённого, преломлённого и падающего лучей при любых углах падения - так называемые формулы Френеля. Из этих соотношений вытекает, что при полном внутреннем отражении интенсивность отражённого луча равна интенсивности падающего. Этот вывод является очевидным, поскольку мы рассматриваем среды, не поглощающие электромагнитные волны. Вектор УмоваПойнтинга $\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H}$ в неоднородной волне параллелен границе раздела сред, так что поток энергии вглубь второй среды равен нулю.

Отметим, что при полном внутреннем отражении сдвиг фаз между отражённой и падающей ЭМ-волнами не равен нулю и зависит от поляризации падающей волны. Вследствие этого изменяется поляризация волны: линейно поляризованная волна после отражения оказывается поляризованной по эллипсу.

Рассмотрим теперь случай, когда вторая среда представляет собой тонкую воздушную прослойку между двумя диэлектрическими пластинками. При полном внутреннем отражении на границе пластинки и воздуха часть электромагнитной энергии всё же достигает второй пластинки с интенсивностью, которая экспоненциально уменьшается с шириной воздушной прослойки по формуле (6.33). Волна проникает в эту пластинку и далее распространяется в ней как обычная однородная плоская волна. При ширине прослойки, стремящейся к нулю, прозрачность воздушного промежутка $T \to 1$, а

коэффициент отражения $R \to 0$. При значительном увеличении ширины прослойки $R \to 1$, а $T \to 0$. Это и есть обычный случай полного внутреннего отражения. В общем случае $R + T = 1$, если поглощение ЭМ-волны отсутствует.

Проникновение электромагнитных волн через менее плотную среду (например, воздушный зазор) при полном внутреннем отражении явление той же природы, что и прохождение частиц через область, где их полная энергия оказывается меньше потенциальной энергии. Это явление изучается в квантовой физике и носит название туннельного эффекта. Классическим примером туннельного эффекта является $\alpha$ распад радиоактивных ядер. По аналогии с этим эффектом прохождение электромагнитных волн через узкий зазор при углах падения, превосходящих угол полного внутреннего отражения, часто называют туннелированием.

**Литература**


1. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. §§1, 26, 27, 33, 35, 63 – 66.
2. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. III, Гл. 6, §6.7; Гл. 7.
3. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. 13, 16, 135, 137, 139.
4. Лебедев И.В. Техника и приборы СВЧ. Т. 2. - М.: Высшая школа, 1972. §§5.7, 5.8.
5. Калитеевский Н.И. Волновая оптика. - М.: Высшая школа, 1978. Гл. II.


## 28.2 Polarization. Crystal Optics

### 28.2.1 1. Types of Polarization, Basic Concepts

Плоские электромагнитные волны являются поперечными, т. е. проекции осциллирующих электрического и магнитного полей на направление распространения такой волны равны нулю. Это непосредственно следует из теоремы Гаусса и теоремы о циркуляции магнитного поля. Пусть, для определённости, плоская электромагнитная волна распространяется в вакууме вдоль оси z. Тогда по теореме Гаусса

$$\operatorname{div} \boldsymbol{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0$$

но поскольку в рассматриваемой волне все производные по $x$ и по $y$ равны нулю, то и $\partial E_z/\partial z = 0$, т. е. для $E_z$ равны нулю все пространственные производные. С другой стороны, из уравнения циркуляции магнитного поля следует, что

$$(\operatorname{rot} \boldsymbol{H})_z = \frac{\partial H_y}{\partial x} + \frac{\partial H_x}{\partial y} = \frac{1}{c}\frac{\partial E_z}{\partial t}$$

т. е. и $\partial E_z/\partial t = 0$ по той же причине, и, таким образом, $E_z$ является константой. Но поскольку мы рассматриваем осциллирующее поле с нулевым средним значением, эта константа равна нулю. Аналогично из теоремы Гаусса для магнитного поля и закона электромагнитной индукции получим $H_z = 0$.

Теперь, выписав из уравнения циркуляции магнитного поля и уравнения электромагнитной индукции выражения для оставшихся $x$ - и $y$ компонент роторов $\boldsymbol{E}$ и

$\boldsymbol{H}$ и воспользовавшись условием $E_z = H_z = 0$, мы увидим, что уравнения Максвелла для электрического и магнитного полей разделились на две независимые системы уравнений, одна из
которых связывает $E_x$ и $H_y$, а другая - $E_y$ и $H_x$:

$$\begin{cases} (\operatorname{rot} \boldsymbol{E})_y = \dfrac{\partial E_x}{\partial z} = -\dfrac{1}{c}\dfrac{\partial H_y}{\partial t} \\ (\operatorname{rot} \boldsymbol{H})_x = -\dfrac{\partial H_y}{\partial z} = \dfrac{1}{c}\dfrac{\partial E_x}{\partial t} \end{cases}$$

и

$$\begin{cases} (\operatorname{rot} \boldsymbol{E})_x = \dfrac{\partial E_y}{\partial z} = \dfrac{1}{c}\dfrac{\partial H_x}{\partial t} \\ (\operatorname{rot} \boldsymbol{H})_y = \dfrac{\partial H_x}{\partial z} = \dfrac{1}{c}\dfrac{\partial E_y}{\partial t} \end{cases}$$

Решениями этих систем являются две независимые плоские волны вида $E_x = \mp H_y = f_x(ct \pm z)$ и $E_y = \pm H_x = f_y(ct \pm z)$, где $f_x$ и $f_y$ - произвольные функции. Важным частным случаем являются монохроматические волны:

$$E_x = H_y = A_x \cos(\omega t - kz + \varphi_x) \tag{7.1}$$

или

$$E_y = -H_x = A_y \cos(\omega t - kz + \varphi_y) \tag{7.2}$$

В любой точке пространства концы векторов $\boldsymbol{E}$ и $\boldsymbol{H}$ в каждой из этих волн движутся по отрезкам прямых линий в плоскости $(E_x, E_y)$, поэтому они называются линейно поляризованными (или плоскополяризованными). По историческим причинам плоскость, параллельно которой направлены вектора $\boldsymbol{H}$, обычно называется плоскостью поляризации, а плоскость, параллельно которой направлены вектора $\boldsymbol{E}$ - плоскостью колебаний.

Сам термин «поляризация» исторически возник из предположения о корпускулярной природе света. Его ввёл в 1808 г. французский физик Этьен Луи Малюс, он представлял световые частицы чем-то вроде маленьких магнитиков и полагал, что процесс поляризации света при различных его взаимодействиях с веществом связан с отбором частиц с определённой ориентацией полюсов.

Если обе описанные выше монохроматические волны (7.1) и (7.2) распространяются одновременно, то концы векторов $\boldsymbol{E}$ и $\boldsymbol{H}$ движутся по эллипсам в плоскости $(E_x, E_y)$. Это наиболее общий тип поляризации - эллиптическая поляризация, все другие типы поляризации могут рассматриваться как его частные случаи. Если полуоси эллипса равны, то такую поляризацию называют круговой. Различают правую и левую эллиптические (и круговые) поляризации. Свет называют правополяризованным, если для наблюдателя, смотрящего навстречу лучу, вектор $\boldsymbol{E}$ вращается по часовой стрелке.

Монохроматические волны, таким образом, всегда являются так или иначе поляризованными. Если же свет квазимонохроматический, т.е. обладает ненулевой шириной спектра и, значит, конечными временем и длиной когерентности (т. е. может рассматриваться как последовательность независимых монохроматических цугов, между которыми происходит случайный сбой фазы), то возможны три случая. В первом поляризация не меняется от цуга к цугу, несмотря на сбой фазы между ними. Такой свет считается поляризованным наравне с монохроматическим. Во втором поляризации цугов меняются случайным и независимым образом. Такой свет считают неполяризованным. И в третьем случае поляризации цугов отличаются, но некоторые поляризации появляются

с большей вероятностью, чем другие. Такой свет считается поляризованным частично, его можно рассматривать как смесь поляризованного и неполяризованного света.

Такая классификация справедлива и для немонохроматического (естественного) света, длина когерентности которого сравнима с длиной волны, однако практическое значение для такого света имеет лишь линейная поляризация.

### 28.2.2 2. Obtaining Polarized Light

В основном поляризованный свет получают, преобразуя при помощи различных устройств естественный свет, хотя существуют явления, при которых его можно получить непосредственно от источника. К таким явлениям относятся излучение анизотропных материалов (люминисценция и тепловое излучение при высоких температурах), излучение веществ во внешних магнитных и электрических полях (эффекты Зеемана и Штарка), излучение частицы, движущейся быстрее скорости света в веществе (излучение Вавилова-Черенкова), тормозное излучение релятивистских электронов и, при некоторых условиях, излучение лазеров.

Устройства, с помощью которых из естественного света получают поляризованный, называются поляризаторами. Они разделяют исходный пучок на две компоненты, ортогональные по типу поляризации, одну из них пропускают, а другую поглощают или отклоняют. Действие их может быть основано на одном из четырёх явлений: дихроизме, двойном лучепреломлении в кристаллах, отражении и преломлении на границе изотропных диэлектриков и поляризации при рассеянии света.

Дихроизм, или анизотропия поглощения, - это способность вещества поглощать свет по-разному в зависимости от его поляризации. Ис торически этот термин [1] (от греч. dchroos - двухцветный), возник изза различной окраски одноосных кристаллов в проходящем свете при взаимно перпендикулярных направлениях наблюдения - вдоль оптической оси и перпендикулярно к ней. На этом явлении основано действие поляроидных плёнок, используемых в лабораторных поляроидах. Это прозрачная полимерная плёнка, в которую внедрено множество игольчатых кристаллов, причём в процессе производства путём растяжения плёнки в одном направлении или применением сильных электрических или магнитных полей добиваются того, что эти кристаллы ориентированы преимущественно в одном направлении. Такая плёнка сильно поглощает свет, плоскость поляризации которого перпендикулярна направлению выравнивания кристаллов, но слабо поглощает свет, плоскость поляризации которого параллельна направлению выравнивания.

Двойное лучепреломление в анизотропных кристаллах - результат зависимости в этих кристаллах коэффициента преломления (и соответственно угла преломления, а также угла полного внутреннего отражения) от поляризации световой волны. На этом явлении основано действие широко применяемых в лабораториях призм Николя, Глана и других схожих по конструкции приборов, где свет пропускается через границу раздела двух сред под углом, который обеспечивает полное внутреннее отражение только для одной из двух ортогональных поляризаций.

Поляризация света при отражении и преломлении на границе изотропного диэлектрика следует из формул Френеля. Так, для энергетических коэффициентов отражения плоских волн с направлениями электрических колебаний, одно из которых перпендикулярно плоскости падения ($\perp$), а другое параллельно ей ($\parallel$), получаем соответственно

$$\rho_\perp = \frac{\sin^2(\varphi - r)}{\sin^2(\varphi + r)}, \quad \rho_\parallel = \frac{\mathrm{tg}^2(\varphi - r)}{\mathrm{tg}^2(\varphi + r)}$$

здесь $\varphi$ - угол падения, а $r$ - угол преломления волны. Отсюда следует, что, если $\varphi+r = 90°$, т. е. угол между направлениями распространения отражённой и преломлённой волн является прямым, коэффициент отражения $\rho_{\|}$обращается в ноль и отражённый свет оказывается полностью поляризованным. Соответствующий угол падения

$$\varphi_{\text{B}} = \text{arctg}(n) \tag{7.3}$$

где $n$ - коэффициент преломления среды, называется углом Брюстера или углом полной поляризации. Физически обращение $\rho_{\|}$в ноль связано с тем, что отражённый свет является суперпозицией излучения молекул среды, дипольные моменты которых колеблются под действием преломлённой волны. Однако если направление колебаний дипольных моментов совпадает с направлением распространения отражённой волны, то это излучение отсутствует, так как мощность излучения электрических диполей в направлении колебаний равна, как известно, нулю.

Преломлённый свет также оказывается поляризован, однако лишь частично при любом угле падения. Соответствующие коэффициенты преломления равны $(1 - \rho_{\perp})$ и $\big(1 - \rho_{\|}\big)$. Но степень поляризации можно увеличить, если использовать не одну пластинку диэлектрика, а стопку последовательно расположенных пластинок.

При рассеянии света на частицах пыли, молекулах газа и даже на атомах в кристаллах рассеянная волна также оказывается поляризованной, причём преимущественное направление электрических колебаний перпендикулярно падающему лучу. Это связано с разным значением проекций дипольных моментов рассеивающих частиц, колеблющихся под действием падающей волны, на различные направления, перпендикулярные распространению рассеянного света. Степень поляризации при рассеянии на углы 90° может быть близка к 100%. Интенсивность рассеянного света в газах, конечно, невелика, причём согласно закону Релея она пропорциональна $1/\lambda^4$. Этот метод поляризации имеет особое значение для рентгеновской области спектра, где все другие методы неприменимы.

## 28.2.3 3. Observation and Analysis of Polarized Light. Malus' Law

Для анализа поляризованного света его пропускают через поляризатор, который в этом случае принято называть анализатором. При прохождении через анализатор линейно поляризованного света интенсивность прошедшей волны, очевидно, пропорциональна квадрату проекции амплитуды исходной волны на разрешённое направление поляризатора, т. е.

$$I = I_0 \cos^2 \varphi \tag{7.4}$$

где $\varphi$ - угол между направлением электрических колебаний анализируемой волны и разрешённым направлением поляризатора. Этот закон был открыт Э. Малюсом и носит его имя, хотя сам Малюс, как уже упоминалось, выводил его исходя из представлений о корпускулярной природе света.

Для частично линейно поляризованного света (т.е. для смеси линейно поляризованного и естественного света) вводят понятие степени поляризации

$$P = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{7.5}$$

[0]1 Помимо используемого в применении к поляроидам термин «дихроичный» имеет ещё два разных смысла, которые не следует путать. Это способность красителей сильно менять цвет при изменении концентрации и наличие двух полос пропускания в разных областях видимого спектра у некоторых фильтров.

где $I_{\max}$ и $I_{\min}$ - соответственно максимальное и минимальное значения интенсивности света, прошедшего через анализатор при различных углах $\varphi$.

С помощью одного анализатора, однако, невозможно отличить неполяризованный свет от поляризованного по кругу, так как в обоих случаях интенсивность прошедшего света не будет меняться при вращении анализатора. Чтобы различать эти варианты, необходимо дополнительно пользоваться устройствами, изменяющими сдвиг фаз между ортогональными поляризациями и, таким образом, способными преобразовать эллиптическую поляризацию в линейную. Таким устройством может быть набор двояколучепреломляющих пластинок, вырезанных параллельно главной оптической оси кристалла, которые помещают перед анализатором. У таких пластинок различаются коэффициенты преломления (и соответственно оптические длины путей при прохождении по ним света) для волн, поляризованных параллельно и перпендикулярно главной оси, что и приводит к изменению сдвига фаз между ними. Степень поляризации теперь можно определить для общего случая произвольно поляризованной волны:

$$P = \frac{I_{\max}(\varphi, \delta) - I_{\min}(\varphi, \delta)}{I_{\max}(\varphi, \delta) + I_{\min}(\varphi, \delta)}$$

где максимальное и минимальное значения интенсивности берутся уже не только при изменении угла поворота анализатора $\varphi$ от 0 до $\pi$, но и при изменении разницы оптических длин путей $\delta$ в пластинке перед анализатором от 0 до $\lambda$. На практике достаточно всего одной пластинки, в простейшем случае $\lambda/4$ (т.е. вносящей разницу оптических длин путей для указанных поляризаций, равную $\lambda/4$). Значения интенсивностей, входящие в формулу для определения степени поляризации, можно рассчитать, померив максимальные и минимальные по $\varphi$ интенсивности при отсутствии такой пластинки и при её наличии перед анализатором.

### 28.2.4 4. Double Refraction

В некоторых кристаллах потенциальные ямы, в которых находятся электроны вблизи узлов решётки, не являются сферическисимметричными. При этом систему координат, оказывается, всегда возможно выбрать так, что для малых отклонений от положения равновесия потенциальная энергия электрона будет иметь вид

$$U = a_x x^2 + a_y y^2 + a_z z^2$$

Если все три коэффициента $a_x, a_y$ и $a_z$ различны, то кристалл называется двуосным, если два коэффициента равны - одноосным, если равны все три коэффициента, то яма является сферически-симметричной, а вещество соответственно оптически изотропным. Наибольшее практическое значение имеют одноосные кристаллы. Положим $a_y = a_z = a_\perp$, $a_x = a_\parallel$. Ось $x$ при этом называется оптической осью кристалла.

Поскольку отклонения от положения равновесия под действием внешнего поля у электронов вдоль разных осей координат обратно пропорциональны соответствующим коэффициентам в выражении для потенциальной энергии, то вектор удельного дипольного момента $\boldsymbol{P}$ и, как следствие, вектор электрической индукции $\boldsymbol{D}$ окажутся в общем случае неколлинеарны вектору напряжённости электрического поля $\boldsymbol{E}$:

$$\boldsymbol{P} = \alpha_\perp \boldsymbol{E}_\perp + \alpha_\parallel \boldsymbol{E}_\parallel, \quad \boldsymbol{D} = \varepsilon_\perp \boldsymbol{E}_\perp + \varepsilon_\parallel \boldsymbol{E}_\parallel \tag{7.6}$$

Здесь $\alpha_\parallel, \alpha_\perp$ - поляризуемости, $\varepsilon_\parallel = 1 + 4\pi\alpha_\parallel, \varepsilon_\perp = 1 + 4\pi\alpha_\perp$ - диэлектрические проницаемости кристалла вдоль и поперёк его оптической оси соответственно.

Распространение электромагнитных волн описывается уравнениями Максвелла, которые в отсутствие электрических зарядов и токов имеют вид

$$\operatorname{rot} \boldsymbol{H} = \frac{1}{c}\frac{\partial \boldsymbol{D}}{\partial t}, \quad \operatorname{rot} \boldsymbol{E} = -\frac{1}{c}\frac{\partial \boldsymbol{H}}{\partial t}$$

Определим, при каких условиях в одноосных кристаллах могут распространяться плоские монохроматические электромагнитные волны. Такие волны в общем случае можно записать в виде

$$\boldsymbol{E} = \boldsymbol{E}_0 e^{i(\omega t - \boldsymbol{kr})}, \quad \boldsymbol{H} = \boldsymbol{H}_0 e^{i(\omega t - \boldsymbol{kr})}, \quad \boldsymbol{D} = \boldsymbol{D}_0 e^{i(\omega t - \boldsymbol{kr})}$$

откуда следует $\operatorname{rot} \boldsymbol{H} = -i\boldsymbol{k} \times \boldsymbol{H}$, $\frac{\partial \boldsymbol{D}}{\partial t} = i\omega \boldsymbol{D}$ и аналогичные выражения для других векторов [2]. Подставив эти выражения в выписанные выше уравнения Максвелла, получим, что $\boldsymbol{D}, \boldsymbol{H}$, и $\boldsymbol{E}$ должны быть связаны соотношениями

$$\boldsymbol{D} = -\frac{c}{\omega} k \times \boldsymbol{H}, \quad \boldsymbol{H} = \frac{c}{\omega} \boldsymbol{k} \times \boldsymbol{E} \tag{7.7}$$

Отсюда видно, что векторы $\boldsymbol{D}, \boldsymbol{H}$, и $\boldsymbol{k}$ взаимно перпендикулярны, т. е. плоские волны поперечны в отношении $\boldsymbol{D}$

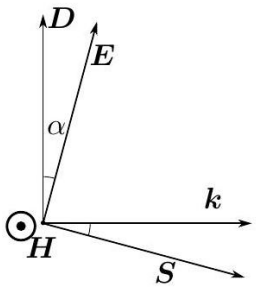



Рис. 7.1.

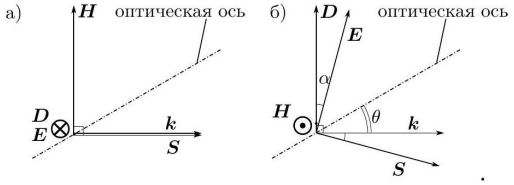
Расположение векторов $\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{k}, \boldsymbol{S}$ в анизотропной среде и $\boldsymbol{H}$, но в общем случае не поперечны в отношении $\boldsymbol{E}$. Кроме того, вектор $\boldsymbol{E}$ должен лежать в одной плоскости с векторами $\boldsymbol{D}$ и $\boldsymbol{k}$, т. к. все они перпендикулярны вектору $\boldsymbol{H}$.

Взаимное расположение векторов $\boldsymbol{D}, \boldsymbol{E}, \boldsymbol{k}$, а также вектора плотности потока энергии (вектора Пойнтинга) $\boldsymbol{S} = \frac{4\pi}{c} \boldsymbol{E} \times \boldsymbol{H}$ показано на рис. 7.1.

Несложный геометрический анализ показывает, что одновременное выполнение условий (7.7) и материального уравнения (7.6) $\boldsymbol{D} = \varepsilon_\perp \boldsymbol{E}_\perp + \varepsilon_\parallel \boldsymbol{E}_\parallel$возможно только в двух случаях (рис. 7.2).

1. Если вектор $\boldsymbol{D}$ перпендикулярен плоскости, в которой лежат оптическая ось кристалла и волновой вектор $\boldsymbol{k}$ (эта плоскость называется главным сечением).

2. Если вектор $\boldsymbol{D}$ лежит в главном сечении.




[02] Нетрудно убедиться, что производная по времени заменяется всюду на $\frac{\partial}{\partial t} \to i\omega$, а оператор градиента $\nabla$ во всех векторных операциях заменяется на выражение $\nabla \to -i\boldsymbol{k}(\operatorname{rot} \boldsymbol{E} \equiv \nabla \times \boldsymbol{E} \to -i\boldsymbol{k} \times \boldsymbol{E}, \operatorname{div} \boldsymbol{E} \equiv \nabla \cdot \boldsymbol{E} \to -i\boldsymbol{k} \cdot \boldsymbol{E})$.

Рис. 7.2. Обыкновенная (а) и необыкновенная (б) волны и положение основных векторов в них. Плоскость рисунка - главное сечение

В первом случае плоская волна называется обыкновенной, а во втором - необыкновенной волной. Во всех остальных случаях электромагнитное поле неизбежно имеет более сложный вид, чем простая плоская монохроматическая волна. Поскольку уравнения Максвелла линейны, в общем случае любое монохроматическое поле в кристалле можно представить в виде суперпозиции обыкновенной и необыкновенной волн.

Обыкновенная и необыкновенная волны распространяются в кристалле с разной скоростью. У обыкновенной волны $\boldsymbol{E}_{\parallel} = 0, \boldsymbol{E} = \boldsymbol{E}_{\perp}$, значит, векторы индукции и напряжённости электрического поля коллинеарны: $\boldsymbol{D} = \varepsilon_{\perp}\boldsymbol{E}_{\perp}$, поэтому вид уравнений Максвелла для неё ничем не отличается от уравнений для плоских волн в изотропных средах. Следовательно, фазовая скорость $v = \omega/k$ этой волны не зависит от направления $\boldsymbol{k}$ и равна

$$v_o = \frac{c}{\sqrt{\varepsilon_{\perp}}} = \frac{c}{n_o} \tag{7.8}$$

где $n_o = \sqrt{\varepsilon_{\perp}}$ - коэффициент преломления для обыкновенной волны [3].

У необыкновенной волны векторы индукции и напряжённости электрического поля в общем случае неколлинеарны, а фазовая скорость такой волны зависит от угла $\theta$ между оптической осью и волновым вектором $\boldsymbol{k}$. Квадрат коэффициента преломления есть уже не отношение модулей $D/E$, а отношение модуля $D = |\boldsymbol{D}|$ и проекции $E_D = E\cos\alpha$ вектора $\boldsymbol{E}$ на направление вектора $\boldsymbol{D}$ (строгий вывод см. ниже):

$$D = \varepsilon E_D, \quad v_e = \frac{c}{\sqrt{\varepsilon}} = \frac{c}{n(\theta)}, \qquad \text{где } \varepsilon = \frac{1}{\frac{\sin^2\theta}{\varepsilon_{\parallel}} + \frac{\cos^2\theta}{\varepsilon_{\perp}}} \tag{7.9}$$

$n(\theta) = \sqrt{\varepsilon}$ - коэффициент преломления необыкновенной волны, зависящий от угла $\theta$ между оптической осью и волновым вектором. Кроме того, у необыкновенной волны вектор Пойнтинга $\boldsymbol{S} = \frac{c}{4\pi}\boldsymbol{E} \times \boldsymbol{H}$ и соответственно групповая скорость не коллинеарны в общем случае волновому вектору $\boldsymbol{k}$, то есть направление переноса энергии и направление смещения поверхностей равных фаз не совпадают.

Плоская монохроматическая волна, попадающая из изотропной среды в анизотропный одноосный кристалл, распадается в общем случае на две взаимно ортогонально поляризованные плоские волны, распространяющихся в общем случае в разных направлениях и с разными скоростями.

Вывод формулы (7.9). Выразим фазовую скорость необыкновенной волны $v = \omega/k$ из соотношений (7.7). С учётом геометрии рис. 7.1 из первого уравнения имеем $v = \omega/k = cH/D$, а из второго $v = cE\cos\alpha/H$, где $\alpha$ - угол между векторами $D$ и $E$. Отсюда, исключая $H$ и выражая угол $\alpha$ через скалярное произведение, находим

$$v = c\sqrt{\frac{E\cos\alpha}{D}} = c\sqrt{\frac{(\boldsymbol{E}\cdot\boldsymbol{D})}{D^2}} \tag{7.10}$$

Это общее соотношение для фазовой скорости электромагнитной волны в анизотропной среде (если $\boldsymbol{D} = \varepsilon\boldsymbol{E}$, то (7.10) переходит в стандартное $v = c/\sqrt{\varepsilon}$.).

Проекции вектора $\boldsymbol{D}$ на оптическую ось $D_{\parallel}$и нормаль к ней в главном сечении $D_{\perp}$ запишем как $D_{\parallel} = \varepsilon_{\parallel}E_{\parallel}$и $D_{\perp} = \varepsilon_{\perp}E_{\perp}$. Тогда

[0]3 Индексы «о»и «е» происходят от английских слов ordinary (обыкновенная) и extraordinary (необыкновенная).

$$(\boldsymbol{ED}) = \varepsilon_{\parallel} E_{\parallel}^2 + \varepsilon_{\perp} E_{\perp}^2 = \frac{D_{\parallel}^2}{\varepsilon_{\parallel}} + \frac{D_{\perp}^2}{\varepsilon_{\perp}}$$

С другой стороны, из геометрии (рис. 7.2б) имеем $D_{\parallel} = D\sin\theta$ и $D_{\perp} = D\cos\theta$, из чего, подставляя полученное выражение для $(\boldsymbol{ED})$ в (7.10), получаем зависимость фазовой скорости необыкновенной волны $v_e$ от угла $\theta$:

$$v_e = c\sqrt{\frac{\sin^2\theta}{\varepsilon_{\parallel}} + \frac{\cos^2\theta}{\varepsilon_{\perp}}} \equiv \frac{c}{n(\theta)} \tag{7.11}$$

Положив $\varepsilon \equiv n^2(\theta)$, убеждаемся, что формулы (7.10) и (7.11) находятся в полном согласии с (7.9).

## 28.2.5 5. Crystal Plates

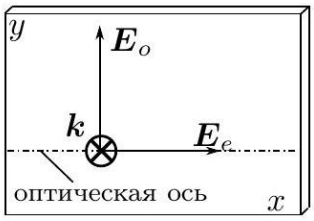



Рис. 7.3

Важным частным случаем является распространение волн в одноосном кристалле перпендикулярно оптической оси. При этом у обыкновенной волны вектор $\boldsymbol{E}$ колеблется в плоскости, перпендикулярной оптической оси, а у необыкновенной - вдоль оптической оси (ось $x$ на рис. 7.3). Показатель преломления для обыкновенной волны равен $n_o = \sqrt{\varepsilon_{\perp}}$, а для необыкновенной $-n_e = \sqrt{\varepsilon_{\parallel}}$(согласно (7.9) при $\theta = 90°$).

Рассмотрим волну, падающую нормально на кристаллическую пластинку, вырезанную так, что оптическая ось лежит в плоскости пластинки. Компоненты электрического поля $E_x$ и $E_y$ падающей волны в кристалле будут распространяться независимо: в качестве необыкновенной и обыкновенной волны соответственно. Причём обыкновенная и необыкновенная волны в кристалле пространственно не разделятся. Таким образом, результатом прохождения пластинки толщиной $h$ будет поле, компоненты которого $E_x$ и $E_y$ наберут дополнительную разность фаз

$$\Delta\varphi = kh\,(n_e - n_o) \tag{7.12}$$

где $k = 2\pi/\lambda$ - волновое число в пустоте.

Такая пластинка может быть использована для управления типом поляризации плоских волн за счёт внесения разницы между оптическимим длинами пути для двух ортогональных поляризаций (а значит, и внесения разности фаз между колебаниями соответствующих проекций $E_x$ и $E_y$). В зависимости от соотношения величины вносимой разницы с длиной волны такие пластинки из кристаллов, вырезанные параллельно оптической оси, называются пластинками $\lambda/4, \lambda/2, \lambda$ и т. д. (разность фаз $\Delta\varphi = \pi/2, \pi, 2\pi$ соответственно). Важно помнить, что у одной пластинки это соотношение является разным для разных длин волн, поскольку сама разность фаз (7.12) зависит от длины волны.

### 28.2.6 6. Electro-Optical Effects

Изотропное вещество можно превратить в анизотропное, поместив его в достаточно сильное внешнее электрическое поле. В этом случае электроны в потенциальных ямах под действием поля смещаются к новому положению равновесия. И если зависимость потенциальной энергии от смещения электрона в направлении поля имеет не только чисто квадратичную зависимость $U = ax^2$, но и слагаемые более высоких порядков: $U = ax^2 + \beta x^3 + \gamma x^4$, то в новом положении равновесия $x_0 \neq 0$ значение второй производной $\partial^2 U/\partial x^2 = 2a + 6\beta x_0 + 12\gamma x_0^2$ и соответственно поляризуемость по оси $x$ изменятся и вещество приобретёт свойства, аналогичные свойствам одноосных кристаллов, включая двойное лучепреломление.

Если $\beta \neq 0$, то эффект оказывается линейным по $x_0$ и соответственно по полю, такое явление называется эффектом Поккельса. Оно возможно только в кристаллах, не обладающих центром симметрии, поскольку отличие $\beta$ от нуля означает, что потенциальная яма несимметрична. Вследствие этой ассиметричности поляризуемость по разным осям меняется по-разному и оптическая ось в общем случае не совпадает с направлением внешнего поля.

Если $\beta = 0$, а $\gamma \neq 0$, то изменение $\varepsilon_{\|}$оказывается квадратичным по полю, в этом случае описываемое явление называется эффектом Керра. Но у эффекта Керра может быть и иная природа, отличная от рассмотренной выше нелинейной поляризации. В жидкостях и газах эффект Керра обусловлен анизотропией самих молекул, т. е. их разной поляризуемостью в разных направлениях. В отсутствие внешнего поля эта анизотропия не проявляется, т.к. молекулы ориентированы хаотически и изотропно, но внешнее поле ориентирует дипольные моменты молекул по полю, и эта анизотропия начинает проявляться. Оптическая ось при этом будет направлена по направлению поля.

### 28.2.7 7. Optical Activity. The Faraday Effect

Некоторые вещества обладают так называемой оптической активностью. Оптическая активность - это способность среды (кристаллов, растворов, паров вещества) вызывать вращение плоскости поляризации проходящего через неё оптического излучения. Природа этого явления связана с определённым типом диссимметрии молекул этих веществ, а именно с их спиральностью. Такие молекулы можно условно представить в виде небольших право- или левозакрученных спиралек. При прохождении через такое вещество электромагнитной волны с круговой поляризацией дипольные моменты его молекул и, как следствие, удельный дипольный момент вещества, диэлектрическая проницаемость и коэффициент преломления зависят от того, совпадают или противоположны направления закрученности спиралей молекул и спирали, которую описывает в пространстве конец вектора напряжённости электрического поля. Если же через оптически активное вещество пропустить линейно поляризованную волну, то её можно представить в виде суперпозиции двух поляризованных по кругу волн с противоположным по направлению вращением вектора $\boldsymbol{E}$, при этом у этих волн будет разный коэффициент преломления и, следовательно, разная оптическая длина пути. В результате на выходе из образца у этих волн появится разница фаз, что при суперпозиции даст линейно поляризованную волну с отличным от входного направлением поляризации. Относительная разница коэффициентов преломления волн с противоположной круговой поляризацией при этом очень мала, порядка отношения размеров молекул к длине волны излучения, но при достаточно больших оптических длинах пути в веществе эффект может быть заметен.

Эффект Фарадея. Вращение плоскости поляризации можно также наблюдать, если поместить вещество во внешнее магнитное поле, направленное вдоль направления

распространения волны.

Для объяснения этого эффекта рассмотрим сначала, как и в случае с оптической активностью, взаимодействие с веществом волны с круговой поляризацией. Под действием электрического поля такой волны электроны движутся по окружностям, при этом со стороны магнитного поля на них действует дополнительная сила, которая, в зависимости от направления вращения, может быть либо центробежной, либо центростремительной. Сила со стороны электрического поля, в зависимости от отношения $\omega/\omega_e$ частоты волны и частоты собственных колебаний электронов в потенциальных ямах, также может быть либо центробежной (при $\omega < \omega_e$), либо центростремительной (при $\omega > \omega_e$), т. е. действие постоянного магнитного поля можно свести к изменению модуля силы со стороны высокочастотного электрического поля волны, причём величина и знак этого изменения зависят от направления поляризации и частоты волны. При этом меняется радиус траектории, т. е. либо увеличивается, либо уменьшается величина удельного дипольного момента и соответственно величины диэлектрической проницаемости и коэффициента преломления. А разница коэффициентов преломления волн с противоположной круговой поляризацией приводит, как и в оптически активных веществах, к вращению плоскости поляризации линейно поляризованной волны.

**освещенность**

???

как меряем?

волна ведь типа тонкая, а тут переход к размазанной по пространству как-то произошел!!!

## 28.3 Radiation in Space

(see another note, maybe QM with detailed discussion of it)

### 28.3.1 Black Body Radiation: Basic Laws

(она же в квантмехе и статфизе)

#### Интегральные характеристики теплового излучения

Обсудим вначале вопрос о том, почему мы называем тепловое излучение равновесным, т.е.

почему электромагнитные волны, испускаемые за счет внутренней энергии тел, находятся в равно- весии с излучающими телами. Окружим излучающее тело оболочкой с идеально отражаю- щей поверхностью.

Отраженное оболочкой излучение, упав на тело, поглотится им (частично или полностью), и будет происхо- дить непрерывный обмен энергией между телом и заполняющим оболочку излучением.

Допустим, что равновесие между телом и излучением нарушено и тело излучает энергии больше, чем поглощает.

Тогда внутренняя энергия тела будет убывать, что при- ведет к понижению температуры.

Это в свою очередь обусловит уменьшение количества излучаемой телом энергии.

Температура тела будет понижаться до тех пор, пока баланс излучаемой и поглощаемой энергий не восстановится.

Совершенно аналогично процесс восстановления равновесия будет происходить и при изменении баланса энергий в другую сторону. Таким образом, способность теплового излучения находиться в равновесии с излучающими телами обусловлена тем, что его интенсивность возрастает при повышении температуры.

К рав- новесным состояниям и процессам применимы законы термоди- намики, поэтому и тепловое излучение должно подчиняться ее общим закономерностям.

Прежде всего рассмотрим, как связаны между собой излучательная и поглощательная способности тел.

Рассмотрим замкнутую полость, в которой имеется равновесное излучение при какой-то температуре.

В силу полной изотропии излучения выражение для плотности потока энергии $\Phi_\omega$ в спектральной полосе $d\omega$ и единице телесного угла $d\Omega$ можно записать в виде

$$\frac{d\Phi_\omega}{d\omega}d\Omega = c\rho(\omega)\frac{d\Omega}{4\pi}, \quad d\Omega = \sin\theta d\theta d\varphi$$

Мы использовали тут обычное соотношение для потока $j = \rho v$, с учетом того, что скорость фотона равна скорости света, а плотность есть плотность энергии равновесного излучения.

Как показано на рис.

6.3, на единицу плоЩади поверхности $\Delta S$ полости в единицу времени под углом $\theta$ в интервале углов $d\theta$ и $d\varphi$ падает излучение, энергия которого

$$d\Phi_\omega = \frac{c}{4\pi}\rho(\omega)\cos\theta\sin\theta d\theta d\varphi d\omega$$

Интегрирование по азимутальному углу $\varphi$ в силу изотропии дает $2\pi$.

Учтем теперь свойства поверхности.

Пусть на элементарную площадку поверхности

тела падает поток излучения $d\Phi_\omega$, частота которого заключена в интервале $d\omega$. Часть этого потока $d\Phi'_\omega$ будет поглощена телом Поглощательной способностью тела называется отношение

$$A(\omega) = \frac{d\Phi'_\omega}{d\Phi_\omega}$$

Если наше тело - абсолютно черное, то отношение $A(\omega)$ интенсивностей поглощенного и падающего на тело излучения равно единице.

Тело, для которого $A(\omega) = \text{const} < 1$, называют серым.

Количество энергии, поглощенное в единицу времени, равно $A(\omega)\Phi_\omega d\omega$ и оно должно при равновесии компенсироваться излучением стенки в том же телесном угле.

Если обозначить через $I(\omega)$ поток энергии, испускаемый единицей поверхности тела в интервале частот $d\omega$ (испускательная способность тела), то баланс излученной и поглощенной энергий единицей поверхности стенки под| углом $\theta$ к ней в интервале углов $d\theta$ определяется равенством

$$\frac{c}{4\pi}\rho(\omega)A(\omega)\cdot 2\pi\cos\theta\sin\theta d\theta d\omega = I(\omega)\cdot 2\pi\sin\theta d\theta d\omega$$

или

$$\frac{I(\omega)}{A(\omega)} = \frac{c}{4\pi}\rho(\omega)\cos\theta$$

**закон излучения Кирхгофа**

Полученное нами соотношение между поглощательной и испускательной способностями тела носит название закона Кирхгофа, который получил его в 1859 г. на основе чисто термодинамических закономерностей.

закон кирхгофа гласит, что, хотя функции $a(\omega)$ и $i(\omega)$ различны для разных материалов, их отношение всегда одно и то же и определяется универсальной функцией частоты и направления.

как следует из закона кирхгофа, интенсивность излучения пропорциональна косинусу угла между нормалью к поверхности и направлением излучения.

эта закономерность известна как закон Ламберта.

из закона кирхгофа следует кажущийся вначале удивительным факт: количество излучаемой энергии пропорционально его поглощательной способности, т.е.

лучше всего излучает абсолютно черное тело.

этот результат кажется странным, ибо мы привыкли, что наиболее яркими являются белые предметы.

все дело в том, что мы видим, как правило, не испускаемое телом излучение, а отраженный и рассеянный солнечный свет, который натной температуре, равновесным.

При комнатных температурах равновесное излучение имеет максимум в области невидимого для нашего глаза инфракрасного диапазона.

тот факт, что чем больше поглощательная способность тела (чем оно «чернее»), тем выше его излучательная способность, качественно легко понять из следующих рассуждений.

поместим, как мы уже это делали раньше, какое-либо нагретое тело в замкнутую полость.

при равновесии излучаемая телом энергия должна быть равна поглошаемой энергии, что эквивалентно закону Кирхгофа так как нам известен закон распределения энергии равновесного излучения по частотам, мы можем найти интегральную излучательную способность абсолютно черного тела, т.е. полную

интенсивность излучения с единицы поверхности, находящейся при заданной температуре.

Для этого надо проинтегрировать выражение (6.27) по частоте и по всем углам:

$$J_0 = \frac{c}{4\pi} \int_0^\infty \rho(\omega) d\omega \int_0^{\pi/2} 2\pi \sin\theta \cos\theta d\theta$$

Интегрирование по углам производится просто:

$$\int_0^{\pi/2} 2\pi \sin\theta \cos\theta d\theta = 2\pi \int_0^{\pi/2} \sin\theta (d \sin\theta) = \pi$$

а интегрирование по частотам сводится к вычислению следуюЩего интеграла:

$$\rho = \int_0^\infty \rho(\omega) d\omega = \frac{\hbar}{\pi^2 c^3} \int_0^\infty \frac{\omega^3 d\omega}{e^{\hbar\omega/k_\text{Б} T} - 1}$$

Сделаем замену переменных $x = \hbar\omega/k_\text{Б}\, T$, и тогда интеграл (6.31) преобразуется к виду

$$\rho = \frac{k_\text{E}^4 T^4}{\pi^2 c^3 \hbar^3} \int_0^\infty \frac{x^3 dx}{e^x - 1}$$

Интеграл в полученном выражении равен $\pi^4/15$ и поэтому

$$\rho = \frac{\pi^2 k^4}{15 c^3 \hbar^3} T^4$$

полная интенсивность излучения с единицы поверхности абсолютно черного тела равна

$$I_0 = \frac{c\rho}{4} = \sigma T^4$$

где $\sigma = \pi^2 k_Б^4 \,/\, (60 c^2 \hbar^3) = 5,67 \cdot 10^{-8}$ Дж/м $^2 \cdot$ с $\cdot$ К$^4$.

Полученное выражение было в в 1879 г.

выведено австрийским физиком Й.

Стефаном (1835-1893) из измерений французских физиков, а в 1884 г.

теоретически получено Больцманом на основе применения к излучению понятий давления и температуры и потому называется законом Стефана-Больцмана, и соответственно константа $\sigma$ называется постоянной Стефана-Больцмана.

интенсивность излучения очень сильно зависит от температуры - как четвертая степень абсолютной температуры.

Закон Стефана-Больцмана легко получить качественно на основе следующих рассуждений.

Так как закон дисперсии для фотонов $E = pc$, то из соотношения неопределенностей следует, что фотон частоты $\omega$ занимает объем порядка

$$(\Delta x)^3 \sim \left(\frac{\hbar}{\Delta p}\right)^3 = \left(\frac{\hbar c}{E}\right)^3 = \left(\frac{c}{\omega}\right)^3$$

Напомним, что как только мы стараемся локализовать частицу в интервале $\Delta x$, появляется неопределенность в ее импульсе $\Delta p_x$, определяемая по соотношению неопределенностей, а значит ее энергия $E \approx \cdot \Delta p c$. Так как энергия фотона равна $\hbar\omega$ и она в среднем по спектру порядка $k_Б\, T$, то плотность энергии равновесного излучения, т.е.

энергия в единице объема, равна

$$\rho \approx \frac{\hbar\omega}{(c/\omega)^3} = \frac{\hbar\omega^4}{c^3} \approx \frac{(kT)^4}{\hbar^3 c^3}$$

Полученная оценка достаточно близка к точной формуле (6.32), отличаясь от нее на множитель $\pi^2/15$.

### 28.3.2 Black Body Radiation: Consequences

## 28.4 Radiation in Media

(see another note, maybe electrodynamics with detailed discussion of it)

### 28.4.1 Vavilov-Cherenkov radiation

(see note on ED for now about it)

**излучение Вавилова-Черенкова в двух словах**

**Давние размышления**

(прикольно, что это у [2] в теме интерференция)

хорошее рассуждение:

что линза вот есть, лучи собираются в фокусе, так неужели там была бы бесконечная интенсивность?

(кстати, почему именно бесконечная?)

но в общем, конечная интенсивность именно из-за интерференции.

тут самые интересные вопросы.

есть среда, в ней движется электрон.

Если фазовая скорость электрона больше фазовой скорости среды, то волны,

исходящие в разные моменты могут приходить одновременно в точку наблюдения. В этом суть этого явления. (какого хрена?)

) какое-то рассуждение из картинки и глубокого понимания, что же такое фазовая скорость в среде?)

и в итоге имеем формулу на угол, при котором придут волны одновременно: $\cos\vartheta = \frac{v}{V} = \frac{1}{n\beta}$.

и они конечно же интерферируют.

и может быть как усиление.

так и погашение.

(потом изучу, почему.)

в итоге из-за интерференции излучение происходит вдоль конуса определяемым углом выше, а в дргуих направлениях излучния нет. !!!!! потом я задумаюсь, почему же это так?

изучим, что же происходит.

полная энергия такая:

$$\frac{dW}{dt} = \frac{e^2}{c^2} \int\limits_{\beta n > 1} \left(1 - \frac{1}{\beta^2 n^2}\right) \omega d\omega$$

какого это хрена так??????????????????? ????????????????
?????????????????? ????

прийти еще можно к распределению энергии по спектру частот!!

черт, я вообще хз.

тут имеется связка с ядерной физикой и физикой высоких энергий, так что плюнем, я ничего не понимаю.

## 28.5 Chemical Action of Light (????)

see Landsberg, page 605, par 188. There are 15 pages about it, I'll read maybe later.

## 28.6 Holography

Приёмники света (глаз человека, фотопластинка, фотоэлемент) реагируют на поток энергии световой волны, т. е. величину, пропорциональную квадрату амплитуды. Информация о фазовой структуре волны (т. е. о форме волнового фронта) оказывается при этом утраченной.

В этом разделе мы рассмотрим способ записи изображения, который позволяет по картине интенсивности восстановить полную информацию о волновом поле.

Пусть некоторый предмет (пешка, изображённая на рис. 3.47) освещается когерентным светом лазера и пусть волна, отражённая предметом (будем называть её далее предметной волной), создаёт в плоскости $z = 0$ световое поле:

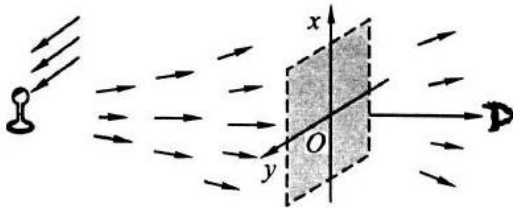
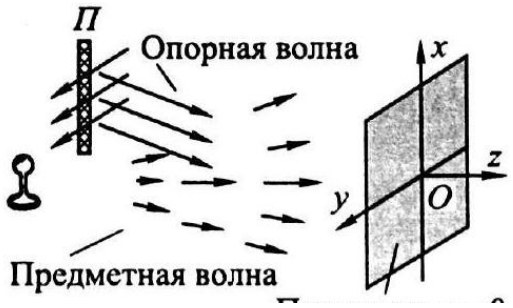


Рис. 3.47. Световое поле предмета

$$f_{\text{П}}(x,y) = a(x,y)e^{i\varphi(x,y)}$$

Установим в плоскости $z = 0$ фотопластинку. Предположим, что функция пропускания фотопластинки после необходимой обработки (проявление, закрепление) пропорциональна интенсивности волны, освещающей пластину во время экспозиции, т. е. пропорциональна величине $I(x,y) = |a(x,y)|^2$. Информация о фазовой структуре волны $\varphi(x,y)$ будет при этом утеряна.

Следуя идее Габора (1948 г.), изменим схему эксперимента: пусть на фотопластинку кроме предметной волны, создающей на фотопластинке поле $f_{\text{п}}(x,y)$, падает ещё некоторая волна с известным распределением амплитуд и фаз колебаний $f_{\text{о}}(x,y) = a_{\text{о}}(x,y)e^{i\varphi_{\text{о}}(x,y)}$, называемая опорной волной (см. рис. 3.48). При этом необходимо обеспечить когерентность предметной и опорной волны - волны должны интерферировать. Например, как показано на рис. 3.48: излучение лазера частично проходит через полупрозрачную пла-




Плоскость $z = 0$

Рис. 3.48. Запись голограммы стинку П, освещая предмет, а частично отражается от неё, создавая

опорный пучок. Когерентность предметной и опорной волн обеспечивается высокой степенью монохроматичности лазерного излучения (разность хода между предметной и опорной волнами меньше длины когерентности для любой точки фотопластинки).

Суммарное поле на голограмме

$$f(x,y) = f_{\text{П}}(x,y) + f_{\text{о}}(x,y)$$

а интенсивность (а следовательно, и функция пропускания фотопластинки после необходимой обработки) есть

$$t(x,y) \propto I(x,y) = |f_{\text{П}}(x,y) + f_{\text{О}}(x,y)|^2$$

Раскрывая квадрат модуля в полученном выражении

$$t \propto |f_{\text{П}}|^2 + |f_{\text{о}}|^2 + f_{\text{П}}f_{\text{о}}^* + f_{\text{П}}^* f_{\text{О}} = a^2 + a_{\text{о}}^2 + 2aa_{\text{о}}\cos(\varphi - \varphi_{\text{о}})$$

убеждаемся, что в функции пропускания (а именно, в слагаемом, связанном с интерференцией опорной и предметной волн) сохранилась информация о фазе предметной волны $\varphi(x,y)$.

Фотопластинка с зарегистрированным на ней результатом интерференции предметной и опорной волны называются голограммой.

## 28.6.1 3.1. Point Source Hologram (Gabor Hologram)

Для уяснения сути метода рассмотрим в качестве предмета точечный источник света $S$, т. е. создадим сферическую предметную волну:

$$f_{\text{П}} = \frac{a_{\text{П}}}{r} e^{ikr}$$

где $r = \sqrt{z_0^2 + x^2 + y^2}$ - расстояние от источника $S$ до точки $(x, y)$ фотопластинки.

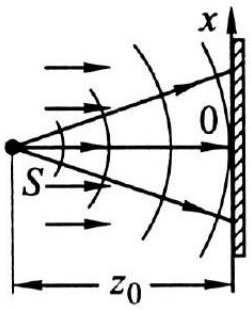



Рис. 3.49. Запись голограммы точечного источника

В качестве опорной волны возьмём плоскую волну, бегущую вдоль оси $z$ и падающую нормально на фотопластинку, расположенную в плоскости $z = 0$ (рис. 3.49) :

$$f_{\text{о}} = ae^{ikz}$$

т. е. опорная волна создаёт во всех точках фотопластинки поле одинаковой амплитуды и фазы.

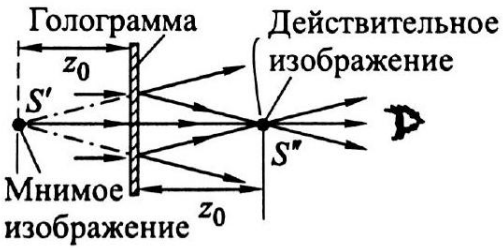
Принимая фазу колебаний в плоскости $z = 0$ равной нулю, запишем $f_{\text{о}} = a$. Кроме того, для упрощения формул будем далее полагать, что амплитуда сферической волны во всех точках фотопластинки равна амплитуде плоской волны, т.е. $f_{\text{п}} \approx ae^{ikr}$. Суммарное поле в таком случае имеет вид

$$f = ae^{ikr} + a.$$

После необходимой фотообработки получаем голограмму с функцией пропускания:

$$t(x, y) \sim I(x, y) = \left|a + ae^{ikr}\right|^2 \tag{3.83}$$

Мы описали первую стадию голографического процесса - процесс записи голограммы. Теперь




необходимо использовать получен-

Рис. 3.50. Восстановление голограммы точечного источника ную голограмму для восстановления (реконструкции) изображения.

Устанавливаем голограмму в плоскости $z = 0$ и освещаем её плоской нормально падающей волной (рис. 3.50). Для упрощения формул будем считать амплитуду волны равной единице, а фазу без ограничения общности равной нулю, т. е. комплексная амплитуда этой волны (её называют восстанавливающей волной) есть $f_{-}(x, y) = 1$.

Тогда на выходе голограммы, в плоскости, примыкающей к ней справа, получим

$$f_+(x,y) = f_-(x,y)t(x,y) = \left|a + ae^{ikr}\right|^2 \tag{3.84}$$

Итак, суммарное поле за голограммой описывается действительной положительной функцией. Исследуем его более детально.

Из (3.84) находим

$$f_+(x,y) = 2a^2 + a^2e^{ikr} + a^2e^{-ikr} = 2a^2(1+\cos(kr)) \tag{3.85}$$

Поле в плоскости $z = 0_+$представляется в виде суммы трёх слагаемых. Соответственно волна в области $z > 0$ есть сумма трёх волн: каждое слагаемое на границе $z = 0_+$ответственно за появление «своей» волны в области $z > 0$.

Первое слагаемое $2a^2$ «отвечает» за появление волны $f_1 = 2a^2e^{ikz}$ - плоской волны, бегущей в области $z > 0$ вдоль оси $z$. Второе слагаемое $f_2 = a^2e^{ikr}$ - волновое поле со сферическим волновым фронтом: именно такое поле создавал в каждой точке плоскости $z = 0$ точечный

источник света, который находился на расстоянии $z_0$ слева от голограммы в процессе записи. Ясно, что и в области $z > 0$ поле $f_2$ порождает расходящуюся сферическую волну, неотличимую от предметной волны (сферической волны, исходящей из точки $S$ в процессе записи голограммы). Поэтому наблюдатель, который находится справа (в области $z > 0$) и смотрит на голограмму как в окошко, «увидит в окне» на расстоянии $z_0$ слева от голограммы светящуюся точку $S'$, т. е. расходящаяся сферически волна кажется наблюдателю исходящей из точки $S'$, хотя никакой светящейся точки на стадии реконструкции изображения нет, - есть только фотопластинка-голограмма, освещаемая плоской волной. Это - мнимое изображение предмета.

Рассмотрим теперь последнее, третье, слагаемое в (3.85): $f_3 = a^2e^{-ikr}$. Волновой фронт в этой волне «вывернут наизнанку» по отношению к волне $f_2$: если в волне $f_2$ колебание в точке $(x,y)$ плоскости $z = 0$ отстаёт по фазе от колебаний в начале координат, то в волне $f_3$ колебание в точке $(x,y)$ опережает по фазе колебание в начале координат на ту же величину. Это слагаемое описывает поле, которое создавала бы в плоскости $z = 0$ сферическая волна, сходящаяся в точку $S''$ на расстоянии $z_0$ справа (рис. 3.50). Ясно поэтому, что граничное поле $f_3$, возникающее в процессе реконструкции, порождает сферическую волну, сходящуюся в точку $S''$. Это - действительное изображение предмета. Его можно наблюдать, расположив на расстоянии $z_0$ за голограммой экран наблюдения, - мы увидим яркую точку, в которую сфокусирована волна.

Описанные закономерности сохраняются в общих чертах при голографировании произвольного предмета: в процессе реконструкции возникают три волны: волна, создающая мнимое изображение, которое находится слева от голограммы, там же, где находился предмет при записи; волна, создающая действительное изображение, которое располагается симметрично справа от голограммы; а также волна, бегущая вдоль оси $z$ и не несущая информации о форме волнового фронта предметной волны.

Зонная решётка Габора. Вернёмся к интерференционной картине (3.85), которая записана на голограмме точечного источника. Используем приближённое выражение

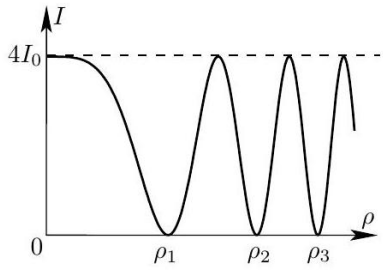
$$r \approx \sqrt{z_0^2 + x^2 + y^2} \approx z_0 + \frac{x^2+y^2}{2z_0} = z_0 + \frac{\rho^2}{2z_0}$$

где $\rho^2 = \sqrt{x^2+y^2}$ - расстояние от начала координат до точки $(x,y)$ в плоскости фотопластинки. Соответственно фаза колебаний в сфериче ской волне в точке $(x,y)$ есть

$$\varphi = kr \approx kz_0 + \frac{k}{2z_0}\rho^2 \tag{3.86}$$

Принимая начальную фазу на оси $z$ в плоскости $z = 0$ нулевой, запишем: $\varphi = \frac{k}{2z_0}\rho^2$. Мы полагаем, таким образом, что колебания опорной (плоской) волны и сферической волны в точке $(x = 0, y = 0)$ синфазны. Тогда получим

$$I(\rho) = 2a^2\left(1 + \cos\frac{k\rho^2}{2z_0}\right) \tag{3.87}$$




Мы видим, что интерференционная кар-

Рис. 3.51. Зависимость $I(\rho)$

тина имеет вид колец, центр которых находится в начале координат. Функция $I(\rho)$ показана на рис. 3.51. Радиус первого (тёмного) кольца $\rho_1$ находим из условия $k\frac{\rho_1^2}{2z_0} = \pi$ (при этом $\cos k\frac{\rho_1^2}{2z_0} = -1$ и $I(\rho_1) = 0$), откуда получаем: $\rho_1 = \sqrt{\lambda z_0}$. Вообще радиусы светлых и тёмных колец находятся по формуле

$$\rho_m = \sqrt{m\lambda z_0} \tag{3.88}$$

(при $m$ нечётном - тёмные кольца, при $m$ чётном - светлые), которая совпадает с выражением, определяющим радиусы тёмных и светлых колец в зонной пластинке Френеля.

Отличие состоит в том, что в нашем случае переход от светлых колец к тёмным происходит плавно. Интерференционная картина (3.85) (голограмма точечного источника) показана на рис. 3.52 и называется зонной решёткой Габора. Как следует из проведённого выше анализа, зонная решётка Габора работает одновременно и как собирающая линза (фокусируя параллельный пучок света в точку $S''$ - действительный фокус зонной пластинки), и как рассеивающая линза, которая преобразует параллельный пучок, падающий на пластинку, в расходящуюся сферическую волну, исходящую из мнимого фокуса $S'$, и как плоскопараллельная пластинка, поскольку часть света за пластинкой сохраняет структуру освещающего параллельного пучка.



Рис. 3.52. Зонная решётка Габора

Сферическая восстанавливающая волна. Рассмотрим ещё одно свойство голограммы точки. Пусть восстанавливающая волна является не плоской, а сферической, расходящейся от точки, расположенной на расстоянии $z_1 > z_0$ от голограммы. Используя параболическое приближение (3.86), запишем комплексную амплитуду падающей волны в плоскости голограммы:

$$f_-(\rho) = e^{-i\frac{k}{2z_1}\rho^2} \tag{3.89}$$

Восстановленное поле в плоскости $z = 0$:

$$f_+(\rho) = f_-(\rho)t(\rho)$$

Используя (3.89) и (3.83), получим (с точностью до несущественных констант)

$$f_+(\rho) = e^{-i\frac{k}{2z_1}\rho^2} + e^{-i\frac{k}{2}\left(\frac{1}{z_1}+\frac{1}{z_0}\right)\rho^2} + e^{-i\frac{k}{2}\left(\frac{1}{z_0}-\frac{1}{z_1}\right)\rho^2}$$

Из этого выражения видно, что расстояние $b_2$ от голограммы до мнимого изображения источника уменьшается, так как

$$\frac{1}{b_2} = \frac{1}{z_0} + \frac{1}{z_1}$$

т. е. это изображение приближается к голограмме. В то же время расстояние $b_3$ от голограммы до действительного изображения увеличивается:

$$\frac{1}{b_3} = \frac{1}{z_0} - \frac{1}{z_1}$$

Это свойство голограммы следует иметь в виду при объяснении изменения масштабов изображений трёхмерного предмета при освещении голограммы сферической волной.

В случае произвольного предмета на голограмме записывается сложный интерференционный узор, который можно рассматривать как совокупность зонных кольцевых решёток Габора, причём каждая из них ответственна за восстановление «своей» точки предмета на стадии реконструкции изображения.

## 28.6.2 3.2. The Resolution of the Hologram

Рассмотрим более детально волну, ответственную за создание действительного изображения. Конечно, сходящаяся сферическая волна не фокусируется строго в точку $S''$: изображением является не точка, а маленькое дифракционное пятнышко. Действительно, плоская восстанавливающая волна, пройдя сквозь голограмму, преобразуется в сферическую волну (мы говорим сейчас о той части волны, которая ответственна за создание действительного изображения), причём сферический волновой фронт ограничен размерами голограммы. Эта сходящаяся сферическая волна (третье слагаемое в (3.85)) не отличается от сферической волны, возникающей за собирающейся линзой, размер $D$ которой равен размеру голограммы. При фокусировке света линзой возникает дифракционное пятно Эйри, размер которого определяется формулой (3.74). Ясно, что и голограмма создаёт изображение - дифракционное пятно, размер которого определяется аналогичной формулой:

$$\Delta x \sim \frac{\lambda}{D} z_0 \tag{3.90}$$

где $z_0$ - расстояние от точечного источника до голограммы в процессе записи, $D$ - размер голограммы. Если записывать на голограмму изображения двух точек, расстояние между которыми меньше $\Delta x$, то при восстановлении изображений возникают два пятна, налагающиеся друг на друга так, что, согласно критерию Рэлея, они оказываются неразрешимыми. Сказанное касается, разумеется, и мнимого изображения точечного источника: наблюдателю оно кажется маленьким пятнышком размера $\Delta x$, находящимся на расстоянии $z_0$ за голограммой.

Важнейшее свойство голограммы состоит в том, что любой её малый участок содержит информацию обо всем объекте: ведь поле в каждой точке голограммы является суперпозицией полей, посылаемых всеми точками предмета (и опорной волной). Другими словами, интерференционная картина на каждом небольшом участке голограммы содержит информацию об амплитуде и фазе колебаний всех точек предмета, поэтому изображение может быть восстановлено с помощью небольшого осколка голограммы, полученной при записи. Разумеется, разрешающая способность определяется размером осколка. [10]

## 28.6.3 3.3. A Diagram with an Inclined Support Beam

При наблюдении голографических изображений, полученных по схеме Габора, действительное и мнимое изображения создают взаимные помехи: изображения $S'$ и $S''$ находятся на одной прямой с точкой, в которой расположен глаз наблюдателя (рис. 3.50). Помехи создаются также возникающей при восстановлении плоской волной, бегущей вдоль оси $z$ и не несущей информации об объекте. В 1962 году Лейт и

10 В реальной голограмме разрешающая способность дополнительно ограничивается такими факторами, как разрешающая способность фотоэмульсии (зернистость), засветка голограммы сторонними источниками (пятна) и пр.

Упатниекс предложили использовать для записи голограммы наклонный опорный пучок.

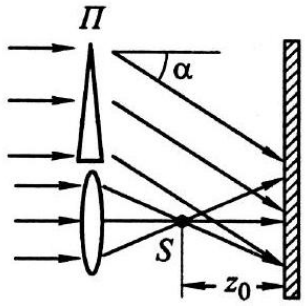



Рис. 3.53. Наклонный опорный пучок

Проследим, каким образом введение опорного наклонного пучка обеспечивает возможность независимого (без взаимных помех) наблюдения действительного и мнимого изображений. На рис. 3.53 «предметом» является «точечный источник» $S$. Наклон опорного пучка обеспечивается призмой П. Поле предметной волны на голограмме есть

$$f_{\Pi}(x,y) = ae^{i\frac{k}{2z_0}\left(x^2+y^2\right)}$$

(мы вновь используем параболическое приближение). Наклонная опорная волна имеет вид $f_0 = ae^{iu_0x}$, где $u_0 = k\sin\alpha$ - пространственная частота опорной волны. Мы полагаем, что волны имеют равные амплитуды и создают в начале координат плоскости $xy$ (плоскость голограммы) синфазные колебания. Функция пропускания проявленной голограммы, пропорциональная интенсивности суммарной волны при записи, имеет вид

$$t(x,y) \sim \left|ae^{iu_0x} + ae^{i\frac{k}{2z_0}\left(x^2+y^2\right)}\right|^2$$

Предоставляем читателю убедиться, что слагаемые, ответственные за появление действительного и мнимого изображения при восстановлении, можно записать в виде:

$$f_3(x,y) \sim be^{i\frac{k}{2z_0}\left((x+z_0\sin\alpha)^2+y^2\right)}, \quad f_4(x,y) \sim be^{-i\frac{k}{2z_0}\left((x+z_0\sin\alpha)^2+y^2\right)}$$

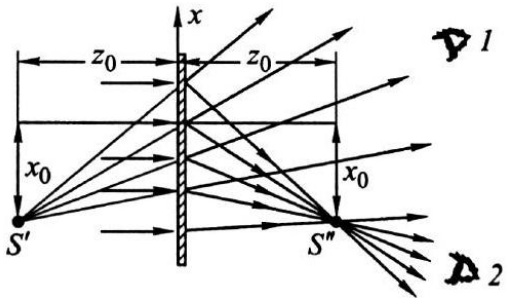



Рис. 3.54. Мнимые и действительные изображения в схеме с наклонным пучком

где $b$ - комплексная константа. Выражение $f_3(x, y)$ представляет собой поле, которое создавал бы в плоскости $z = 0$ точечный источник, находящийся в точке с координатами $x_0 = -z_0 \sin\alpha, y = 0$, на расстоянии $z_0$ слева от голограммы. Это мнимое изображение предмета $S'$ (рис. 3.54). Соответственно $f_4(x, y)$ описывает поле сходящейся в точку $S''$ сферической волны. Эта точка находится справа от голограммы на расстоянии $z_0$ и смещена относительно оси $z$ на расстояние $x_0 = -z_0 \sin\alpha$. Это действительное изображение. Мы видим, что введение наклонного опорного пучка приводит к сдвигу мнимого и действительного изображений таким образом, что наблюдатель может рассматривать их из разных положений (1 и 2 на рис. 3.54), в которых они не создают взаимных помех. Разумеется, угол наклона опорного пучка $\alpha$ (т. е. его пространственная частота) должен быть выбран правильно: чем шире пространственный спектр предметной волны, тем больше должен быть угол наклона опорного пучка, при котором волны, ответственные за создание действительного и мнимого изображений, не создают взаимных помех.

**Литература**

1. Кингсеп А.С., Локшин Г.Р., Олъхов О.А. Основы физики. Курсы общей физики: учебн. в 2 т. Т. 1. Механика, электричество и магнетизм, колебания и волны, волновая оптика / под ред. А.С. Кингсепа. - М.: Физматлит, 2001. Ч. 3, гл. 8.

2. Сивухин Д.В. Общий курс физики. Т. IV. Оптика. - М.: Физматлит, 2005. Гл. IV.

3. Ландсберг Г.С. Оптика. - М.: Физматлит, 2003. Гл. VIII-XI.

4. Бутиков Е.И. Оптика. - М.: Высшая школа, 1986. Гл. 7.

5. Борн М., Волъф Э. Основы оптики. - М.: Наука, 1973. Гл. 8.

### 28.6.4 Connection of Holography in Optics and Holography in High Energy Physic

(??? still, it is a good question, even though few people really want to understand it)

## 28.7 Other Effects

(мб потом выше их вставлю, пока что это просто как отдельные эффекты и все, хочу - посмотрю, нет - забью)

### 28.7.1 Photonic Crystals

(пока не до них)
(?? что это вообще такое???)

**Обзор**

**О применениях**

**используются в оптических волокнах**

потому что волокна без них при изгибах портятся, а с фотонными кристаллами есть возможность сгибать из потерь света.
"hollow fiber"

**используются в волноводах**

(???)

### 28.7.2 Luminescence by Landsberg (???)

(??? if needed, I'll look at it in his book)

### 28.7.3 Optical Tweezers (?)

(разберу потом, интересно, просто не до этого всегда было.)

## 29 Special optical devices and applications

### 29.1 Modeling of Optical Phenomena

то, что самое интересное, очень интересно было бы всё это моделировать.

### 29.2 Analysis of Images

(??? how to do it???)

**фурье анализ**

### 29.3 3D computer graphics physical fundamentals

(будет время свободное и подумаю про это - допишу)

#### 29.3.1 Creation a 3D Graphics Program

Это позволяет и в консоле как анимировать разные простейшие объекты, так и запускать даже игры.
Также моделировать фракталы.
и кто знает, что еще?
так что обсудим такое приложение оптики.

**О создании 3д графики**

**О создании 3д графики**

**О создании 2д графики**

**О создании 3д графики в консоли**

(такое интересное дополнение, почему нет)

**О рейкастинге**

И там классные моделирования.

**О реймарчинге**

реймарчинг - быстрое моделирование.
Популярная тема.

### 29.3.2 Ray tracing

(известный метод)

## 29.4 Photonic Crystals

(??? I'll read later about it if needed. see my special note about it)
https://www.youtube.com/watch?v=TJcB2cJ3_k4

## 29.5 Attosecond pulses

(??? I'll read later about it if needed)

# 30 Exotic Theoretical Models and Methods

## 30.1 Exotic Theoretical Methods

(??? I'll write it later, I don't know, which are important and which methods are dead?)

### 30.1.1 $S$-matrix Method for Optics (!?!??)

**Scattering matrix from "Anomalies in Light Scattering" by Alex Krasnok**

(ссылки на статью под номером: Quasinormal-Mode Expansion of the Scattering Matrix Filippo Alpeggiani,1,2 Nikhil Parappurath,1 Ewold Verhagen,1 and L. Kuipers1,2, там те же формулы, так что прописать их по идеее все-таки нужно.)

We assume a linear scatterer is situated in a dielectric medium ($\varepsilon_h$) and described by an inhomogeneously distributed complex-valued relative dielectric permittivity $\varepsilon_s(\omega, \mathbf{r})$. We can define $\Delta\varepsilon(\omega, \mathbf{r}) = \varepsilon_s(\omega, \mathbf{r}) - \varepsilon_h$ and separate the total electric field into the incident $\mathbf{E}_i$ and scattered $\mathbf{E}_s$ components, $\mathbf{E} = \mathbf{E}_i + \mathbf{E}_s$. The total electric field satisfies the Helmholtz equation

$$\nabla \times \nabla \times \mathbf{E}(\omega, \mathbf{r}) - [\Delta\varepsilon(\omega, \mathbf{r}) + \varepsilon_h] \frac{\omega^2}{c^2} \mathbf{E}(\omega, \mathbf{r}) = 0,$$

which stems from Eqs. (1) in the frequency-domain. This equation is equivalent to the Fredholm integral equation for the scattered field

$$\mathbf{E}_s(\mathbf{r}) = \frac{\omega^2}{c^2} \int \hat{G}\left(\omega, \mathbf{r}, \mathbf{r}'\right) \Delta\varepsilon\left(\omega, \mathbf{r}'\right) \left[\mathbf{E}_i\left(\mathbf{r}'\right) + \mathbf{E}_s\left(\mathbf{r}'\right)\right] d^3\mathbf{r}'$$

Here, $\hat{G}\left(\omega, \mathbf{r}, \mathbf{r}'\right)$ is the Green's function of the Helmholtz equation in the host medium[30]. In the operator notation Eq. (3) is written as $\mathbf{E}_s = \hat{L}\left(\mathbf{E}_i + \mathbf{E}_s\right)$, where $\hat{L}$ is the integral operator. Formally, the solution of this equation can be written with the use of the scattering operator $\hat{S}$:

$$\mathbf{E}_s = \hat{S}\mathbf{E}_i = (\hat{I} - \hat{L})^{-1}\hat{L}\mathbf{E}_i.$$

The above equation is the formal solution of the scattering problem. To get a deeper understanding of the mechanics underlying various resonant effects, we need to have a closer look at the structure of $\hat{S}$. In practice, it is convenient to represent the incident and scattered waves in a specific basis of incoming and outgoing waves, or channels, as depicted in Fig. 1(a):

$$\mathbf{E}_i = \sum_p s_p^+ \mathbf{E}_p^+, \mathbf{E}_s = \sum_p s_p^- \mathbf{E}_p^-.$$

The set of amplitudes for the incoming and outgoing waves is denoted as $\mathbf{s}^+ = \left\{s_1^+, s_2^+, \ldots\right\}$ and $\mathbf{s}^- = \left\{s_1^-, s_2^-, \ldots\right\}$, respectively. Hence, in the basis of these channels, Eq. (4) can be rewritten as

$$\mathbf{s}^- = \hat{S}\mathbf{s}^+,$$

where $\hat{S}$ is the scattering matrix of the system[31][32]. Essentially, Eq. (6) represents the matrix formulation of the scattering problem: amplitudes of the outgoing waves are linear combinations of the incoming wave amplitudes, and this relationship is established by the scattering matrix, as depicted in Fig. 1(a). Usually, the channels are normalized such that $\left|s_p^+\right|^2$ and $\left|s_p^-\right|^2$ correspond to the energy flux of the incoming and outgoing waves in a channel $p$.

Using the definition of channels in Eq. (5), we can introduce scattering $S_{pq}$ and reflection $S_{pp} \equiv S_p$ coefficients as

$$S_{pq} = \frac{s_q^-}{s_p^+}, S_p = \frac{s_p^-}{s_p^+},$$

which describe scattering from a channel $p$ to a channel $q$ and reflection to the same channel $p$, respectively. For example, in the case of extended flat objects, considered in Section (2), channels $p$ and $q$ are simply incident ("1"), reflected ("2"), and transmitted ("3") plane waves. Therefore, Eqs. (7) define the well-known Fresnel's reflection $\left(r \equiv S_{12} = s_2^-/s_1^+\right)$and transmission $t \equiv S_{13} = s_3^-/s_1^+$coefficients. In the case of 2D or 3D scatterers, see Section (5), the channels may be cylindrical or spherical harmonics. Equation (7) allows us to introduce energy conservation relations, $\left|S_{pq}\right|^2, \left|S_p\right|^2 \leq 1$, which apply to any passive (namely, no generation of energy) system. Generally, this statement is not valid for $\mathcal{PT}$-symmetrical systems, discussed in Section (3), since they involve active elements.

The channels introduced in Eq. (5) exist outside the system and represent freely propagating solutions of the wave equation in the absence of the scattering object. The specific choice of channels often depends on the exact geometry of the system. For a planar slab and normal incidence, for example, the basis of normally incident plane waves is the obvious choice, whereas, for a spherical particle, it is convenient to use a basis of vector spherical harmonics. 1.2 Hamiltonian approach It is instructive to introduce the effective Hamiltonian approach, which provides a bottom-up scheme to construct the S-matrix of a system and also highlights the essential features and limitations of the coupled-mode theory[33-37] for wave scattering in

various photonic structures[34,38-40]. To some extent, this effective Hamiltonian approach can be considered an advanced form of the standard coupled-mode theory, taking into account the dispersive properties of the channels[41]. Let us consider, to be specific, an optical scatterer described by the effective Hermitian Hamiltonian $\hat{H}_0$, which is coupled to $M$ propagating channels labeled by $p = 1, 2, \ldots, M$. In the modal space of the scatterer, the full Hamiltonian $\hat{H}_0$ is given as a $N \times N$ Hermitian matrix with $N$ being the number of supported modes.

The unperturbed portion of $\hat{H}_0$ is typically associated with an isolated ideal scatterer with canonical geometry, for which the modes are analytically solvable. Correspondingly, the diagonal terms of the Hamiltonian $\hat{H}_0$ represent the resonant frequencies of the normal modes in this basis, while the off-diagonal ones denote the coupling between them due to imperfections, such as deformations, defects, etc.[42]. On the other hand, the propagating channel $p$ is associated with the $p$-th waveguide coupling one of the impinging channels to the scatterer.

For waveguides fulfilling the Helmholtz equation, Ref.[41] develops a general treatment of the corresponding Hamiltonian model. Here, for simplicity we consider Coupled-Resonator Optical Waveguides (CROWs) satisfying the Schrodinger-like equation[43]. We set the distance between the nearest-neighbor resonators to be $a_0 = 1$ with dispersion $\omega = 2c_p \cos k_p$ for the $p$-th channel, after suitably fixing the reference frequency, and $c_p < 0$ is the coupling strength between its nearest sites/resonators. Generally, within the channel $p$, the field profile of a monochromatic wave can be written in terms of counter-propagating waves as

$$\psi_p(x,t) = \frac{s_p^+}{\sqrt{v_{g,p}}} e^{-i\omega rt + ik_p x} + \frac{s_p^-}{\sqrt{v_{g,p}}} e^{-i\omega v - ik_p x},$$

where the group velocity $v_{g,p} \equiv \partial\omega/\partial k_p = -2c_p \sin k_p$, $k_p \in (0, \pi)$ and the integer $x = -1, -2, \ldots, -\infty$ labels the sites in each waveguide. In Eq. (8), $\left|s_p^{\pm}\right|^2$ are normalized to be the incoming/outgoing flux for each channel $p$.

If $M$ channels at $x = -1$ are coupled through the scatterer with $N$ modes, a real $M \times N$ bare coupling matrix $-\hat{W}$ describes the process. Correspondingly, Schrodinger-like equations describing this setup can be written as

$$\begin{aligned} i\frac{d}{dt}\mathbf{a}(t) &= \hat{H}_0\mathbf{a}(t) - \hat{W}^T\boldsymbol{\psi}(-1,t), \\ i\frac{d}{dt}\boldsymbol{\psi}(-1,t) &= \hat{c}\psi(-2,t) - \hat{W}\mathbf{a}(t), \end{aligned}$$

where $\boldsymbol{\psi}(x,t) = (\psi_1, \psi_2, \cdots, \psi_M)^T$, $\hat{c} = \text{diag}\{c_1, c_2, \cdots c_M\}$ and the vector $\mathbf{a}(t) = (a_1, a_2, \cdots, a_N)^T$ represents the set of mode amplitudes, which are normalized such that $|a_n|^2$ corresponds to the energy in the $n$-th mode. Note that Eq. (9a) can be considered a definition of the Hamiltinian $\hat{H}_0$. We proceed to derive the scattering matrix $\hat{S}$ based on Eq. (9). First, we derive the coupledmode equation for the wave scattering from Eq. (9). Essentially setting $\mathbf{a}(t) = \mathbf{a}(0)e^{-\text{lat}}$ and plugging Eq. (8) into Eq. (9), we get the coupled-mode equations

$$\begin{aligned} -i\omega\mathbf{a}(0) &= -i(\hat{H} - i\hat{\Gamma})\mathbf{a}(0) - \hat{K}^T\mathbf{s}^+, \\ \mathbf{s}^- &= -\mathbf{s}^+ - \hat{K}\mathbf{a}(0), \end{aligned}$$

with the effective Hamiltonian

$$\hat{H} = \hat{H}_0 - \frac{1}{2}\hat{K}^T \cot\hat{k}\hat{K}$$

where $\Gamma = \frac{1}{2}\hat{K}^T\hat{K}$[44], and $\hat{K} \equiv \hat{c}^{-1}\sqrt{\hat{v}_g}\hat{W}$ is the normalized $M \times N$ coupling matrix, with $\hat{v}_g \equiv \text{diag}\{v_{g,1}, v_{g,2}, \cdots, v_{g,M}\}$ being the group-velocity matrix of the channels. The scattering

matrix $\hat{S}$ can be obtained from Eq. (10) by eliminating a(0) and using the definition of $\hat{S}$. Specifically, we have

$$\hat{S} = -\hat{I}_M + i\hat{K}\frac{1}{\omega - \hat{H} + i\hat{\Gamma}}\hat{K}^T,$$

This expression explicitly contains the scattering matrix $\hat{S}$ of an object expressed in terms of its Hermitian Hamiltonian $\hat{H}_0$, coupling operator $\hat{K}^T$, and the incident frequency $\omega$ of the monochromatic wave. Interestingly, the structure of the scattering matrix Eq. (12) is general, irrespective of the model for the channels [45,46], and can be extended even more generally starting from the vector Helmholtz equation Eq. (2) (see Ref. [47] and references therein). We also note that although Eq. (12), as well as the coupled-mode theory, is derived in the weak coupling regime, it may be generalized to strongly coupled systems by changing the basis to "dressed states" [39]. Useful distribution information about scattering singularities can be extracted from Eq. (12) by assuming proper ensembles of the random matrix $\hat{H}_0$ based on the preserved symmetry class, such as the Gaussian orthogonal/unitary ensemble (GOE/GUE) for isolated scatterers with/without time-reversal symmetry[48,49]. Nevertheless, the information required for the reconstruction of the scattering matrix $\hat{S}$ based on the Hamiltonian $\hat{H}_0$ and the coupling operator $\hat{K}$ may be overwhelming. Therefore, it is meaningful to develop an approach for the reconstruction of $\hat{S}$ using the minimal available knowledge, say the complex eigenfrequencies and the far-field properties of the quasinormal modes (or resonant states) of the system, see Ref.[50-52].

The quasinormal modes (QNMs) of photonic structures are defined as the solution of Eq. (9) with outgoing radiation boundary conditions [ 51, 53, 54], whose completeness has been proven for simple geometries (e.g., spherically symmetric objects) and is a conjecture in other scenarios[52,55-57]. The QNMs are associated with the poles of the scattering matrix in Eq. (12). To see this clearly, setting $\mathbf{s}^+ = 0$ in Eq. (10a), corresponding to outgoing boundary conditions, we get a set of complex eigenfrequencies of the quasinormal modes $\tilde{\omega}_n$ and a corresponding eigenset of QNMs profiles $\tilde{\mathbf{a}}_n(0), n = 1, 2, \cdots, N$. The complex eigenfrequencies $\{\tilde{\omega}_n\}$ turn out to be the poles of the scattering matrix $\hat{S}$ in Eq. (12), i.e., $\det\left(\tilde{\omega}_n - \hat{H} + i\Gamma\right) = 0$. Moreover, the asymptotic behavior of the QNMs in the outgoing channels is $\tilde{\mathbf{b}}_n = -\hat{K}\tilde{\mathbf{a}}_n(0)$, which defines the structure of radiation leaking from resonant states into free space.

Next, we can reconstruct the scattering matrix $\hat{S}$ based on its quasinormal mode expansion, following Ref.[50]. It is important to point out the implicit assumption in optical systems that tacitly ignores the frequency-dependence of the Hermitian Hamiltonian $\hat{H}$ and coupling operators $\hat{K}$. This assumption is the analog of the so-called wideband approximation in condensed-matter physics[58]. For the reconstruction, the required information can be written compactly as the $N$ dimensional diagonal matrix $\tilde{\Omega} = \mathrm{diag}\,(\tilde{\omega}_1, \tilde{\omega}_2, \cdots, \tilde{\omega}_N)$ for the complex eigenfrequency spectrum and the $M \times N$ matrix $\tilde{B} = \left(\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \cdots, \tilde{\mathbf{b}}_N\right)$ for the far-field properties of the QNMs. First, we notice that

$$(\hat{H} - i\hat{\Gamma})\tilde{A} = \tilde{A}\tilde{\Omega}$$
$$\tilde{B} = -\hat{K}\tilde{A}$$

where the $N \times N$ matrix $\tilde{A} = (\tilde{\mathbf{a}}_1(0), \tilde{\mathbf{a}}_2(0), \cdots, \tilde{\mathbf{a}}_N(0))$ with the columns being the profile of the QNMs inside the scatterer. The above relations allow us to rewrite the S-matrix in Eq. (12) as

$$\hat{S} = -\hat{I}_M + i\tilde{B}\frac{1}{\omega - \tilde{\Omega}}\Lambda^{-1}\tilde{B}^T,$$

where $\Lambda \equiv \tilde{A}^T\tilde{A} = \mathrm{diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_N)$ is a diagonal matrix, which follows from the reciprocity of the structure, i.e., mathematically $(\hat{H} - i\hat{\Gamma})^T = \hat{H} - i\hat{\Gamma}$. Next, we determine the matrix $\Lambda$ (or specifically its diagonal entries $\lambda_N$) from the given information of the quasinormal-modes $\tilde{\Omega}$

and $\tilde{B}$. Since the coupling operator $\hat{K}$ is a real matrix, we have the identity $\tilde{B}^* \left(Q^*\right)^{-1} \Lambda = \tilde{B}$, where $Q \equiv \tilde{A}^+ \tilde{A}$ with components $Q_{nn'} = i\frac{\tilde{b}_n^+ \tilde{b}_{n'}}{\tilde{\omega}_n^* - \tilde{\omega}_{n'}}$, which is determined completely from the available information[50]. For convenience, we can define $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) = \tilde{B}^* \left(Q^*\right)^{-1}$ with the vector $\mathbf{x}_n$ as the columns of the latter matrix, which allows us to rewrite the last equation equivalently as $\mathbf{x}_n \lambda_n = \tilde{\mathbf{b}}_n$, where $n = 1, 2, \cdots, N$. Under the present formalism, this last equation is overdetermined but consistent to be used to determine $\lambda_N$. Considering potential numerical or measurement errors, a unique solution of $\mathbf{x}_n \lambda_n = \tilde{\mathbf{b}}_n$ in the least square sense can be given, which reads $\lambda_n = \mathbf{x}_n^+ \tilde{\mathbf{b}}_n / \mathbf{x}_n^+ \mathbf{x}_n$. Finally, the desired quasinormal-mode expansion of the scattering matrix follows[50]

$$\hat{S} = -\hat{I}_M + i \sum_{n=1}^{N} \frac{x_n^+ x_n}{x_n^+ \tilde{\mathbf{b}}_n} \frac{\tilde{\mathbf{b}}_n \tilde{\mathbf{b}}_n^T}{\omega - \tilde{\omega}_n},$$

where the vector $\mathbf{x}_n$ can be written explicitly as $\mathbf{x}_n = \sum_{n'} \left(Q_{n'n}^{-1}\right) \tilde{\mathbf{b}}_{n'}^*$. The material gain/loss can be easily incorporated into the previous formalism by setting $\hat{H}_0 \to \hat{H}_0 - i\Gamma_{\text{nr}}$ in Eq. (12), where $\Gamma_{nr}$ is a Hermitian decay matrix. When the non-radiative gain/loss is small compared to the frequency of the mode, the diagonal non-radiative gain/loss terms $\Gamma_{\text{nr}} = \text{diag}\left(\gamma_{\text{nr},1}, \gamma_{\text{nr},2}, \cdots, \gamma_{\text{nr},N}\right)$ account for the gain ($\gamma_{\text{nr},n} < 0$) and loss ($\gamma_{\text{nr},n} > 0$) of the modes. Correspondingly, the matrix components of $Q$ in $(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N) = \tilde{B}^* \left(Q^*\right)^{-1}$ will become $Q_{nr'} = i\frac{\tilde{b}_n^+ \tilde{b}_{n'}}{\tilde{\omega}_n^* - i\gamma_{\text{nr,n}} - \tilde{\omega}_{n'} - i\gamma_{\text{nr},n'}}$[50]

It is also important to stress that the quasinormal-mode expansion Eq. (15) does not depend on the normalization of the far-fields $\tilde{\mathbf{b}}_n$. Also, in the above theoretical exploration, an explicit phase relation is fixed between the input and output fields to eliminate redundant degrees of freedom. In other words, we have $\hat{C} = -\hat{I}_M$ for the direct coupling operator $\hat{C}$ between the incoming and outgoing channels in the absence of the scatterer [38,50]. This choice ensures that the incoming wave will be completely reflected when disconnected from the scatterer. Note that the direct coupling operator accounts for the relative phase of the input and output waves and it is not related to QNMs or poles. Interested readers may resort to Ref.[59] for other choices and relevant discussions.

Finally, it is worthwhile to briefly mention that the Hamiltonian approach for optical systems can be extended to time-modulated systems[60], hold the promise for Floquet scattering engineering [61, 62], and may facilitate the study of the reconfigurable control of scattering anomalies like Refs. [63, 64]. However, the exploration along this line is beyond the context of this paper. 1.3 Poles and zeros of the S-matrix Generally, the S-matrix $\hat{S}$ can be reduced to its diagonal form $\hat{S}_D = \text{diag}\left(d_1, d_2, \cdots, d_M\right)$ after diagonalization. Each eigenvalue $d_m(\omega)$ is associated with its incoming eigenvector $\mathbf{s}_m^+$

$$\hat{S}(\omega)\mathbf{s}_m^+ = d_m(\omega)\mathbf{s}_m^+.$$

In the case of scatterers without material gain/loss, i.e., when the permittivity and permeability are real, the scattering eigenvalues $d_m(\omega)$ for real $\omega$ are unimodular, i.e., $|d_m(\omega)| = 1$. This corresponds to a unit circle in the complex domain. This constraint is imposed by the unitarity of the scattering matrix, i.e., $\hat{S}\hat{S}^\dagger = \hat{I}_M$, which is associated with a Hermitian Hamiltonian $\hat{H}_0 = \hat{H}_0^\dagger$ of the scatterer[65,66]. The physical meaning of these eigenvalues becomes apparent when we present them in the exponential form $d_m(\omega) = \exp\left[2\iota\delta_m(\omega)\right]$. Indeed, such representation shows that an incident field corresponding to the eigenvector $\mathbf{s}_m^+$upon scattering on the system experiences phase shift $\delta_m(\omega)$, corresponding to the incident eigenvector being multiplied by its eigenvalue $d_m$.

When the frequency $\omega$ is analytically continued in the complex plane, the scattering eigenvalues $d_m(\omega)$ can take any values between $\infty$ and 0. These limit values correspond to the poles

and zeros of the S-matrix eigenvalues, respectively. The zeros of the S-matrix are different from the "scattering zeros" (absence of scattered waves), as discussed in Section (1.4). The poles refer to the QNMs, as discussed before, while the zeros are associated with vanishing outgoing fields (perfect absorbing regimes), which will be discussed in Section (2). For now, we point out that, for lossless scatterers with $\hat{H}_0 = \hat{H}_0^\dagger$, poles $\omega_{\mathrm{r},n} = \tilde{\omega}_n$ and zeros $\omega_{\mathrm{pa},n}$ occur in complex conjugate pairs, i.e., $\omega_{\mathrm{pa},n} = \omega^*_{\mathrm{D},n^*}$ Indeed, the zeros $\omega_{\mathrm{p},n}$ satisfy the equation $\det\left(\omega_{\mathrm{pa},n} - \hat{H} - i\Gamma\right) = 0$. Moreover, after a Hermitian transpose operation, this equation for the zeros becomes $\det\left(\omega^*_{\mathrm{pa},n} - \hat{H}^\dagger + i\Gamma\right) = 0$, which, for a Hermitian Hamiltonian $\hat{H}$ is the secular equation for the poles, implying that the poles $\omega_{\mathrm{r,n}} = \omega^*_{\mathrm{pen}}$ for Hermitian systems, as expected from time-reversal symmetry considerations.

A fundamental property of the scattering matrix eigenvalues $d_n$, which is going to be important for further discussion, is that they can be expanded into a special product via the Weierstrass factorization theorem [26,67]:

$$d_n(\omega) = A_n \exp\left(iB_n\omega\right) \prod_m \frac{\omega - \omega_{\mathrm{p,m}}}{\omega - \omega_{\mathrm{r,m}}},$$

where $A_n$ and $B_n$ are constants, and the product is taken over all resonances which match the symmetry of $d_n$. These expressions include only information about the positions of poles $\omega_{\mathrm{rs,m}}$ and zeros $\omega_{\mathrm{pam}}$ in the complex plane. Therefore, the Weierstrass expansion shows that the scattering process is fully described only by the positions of poles $\omega_{\mathrm{r,m}}$ and zeros $\omega_{\mathrm{pa,m}}$. This fact makes it possible to describe and classify a variety of scattering phenomena in terms of the position of zeros and poles. It is worth noting that Eq. (17) works only for meromorphic functions, when any finite area of the complex frequency plane contains a finite number of poles [52, 68, 69]. The scattering process is fully described by the position of poles and zeros. This makes possible to describe, classify, and predict a variety of scattering phenomena. We note that this technique for scattering calculations based on the S-matrix zeros and poles has been widely used in the study of gratings, especially in the context of Wood's anomalies [23, 24, 70 − 75]. These features arise when new diffraction channels open (with an increase in frequency or an increase in the structure period), accompanied by the appearance of a diffracted beam parallel to the surface of the grating. In this paper, we do not discuss Wood's anomalies [76], and we refer the interested readers to several comprehensive works [77-79]. 1.4 Scattering zeros and poles Besides poles and zeros of the $\hat{S}$ matrix, we can use poles and zeros of other physical quantities. For example, the reflection coefficient $(R(\omega) = |r(\omega)|^2)$ also has singularities in the complex frequency plane. Because a pole originates as an eigenfrequency of a QNM, $R(\omega)$ shares the same poles with the $\hat{S}$ matrix eigenvalues. However, in general the scattering zeros (sometimes called reflectionless scattering modes) do not coincide with the zeros of the $\hat{S}$-matrix, since they correspond to the absence of scattered waves, whereas the outgoing fields in the channels are unperturbed incident waves. In the scenario of scattering by extended objects, a scattering zero on the a real frequency axis corresponds to the well-known tunneling at a Fabry-Perot resonance (vanishing reflection along with total transmission) [80], tunneling through epsilon-near-zero PTsymmetric layered structures [81,82], and epsilon-near-zero structures [83]. Interestingly, unlike the zeros of the S-matrix, the position of scattering zeros depends on the excitation arrangement. Generally, it depends on signals in all excitation channels, their amplitudes and phases. This provides a great opportunity for coherent all-optical scattering control [84]. For example, two-port coherent excitation can result in coalescencing of a pole and a zero of the scattering coefficient with BIC-like scattering lines in reflection [84].

Scattering zeros of 3D structures play an even more important role because they give rise to a plethora of unusual scattering phenomena, including cloaking, anapole [85], Fano resonance, and coupled-resonator induced transparency [86]. For example, let us consider a small spherical

resonator. For analysis of such particle, one can use the Mie scattering theory, which yields, for example, the equation for the scattering cross-section (SCS) [87,88], containing scattering coefficients in the form $c_l = -U_l/\left(U_l - jV_l\right)$ [89]. Here, the quantities $U_l^{TE,TM}$ and $V_l^{TE,TM}$ are defined by a combination of spherical Bessel and Neumann functions [90]. These amplitudes have scattering zeros when $U_1 = 0$ giving rise to cloaking [91] or anapoles [85]. Such nonscattering states are of great interest in modern photonics and have been employed for boosting nonlinear effects in dielectric nanoparticles, such as third harmonic generation and four-wave mixing [9295]. In its turn, the scattering poles are realized when the denominator turns to zero, $U_1^i - jV_1^i = 0$. The pole and zero of the 3D structure can also coalesce at the real axis by tailoring the structure geometry with the realization of the bound state in the continuum (BIC) scattering regime [96]. 1.5 Q-factor In scattering problems, QNMs of a system manifest themselves as resonant enhancement of scattering/extinction cross-section, reflection/transmission coefficients, etc. at real frequencies. Since the energy leaks out from the scatterer and gets dissipated, each specific quasinormal mode $l$ is characterized by a complex frequency $\omega_l = \mathrm{Re}\left[\omega_l\right] + i\,\mathrm{Im}\left[\omega_l\right]$, with the position at the real axis $\mathrm{Re}\left[\omega_l\right]$ and a finite lifetime $\tau_l = 1/\gamma_l = -1/\,\mathrm{Im}\left[\omega_l\right]$, where $\gamma_l$ is the decay rate and determines the resonance linewidth at its half-maximum[54]. A resonance is usually described by quality ($Q$) factor, which is defined as

$$Q = -\frac{\mathrm{Re}\left[\omega_l\right]}{2\,\mathrm{Im}\left[\omega_l\right]}$$

Factor 2 occurs in the denominator because the energy decays with the decay rate $2\gamma$. In general, the cavity decay rate includes both radiative and absorptive parts $\gamma = \gamma^{\mathrm{rad}} + \gamma^{\mathrm{loss}}$, such that the quality factor can be expressed as $Q^{-1} = Q_{\mathrm{rad}}^{-1} + Q_{\mathrm{loss}}^{-1}$ [3].

The Q-factor in the form of Eq. (18) requires an important note. By its definition, the Qfactor defines the decay rate of a resonant state, or the number of oscillations of the resonant state before it decays by a factor of 2.7. Thus, Eq. (18) is applicable to a single QNM, uncoupled from other resonant nodes of the system (orthogonal to them). Otherwise, turning-off of a monochromatic excitation causes its spectral broadening and inevitable excitation of other (not orthogonal) QNMs.

Interference of different QNMs plays a crucial role in various effects, e.g., Fanoresonances, anapoles, bound states in the continuum (BIC), and others. Hence, although many papers report "high-Q Fano resonances", the definition of Q-factor in these coupled resonance phenomena requires special handling. For example, an optical Fano resonance [as well as anapoles, see Section (5)] is caused by interference of radiation from at least two QNMs, bright and dark ones, in the far-field [97,98]. The coupling of these nonorthogonal modes can lead to speeding up the decay rate of the dark mode and, hence, the temporal dynamics are no longer directly related with the Q-factor of the individual modes. 1.6 S-matrix symmetry properties We summarize here the general requirements imposed on the S-matrix by different physical realizations and symmetries.

Reciprocity implies that if a signal can travel a certain path, the same signal should be able to travel back with the same transmission coefficient[99-104]. If the system is reciprocal, its scattering operator is symmetric: $\hat{S} = \hat{S}^T$. For example, for a 2-port structure it provides equal transmission in any direction, $t_{12} = t_{21}$. For a 3-port system, this requirement yields, $t_{12} = t_{21}$, $t_{13} = t_{31}$, and so on.

Reciprocity has a great deal of importance in this discussion, because it poses fundamental constraints on the electromagnetic response of a system [35] and on the inverse design of optical devices [105]. However, there are several practical situations (e.g., isolators, filters, full-duplex communication systems) in which it is advantageous to break reciprocity. Reciprocity can be broken $\left(\hat{S} \neq \hat{S}^T\right)$ via the use of magnetized materials [106], optical nonlinearity in spatially asymmetric structures [107], or employing time modulation [104]. The holy grail in this field

is a so-called ideal isolator, a device allowing signals to propagate outwards from the source and preventing reflections towards the source. This functionality can be generalized to a 3-port system, for which the S-matrix reads as

$$\hat{S} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Photonic topological insulators [108,109] are artificially designed materials supporting nonreciprocal one-way edge states propagating around sharp bends or other types of defects without backscattering. This fascinating property has been recently connected with the physics of nonreciprocal electromagnetic [110] and acoustic structures [111].

Hermitian scatterers with $\hat{H}_0 = \hat{H}_0^\dagger$. By a Hermitian scatterer, we denote a system without material gain/loss, and thereby we use the fact that radiation to the propagating channels does not spoil the flux conservation of the stationary scattering process. However, one has to bear in mind that the Hamiltonian of a scatterer in an open set-up is non-Hermitian with complex-valued QNMs when radiation losses are included[112]. For systems without material gain/loss, i.e., when the permittivity and permeability are real everywhere, $\hat{S}$ is unitary at the real frequency, $\hat{S}^+\hat{S} = \hat{I}$ and its eigenvalues obey $|d_m(\omega)|^2 = 1$ [see Eq. (16)] corresponding to each scattering channel. For such systems, the frequencies of poles and zeros are restricted to specific parts of the complex frequency plane. QNMs (resonant states) of a passive system must decay in time, due to leakage of the energy from the system into free space. With the time dependence defined through the factor $\exp(-i\omega t)$, the attenuation corresponds to the complex frequencies $\omega_{\mathrm{r},n}$ in the lower half-plane. Correspondingly, the frequencies of perfectly absorbing modes (S-matrix zeros) $\omega_{\mathrm{pa},n} = \omega_{\mathrm{r},n}^*$ sit symmetrically in the upper half-plane, Figs. 2(a,b). One exception is a bound state in the continuum (BIC), Fig. 2(b), which is a QNM of the scatterer with a real frequency in the continuum of unbounded radiation modes and can exist in lossless or $\mathcal{PJ}$-symmetrical[113-115] structures infinite in at least one direction [116] or for extreme values of parameters[96,117-119], see Section (4).

When the coupling $\hat{K}$ between the scatterer and the channels decreases, the pairs of poles and zeros approach the real axis symmetrically. Finally, in the limit of vanishing coupling, they coincide with the eigenfrequencies of the closed system Hamiltonian $\hat{H}_0$[120], Fig. 2(a). The discussion up to now also holds for systems with time-reversal symmetry $\mathcal{T}$, which corresponds to real symmetric Hamiltonians (a restricted class of Hermiticity).

**Non-Hermitian scatterers with $\hat{H}_0 \neq \hat{H}_0^+$.**

The scenario changes when we consider wave scattering due to scatterers described by non-Hermitian Hamiltonians $\hat{H}_0 \neq \hat{H}_0^+$. The strength of the non-Hermiticity can be controlled by the material gain/loss parameters. Material loss and gain break time-reversal symmetry $\mathcal{T}$, but preserve reciprocity. In the case of lossy ($|d_m(\omega)|^2 < 1$, for each channel) scatterers, the complex zeros move down from the upper half-plane as dissipation increases from zero. At some critical value of dissipation, the zero crosses the real axis at some frequency $\omega_{\mathrm{CPA}}$ leading to perfect absorption of an appropriate coherent incoming waveform, Fig. 2(c). This perfect waveform is determined by the eigenstate of $\hat{S}(\omega_{\mathrm{CPA}})$ corresponding to the zero eigenvalue. This coherent perfect absorber (CPA) is the time-reversed counterpart of a laser for gain media[28,121]. Lasing occurs when a pole touches the real axis at the threshold value of gain, corresponding to an eigenvector of the $\hat{S}$ matrix with an infinite eigenvalue, Fig. 2(c). The interplay between the material gain and loss inside the scatterer can result in novel wave transport phenomena. We consider perfect absorption, CPA, VPA, and lasing effects in Section (2).

**$\mathcal{PT}$-symmetric scatterers.**

The evolution of an isolated $\mathcal{PT}$ - symmetric system is governed by a $\mathcal{PT}$-symmetric Hamiltonian, which is symmetrical upon reversal of space $\mathcal{P}$ and time $\mathcal{T}$. The $PT$-symmetric optical system can support balanced absorption and amplification, hence respond as a lossless/gainless scatterer, as long as the relative permittivity of the medium satisfies $\varepsilon(\mathbf{r}) = \varepsilon^*(-\mathbf{r})$ [122]. $\mathcal{PT}$-symmetric Hamiltonians support a phase transition in the eigenvalue spectrum from entirely real to (partially or completely) complex ones, which occurs at the so-called exceptional points (EPs)[123-132]. EPs are specific to non-Hermitian systems and emerge when two (or more) eigenvalues and their corresponding eigenstates coalesce so that the Hamiltonian becomes defective. In the interesting subclass associated with wave scattering due to $\mathcal{PT}$-symmetric scatterers, like in the Hermitian case, poles and zeros also occur in complex conjugate pairs, as seen from the interchange between the secular equation for the poles and zeros after performing the combined $\mathcal{P}$-symmetry operation. Nevertheless, as the strength of nonHermiticity varies, pairs of pole and zero move symmetrically in the complex frequency plane, Fig. 2(d). Eventually, a pole and a zero of the $\hat{S}$ matrix may coalesce on the real axis, and a $CPA$ - laser[133-135] regime emerges. In this regime, the system behaves simultaneously as a laser oscillator and a CPA. We consider scattering in $\mathcal{PT}$-symmetric systems in Section (3).

**Quantum noise and the S-matrix approach**

The properties of the S-matrix in the case of $\mathcal{P}$-symmetric systems require special handling. It is known that active systems typically deal with quantum noise associated with enhanced spontaneous emission processes. Noise can be taken into account based on the so-called Langevin forces obeying the fluctuation-dissipation theorem [136]. In the presence of quantum noise, the input and output waves are treated as bosonic operators ($\hat{\mathbf{s}}^+$ and $\hat{\mathbf{s}}^-$), with the canonical bosonic commutation relations, $[\hat{\mathrm{s}}_{\mathrm{n}}^+(\omega), \hat{\mathrm{s}}_{\mathrm{m}}^+(\omega')] = \delta_{\mathrm{nm}}\delta(\omega - \omega')$ , $[\hat{\mathrm{s}}_{\mathrm{n}}^-(\omega), \hat{\mathrm{s}}_{\mathrm{m}}^-(\omega')] = \delta_{\mathrm{nm}}\delta(\omega - \omega')$, and the input-output equation Eq. (6) generalizes to [137]

$$\hat{\mathbf{s}}^- = \hat{S}^+\hat{\mathbf{s}}^+\hat{L}\hat{\mathbf{b}},$$

where, $\hat{L}$ is the Langevin operator, and the operator $\hat{\mathbf{b}}$ represents quantum fluctuations in the system. This approach has been extensively used to describe absorbing and amplifying systems [136, 138]. The fluctuation-dissipation theorem yields $\hat{L}\hat{L}^+ = \hat{S}\hat{S}^+ - 1$[138, 139].

Based on this general theory, in the next Sections, we analyze various scattering anomalies corresponding to some of the extreme scenarios captured by the scattering matrix poles and zeros. 2. Scattering by extended objects We will begin our description of scattering anomalies looking into structures extended in one or two dimensions, such as slabs, multilayered structures, and waveguides. Scattering of an incident plane wave or a waveguide mode can be usually reduced to only one or two channels (or four channels for oblique incidence), which makes the problem simpler. In this case, scattering can be characterized by reflection and transmission coefficients, which capture all possible scattering anomalies exhibited by a system.

**Perfect absorption (PA)**

We have discussed above how the scattering behavior of any linear structure is entirely determined by two basic kinds of special points in the complex frequency plane, zeros and poles of the scattering matrix or its eigenvalues. As we established above, the first type corresponds to perfectly absorbing solutions, while the latter corresponds to QNMs. In a reciprocal Hermitian system, the poles and zeros of $\hat{S}$ always exist in pairs, and their positions are related via complex conjugation. Specifically, poles are located in the lower plane [we use the time convention $\exp(-i\omega t)$ ], while zeros are located in the upper plane, which corresponds to exponential attenuation and divergence in time, respectively.

The presence of losses or gain breaks time-reversal symmetry $\mathcal{T}$ (but preserves reciprocity). An appropriate combination of geometrical design and a specific amount of material loss can push one of the zeros to the real axis, enabling a perfectly absorbing mode at a real frequency, i.e., the possibility of totally absorbing the impinging wave without any scattering for monochromatic excitation. Vice versa, adding gain to the system can push poles towards the real axis until the trajectory of a certain pole intersects it at the lasing threshold, giving rise to self-sustained oscillations of the electromagnetic field without incident field at the same frequency [140, 141]. Note that the presence of a pole in the upper plane turns the system in an unstable regime with the corresponding QNM divergent in time[142].

In most perfect absorbers, the incident energy is delivered via a single channel, e.g., by a plane wave incident on one side of the absorber. This is the case of one-port perfect absorbers, such as Salisbury and Dallenbach screens[143,144]. More recently, diffraction gratings and metasurfaces were shown to enable perfect absorption with a one-side incidence[145-151]. From the fundamental perspective, perfect absorption in all such devices occurs through destructive selfinterference, Fig. 3(a) [152-154]. To enable destructive interference, an absorbing material should be placed above a mirror or covered by an anti-reflective coating[155]. This requirement of selfinterference imposes the so-called Rozanov limit on the absorber thickness[156].

Without self-interference, perfect absorption in one-port structures usually cannot be achieved. Consider, for example, a p- (s-) polarized plane wave incident from vacuum at an angle $\theta$ on a half-space of isotropic material with permittivity $\varepsilon$. The amplitude of the reflected wave in this case reads

$$r_p = \frac{\cos\theta - \frac{1}{\varepsilon}\sqrt{\varepsilon - \sin^2\theta}}{\cos\theta + \frac{1}{\varepsilon}\sqrt{\varepsilon - \sin^2\theta}},$$

$$r_s = \frac{\cos\theta - \sqrt{\varepsilon - \sin^2\theta}}{\cos\theta + \sqrt{\varepsilon - \sin^2\theta}}.$$

In the lossless scenario, at the Brewster angle, the reflection of a p-polarized plane wave vanishes. The presence of loss forbids perfect transmittance due to the impedance mismatch between air and the lossy medium, whose normalized impedance $Z = \frac{1}{\varepsilon}\sqrt{\varepsilon - \sin^2\theta_i}$ has a non-vanishing imaginary part associated with the electric loss in the dielectric, and making perfect absorption impossible.

Thus, achieving one-sided perfect absorption without backing mirrors is challenging. This requires excitation of both electric and magnetic currents within an absorbing conductive sheet, creating destructive interference similar to a Huygens source[153]. There are non-trivial examples of one-port systems that allow perfect absorption to occur without self-interference. One such example is a half-space made of an anisotropic medium with the optical axis being normal to the surface. In that case, even in the presence of losses, the reflection of a p-polarized plane wave can become zero, giving rise to complete absorption of the incident wave inside the lossy material. Such interference-free total absorption was demonstrated in mid-IR for hexagonal boron nitride[157] and theoretically predicted in asymmetric hyperbolic media[158]. Photonic structures lacking $C_2^z$ symmetry, i.e., 180-degree rotation around the out-of-plane axis, were also shown to exhibit nearly perfect one-sided absorption[159].

**Coherent perfect absorption (CPA)**

In the above examples of structures exhibiting zeros of S-matrix eigenvalues (perfect absorption), incident energy was assumed to enter a system via one specific channel. That situation corresponds to the most practical case of an absorber illuminated, e.g., by the Sun or a lab light source, from one side. Coherent perfect absorber (CPA) concept generalizes this scenario

to an arbitrary incoming waveform $\mathbf{s}^+$, consisting of two or more waves, such as waves incident on opposite faces of an open slab or film [28, 121, 160, 161]. To achieve perfect absorption, one must choose appropriate values for system parameters and the operating wavelength-as well as the suitable input waveform $\mathbf{s}^+$, including the intensities and relative phases of the input beams. Conversely, by adjusting the relative phase of the inputs, one can switch to the regime of suppressed absorption, caused by the constructive interference of exiting waves[28,161,162]. Interestingly, this regime corresponds to the time-reversed operation of a laser (anti-laser) [28,163-166]. We caution, however, against a direct literal understanding of this relationship, because it does not take into account nonlinear processes and spontaneous emission, which are essential in the laser operation. Note also that the CPA can be considered as the generalization of the concept of critical coupling [33, 167].

The mathematical condition for coherent perfect absorption is that output field amplitudes vanish for a certain non-zero set of input amplitudes $\hat{S}^+_{\mathbf{sPA}^+} = 0$. For this to happen, at least one of the eigenvalues of $\hat{S}$ must be zero, just like in the case of a single-port absorber. The difference lies in the structure of the "CPA eigenmode" eigenvector $\mathbf{s}_{\mathrm{CPA}}$, which can have several non-zero components meaning that the system should be illuminated in more than one channel to achieve coherent perfect absorption, Fig. 3(c).

To illustrate the phenomenon of coherent perfect absorption, we consider the simplest case of a two-port planar structure illuminated from the two sides. Ignoring the polarization degree of freedom, the inputs and outputs are related via (for normal incidence)

$$\begin{pmatrix} s_1^- \\ s_2^- \end{pmatrix} = \hat{S} \begin{pmatrix} s_1^+ \\ s_2^+ \end{pmatrix}, \hat{S} = \begin{pmatrix} r_{11} & t_{12} \\ t_{21} & r_{22} \end{pmatrix}.$$

Here $s_i^+$and $s_i^-$respectively denote the input and output wave amplitudes in the i-th channel; $r_{ii}$ are the reflection coefficients in each port, and $t_{ij}$ are the transmission coefficients. If $\hat{S}$ obeys reciprocity[101], then $t_{12} = t_{21} = t$. If, also, the two-port structure is symmetric under mirror reflection that exchanges the two ports, then $r_{11} = r_{22} = r$ and the two eigenvalues of $\hat{S}$ are then $r \pm t$. The corresponding eigenvectors are $\mathbf{s}^+_{\mathrm{CPA}} \propto (1, \pm 1)$, representing symmetric and antisymmetric inputs of equal intensity. Now, if $r \pm t = 0$ for a planar structure at some frequency, then illumination of it with the corresponding CPA eigenmode will lead to complete absorption. The absorption in a CPA turns out to be highly sensitive to variations in the conditions of illumination, in particular in the relative phase between the components of the exciting field. To characterize a CPA, it is convenient to define the joint absorption coefficient [162]

$$\mathcal{A} \equiv 1 - \frac{\left|s_1^-\right|^2 + \left|s_2^-\right|^2}{\left|s_1^+\right|^2 + \left|s_2^+\right|^2},$$

Here, $\left|s_1^+\right|^2 + \left|s_2^+\right|^2$ is proportional to the total input intensity and $\left|s_1^-\right|^2 + \left|s_2^-\right|^2$ to the total output intensity; thus a CPA has $\mathcal{A} = 1$. Fig. 3(c, right panel) features a typical plot of $1 - \mathcal{A}$ for the symmetric (blue curve) and anti-symmetric (orange curve) modes of a uniform dielectric slab near a CPA frequency. CPA occurs whenever one of the curves falls to zero. An intriguing feature of this plot is that at the CPA frequency, while one of the incident modes yields total absorption, the other mode strongly scatters due to constructive interference outside the system, thus suppressing the absorption.

Zeros of different eigenvalues of $\hat{S}$ typically occur at distinct frequencies, meaning that if a certain incident waveform $\mathbf{s}_{\mathrm{CPA}}$ experiences coherent absorption, the orthogonal waveform does not. However, there are examples of two-port mirror-symmetric structures with doubly degenerate CPA mode[ 168,169], in which both eigenvalues of $\hat{S}$ can be zero at a certain frequency (double zero). Note that the coalescence of two (or more) poles leads generically to an exceptional point (EP) [122, 126, 170], as discussed in Section (3). The absorption becomes

insensitive to the relative phase and amplitudes of the two input waves so that even a single input wave incident on either port would be perfectly absorbed. Such a system presents an example of one-sided perfect absorption without backing mirrors.

CPAs have been studied and realized in a wide variety of geometries, including planar structures (slabs, diffraction gratings, metasurfaces, and thin films) [28, 160, 161, 171] and guidedmode structures[172-175] compatible with integrated nanophotonics applications. This approach has been utilized for controlling scattering and radiation from optical nanostructures and nanoantennas [175,176], and also for boosting wireless power transfer between two antennas [177]. Different techniques have been devised for increasing or decreasing the absorption bandwidth[178-180], and for achieving perfect absorption under strongly-asymmetric beam intensities[181]. Also, CPA has been extended to systems exhibiting strong light-matter interaction[182]. A recent study demonstrates CPA in a disordered medium implementing a random anti-lasing regime [164].

The operation of CPAs in the quantum optical case, when a few-photon state represents an incident field, has been also studied[183-185]. For example, in Ref.[185] absorption of twophoton states with 40% efficiency has been demonstrated. The framework of the electromagnetic $\hat{S}$ matrix does not accurately capture this situation, since a lossy structure (beam splitter) may induce transitions between few-photon and single-photon states, which are unmonitored by the electromagnetic $\hat{S}$ matrix.

The phenomenon of coherent perfect absorption may be useful in a variety of applications, ranging from all-optical data processing to enhanced photocurrent generation. CPAs can be highly sensitive to the parameters of the input waves, enabling attractive opportunities for the flexible control of light scattering and absorption [161, 186, 187]. In a two-port CPA, for instance, one input can be regarded as a signal beam, and the other as a control beam. A coherent control beam serves as a resource enabling efficient control of optical absorption, even in the linear regime. This opens the door to novel low-energy logical circuits and small signal amplifiers[186,187] without the use of non-linear materials that rely on high intensity of incident radiation to achieve a nonlinear response. Moreover, coherent perfect absorption can be employed for the processing of binary images and pattern recognition[188]. Coherent absorption effects may prove very useful for improving photocurrent generation in photoelectrochemical systems[189]. Finally, the ability to control and remotely trigger coherent perfect absorption of single- or few-photon states[183,184] may be promising for quantum information processing.

Recently, these studies on coherent all-optical tuning and manipulation have inspired research efforts studying the possibility of boosting the performance of antenna systems. It has been shown that the same principle that underlies the operation of CPAs can be employed to improve the efficiency of antenna energy transfer, and as a result enhance the efficiency of wireless power transfer (WPT) systems [190]. More specifically, by coherently exciting the receiving antenna with an auxiliary signal, tuned in syne with the impinging signal from the transmitting antenna, it is possible to largely enhance the robustness and overall efficiency of the system. This additional signal improves energy transfer through constructive interference with the impinging wave, compensating any imbalance in the antenna coupling, without having to modify the load. In another work, it has been explored that coherent excitation of an antenna by two dipole sources allows controlling the multipoles resulting in boosted superradiance and enhanced directivity [176]. Interestingly, it makes possible to excite nonradiative field configurations, and turning them on/off on demand. Remarkably, this approach also allows reducing the quenching effect and enhancing the radiation efficiency without changing the antenna's geometry via constructive or destructive interference of the excited QNMs. 2.3. Virtual perfect absorption (VPA) and virtual gain In the previous section, we have discussed the perfect absorption effect for harmonically oscillating incident waves, which can occur when an S-matrix zero is located at a real frequency $\omega$. However, in the case of lossless objects

obviously, the S-matrix zeros [unlike scattering zeros, see Section (1.4)] cannot arise at any real frequency (except for bound states in the continuum, where the zeros are canceled out by the complex conjugate pole) and are restricted to complex frequencies. These frequencies correspond to field solutions $\mathbf{E}_{\text{zero}}(r,t)$ with time dependence described by an exponentially growing amplitude $e^{\omega^* t}$, such that

$$\hat{S}\mathbf{E}_{\text{nero}} = 0$$

In light of this fact, the following question arises: can we somehow engage these complex modes to mimic perfect absorption in lossless systems?

It turns out that, instead of adding loss to the system to push a complex zero towards the real axis, we can tailor the incident field in time, such that its temporal profile matches the exponentially diverging mode associated with a complex zero over a finite interval of time, giving rise to virtual perfect absorption[191]. During this transient, the scattering from the structure totally vanishes, as if the structure were perfectly absorbing, despite the absence of material loss, Fig. 3(d). Instead, the energy is efficiently stored inside the open cavity without scattering. When the exponentially growing incident field is stopped, it gives rise to the release of the energy stored in the system. This behavior can be interpreted as transient destructive interference of the scattered waves with the waves that would leak from the cavity.

The effect of virtual absorption can be reproduced in any lossless electromagnetic structure, provided that the spectral position of its zero is known, and the corresponding spatial and temporal profile of the incident field can be created. It is also robust to inevitable material dissipation and material dispersion because the position of poles and zeros entails information about this dispersion. Finally, this effect is not limited to optics. Indeed, the scattering matrix formalism is equally applicable to acoustics[192] and elastodynamics and single-particle quantum mechanical problems. Since the Hamiltonians of quantum mechanical systems are naturally Hermitian, such systems have zeros of S-matrix in the upper half-plane of the complex energies. Therefore, properly shaping the temporal wavefunction of an incident particle[193] is expected to enable efficient storing of the wavefunction inside a potential without scattering. This observation suggests that this effect can be employed for efficient low-energy light storage and release on demand. The idea of coherent virtual absorption has been realized experimentally with elastodynamic waves [192].

Finally, if in the transient the incoming signal exponentially decays faster than the cavity decay rate, the output signal [along with the energy and momentum flux] can be larger than the input one. In this case, the stored energy becomes effectively "negative", as one would have in a system with real gain. This effect, which we refer to as "virtual gain", has been recently utilized for tailoring the optical pulling force in paraxial fields where otherwise the radiation force is always positive due to the momentum conservation law. Somewhat related, a complex excitation with negative imaginary frequency has been also suggested to improve flat imaging devices [194].

**Unconventional lasing scenarios**

Non-Hermitian photonic structures involving optical gain and loss, often result in anomalous scattering behaviors that are absent in passive systems[122,127,195-199]. Given the ubiquity of gain and loss processes in various electromagnetic structures and devices, such as lasers, absorbers, and isolators, systematic studying their optical response via the scattering poles and zeros is highly desirable.

Poles of the scattering matrix hitting the real-frequency axis correspond to lasing [141, 200 - 203]. From the classical electromagnetics perspective, lasing can be viewed as the time-reversed process of perfect absorption, i.e., emission of radiation from an otherwise non-illuminated structure[28,140]. This picture does not account for optical or electrical pumping of the gain

medium that provides population inversion and amplification of light at the emission frequency. Instead, fixed population inversion is simply assumed in this picture, which is reflected in the negative imaginary part of the gain medium permittivity [203-208]. For a rigorous analysis of a laser, this approach is not sufficient, and the rate equations [209-211] or Maxwell-Bloch equations [208,212] can be used.

An important difference between lasing and perfect absorption in terms of the poles and zeros analysis is that, when a zero moves below the real frequency axis as more absorption is added to the system, the perfectly absorbing response weakens, i.e., absorption is maximized for a specific value of loss that balances radiation loss at the critical coupling condition. In contrast, adding gain to a lasing system implies in most cases a further increase in the amplitude of its emission. In a more rigorous description of lasers based on Maxwell-Bloch equations[208,212], the population inversion of the gain medium itself depends on the electric fields in the system, which makes the problem nonlinear[213,214]. For this reason, linear description of a system with gain is only applicable below the lasing threshold. Nevertheless, the linear theory still allows one to evaluate whether the system can operate as a laser at all. If at least one scattering pole lies in the upper half of the complex plane (or at the real axis), lasing oscillations arise.

In a typical representation of the laser operation, adding gain into the system shifts the poles towards the real axis. Adding loss, on the contrary, moves poles away from the real axis making lasing harder to reach. Such a naive picture originating from the single-mode laser rate equations, however, turns out to be not always correct. In particular, a non-uniform increase of gain may lead to the death of laser generation[215,216]. The first discovery of this counterintuitive gain-induced laser turn-off was reported in[215], where the authors theoretically studied lasing from two coupled slabs containing gain medium pumped non-uniformly. First, one cavity of the system was pumped until it started lasing. Then, the second cavity was pumped while the pumping (and therefore gain) of the first one kept unchanged. It turned out that, upon pumping of the second cavity, the output laser emission began to drop until the laser eventually turned off, Fig. 4(a). Such unusual behavior is induced by the presence of exceptional points (EPs)[217], at which two scattering poles of the systems coalesce. We stress that this effect, although being related to the non-linear threshold dynamics of a laser, can be totally captured by linear equations. Laser turn-off was experimentally demonstrated in a configuration involving a pair of coupled microdisk lasers[218] and in a pair of active RLC-circuits[142].

Not only the increase of gain can lead to laser turn-off, but, vice versa, increase of loss can lead to the onset of lasing. Similarly to gain-induced laser turn-off, this effect may be caused by EPs in the spectra of systems with the non-uniform loss[219]. Such behavior was also predicted for homogeneous structures without EPs, but with dispersive absorbing medium[220]. An increase of the loss parameter in this case not only increases the rate of energy dissipation but also changes the effective refractive index of the cavity, potentially pulling the scattering pole into the region of the gain medium and enabling lasing.

In general, for a real-valued EP (two poles coalesce at the real axis), the variation of either gain or loss (or both of them) can cause EP breaking with the formation of two QNMs, lossy and "gainy" (i.e., located in the upper plane). An analysis involving nonlinearities (e.g., a saturable gain and a saturable loss) shows that imbalances between the saturation intensities cause a bifurcation in the vicinity of the exceptional point [221]. Such bistable lasers are interesting for possible applications as optical "flip-flop" memory devices [222-224].

For our example of a slab uniformly filled with a gain medium, the eigenvectors of the lasing modes have the form $\mathbf{a_{us}} = (1, \pm 1)$ corresponding to symmetric and anti-symmetric eigenmodes. Such modes produce the equal distribution of emission intensity leaving the system from the left and right sides. In fact, any reciprocal slab without ideal mirrors (such that $t \neq 0$) produces non-zero emission at both sides at the lasing threshold. However, adding

magneto-optical material to the design of a mirrorless multilayered structure has been shown to enable unidirectional lasing, Fig. 4(b)[225]. The coexistence of non-reciprocity and slow light modes leads to giant asymmetry of transmission, resulting in unidirectional mirrorless lasing not available in reciprocal active cavities. Another approach towards unidirectional singularities based on PTsymmetry was proposed in[226,227], as discussed in more detail in the next section.

To reduce radiation losses and hence reduce lasing threshold one can engineer the structure supporting resonant dark modes [228,229]. For example, a low lasing threshold has been demonstrated for collective resonances in 2D plasmonics arrays [230,231]. In an extreme scenario, we can realize bound states in the continuum (BICs), which are QNMs uncoupled from the continuum of radiation modes [118,232-234]. BICs have no radiation losses, and hence the Q- factor of a structure without no material loss can be unlimitedly large. Therefore, such structures are promising for low-threshold lasers [233, 235]. BICs are discussed in more detail in Section 4. Finally, we note that this criterium of a laser regime, i.e. the presence of a pole (of scattering amplitude or S-matrix) at the real frequency axis corresponds to the so-called classical condition[33,140]. For nanolasers, or lasers having dimensions less than the radiation wavelength, there is the quantum threshold condition [236,237], which requires the number of stimulated photons in the mode to overcome the number of spontaneous ones[236]. The spontaneous emission also causes the spectral broadening of the laser regime [139], so that the spectral width is not $\delta$ like but rather broadened.

## 30.2 PT-symmetry in optics (!!!?!?!??!?)

(статья пухова и др. мб на самом деле там много всего??? пока пусть как заготовка будет)

A A Zyablovsky, A P Vinogradov, A A Pukhov, A V Dorofeenko, A A Lisyansky DOI: 10.3367/UFNe.0184.201411b.1177 Contents 1. Introduction 1063 2. PT symmetry. Basic concepts and definitions 1064 2.1 Parity and time reversal operators; 2.2 Necessary and sufficient condition for real eigenvalues of a PTsymmetric Hamiltonian; 2.3 Phase transition in PT-symmetric systems 3. PT-symmetric optical systems 1067 3.1 PT-symmetry concept for optical systems; 3.2 Restrictions imposed by the causality principle. Kramers±Kronig relations and the possibility of PT-symmetric optical systems existing in a finite frequency range 4. Two-dimensional PT-symmetric optical systems 1068 4.1 Analogy between the two-dimensional Helmholtz equation and the one-dimensional SchroÈdinger equation; 4.2 Phase transition in two-dimensional PT-symmetric optical systems; 4.3 'Loss-induced transparency'; 4.4 Refraction asymmetry in the PT-symmetric optical system 5. One-dimensional PT-symmetric optical systems 1072 5.1 Phase transition in a one-dimensional optical system; 5.2 Lasing in PT-symmetric optical systems and the 'hidden' phase transition in a one-dimensional optical system 6. 'Quasi-diode' PT-symmetric optical systems. The reciprocity principle

## 30.3 Hamiltonian Approach to Optics (????!!!)

(выпишу из статьи турока что-нибудь мб. также в статье Краснока это используется, и я пока что совсем плохо понимаю это!!)

(пока пара структура от статьи краснока, мб в ней что-то и есть, но вроде чисто отсылки на нормальную литературу.)

### Hamiltonian Approach

It is instructive to introduce the effective Hamiltonian approach, which provides a bottom-up scheme to construct the S-matrix of a system and also highlights the essential features

and limitations of the coupled-mode theory [36–40] for wave scattering in various photonic structures [37,41–43]. To some extent, this effect

Hamiltonian approach can be considered an advanced form of the standard coupledmode theory, taking into account the dispersive properties of the channels [44]. Let us consider, to be specific, an optical scatterer described by the effective Hermitian Hamiltonian $\hat{H}_0$, which is coupled to $M$ propagating channels labeled by $p = 1, 2, \ldots, M$. In the modal space of the scatterer, the full Hamiltonian $\hat{H}_0$ is given as a $N \times N$ Hermitian matrix with $N$ being the number of supported modes.

The unperturbed portion of $\hat{H}_0$ is typically associated with an isolated ideal scatterer with canonical geometry, for which the modes are analytically solvable, Fig. 1(b). Correspondingly, the diagonal terms of the Hamiltonian $\hat{H}_0$ represent the resonant frequencies of the normal modes in this basis, while the off-diagonal ones denote the coupling between them due to imperfections, such as deformations and defects [45]. On the other hand, the propagating channel $p$ is associated with the $p$ th waveguide coupling one of the impinging channels to the scatterer.

For waveguides fulfilling the Helmholtz equation, Ref. [44] develops a general treatment of the corresponding Hamiltonian model. Here, for simplicity, we consider coupled-resonator optical waveguides (CROWs) satisfying the Schrodinger-like equation [46]. We set the distance between the nearest-neighbor resonators to be $a_0 = 1$ with dispersion $\omega = 2c_p \cos k_p$ for the $p$ th channel, after suitably fixing the reference frequency, and $c_p < 0$ is the coupling strength between its nearest sites/resonators. Generally, within the channel $p$, the field profile of a monochromatic wave can be written in terms of counterpropagating waves as

$$\psi_p(x,t) = \frac{s_p^+}{\sqrt{v_{g,p}}} e^{-i\omega t + ik_p x} + \frac{s_p^-}{\sqrt{v_{g,p}}} e^{-i\omega t - ik_p x},$$

where the group velocity $v_{g,p} \equiv \partial\omega/\partial k_p = -2c_p \sin k_p, k_p \in (0, \pi)$ and the integer $x = -1, -2, \ldots, -\infty$ labels the sites in each waveguide. In Eq. (8), $\left|s_p^{\pm}\right|^2$ are normalized to be the incoming/outgoing flux for each channel $p$.

If $M$ channels at $x = -1$ are coupled through the scatterer with $N$ modes, a real $M \times N$ bare coupling matrix $-\hat{W}$ describes the process. Correspondingly, Schrodinger-like equations describing this setup can be written as

$$i\frac{d}{dt}\mathbf{a}(t) = \hat{H}_0\mathbf{a}(t) - \hat{W}^T\mathbf{\Psi}(-1,t)$$
$$i\frac{d}{dt}\mathbf{\Psi}(-1,t) = \hat{c}\mathbf{\Psi}(-2,t) - \hat{W}\mathbf{a}(t)$$

where $\mathbf{\Psi}(x,t) = (\psi_1, \psi_2, \ldots, \psi_M)^T$, $\hat{c} = \text{diag}\{c_1, c_2, \ldots c_M\}$, and the vector $\mathbf{a}(t) = (a_1, a_2, \ldots, a_N)^T$ represents the set of mode amplitudes, which are normalized such that $|a_n|^2$ corresponds to the energy in the $n$th mode. Note that Eq. (9a) can be considered a definition of the Hamiltonian $\hat{H}_0$.

We proceed to derive the scattering matrix $\hat{S}$ based on Eq. (9). First, we derive the coupled-mode equation for the wave scattering from Eq. (9). Essentially setting $\mathbf{a}(t) = \mathbf{a}(0)e^{-i\omega t}$ and plugging Eq. (8) into Eq. (9), we get the coupled-mode equations,

$$-i\omega\mathbf{a}(0) = -i(\hat{H} - i\hat{\Gamma})\mathbf{a}(0) - \hat{K}^T\mathbf{s}^+$$
$$\mathbf{s}^- = -\mathbf{s}^+ - \hat{K}\mathbf{a}(0)$$

with the effective Hamiltonian

$$\hat{H} = \hat{H}_0 - \frac{1}{2}\hat{K}^T \cot\hat{k}\hat{K},$$

where $\Gamma = \frac{1}{2}\hat{K}^T\hat{K}$[47], and $\hat{K} \equiv \hat{c}^{-1}\sqrt{\hat{v}_g}\hat{W}$ is the normalized $M \times N$ coupling matrix, with $\hat{v}_g \equiv \text{diag}\{v_{g,1}, v_{g,2}, \ldots, v_{g,M}\}$ being the group velocity matrix of the channels, and the matrix $\hat{k} \equiv \text{diag}\{k_1, k_2, \ldots k_M\}$ consists of wavenumbers. The scattering matrix $\hat{S}$ can be obtained from Eq. (10) by eliminating $\mathbf{a}(0)$ and using the definition of $\hat{S}$. Specifically, we have

$$\hat{S} = -\hat{I} + i\hat{K}\frac{1}{\omega - \hat{H} + i\hat{\Gamma}}\hat{K}^T$$

This expression explicitly contains the scattering matrix $\hat{S}$ of an object expressed in terms of its Hermitian Hamiltonian $\hat{H}_0$, coupling operator $\hat{K}^T$, and the incident frequency $\omega$ of the monochromatic wave. Interestingly, the structure of the scattering matrix Eq. (12) is general, irrespective of the modelfor the channels [48, 49], and can be extended even more generally starting from the vector Helmholtz equation Eq. (2) (see Ref. [50] and references therein). We also note that although Eq. (12), as well as the coupled-mode theory, is derived in the weak coupling regime, it may be generalized to strongly coupled systems by changing the basis to "dressed states" [42]. Useful distribution information about scattering singularities can be extracted from Eq. (12) by assuming proper ensembles of the random matrix $\hat{H}_0$ based on the preserved symmetry class, such as the Gaussian orthogonal/unitary ensemble (GOE/GUE) for isolated scatterers with/without time-reversal symmetry [51, 52]. Nevertheless, the information required for the reconstruction of the scattering matrix $\hat{S}$ based on the Hamiltonian $\hat{H}_0$ and the coupling operator $\hat{K}$ may be overwhelming. Therefore, it is meaningful to develop an approach for the reconstruction of $\hat{S}$ using the minimal available knowledge, say the complex eigenfrequencies and the far-field

properties of the quasi-normal modes (QNMs) (or resonant states) of the system (see Ref. [53-55]).

The QNMs of photonic structures are defined as the solution of Eq. (9) with outgoing radiation boundary conditions [54, 56, 57], whose completeness has been proven for simple geometries (e.g., spherically symmetric objects) and is a conjecture in other scenarios [55,58-60]. The QNMs are associated with the poles of the scattering matrix in Eq. (12). To see this clearly, setting $\mathbf{s}^+ = 0$ in Eq. (10a), corresponding to outgoing boundary conditions, we get a set of complex eigenfrequencies of the QNMs $\tilde{\omega}_n$ and a corresponding eigenset of QNMs profiles $\tilde{\mathbf{a}}_n(0), n = 1, 2, \ldots, N$. The complex eigenfrequencies $\{\tilde{\omega}_n\}$ turn out to be the poles of the scattering matrix $\hat{S}$ in Eq. (12), i.e., $\det\left(\tilde{\omega}_n - \hat{H} + i\Gamma\right) = 0$. Moreover, the asymptotic behavior of the QNMs in the outgoing channels is $\tilde{\mathbf{b}}_n = -\hat{K}\tilde{\mathbf{a}}_n(0)$, which defines the structure of radiation leaking from resonant states into free space.

Next, we can reconstruct the scattering matrix $\hat{S}$ based on its QNM expansion, following Ref. [53]. It is important to point out the implicit assumption in optical systems that tacitly ignores the frequency dependence of the Hermitian Hamiltonian $\hat{H}$ and coupling operators $\hat{K}$. This assumption is the analog of the so-called wideband approximation in condensed-matter physics [61]. For the reconstruction, the required information can be written compactly as the $N$-dimensional diagonal matrix $\tilde{\Omega} = \text{diag}\,(\tilde{\omega}_1, \tilde{\omega}_2, \ldots, \tilde{\omega}_N)$ for the complex eigenfrequency spectrum and the $M \times N$ matrix $\tilde{B} = \left(\tilde{\mathbf{b}}_1, \tilde{\mathbf{b}}_2, \ldots, \tilde{\mathbf{b}}_N\right)$ for the far-field properties of the QNMs. First, we notice that

$$(\hat{H} - i\hat{\Gamma})\tilde{A} = \tilde{A}\tilde{\Omega}$$

$\tilde{B} = -\hat{K}\tilde{A}$ (13b) where the $N \times N$ matrix $\tilde{A} = (\tilde{\mathbf{a}}_1(0), \tilde{\mathbf{a}}_2(0), \ldots, \tilde{\mathbf{a}}_N(0))$ with the columns being the profile of the QNMs inside the scatterer. The above relations allow us to rewrite the S-matrix in Eq. (12) as

$$\hat{S} = -\hat{I} + i\tilde{B}\frac{1}{\omega - \tilde{\Omega}}\Lambda^{-1}\tilde{B}^T$$

where $\Lambda \equiv \tilde{A}^T \tilde{A} = \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_N)$ is a diagonal matrix, which follows from the reciprocity of the structure, i.e., mathematically $(\hat{H} - i\hat{\Gamma})^T = \hat{H} - i\hat{\Gamma}$. Next, we determine the matrix $\Lambda$ (or specifically its diagonal entries $\lambda_N$) from the given information of the QNM $\tilde{\Omega}$ and $\tilde{B}$. Since the coupling operator $\hat{K}$ is a real matrix, we have the identity $\tilde{B}^* \left(Q^*\right)^{-1} \Lambda = \tilde{B}$, where $Q \equiv \tilde{A}^+ \tilde{A}$ with components $Q_{nn'} = i\frac{\tilde{b}_n^+ \tilde{b}_{n'}}{\tilde{\omega}_n^* - \tilde{\omega}_{n'}}$, which is determined completely from the available information [53]. For convenience, we can define $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \tilde{B}^* \left(Q^*\right)^{-1}$ with the vector $\mathbf{x}_n$ as the columns of the latter matrix, which allows us to rewrite the last equation equivalently as $\mathbf{x}_n \lambda_n = \tilde{\mathbf{b}}_n$, where $n = 1, 2, \ldots, N$. Under the present formalism, this last equation is overdetermined but consistent to be used to determine $\lambda_N$. Considering potential numerical or measurement errors, a unique solution of $\mathbf{x}_n \lambda_n = \tilde{\mathbf{b}}_n$ in the least square sense can be given, which reads $\lambda_n = \mathbf{x}_n^+ \tilde{\mathbf{b}}_n / \mathbf{x}_n^+ \mathbf{x}_n$. Finally, the desired QNM expansion of the scattering matrix follows [53]:

$$\hat{S} = -\hat{I}_M + i \sum_{n=1}^{N} \frac{x_n^+ x_n}{x_n^+ \tilde{\mathbf{b}}_n} \frac{\tilde{\mathbf{b}}_n \tilde{\mathbf{b}}_n^T}{\omega - \tilde{\omega}_n},$$

where the vector $\mathbf{x}_n$ can be written explicitly as $\mathbf{x}_n = \sum_{n'} \left(Q_{n'n}^{-1}\right)^* \tilde{\mathbf{b}}_{n'}^*$. The material gain/loss can be easily incorporated into the previous formalism by setting $\hat{H}_0 \to \hat{H}_0 - i\Gamma_{\mathrm{nr}}$ in Eq. (12), where $\Gamma_{\mathrm{nr}}$ is a Hermitian decay matrix. When the nonradiative gain/loss is small compared to the frequency of the mode, the diagonal nonradiative gain/loss terms $\Gamma_{\mathrm{nr}} = \mathrm{diag}\,(\gamma_{\mathrm{nr},1}, \gamma_{\mathrm{nr},2}, \ldots, \gamma_{\mathrm{nr},N})$ account for the gain ($\gamma_{\mathrm{nr},n} < 0$) and loss ($\gamma_{\mathrm{nr},n} > 0$) of the modes. Correspondingly, the matrix components of $Q$ in $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \tilde{B}^* \left(Q^*\right)^{-1}$ will become $Q_{nn'} = i\frac{\tilde{b}_n^+ \tilde{b}_{n'}}{\tilde{\omega}_n^* - i\gamma_{\mathrm{nr},n} - \tilde{\omega}_{n'} - i\gamma_{\mathrm{nr},n'}}$ [53]. It is also important to stress that the QNM expansion Eq. (15) does not depend on the normalization of the far-fields $\tilde{\mathbf{b}}_n$. Also, in the above theoretical exploration, an explicit phase relation is fixed between the input and output fields to eliminate redundant degrees of freedom. In other words, we have $\hat{C} = -\hat{I}_M$ for the direct coupling operator $\hat{C}$ between the incoming and outgoing channels in the absence of the scatterer [41, 53]. This choice ensures that the incoming wave will be completely reflected when disconnected from the scatterer. Note that the direct coupling operator accounts for the relative phase of the input and output waves, and it is not related to QNMs or poles. Interested readers may resort to Ref. [62] for other choices and relevant discussions. Finally, it is worthwhile to briefly mention that the Hamiltonian approach for optical systems can be extended to time-modulated systems [63], holds the promise for Floquet scattering engineering [64, 65], and may facilitate the study of the reconfigurable control of scattering anomalies like Refs. [66,67]. However, the exploration along this line is beyond the context of this paper.

### Poles and Zeros of the S-Matrix (!?!?!?!?!?!?)

(вот это так я и не преодолел, пропишу потом!!! как там их искать!!! отдельно про это у других разделах укажу!!!)

Generally, the S-matrix $\hat{S}$ can be reduced to its diagonal form $\hat{S}_D = \mathrm{diag}\,(d_1, d_2, \ldots, d_M)$ after diagonalization. Each eigenvalue $d_m(\omega)$ is associated with its incoming eigenvector $\mathbf{s}_m^+$

**Scattering Matrix**

**Hamiltonian Approach**

**Poles and Zeros of the S-Matrix**

**Scattering Zeros and Poles**

**Q-Factor**

**S-Matrix Symmetry Properties**

**Quantum Noise and the S-Matrix Approach**

**уравнение гамильтона для оптики**

Механика материальных частиц описывается уравнением Гамильтона-Якоби

$$\mathbf{p} = \frac{\partial S}{\partial \mathbf{r}}, \quad \mathcal{H} = -\frac{\partial S}{\partial t} \tag{30.1}$$

В то же время геометрическая оптика описывается уравнениями:

$$\mathbf{k} = \frac{\partial \psi}{\partial \mathbf{r}}, \quad \omega = -\frac{\partial \psi}{\partial t} \tag{30.2}$$

Видим, что уравнения имеют одну и ту же структуру, разве что параметры разные: волновой вектор $\mathbf{k}$ в геометрической оптике играет роль импульса частицы $\mathbf{p}$ в механике, а частота $\omega$ - роль функции Гамильтона $\mathcal{H}$, то есть энергии частицы. Заметим, что связь частоты с волновым вектором $k = \omega/c$ такая же, как и энергии с импульсом: $p = E/c$.

Таким образом, зная уравнения Гамильтона из механики:

$$\dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{r}}, \ \dot{\mathbf{r}} = \frac{\partial H}{\partial \mathbf{p}},$$

и зная аналогию между частицей и волной, можно написать уравнение для лучей

$$\dot{\mathbf{k}} = -\frac{\partial \omega}{\partial \mathbf{r}}, \ \dot{\mathbf{r}} = \frac{\partial \omega}{\partial \mathbf{k}} \tag{30.3}$$

В пустоте $\omega = ck$, так что $\vec{v} = c\vec{n}$, то есть мы имеем прямые линии.

## 30.4 Ray transfer matrix analysis (?)

(такая есть, и все на том, вряд ли воспользуюсь ей когда-то)
(see wiki)

## 30.5 Integrable Models in Optics

?????

связь с интегрируемыми системами - хз какая.

ну так вот, когда речь о распространении волны в среде - есть диффуры, и определенная часть отвечает за дисперсию и диссипацию.

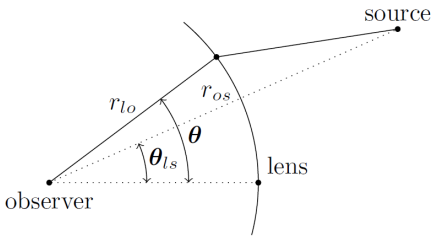
## 30.6 Optics in Curved Spacetime and Gravitational Lensing: Basics

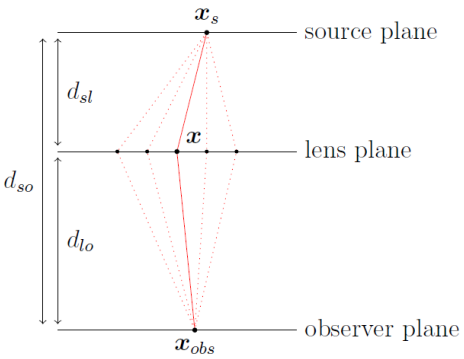
(коснулся, ушел в доработку структур, да и забил, мб когда-то вернусь, но пока не факт. по идее тут для линзирования заготовки будут, а выше на них будет отсылка где-то)

## 30.6.1 Basic Setup of Gravitational Lensing

(тут может быть очень насыщенный раздел с самыми разными подходами, геометрическими, волновыми, еще много писать это нужно!)

(!!!??? по идее в 2й части будут все указания, чтобы заниматься этим разделом!! проверю, чтобы на самом деле там нормальная была теория дифракции!!)




### Оптические законы в искривленных пространствах

соберем тут теорию, которую далее будем использовать нпр в гравитационном линзировании.

### Мотивация изучать оптику в искривленных пространствах

пропишем подробно, зачем это делать, потому что многим казаться может, что это фигня.

вроде нет, объясню потом, почему

### гравитационное линзирование

тут связка с научкой.

мб увеличу раздел

### оптика на ранних стадиях вселенной

интересно, что там было, как свет распространялся?

а пока что вообще хз

### реликтовое излучение (?)

(полная подготовка к нему!)

**Простой вывод уравнения линзы**

Приведем очень короткий вывод уравнения линзы и укажем самые ключевые его особенности.

(!?!?!? пока так на пальцах всё еще не могу это вывести, проблема!!!)

**Получение формулы линзы в двух словах**

Есть достаточно простой способ получить основные формулы линзирования, который обсудим в этом разделе. Другому,более фундаментальному выводу, использующему свойства континуального интеграла и переходу к геометрической оптике в волновом уравнении посвящен следующий раздел.

Рассмотрим ситуацию, изображенную на рис. **??**. В нашей модели есть наблюдатель (observer), источник света (source), находящийся в другой части вселенной, а также линза (lens), представляющая собой тонкий сферический слой, который как раз и обеспечивает увеличение или уменьшение интенсивности.

Рассмотрим движение когерентного монохроматического света с частотой $\omega$ с точки зрения квантовой механики, с помощью интеграла по траекториям при постоянной энергии $E = \hbar\omega$.

Амплитуда, имеет знакомый вид из первой главы интеграла по траекториям

$$\Psi[\boldsymbol{\beta}] = \int e^{iS/\hbar},$$

где действие следует записать согласно принципу Мопертьюи,

$$\mathcal{S} = \int \boldsymbol{p} \cdot \boldsymbol{dr}.$$

Учтем в действии вид простейшей метрики

$$ds^2 = -(1 + 2U(\boldsymbol{r}))dt^2 + (1 - 2U(\boldsymbol{r}))d\boldsymbol{r}^2,$$

в котором $U(r)$ - гравитационный потенциал. Сделать это можно, выразив импульс при учете движения фотона, которое в гравитационном поле имеет вид:

$$\boldsymbol{p}^2/(1 - 2U) + m^2 = E^2/(1 + 2U)$$

при $m \to 0$. Поэтому в первом порядке по $U$ действие становится:

$$\mathcal{S} \approx E \int |\boldsymbol{dr}|(1 - 2U(\boldsymbol{r})).$$

Разобьем действие на два слагаемых, первый $\mathcal{S}_1 \approx E \int |\boldsymbol{dr}|$ - это по сути увеличение евклидовой длины пути света, прошедшего линзу в сравнении с прямой длины от источником к наблюдателю. Если рассмотреть точку $A$ на поверхности линзы, то в приближении малых улов $|\boldsymbol{\theta} - \boldsymbol{\theta_i}| \int d\boldsymbol{r}$, где $\boldsymbol{\theta_i}$ - координаты $i$-го линзирующего объекта, можно записать с помощью несложных геометрических выкладок:

$$r_{sAo} - r_{so} = r_{sA} - (r_{so} - r_{lo}) = \sqrt{(r_{so} - r_{lo})^2 + r_{lo}r_{os}\,|\boldsymbol{\theta} - \boldsymbol{\theta_i}|^2} - (r_{so} - r_{lo}) = \frac{1}{2}\frac{r_{lo}r_{os}}{(r_{so} - r_{lo})}\,|\boldsymbol{\theta} - \boldsymbol{\theta_i}|^2\,.$$

В итоге первый вклад запишется в виде:

$$\mathcal{S}_1 \approx E \int |d\boldsymbol{r}| \approx \frac{r_{lo}r_{os}}{(r_{os} - r_{lo})}\,|\boldsymbol{\theta} - \boldsymbol{\theta_i}|^2\,.$$

Второй вклад в действие можно записать, перейдя к конкретной метрике: возьмем гравитирующий потенциал $U = -GM_i/\left|r - r_i\right|$ для массы $M_i$ в координате $r_i$ для малых $|\boldsymbol{\theta} - \boldsymbol{\theta_i}|$, только логарифм, поэтому действие станет в виде:

$$\mathcal{S}_2 \approx -E \int |d\boldsymbol{r}| 2U(\boldsymbol{r}) \approx 2GM_i \ln|\boldsymbol{\theta} - \boldsymbol{\theta_i}|\,,$$

с точностью до константы.

Удобно было бы сделать переобозначения, учитывающие геометрию, чтобы остаться с параметрами, которые играют ключевую роль, поэтому выразим это всё в известных углах Эйнштейна $\theta_E = \sqrt{4GM\left(r_{os} - r_{lo}\right)/\left(r_{lo} r_{os}\right)}$, где $M$ - общая масса. То есть перейдем к безразмерным угловым координатам $x, \beta$ следующим образом: $\boldsymbol{\theta} \equiv \theta_E \boldsymbol{x}$ и $\boldsymbol{\theta}_{ls} \equiv \theta_E \boldsymbol{\beta}$. Задача нормировки амплитуды для обеспечения унитарности тривиальна и решается введением фактора $\frac{\Omega}{2\pi i}$. По итогу амплитуда запишется в виде

$$\Psi[\boldsymbol{\beta}] = \frac{\Omega}{2\pi i} \int\limits_{\boldsymbol{x}} e^{i\Omega\left(\frac{1}{2}|\boldsymbol{x}-\boldsymbol{\beta}|^2 + \alpha \ln|\boldsymbol{x}-\boldsymbol{x_i}|\right)},$$

где $\Omega \equiv 4GM\omega$, здесь и далее будет использовано обозначение интеграла $\int\limits_{\boldsymbol{x}} \equiv \int d^2\boldsymbol{x}$ для укорочения записи. Фазовое слагаемое $\varphi(\boldsymbol{x})$ зависит от параметра $\alpha$, учитывающего геометрию и массу системы, в произвольном случае его можно находить определенные поправки к потенциалу одинокой массы. Для бинарных систем можно записывать фазовое слагаемое просто как взвешенную сумму нескольких потенциалов, например, по аналогии с случаем, рассмотренным выше, оно может иметь вид $\varphi(\boldsymbol{x}) = -\sum\limits_i f_i \log|\boldsymbol{x} - \boldsymbol{x}_i|$.

**Некоторые свойства амплитуды**

Для лучшего понимания полученных выражений а также дальнейших вычислений обсудим некоторые их свойства. Если гравитирующие массы перемещаются по небу с помощью параметра $t$, то $x_i \to x_i + \chi t$, соответственно выражение для амплитуды также будет меняться : $\Psi[\beta] \to \Psi[\beta - \chi t]$.

Обозначим как $A$ площадь сферы с центром в источнике света, на которой линза. Полная интенсивность, дошедшая до линзы $\int\limits_{\boldsymbol{\beta}} |\Psi[\boldsymbol{\beta}]|^2 = A$, не зависит от $\varphi(x)$. Она унитарная, что соответствует сохранению полного потока энергии.

Можно сделать пару аналитических преобразований для поиска интенсивности. Умножим $\Psi[\boldsymbol{\beta}]$ на его комплексно сопряженное выражение, с $x \to x'$, и сделаем замены $\boldsymbol{u} = \frac{1}{2}(\boldsymbol{x}' + \boldsymbol{x})$, $\boldsymbol{\Delta} = \frac{1}{2}(\boldsymbol{x}' - \boldsymbol{x})$, получим

$$I = \frac{\Omega^2}{4\pi^2} \int\limits_{\boldsymbol{u},\boldsymbol{\Delta}} e^{-i\Omega(\varphi(\boldsymbol{u}+\boldsymbol{\Delta}) - \varphi(\boldsymbol{u}-\boldsymbol{\Delta}) + 2\Delta\cdot(\boldsymbol{u}-\boldsymbol{\beta}))}.$$

Разложение в ряд тейлора экспоненты по $\boldsymbol{\Delta}$ и изменение масштаба $\boldsymbol{\Delta} \to \boldsymbol{\Delta}/\Omega$ дает линейный, независимый от $\Omega$ член, который преобладает при больших $\Omega$. Для удобства, обозначим как ранее, Переобозначим $\boldsymbol{u}$ как $\boldsymbol{x}$, теперь получим:

$$I(\boldsymbol{\beta}) \approx \int\limits_x \delta(\nabla\varphi(\boldsymbol{x}) + \boldsymbol{x} - \boldsymbol{\beta}) = \sum_s \frac{1}{\left|M\left(x_s(\boldsymbol{\beta})\right)\right|}.$$

где $M(\boldsymbol{x}) = \det\left(\delta_{ij} + \partial_i\partial_j\varphi(\boldsymbol{x})\right), i, j = 1, 2$, а сумма пробегает решения $\boldsymbol{x}_s(\boldsymbol{\beta})$ уравнения

$$\nabla\varphi(\boldsymbol{x}) + \boldsymbol{x} - \boldsymbol{\beta} = 0,$$

то есть уравнения на стационарные точки амплитуды $I$. При приближении больших $\Omega$, интенсивность сохраняет нормированность на единицу, как можно видеть, проинтегрировав левую часть равенства по всем $\boldsymbol{\beta}$.

В первом порядке по $\varphi$ уравнение имеет решения $\boldsymbol{x} \approx \boldsymbol{\beta}$. Далее $D(\boldsymbol{\beta}) \approx 1 + \nabla^2\varphi(\boldsymbol{\beta})$ и уравнение на стационарные точки становится подобием уравнения Пуассона $-\nabla^2\varphi(\boldsymbol{\beta}) \approx I(\boldsymbol{\beta}) - 1$. Получается, что в высокочастотном случае в слабом линзирующим режиме можно получить фазу линзирования из интенсивности.

Для гравитационных линз существует дальнейшее упрощение, а именно, если $I(\boldsymbol{\beta}) - 1$ можно аппроксимировать суммой дельта-функций, которые выбираются чтобы совпасть с их мультипольными моментами, решение для $\varphi$, которое представляет из себя сумму логарифмов, будет хорошо аппроксимировано точным нелинейным решением. Далее при больших $\Omega$ таким же методом итерационного алгоритма можно найти $\varphi$ по амплитуде.

Более того, есть методы, по которым также может быть найден $\Omega$, о них можно подробнее почитать в [? ]. Факт, который позволяет это сделать, состоит в том, что амплитуду можно рассматривать как унитарное преобразование фазы $e^{i\Omega\varphi(\boldsymbol{x})}$ вида

$$(\Omega/2\pi i)\int_{\boldsymbol{x}} e^{i\frac{1}{2}\Omega(\boldsymbol{x}-\boldsymbol{\beta})^2}.$$

В таком случае обратное преобразование $i\int_{\boldsymbol{\beta}} e^{-i\frac{1}{2}\Omega(\boldsymbol{z}-\boldsymbol{\beta})^2}$. Применяя его к $\Psi[\boldsymbol{\beta}]$, получаем $e^{i\Omega\varphi(\boldsymbol{z})}$, с полностью равномерной интенсивностью для любых $\varphi(\boldsymbol{z})$. Здесь видно, что амплитуда линзы способна восстановить $\Omega$.

### Фундаментальный вывод уравнения линзы

Более фундаментальный вывод уравнения линзы состоит в рассмотрении волнового уравнения и решения его методом континуального интеграла, хорошо изученных в прошлой части. Произведем его и далее будем рассматривать линзирование в рамках этого формализма. В данном разделе будем пользоваться системой единиц $c = 1$.

#### Описание модели

Рассмотрим точечный источник, излучающий монохроматические волны частоты $\omega$, часть которых распространяются через гравитационную линзу и достигают наблюдателя на далеком расстоянии. Данная модель изображена на рис. (30.1).

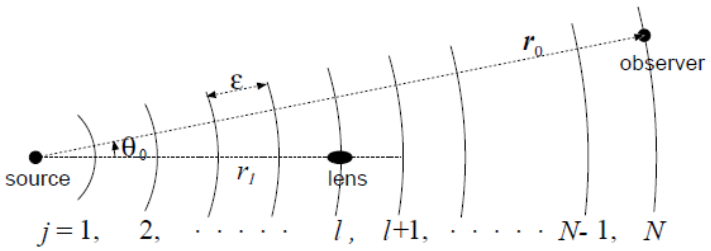



Figure 30.1: Модель геометрии гравитационного линзирования

Для данной модели пространство разбирается на участки длиной $\varepsilon$, в остальном геометрия нас из прошлой части уже хорошо знакома.

Метрика пространства-времени имеет хорошо известный вид:

$$ds^2 = -(1 + 2U)dt^2 + (1 - 2U)d\vec{r}^{\,2}.$$

**Вывод формулы линзы через континуальный интеграл**

Запишем волновое уравнение в терминах фактора усиления амплитуды (**??**):

$$\frac{\partial^2 F}{\partial r^2} + 2i\omega\frac{\partial F}{\partial r} + \frac{1}{r^2}\nabla_\theta^2 F, = 4\omega^2 U F$$

где $\nabla_\theta^2 = \partial^2/\partial\theta^2 + \theta^{-1}\partial/\partial\theta + \theta^{-2}\partial^2/\partial\varphi^2$.

Предполагая, что

$$\omega/|\partial \ln F/\partial r| \sim)\text{масштаб изменнеия}F))\text{длина волны}) \gg 1,$$

пренебрежем членом $\partial^2 F/\partial r^2$ в сравнении со вторым. Третий член отбросим ввиду сферической симметричности задачи.

Тогда мы получаем уравнение

$$2i\omega\frac{\partial F}{\partial r} = 4\omega^2 U F,$$

которое имеет вид уравнения Шредингера, разве что " время" это $r$, "масса частицы" - $\omega$, и "зависящий от времени потенциал" это $2\omega U(r,\theta)$.

Уравнение Шредингера известно тем, что решение формально можно записать, используя лагранжиан, соответствующий данной модели:

$$L(r,\boldsymbol{\theta},\dot{\boldsymbol{\theta}}) = \omega\left(\frac{1}{2}r^2|\dot{\boldsymbol{\theta}}|^2 - 2U(r,\boldsymbol{\theta})\right),$$

где $\dot{\boldsymbol{\theta}} = d\boldsymbol{\theta}/dr$, через формализм интеграла по траекториям:

$$F\left(\vec{r_0}\right) = \int \mathcal{D}\boldsymbol{\theta}(r)\exp\left\{i\int\limits_0^{r_0} dr L[r,\boldsymbol{\theta}(r),\dot{\boldsymbol{\theta}}(r)]\right\}$$

Это выражение используется следующим образом:

i) Следует подобрать конкретный вид функции $\theta(r)$, представляющей собой путь от источника к наблюдателю на $\vec{r_0} = (r_0,\theta_0)$;

ii) Вычислить интеграл в курсивных скобках по этому выбранному пути, таим образом, добыв "фазу" как функцию от $\theta(r)$;

iii) Суммировать все эти фазы для всевозможных $\theta(r)$.

Для осуществления этого, подобно тому, как делалось в типичном континуальном интеграле, разделим расстояние $r_0$ между источником и наблюдателем на малые отрезки $\epsilon = r_0/N$ за счет бесконечно большого числа $N$ и выберем путь $\theta(r)$ за счет введения координат $\boldsymbol{\theta}_j = \boldsymbol{\theta}\left(r_j\right)(j = 1,\cdots,N)$ на $j$-й сфере радиуса $r_j = j\epsilon$. Пусть линза будет находиться на $l$-й сфере (плоскость линзы), то есть $r_l = l\epsilon$ есть расстояние между источником и плоскостью линзы.

Введем "приближение тонкой линзы", формально за счет вида записи потенциала как $U(r,\theta) = \frac{1}{2}\delta\left(r - r_l\right)\hat{\psi}(\theta)$. То есть через новый параметр

$$\hat{\psi}(\boldsymbol{\theta}) = 2\int\limits_0^{r_0} dr U(r,\boldsymbol{\theta}).$$

Такое приближение не означает, что линза бесконечно тонкая, но означает, что пути, дающие вклады в фазовый интеграл хорошо аппроксимируются за счет постоянного вектора $\boldsymbol{\theta}(r) \simeq \boldsymbol{\theta}_l$ в узкой области $U(\vec{r}) \neq 0$ в сравнении с большими $r_0$ и $r_l$.

Теперь у нас уходит интегрирование по потенциалу, а также континуальный интеграл возвращается в свою дискретную аппроксимацию:

$$F\left(\vec{r}_0\right)=\left[\prod_{j=1}^{N-1}\int\frac{d^2\theta_j}{A_j}\right]\exp\left\{i\omega\left[\epsilon\sum_{j=1}^{N-1}\frac{r_jr_{j+1}}{2}\left|\frac{\boldsymbol{\theta}_j-\boldsymbol{\theta}_{j+1}}{\epsilon}\right|^2-\hat{\psi}\left(\boldsymbol{\theta}_l\right),\right]\right\}$$

здесь множители нормировки $A_j=2\pi i\epsilon/\left(\omega r_jr_{j+1}\right)$, то есть они обеспечивают $F=1$ при $\hat{\psi}=0$. Из первой части нам известно, что такой интеграл можно сделать двумерным, так что сразу перейдем к приятному виду:

$$F\left(\vec{r}_0\right)=\frac{\omega}{2\pi i}\frac{r_lr_0}{r_{l0}}\int d^2\boldsymbol{\theta}_l\exp\left\{i\omega\left[\frac{r_lr_0}{2r_{l0}}\left|\boldsymbol{\theta}_l-\boldsymbol{\theta}_0\right|^2-\hat{\psi}\left(\boldsymbol{\theta}_l\right)\right]\right\}\tag{30.4}$$

где $r_{l0}$ расстояния между линзой и наблюдателем.

Таким образом, за счет приближения тонкой линзы мы можем ограничить произвол функций $\boldsymbol{\theta}(r)$ теми, пути которых идут прямо от источника в "точку отклонения" $\theta_l$ и снова пойти прямо от $\theta_l$ к наблюдателю. Первый член квадратных скобок это разница длин пути между прямым путем и от источника к наблюдателю и отклоненным путем и отклоненным, проходящим через плоскость линзы в $\theta_l$. Второй член это задержка времени, получающаяся за счет гравитационного поля вокруг линзы.

**Итоговые уравнения**

Перепишем фактор усиления амплитуды 30.4 в более удобном виде за счет введения характерного углового масштаба, знакомого нам угла Эйнштейна $\theta_E=\left(\frac{4GMr_{l0}}{r_lr_0}\right)^{1/2}\sim 3\mu\,\mathrm{arcsec}\left(\frac{M}{M_\odot}\right)^{1/2}\left(\frac{r_lr_0/r_{l0}}{\mathrm{Gpc}}\right)^{-1/2}$ и, определив безразмерные величины

$$x=\boldsymbol{\theta}_l/\theta_E,\quad \beta=\boldsymbol{\theta}_0/\theta_E,\quad w=\frac{r_lr_0}{r_{l0}}\theta_E^2,\omega$$

$$\psi(\boldsymbol{x})=\frac{r_{l0}}{r_lr_0}\theta_E^{-2}\hat{\psi}\left(\theta_E\boldsymbol{x}\right),\quad T(\boldsymbol{x},\boldsymbol{\beta})=\frac{1}{2}|\boldsymbol{x}-\boldsymbol{\beta}|^2-\psi(\boldsymbol{x}),$$

получим вид, который будет использоваться во всех моделях в этой работе ниже:

$$F(w,\boldsymbol{\beta})=\frac{w}{2\pi i}\int d^2\boldsymbol{x}\exp[iwT(\boldsymbol{x},\boldsymbol{\beta})].$$

Из нее легко вычислить сам фактор усиления, как $|F|^2$. Таким образом, видна зависимость от частоты $w$ и позиции источника $\beta$. Отдельным достоинством этого метода, является то, что скалярный волновой анализ применим для волн любой поляризации, если вращением её направления можно пренебречь.

Если линзирующий объект сферически симметричный, то $\psi(\boldsymbol{x})$ зависит только от $x=|\boldsymbol{x}|$ и величина $F\left(\vec{r}_0\right)$ может быть выражена через функцию Бесселя $J_0$ как

$$F(w,\beta)=-iwe^{iw\beta^2/2}\int\limits_0^\infty dxxJ_0(wx\beta)\exp\left\{iw\left[\frac{1}{2}x^2-\psi(x)\right]\right\}.$$

**Геометрическое описание точечного источника**

Зачастую удобнее работать просто с геометрическим подходом, забыл в волном, именно это и использовалось учеными до развитого более общего подхода. В принципе, можно было бы об этом сказать и в первой главе, потому что для перехода к этому пределу у нас всё готово было еще тогда. Передел геометрической оптики по сути является переходом от

$$F(w,\boldsymbol{\beta}) = \frac{w}{2\pi i}\int d^2\boldsymbol{x}\exp[iwT(\boldsymbol{x},\boldsymbol{\beta})]$$

к более простому, в то же время отличному выражению вида

$$F = \sum_j \left|\gamma\left(\boldsymbol{x}_j\right)\right|^{1/2}\exp\left[iwT\left(\boldsymbol{x}_j\right) - i\pi n_j\right],$$

где $\gamma(\boldsymbol{x}) = 1/\det\left[\partial_a\partial_b T(\boldsymbol{x})\right] = 1/\det\left[\delta_{ab} - \partial_a\partial_b\psi(\boldsymbol{x})\right]$, который мы и рассмотрим сейчас подробнее, дав все минимальные обоснования.

Представим функцию временной задержки двумерной поверхностью $T(\boldsymbol{x}) = \frac{1}{2}|\boldsymbol{x} - \boldsymbol{\beta}|^2 - \psi(\boldsymbol{x})$ над $x$-плоскостью для фиксированного $\beta$. В случае коротких длин волн $w \to \infty$, подынтегральное выражение в факторе усиления $F(w,\boldsymbol{\beta}) = \frac{w}{2\pi i}\int d^2\boldsymbol{x}\exp[iwT(\boldsymbol{x},\boldsymbol{\beta})]$ сильно осциллирующая функция, поэтому величина $F$ получается в основном за счет вкладов стационарных точек поверхности $T(\boldsymbol{x})$. Запишем

$$\boldsymbol{\partial}T(\boldsymbol{x}) = 0,$$

что есть уравнение для стационарных точек или уравнение на обнуление отображения лагранжа, или, в координатах,

$$\boldsymbol{\beta} = \boldsymbol{x} - \boldsymbol{\partial}\psi(\boldsymbol{x}) \quad \text{или} \quad \boldsymbol{\theta}_0 = \boldsymbol{\theta}_l - \frac{r_{l0}}{r_0}\boldsymbol{\alpha}\left(\boldsymbol{\theta}_l\right)$$

где $\boldsymbol{\partial} = \partial/\partial\boldsymbol{x}$ и $\boldsymbol{\alpha}\left(\boldsymbol{\theta}_l\right) = r_l^{-1}\partial\hat{\psi}/\partial\boldsymbol{\theta}_l$.

Эти стационарные точки соответствуют изображения геометрической оптики, и условие $\partial T(x) = 0$ выражает принцип геометрический Ферма. То есть по сути мы получили уравнение линзы, которое определяет координаты изображения $\boldsymbol{x}$)или $\boldsymbol{\theta}_l$) для данных координат источника $\boldsymbol{\beta}$)или $\boldsymbol{\theta}_0$) с углом отклонения $\alpha$ от нулевых геодезических.

Для более "сильной" линзы, потенциал $\psi(\boldsymbol{x})$ таков, что число изображений более, чем одно.

Разложим $T(\boldsymbol{x})$ в окрестности $j$-го изображения $\boldsymbol{x}_j(\boldsymbol{\beta})$ как

$$T(\boldsymbol{x}) = T(\boldsymbol{x}_j) + \frac{1}{2}\sum_{ab}(\boldsymbol{x}-\boldsymbol{x}_j)_a(\boldsymbol{x}-\boldsymbol{x}_j)_b\partial_b T(\boldsymbol{x}_j) + \frac{1}{6}\sum_{abc}(\boldsymbol{x}-\boldsymbol{x}_j)_a(\boldsymbol{x}-\boldsymbol{x}_j)_b(\boldsymbol{x}-\boldsymbol{x}_j)_c\partial_b\partial_c T(\boldsymbol{x}_j) + \cdots$$

где индексы $abc\ldots$ пробегают значения 1 и 2. Предел геометрической оптики, означает большие частоты $w$, то есть можно записать условие

$$w\left|\partial^2 T\right|^3 \gg \left|\partial^3 T\right|^2, \quad w\left|\partial^2 T\right|^2 \gg \left|\partial^4 T\right|, \quad \cdots,$$

можно пренебречь третьим и более высокими членами разложения.

Учитывая все изображения $j$ с координатами $\boldsymbol{x}_j$, можно записать гауссов интеграл в приближении геометрической оптики как

$$F = \sum_j \left|\gamma\left(\boldsymbol{x}_j\right)\right|^{1/2}\exp\left[iwT\left(\boldsymbol{x}_j\right) - i\pi n_j\right]$$

где

$$\gamma(\boldsymbol{x}) = 1/\det\left[\partial_a\partial_b T(\boldsymbol{x})\right] = 1/\det\left[\delta_{ab} - \partial_a\partial_b\psi(\boldsymbol{x})\right]$$

и $n_j = 0, 1/2, 1$ - когда $x_j$ это минимум, седловая точка и максимум $T(x)$, соответственно.

Уравнение для $F$ выражает, что наблюдаемая волна это суперпозиция волн от каждого изображения, с соотношением амплитуд $\left|\gamma\left(\boldsymbol{x}_j\right)\right|^{1/2}$ и фазовым сдвигом $wT\left(\boldsymbol{x}_j\right) - \pi n_j$.

Усиление интенсивности волны принимает вид:

$$|F|^2 = \sum_j |\gamma_j| + 2\sum_{j<k} \left|\gamma_j\gamma_k\right|^{1/2} \cos\left(wT_{jk} - \pi n_{jk}\right) \tag{30.5}$$

где $\gamma_j = \gamma\left(\boldsymbol{x}_j\right)$, $T_{jk} = T(\boldsymbol{x}_k) - T(\boldsymbol{x}_j)$ это средняя временная задержка между $k$-м и $j$-м изображениями, и $n_{jk} = n_k - n_j$.

Обратите внимание, что в геометрической оптике коэффициент увеличения размера изображения для бесконечно малого источника задается как якобиан линзирующего отображения $\boldsymbol{\beta} \mapsto \boldsymbol{x}$.

$$\gamma(\boldsymbol{x}) = 1/\det[\boldsymbol{\partial} \otimes \boldsymbol{\beta}] = 1/\det[\delta_{ab} - \partial_a\partial_b\psi(\boldsymbol{x})]$$

Исходя из сохранения поверхностной яркости в геометрической оптике, это увеличение изображения равно усилению интенсивности источника. Поэтому первый член в уравнении (30.5) содержит результат геометрической оптики, в котором величина $|\gamma_j|$ это фактор усиления от $j$-го изображения. Второй член выражает “интерференцию” между изображениями, которое создает различные интерференционные картины. Уравнение (30.5) по сути есть квазиклассическая аппроксимация квантовой механики.

**Насколько вообще нужно заниматься каустиками и этой теорией?**

(это важный вопрос. может, это всё - вообще не то направление? пока оставлю его.)

### 30.6.2 Observations of Gravitational Lensing and Phenomena

(what is here and what is observed)

### 30.6.3 Wave Optics of Gravitational Lensing

(what is here and what is observed)

### 30.6.4 Some Typical Models of GL

(add some typical, from Wieneberg or astronomers..)

**Волновая модель одиночной линзы**

Прежде чем изучать сложные модели, проиллюстрируем нашу теорию на самом тривиальном случае - случае одиночной массы. И хотя этот чисто иллюстративный, ведь, как будет показано далее, даже небольшое возмущение способно в корне изменить полученный график интенсивности, уже изложенные ниже методы могут предсказывать такие известные явления, как кольцо или крест Эйнштейна, на которых в этой работе останавливаться нет возможности.

Начнем анализ этой модели с вида фактора усиления для сферически симметричного случая

$$F(w,\boldsymbol{\beta}) = -iwe^{iw\boldsymbol{\beta}^2/2}\int\limits_0^\infty dxxJ_0(wx\boldsymbol{\beta})\exp\left\{iw\left[\frac{1}{2}x^2-\psi(x)\right]\right\} \tag{30.6}$$

Подставим это в выражение $\hat{\psi}(\boldsymbol{\theta}) = 2\int\limits_0^{r_0} drU(r,\boldsymbol{\theta})$, через потенциал

$$U(r,\theta) = -GM/\left(r^2-2r_l r\cos\theta+r_l^2\right)^{1/2},$$

где $M$ обозначает ее массу.

и, используя приближение $\theta \ll 1$, получаем линзирующий потенциал в виде $\hat{\psi}\left(\theta_l\right) = 4GM\ln\theta_l +$ const, где постоянный член не важен, так как он вклад дает во все фазы $F$.

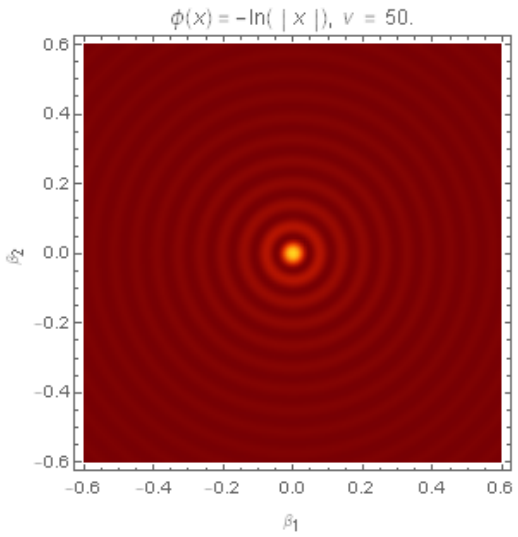
Масштабируя углы по углу Эйнштейна $\theta_E$, подобно тому, как это было в прошлом разделе, получаем выражения

$$\psi(x) = \ln x, \quad w = 4GM\omega \sim 10^5\left(M/M_\odot\right)(\nu/\text{GHz}).$$

Заметим, что $w \sim$ (гравитационный радиус от массивной линзы) / (длина волны) - это большая величина в типичных задачах астрофизики. Подставляя эти величины в (30.6), несложно разглядеть решение в виде известной гипергеометрической функции ${}_1F_1$, то есть в итоге для фактора усиления имеем

$$|F|^2 = \frac{\pi w}{1-e^{-\pi w}}\left|{}_1F_1\left(\frac{1}{2}iw,1;\frac{1}{2}iw\boldsymbol{\beta}^2\right)\right|^2. \tag{30.7}$$

В пределе больших длин волн $w \to 0$, фактор усиления равен единице, и усиление не происходит, то есть по сути волна игнорирует линзу. В противоположном пределе, когда $w \gg 1$, фактор усиления точно оценивается квазиклассическим приближением, которое работает, пока $\beta \gg 1/w$, и $|F|^2 \simeq \pi w J_0^2(w\beta)$ справедливо, пока $\beta \ll w^{-1/2}$.




Функция временной задержки $T(\boldsymbol{x}) = \frac{1}{2}|\boldsymbol{x}-\boldsymbol{\beta}|^2 - \ln x$ имеет две стационарных точки, точку минимума $\boldsymbol{x}_1$ и стационарную точку $\boldsymbol{x}_2$:

$$\boldsymbol{x}_{1,2} = \frac{\boldsymbol{\beta}}{2\beta}\left(\beta \pm \sqrt{\beta^2+4}\right), \quad \gamma_{1,2} = \frac{1}{2} \pm \frac{\beta^2+2}{2\beta\sqrt{\beta^2+4}}$$

Эти точки соответствуют двум классическим лучам, один из которых идет через $\boldsymbol{x}_1$, на той же сторона, что и источник, вне кольца Эйнштейна $x = 1$. Этот луч дает решающий вклад в интенсивность в режиме слабого линзирования. Другой луч идет через $\boldsymbol{x}_2 = -\boldsymbol{x}_1/x_1^2$, на обратной стороне от линзы и внутри круга Эйнштейна. Функция временной задержки, соответственно, принимает вид

$$T_{12} = \frac{1}{2}\beta\sqrt{\beta^2+4} + \ln\frac{\sqrt{\beta^2+4}+\beta}{\sqrt{\beta^2+4}-\beta}, \quad n_{12} = \frac{1}{2}$$

И в итоге квазиклассическое приближение решения волнового уравнения для точечной линзы запишется как:

$$|F|^2 = \frac{\beta^2+2+2\sin(wT_{12})}{\beta\sqrt{\beta^2+4}} \tag{30.8}$$

В случае $\beta \ll 1$, получаем выражение

$$|F|^2 \simeq [1+\sin(2w\beta)]/\beta.$$

Член с синусом выражает интерференцию от двух изображений, за счет него появляются круговые дифракционные кольца.

Критическая кривая радиуса $x = 1$ формирует знаменитое кольцо Эйнштейна в плоскости линзы, а каустика в этой модели это всего лишь точка $\beta = 0$ в плоскости источника, в которой $|F|^2$ в геометрическом подходе, согласно (30.8), расходится. Понятно, что в настоящем волновом описании максимальное значение $|F|^2$ далеко не бесконечность, а $\pi w/(1-e^{-\pi w})$, ведь выражение для него нам известно (30.7). В зависимости от частоты, максимальное усиление $\simeq \pi w$ по порядку пропорционально отношению гравитационного радиуса к длине волны, заметно больше, чем единица.

**О линзировании бинарными системами**

**Типичная модель и результаты**

**Интерпретация результатов**

(помню, что отдельно это нужно прописать, иначе херня какая-то)

**О реальных наблюдениях линзирования бинарных систем (!?!?)**

(очень важно указать, насколько соотносится эта теория с тем, что в реальности?)

### 30.6.5 Other about GL

(тут указание на многое, что буду писать в записи по линзированию.)

# 31 Catastrophe Optics: Morphologies of Caustics and Their Diffraction Patterns by Berry, Upstill

## 31.1 § 8. Concluding Remarks

Catastrophe optics belongs in a more general context, in which wave motion is viewed, unconventionally, in terms of the contrast and interplay between the morphologies of three extreme regimes (Berry [1979a]). Firstly, if the wavelength $\lambda$ is small in comparison with scales

of variation of diffracting objects or refracting media, the wavefield is dominated by the caustics and associated diffraction patterns which have been the subject of this review. Secondly, when waves propagate in environments which can be modelled by a hierarchy of scales extending to the infinitely small, caustics cannot occur and the limit $\lambda \to 0$ is not geometrical optics; Mandelbrot [1977] calls such hierarchical structures fractals, and Berry [1979b] calls the corresponding waves diffractals. And thirdly, when waves are explored on the scale of $\lambda$, the principal features are wavefronts, which are dominated by their singularities, in the form of lines in space; these are the wavefront dislocations described in Appendix 1.

Finally, we emphasize that catastrophe optics, although founded on differential topology, is now firmly part of physics. Its development has been characterized by an intimate combination of observation, experiment, analysis and computation. This has been especially fertile in studies of diffraction catastrophes (§ 3), liquid droplet optics (§5) and caustic networks (§6). The successes so far achieved amply justify the injection into optics of concepts like structural stability and the unfolding of a singularity, which must surely be unfamiliar and strange to most physicists. Catastrophe optics is bringing within the compass of our understanding whole realms of phenomena with the pleasant property, uncommon in physics nowadays, that they are accessible not only in the laboratory but also in nature, where they can be seen by everybody.

## 31.2 § 1. Introduction

As light propagates, it is common for focusing to occur. In the most elementary picture, all the rays pass through a focal point. It is exceptional, however, for this to correspond to reality: what happens typically is that the rays envelop a caustic surface in space. The light intensity is greatest near caustics, and they dominate optical images whenever the wavelength is small enough for geometrical optics to form the basis of a valid approximation to the wavefield.

In spite of their importance, caustics have been awkward to fit into optical theory, for two main reasons. Firstly, their forms are very intricate and until recently were only classified as perturbations from a perfect focus, i.e. as lens aberrations. Secondly, the simplest short-wave theory, in which the rays of geometrical optics are endowed with amplitude and phase and allowed to interfere where they cross, fails on caustics by predicting unphysical intensity divergences. In other words, the theory goes wrong in precisely the most important places - where the light is brightest!

This unsatisfactory state of affairs has been changed dramatically by the injection into optics of a new branch of mathematics called catastrophe theory, invented by THOM [1972] and developed most notably by Arnol'd [1975]. It is the purpose of this review to show how the classification of caustics as catastrophes is leading us to understand a wide variety of optical phenomena. The resulting catastrophe optics is unfamiliar and unorthodox, cutting across traditional categories within the subject. In large measure this new point of view arises from one concept which plays a crucial role in the mathematics, namely structural stability. We shall show that this enables us to classify those caustics whose topology survives perturbation. Such stable caustics abound in nature in circumstances where there is no special symmetry, for example when caustics are formed as a result of refraction or reflection of light by a wavy water surface. The perfect point focus so sought after in optical technology is in this mathematical sense highly unstable, a fact borne out
by the difficulty of producing it. Thus catastrophe optics is complementary to traditional optics, where the emphasis is on solving ray or wave equations in highly symmetric situations.

At the outset we wish to dispel some possible misconceptions. Firstly, the fact that catastrophe theory is rooted in topological (i.e. qualitative) concepts does not make it imprecise; we shall show that it is perfectly possible to make quantitative statements in catastrophe optics. Secondly, the fact that the fundamental theorem is a classification does not make its applica-

tion to optics purely taxonomic; we shall show how to use the theory to make surprising and testable predictions. And thirdly, although some invocations of catastrophe theory in science have been criticised (Zahler and Sussman [1977]), these criticisms do not apply (and were not intended to apply) to catastrophe optics, which has a firm mathematical foundation.

We begin in § 2 by explaining how the physics of stable caustics, which is concerned with envelopes of families of rays, is isomorphic to the mathematics of catastrophe theory, which is concerned with stable singularities of gradient maps. The precise correspondence occurs because rays obey the Hamilton-Fermat principle of stationary time. The stable caustics have the same structure as the catastrophes, and we illustrate them by simple examples.

It is helpful to think of the catastrophes as elemental "atomic forms". The corresponding caustics have a delicate "subatomic" structure of diffraction fringes, and they can link up into elaborate "macroscopic" networks whose details are often not resolved. Catastrophe optics thus embodies a hierarchy of scales. Our subsequent examples will range from the topology of the zeroes of diffraction patterns on scales of a few wavelengths to the topology of caustic networks on scales of metres. The catastrophes themselves occupy a sort of mesoscale, organising finer and grosser structures.

In §3 we step down to the microscale, to explore the diffraction catastrophes that both clothe and underlie caustics. Each structurally stable caustic has its characteristic diffraction pattern, whose wave function has an integral representation in terms of the standard polynomial describing the corresponding catastrophe. We shall explain how these integrals arise as short wave asymptotic solutions of the wave equation. The diffraction catastrophes constitute a new hierarchy of functions, different from the special functions of analysis. Several are understood in detail; theory and experiment agree down to the finest scales. Important
aspects of diffraction near caustics are embodied in a series of universal scaling laws, giving exponents in power laws governing the increase in intensity and the decrease in diffraction fringe dimensions as the wavelength tends to zero.

Often, waves propagate paraxially in a uniform medium, from an initial wavefront which is a gentle deformation of a plane. In §4 we consider paraxiality by describing the relevant differential geometry of wavefronts and their normals. Usually, textbooks confine themselves to second order properties of surfaces; we shall find it necessary to discuss umbilic points (where both principal curvatures are equal), whose classification involves third order properties.

In §5 we apply catastrophe optics to irregular liquid droplet lenses. The existence of these droplets is a consequence of surface tension, but gravity affects their shape and can act as a parameter changing the topology of the caustics they produce by refraction (or reflection).

In § 6 we step up to the macroscale, to explore the caustics produced by an extended irregular wavefront such as emerges from the wavy surface of sunlit water or from randomly undulating glass (as used for bathroom windows). The gross structure of these caustic patterns is unintelligible until it is realised that they are made up of poorly resolved catastrophes. Once this is appreciated, it is possible to design and carry out experiments to reveal the complete caustic structure predicted theoretically.

The newest application of catastrophe optics is to random short waves, whose statistical properties are determined by the random caustic structure. In §7 we show how some of these statistics can be calculated. The deepest results concern the wavelength dependence of the intensity moments of twinkling starlight, which involve the scaling laws of § 3 and depend on the whole infinite hierarchy of catastrophes.

In Appendix 1 we give a brief account of the complementary subject of singularities of wavefronts. These can be of two very different sorts: wavefront dislocations, which are singularities of the phase of the wave function, constituting structurally stable features of diffraction catastrophes, and Legendre singularities, which are singularities of the (fictitious) surfaces normal to the rays of geometrical optics.

In Appendix 2 we give a brief account of the connection between the catastrophe classification of caustics and the traditional classification of lens aberrations.

We have tried to make this review self-contained, so far as is possible. Readers will find catastrophe theory expounded in the excellent text by

Poston and Stewart [1978] which complements the original, often rather speculative work of Thom [1972]; more mathematical treatments are given by Arnol'd [1975] and Zeeman [1977]. The relevant asymptotics of wave equations were discussed before catastrophe theory by Maslov [1965] and Ludwig [1966], and in terms of catastrophe theory by Duistermat [1974] and Gullemin and Sternberg [1977]. Catastrophe theory and some physical applications have been reviewed by Golubitsky [1978] and Poston [1980]. More elementary accounts of catastrophe optics are given by Berry [1976, 1978a, 1979a].

## 31.3 § 2. Catastrophe Theory and Stable Caustics

### 31.3.1 2.1. Caustics Are Singularities of Gradient Maps

In optics what matters is not so much individual rays as families of rays filling regions of space. The singularities of ray families are the caustics that this review will be concerned with. To prepare for the application of catastrophe theory we need to recall some standard properties of ray families. Let the rays (and waves) propagate in an inhomogeneous medium with refractive index $n = n(\mathscr{R})$ at the point $\mathscr{R}$. For simplicity of exposition we consider $n$ independent of direction and time and consider only monochromatic light whose waves have frequency $\omega$. This will be sufficient for later examples, and in any case the generalisation of catastrophe optics to anisotropic, time-varying, dispersive media involves no essentially new ideas.

A ray through the point $\mathscr{R}$ is specified by its direction. It will prove convenient to express this in terms of the wave vector $\boldsymbol{k}$ whose length $k$ is given by the dispersion relation

$$k = n\omega/c \equiv \kappa n \tag{2.1}$$

where $c$ is the speed of light, and $\kappa$ the wave number, in vacuo. If $\tau$ measures arc length along the ray, the unit vector $\hat{\boldsymbol{k}}$ is

$$\hat{\boldsymbol{k}} \equiv \boldsymbol{k}/k = \mathrm{d}\mathscr{R}/\mathrm{d}\tau. \tag{2.2}$$

Of crucial importance to catastrophe optics is the fact that several rays in a family may pass through the point $\mathscr{R}$. These can be labelled by a suffix $\mu$; the wave vector on the $\mu$ th path will be denoted by $\boldsymbol{k}_\mu$. The $\boldsymbol{k}_\mu$ make up vector fields over $\mathscr{R}$. By virtue of their origin in a wave equation

(§ 3) these vector fields must be derivable from scalar fields $\mathscr{S}_\mu = \mathscr{S}_\mu(\mathscr{R})$, defined (up to constants) by

$$k_\mu = \kappa \nabla \mathscr{S}_\mu \tag{2.3}$$

(this implies that the $\boldsymbol{k}_\mu$ are irrotational). $\mathscr{S}_\mu$ are the branches of a multivalued action function $\mathscr{S}$ which from (2.1) satisfies the HamiltonJacobi equation

$$|\nabla \mathscr{S}|^2 = n^2 \tag{2.4}$$

(Lanczos [1966]). The action surfaces $\mathscr{S}$ = constant are the wavefronts of geometrical optics (Appendix 1); from (2.3) and (2.2) they are orthogonal to the rays, which therefore form a normal congruence (Born and Wolf [1975] chap. 3).

The importance of the multivaluedness of $\mathscr{S}$ lies in the fact that it is precisely on a caustic that two or more of its branches join. To see this, consider a point $\mathscr{R}$ lying on a caustic of the ray family (Fig. 2.1a). By definition, a caustic is an envelope of the family, so as $\mathscr{R}$ is approached (from the bright side of the caustic) two or more rays become parallel and hence by (2.2) their $\boldsymbol{k}_\mu$ coincide. By (2.3) this shows that $\mathscr{R}$ is a branch point of $\mathscr{S}$. Also shown in Fig. 2.1 are the wavefronts, whose branches touch the caustic in cusps (Appendix 1). Fig. 2.1b shows a more complicated case where the caustic itself is cusped. Outside the caustic there is one ray through each point, and $\mathscr{S}$ is single valued. Inside the caustic there are three rays and $\mathscr{S}$ is three valued. Each arm of the caustic curve is an envelope where rays touch in pairs, while at the cusp point itself three rays touch.

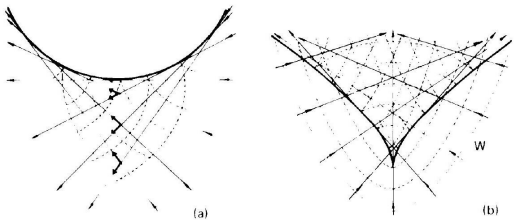



Fig. 2.1. Rays (full lines), wavefronts (dashed lines) and caustics (bold lines). (a) Smooth caustic curve at each of whose points two rays touch: wave vectors $\boldsymbol{k}$ are shown by bold arrows: (b) cusped caustic formed by propagation of initial wavefront $W$.}

Sometimes infinitely many rays pass through each point in some region. Such behaviour is very important in mechanics, and corresponds to stochastic motion (Berry [1978b]); but it is not commonly encountered in optics and moreover does not involve catastrophe theory. Therefore we shall not consider this case here.

In catastrophe theory the multivaluedness of $\mathscr{S}$ is tamed by embedding this function in a single-valued function with extra variables, in the following manner. To find the action $\mathscr{S}$ for a ray family, eq. (2.4) must be solved under some boundary condition. A general boundary condition is to specify $\mathscr{S}$ on a surface $\sum$ (not necessarily a wavefront). If spatial position on $\Sigma$ is denoted by $\mathscr{R}_0$, this boundary condition is

$$\mathscr{S} = \mathscr{S}(\mathscr{R}_0) = \text{ given on } \Sigma \tag{2.5}$$

The corresponding solution of (2.3) is obtained in a convenient from from the following version of Fermat's principle of extremal time (Born and Wolf [1975] chap. 3, Lanczos [1966]).

Consider all paths $P$ consistent with the given conditions, i.e. starting on $\Sigma$ and passing through $\mathscr{R}$ (Fig. 2.2), and specify each $P$ by position as a function of arc length, $\mathscr{R}(\tau)$. For each $P$ form the path integral
i.e.

$$\phi(\mathscr{R}) = \mathscr{S}(\mathscr{R}_0) + \frac{1}{\kappa}\int_{\mathscr{R}_0}^{\mathscr{R}} k(\mathscr{R}(\tau)) \cdot \mathrm{d}\mathscr{R},$$
$$\phi(\mathscr{R}) = \mathscr{S}(\mathscr{R}_0) + \int_0^{\tau(\mathscr{R})} n(\mathscr{R}(\tau))\mathrm{d}\tau, \tag{2.6}$$

which is simply the optical distance along $P$ from $\Sigma$ to $\mathscr{R}$, i.e. the time multiplied by $c$. Then the branches of the action $\mathscr{S}$ are the extreme values of $\phi$ in the space of paths, i.e. the values of $\phi$ for paths $P$ satisfying

$$\delta\left[\mathscr{P}\left(\mathscr{R}_0\right)+\int_0^{\tau(\mathscr{R})} n(\mathscr{R}(\tau))\mathrm{d}\tau\right]=0 \tag{2.7}$$

where $\delta$ denotes first order path variations. The extreme paths are the rays through $\mathscr{R}$. If the ray family has no caustics between $\Sigma$ and $\mathscr{R}$, $\mathscr{S}$ is single valued; otherwise $\mathscr{S}$ is multivalued.

We remark in passing that the Euler-Lagrange equations from (2.7) yield the ray equations

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\left(n(\mathscr{R}(\tau))\frac{\mathrm{d}\mathscr{R}(\tau)}{\mathrm{d}\tau}\right)=\nabla n(\mathscr{R}(\tau)). \tag{2.8}$$

These show how refractive index variations alter the ray direction $\mathrm{d}\mathscr{R}/\mathrm{d}\tau$; the equations reduce to Snell's law if the medium is stratified.

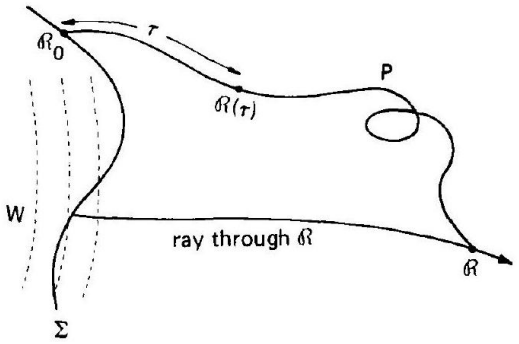



Fig. 2.2. A path $P$ from the point $\mathscr{R}_0$ on an initial surface $\Sigma$, together with a ray from $\Sigma$ to $\mathscr{R}$, in a medium of refractive index $n(\mathscr{R})$. Wavefronts $W$ near $\Sigma$ are shown as dashed lines.}

Now, although $\mathscr{S}$ may be multivalued (and in cases of interest, where there are caustics, it will be), the function $\phi(\mathscr{R})$ is single valued for a given path $P$. Therefore by specifying $P$ with variables $s=(s_1,s_2\cdots)$ the variational equation (2.7) can be expressed as the vanishing of the gradient of a single-valued function $\phi=\phi(s;\mathscr{R})$ :

$$\partial\phi/\partial s_i=0 \quad \forall i, \tag{2.9}$$

where $i$ is an index labelling the separate variables $s_1,s_2\cdots$. By these gradient conditions, $s$ and hence $P$ are fixed for those rays $\boldsymbol{k}$ passing through $\mathscr{R}$.

The $s$ can be chosen in many ways. For example, the path $P$ from $\mathscr{R}_0$ to $\mathscr{R}$ can be divided into infinitely many segments (Fig. 2.3) with three of the $s$ assigned to the starting point of each. Alternatively, the function $\mathscr{R}(\tau)$ could be Fourier-analyzed for each $P$, and $s$ taken as the Fourier coefficients. These procedures result in infinitely many $s$.

However, it is possible to resolve the multivaluedness of $\mathscr{S}$ by choosing paths with a finite number of $s$; then, of course, the segments of $P$ must themselves be rays. In the simplest procedure of all, $P$ is chosen with a single segment, namely the ray from $\mathscr{R}_0$ to $\mathscr{R}$ (Fig. 2.4), and only two variables $s=(s_1,s_2)$ are required, corresponding to coordinates of $\mathscr{R}_0$ on

$\Sigma$. This will result in a single-valued path integral $\phi\left(s_1,s_2;\mathscr{R}\right)$, provided only one ray from $\mathscr{R}_0$ passes through $\mathscr{R}$ - in other words, provided there are no caustics between $\mathscr{R}_0$ and $\mathscr{R}$ in the family of rays emanating from $\mathscr{R}_0$, a condition certainly satisfied for the important case of homogeneous media where the rays are straight, and also in refracting media if $\mathscr{R}$ is not too far from $\Sigma$.

In a sufficiently extended refracting medium this procedure may fail by producing a multivalued $\phi$. To deal with such cases an alternative method must be employed to construct a

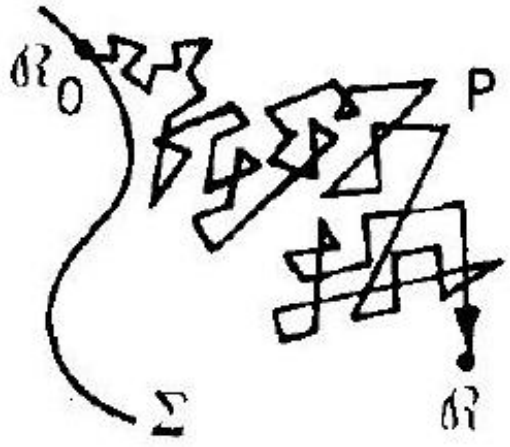



Fig. 2.3. Path $\boldsymbol{P
$ with many segments, from $\mathscr{R}_0$ on $\Sigma$ to $\mathscr{R}$.}

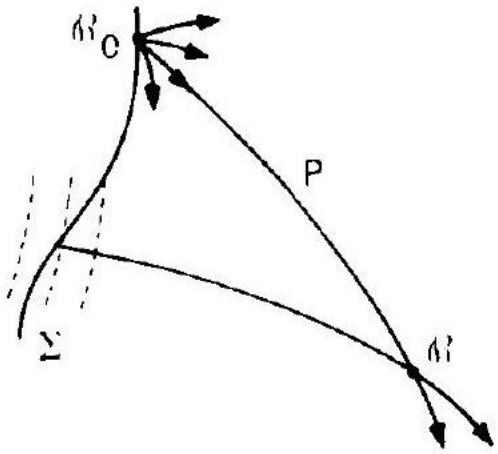



Fig. 2.4. Path $P$ through $\mathscr{R}
$ in the family of rays emanating from $\mathscr{R}_0$ on $\Sigma$, together with an extreme ray from $\Sigma$ to $\mathscr{R}$. Wavefronts near $\Sigma$ are shown as dashed lines.}

single-valued $\phi$; one such method, involving Lagrangian manifolds in the phase space $\mathscr{R}, \boldsymbol{k}$, was devised by Maslov [1965] (see also Kravtsov [1968]), and will be briefly discussed in § 3.2.

The conclusion from this discussion is that there are a number of ways of representing the branches $\mathscr{S}_\mu$ of a multivalued action function $\mathscr{S}(\mathscr{R})$ as the extreme values of a single-valued function $\phi(s;\mathscr{R})$. It helps to think of $\phi$ for fixed $\mathscr{R}$ as the height of a landscape above a plane with coordinates $s$. Then the coordinates $s^\mu(\mathscr{R})$ of the extrema (hilltops, minima, saddles) satisfying (2.9) correspond to the rays, and the heights

$$\mathscr{S}_\mu = \phi\left(s^\mu(\mathscr{R});\mathscr{R}\right) \tag{2.10}$$

give the actions.

In the language of catastrophe theory, $\phi(s;\mathscr{R})$ is a generating function, and (2.9) defines a gradient map from $\mathscr{R}$ to $s$. The $s$ are called state variables. The components of $\mathscr{R}$ are quantities on which the rays depend, and appear as arguments in the action $\mathscr{S}$ (eq. (2.10)); there may be other quantities on which the rays depend, like parameters describing the form of $\Sigma$ or the action on it (e.g. angles specifying the orientation of a refracting surface relative to an incident beam, or the time in a medium

that is changing). All such quantities are termed control parameters and will henceforth be denoted by $C = (C_1, C_2 \cdots)$, and the generating function written as $\phi(s; C)$.

Now let $C$ vary, for example by changing the observation point $\mathscr{R}$. The extrema $s^\mu(C)$ (rays) will change. When two or more extrema coalesce, the map (2.9) is singular, and $\phi$ is stationary to higher than first order. But we have already seen that coalescence of rays occurs on caustics; therefore caustics correspond to singularities of gradient maps.

The condition for $C$ to lie on a caustic is obtained in terms of the generating function as follows: higher order stationarity implies the existence of a set of displacements $\mathrm{d}s_i$, away from the extrema $s^\mu$, for which (2.9) continues to hold, i.e. for which

$$\sum_i \mathrm{d}s_i \frac{\partial^2 \phi}{\partial s_i \partial s_j} = 0. \tag{2.11}$$

The condition for these homogeneous equations to have a solution, and hence for $C$ to lie on a caustic, is that the Hessian

$$\mathscr{H}[\phi] \equiv \det \frac{\partial^2 \phi}{\partial s_i \partial s_i} = 0 \tag{2.12}$$

at points $s^\mu(C)$ where (2.9) holds.

## 31.3.2 2.2. The Classification of Catastrophes

In catastrophe optics we start from a smooth function $\phi$ of state variables $s$ and control parameters $C$, construct the gradient map (2.9), and enquire about the form of its singularities, defined by eq. (2.12), in the control space with coordinates $C$. We have seen that these singularities are the caustics.

We seek a description of the structurally stable singularities. To explain this important concept, consider the singularity $S_1$ produced by a generating function $\phi_1$. Now let $\phi_1$ be perturbed to $\phi_2$ (in optics this perturbation might be produced by altering the initial surface $\Sigma$, or the propagation medium); the singularity will change to $S_2$. Structural stability means that $S_1$ and $S_2$ are related by a diffeomorphism of $C$, that is a smooth reversible change of control parameters. Alternatively, imagine $S_1$ in a control space made of rubber; then it is structurally stable if it can be
transformed into $S_2$ by smooth deformation of the rubber (without tearing or self-overlap or infinite strain).

Structural stability is typical, provided the dimensionality of control space does not exceed seven (in which case the definition of structural stability must be modified somewhat - see below). This follows from the mathematics (Zeeman [1977]) where it is shown that unstable singularities (e.g. point foci) come from a subset of measure zero in the space of all generating functions. In other words, structural stability is a generic property of caustics.

Structurally stable singularities are divided into equivalence classes; all members of a class can be deformed into one another by diffeomorphism. It is these equivalence classes of singularity that constitute the catastrophes. Each is described by a standard polynomial, or normal form, henceforth denoted by $\Phi$, for the generating function $\phi$. In its simplest version, the main classification theorem of catastrophe theory is a list of these standard polynomials. Each polynomial gives a description, up to diffeomorphism, of all singularities in its equivalence class.

The catastrophes are classified by their codimension $K$.$K$ is the dimensionality of control space (i.e. the number of parameters $C$ ) minus the dimensionality of the singularity itself. Alternatively, $K$ is the dimensionality of the subspace that must be explored to encounter the singularity. For example, a caustic point in one space dimension, a caustic curve in two dimensions, and a caustic surface in three dimensions, all have codimension 1.

All catastrophes with $K \leq 4$ (the "elementary catastrophes") are listed in Table 1, with the original names due to Thom [1972] and also the symbols of Arnol'd [1975]. The subscripts on Arnol'd's symbols denote the multiplicity of the catastrophe, i.e. the number of extrema of $\Phi$ (rays)

Table 1
Standard polynomials $\Phi$ for the elementary catastrophes with codimension $K \leq 4$

| Name | Symbol | $K$ | $\Phi(s;C)$ |
|---|---|---|---|
| fold | $A_2$ | 1 | $s^3/3 + Cs$ |
| cusp | $A_3$ | 2 | $s^4/4 + C_2s^2/2 + C_1s$ |
| swallowtail | $A_4$ | 3 | $s^5/5 + C_3s^3/3 + C_2s^2/2 + C_1s$ |
| elliptic umbilic | $D_4^-$ | 3 | $s_1^3 - 3s_1s_2^2 - C_3(s_1^2 + s_2^2) - C_2s_2 - C_1s_1$ |
| hyperbolic umbilic | $D_4^+$ | 3 | $s_1^3 + s_2^3 - C_3s_1s_2 - C_2s_2 - C_1s_1$ |
| butterfly | $A_5$ | 4 | $s^6/6 + C_4s^4/4 + C_3s^3/3 + C_2s^2/2 + C_1s_1$ |
| parabolic umbilic | $D_5$ | 4 | $s_1^4 + s_1s_2^2 + C_4s_2^2 + C_3s_1^2 + C_2s_2 + C_1s_1$ |

which coalesce at the most singular point. For a proof of the classification theorem, see Zeeman [1977]; for a detailed account of the ideas underlying the proof, see Poston and Stewart [1978].

Each polynomial consists of two parts: a germ, involving only the state variables $s$, and unfolding terms that also involve the controls $C$ (which appear linearly). The germ describes the most singular extremum of the $\Phi$ landscape, produced by setting $C = 0$. The unfolding terms describe how the degenerate extremum generifies into a collection of $K + 1$ ordinary extrema as the controls are altered.

Fig. 2.5 shows the topology in control space of the catastrophes in Table 1. The way the caustics unfold away from their singularities is evident: the fold point (Fig. 2.5a) is an isolated singularity; the cusp point (Fig. 2.5b) unfolds into two fold curves; the swallowtail point (Fig. 2.5c) unfolds into two cusped edges (ribs) connected by three fold surfaces with a line of self-intersection; the elliptic umbilic point (Fig. 2.5d) unfolds symmetrically into three ribs joined by three fold surfaces, etc. The phenomenon of higher catastrophes unfolding into configurations of lower ones is general (Arnol'd [1975]).

These catastrophes give the elemental "atomic forms" that stable caustics can adopt. It is important to appreciate that the classification is complete: for $K \leq 4$ all forms other than those in Table 1 are unstable. In this sense the theorem is surprising, for it is far from obvious that in a plane, for example, a cusp point is stable whereas a finite-angled corner is not. In what follows we shall give abundant illustration of the usefulness of the theorem in optics. For the most part we shall confine our attention to Thom's original list as in Table 1 and Fig. 2.5, but we shall occasionally venture into Arnol'd's [1975] enormous extension of the list for $K > 4$. The novel feature of this extension is that for $K > 7$ it contains catastrophes which are stable not under diffeomorphism but only under homeomorphism, which is a weaker form of equivalence; the optical significance of this difference is being studied in this laboratory (see §§5 and 6).

The polynomials $\Phi$ in Table 1 are written in terms of the fewest state variables $s$ necessary to describe the singularity. This is the corank of the catastrophe. The cuspoids $A_{2-5}$ have corank 1, and the umbilics $D_4^{\pm}, D_5$ have corank 2 . It is always possible to add extra state variables $s$, in terms that are at most quadratic, without affecting the singularity; the extrema of $\phi$ are non-degenerate in these extra $s$-directions, and it is easy to verify directly that the new terms leave the Hessian equation (2.12) unaltered.

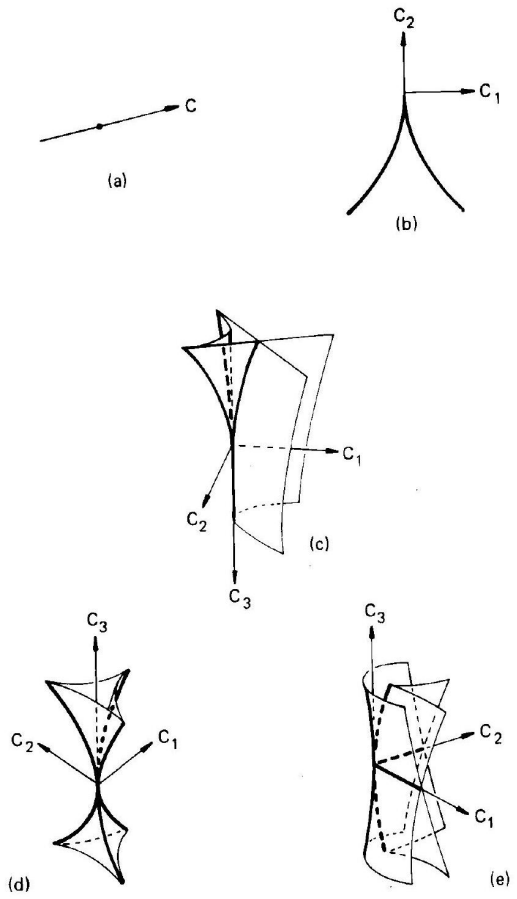



Fig. 2.5. Topology (in control space) of the elementary catastrophes: (a) fold; (b) cusp; (c) swallowtail; (d) elliptic umbilic; (e) hyperbolic umbilic; (f) butterfly; (g) parabolic umbilic.

### 31.3.3 2.3. Examples of the Elementary Catastrophes in Optics

In the rainbow (Descartes [1637], Humphreys [1940], Tricker [1970], Khare and Nussenzveig [1974], Nussenzveig [1969a,b, 1977]), parallel sunrays hit a raindrop (Fig. 2.6a), suffer two refractions and one internal reflection, and are viewed from afar. Because of the rotational symmetry there is only one control parameter $C$, which can be taken as the angle of deflection of the rays, and there need be only one state variable $s$, which can be taken as the latitude of the original point of incidence. Elementary optical laws show the graph of $s(C)$ (Fig. 2.6b) to be multivalued: there are two rays if $C > C_0 = 138°$ (for orange light) and none if $C < C_0$. However, the graph of $C(s)$ is single valued and has a minimum at $s = 59°$. The rays form a directional caustic at $C_0$ and this is a fold catastrophe, which is indeed the only stable singularity with codimension 1. In space the caustic surface is asymptotic to a cone with semiangle

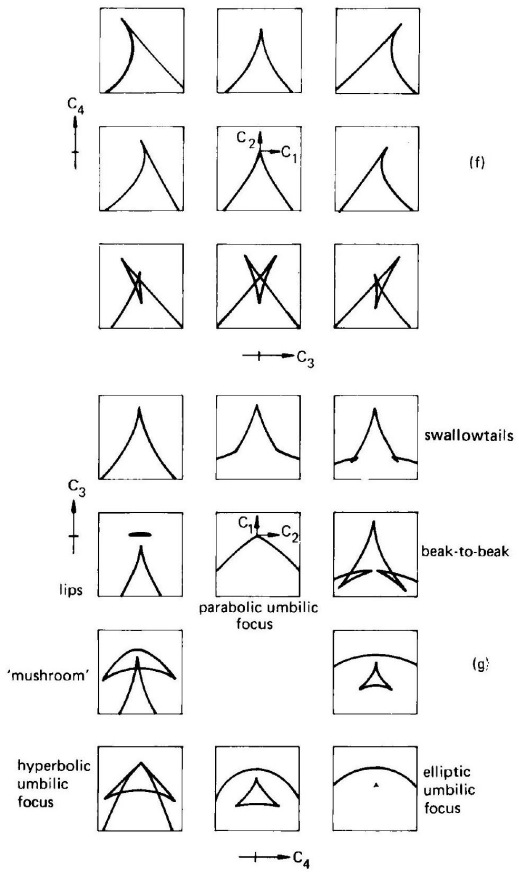



Fig. 2.5 (contd.)

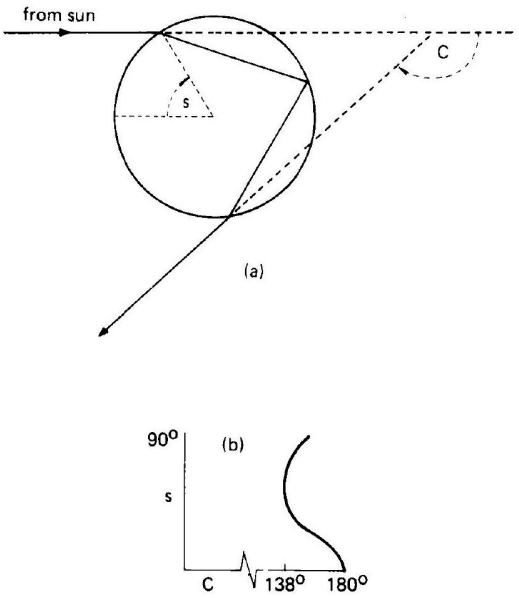



Fig. 2.6. (a) State variable $s$ and control parameter $C$ for a ray contributing to the rainbow; (b) graph of $s$ against $C$, showing multivaluedness.

$180° - C_0 = 42°$; very close to the drop, the caustic forms a more complicated surface (McDonald [1963]).

In the sparkling of sunlight on the sea (Fig. 2.7) (Longuet-Higgins [1960a]), the eye sees a number of reflected images of the sun, corresponding to specular points, where the wavy water surface is so oriented as to reflect rays into the eye. Now the state variables may be taken as two coordinates $s = (s_1, s_2)$ on the water surface, and the control parameters $C$ are the time

$t$ of observation, and the positions $\mathscr{R}_{\mathrm{E}}$ and $\mathscr{R}_{\mathrm{S}}$ of the eye and sun. The generating function $\phi\,(s_1, s_2; t, \mathscr{R}_{\mathrm{E}}, \mathscr{R}_{\mathrm{S}})$ is simply proportional to the distance from the sun via $s$ on the sea to the eye along straight paths, and it is then obvious by elementary reasoning that the extrema occur at the specular points $s$. Over short times, $\mathscr{R}_{\mathrm{E}}$ and $\mathscr{R}_{\mathrm{S}}$ may be considered fixed, so that the only effective control is $t$, which parameterises the changing form of the water surface. Therefore only fold catastrophes occur stably, and correspond to annihilations and births of pairs of images at particular instants, called "twinkles". The vanishing of the determinant (2.12) at a twinkle means that the water surface has momentarily not merely the right slope to direct light into the eye but also the right curvature to focus it there; in space there is a moving caustic surface which passes through the eye in a twinkle.

The inner surface of an illuminated tea-cup produces a caustic surface

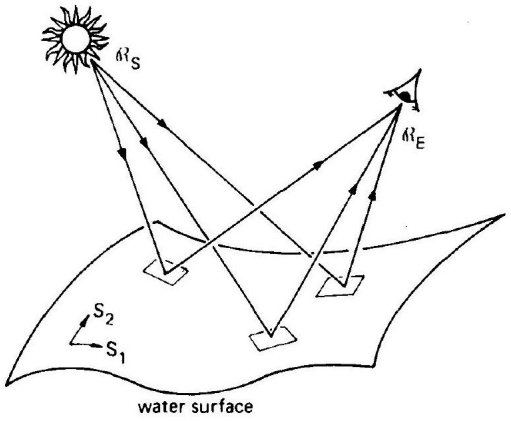



Fig. 2.7. Three sunrays from specular points $s_{1}, s_{2}$ on a water surface. The total length of each ray is proportional to one of the values of the action function $\mathscr{S}$ for the family of rays from the sun to the eye.}

in space, whose intersection with the tea surface is visible by diffuse reflection as a caustic curve containing a cusp point (Fig. 2.5b). This is stable because there are two controls, corresponding to coordinates on the tea, and the stable catastrophes in two dimensions are just fold curves and cusp points.

For a more systematic illustration of the elementary catastrophes we use photographs of caustics produced by reflection from plastic film held in a rigid framework and formed into a curved surface by suction. In the plane of a photograph the only stable caustics are fold curves and cusp points, as in Fig. 2.8. These are sections of fold surfaces and cusped edges (ribs) in three-dimensional space. If a rib is tangent to one of a sequence of plane sections, this gives rise to characteristic lips (Fig. 2.9a) or beak-to-beak events (Fig. 2.9b), depending on the sense in which the rib curves relative to the caustic sheets. These events are not catastrophes of codimension three, because spatial points where they occur are not unique but depend on the orientation of the sequence of plane sections.

However, catastrophes with $K = 3$ do occur stably at isolated points in the space filled by light reflected by the plastic film, and characteristic sections of unfoldings, close to these points, are shown on Figs. 2.10(a-c)

(cf Figs. 2.5c – e ). Altering the shape of the plastic film gives extra control parameters enabling higher catastrophes to be explored. Fig. 2.10d shows an unfolding of the butterfly, and Figs. 2.10 e and f two unfoldings of the parabolic umbilic, produced in this way (cf Figs. 2.5 f and g).

Beautiful catastrophes in focused electron beams, produced by altering the controls of an electron microscope, can be seen in papers by Scheffels, Hahn and Lenz [1953], Liesegang [1953],

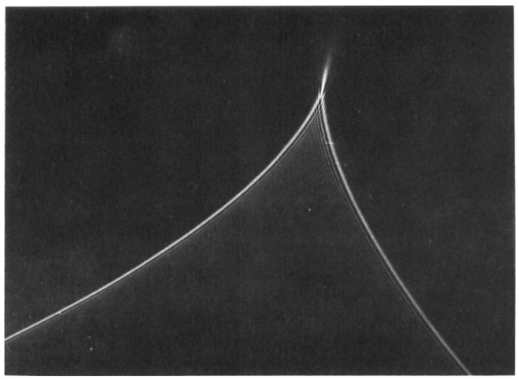


Fig. 2.8. Two fold caustics meeting at a cusp point (courtesy of P. N. Kesterton and D. B. White).

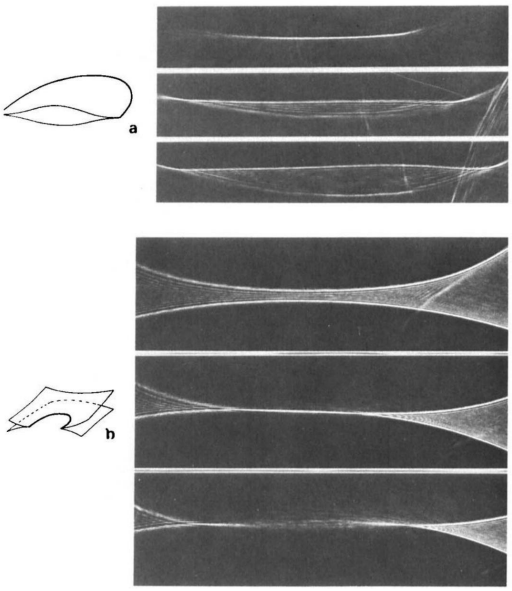



Fig. 2.9. Non-catastrophic events produced when a rib (cusped edge) is tangent to a plane of observation: (a) lips; (b) beak-to-beak (courtesy of P. N. Kesterton and D. B. White).

Bartz, Weissenberg and Wiskott [1956] and Schleich, Hoffmeister, Koops and Lenz [1968] (see also Glaser [1956]).

276

## 31.4 § 3. Diffraction Catastrophes

### 31.4.1 3.1. Asymptotic Solution of the Wave Equation

We shall discuss diffraction in terms of the scalar Helmholtz equation:

$$\nabla^2\psi(\mathscr{R}) + \kappa^2 n^2(\mathscr{R})\psi(\mathscr{R}) = 0, \tag{3.1}$$

where $\psi(\mathscr{R})$ is a complex scalar wave function. Catastrophe optics is concerned with the approach to the short-wave limit $\kappa \to \infty$. If the propagation medium is homogeneous ( $n =$ constant) the use of scalar theory has been rigorously justified (Green and Wolf [1953], Wolf [1959, 1960]). For inhomogeneous media the equations for each component of the electromagnetic field vectors (Born and Wolf [1975] chap. 1) are not of the form (3.1) but contain extra

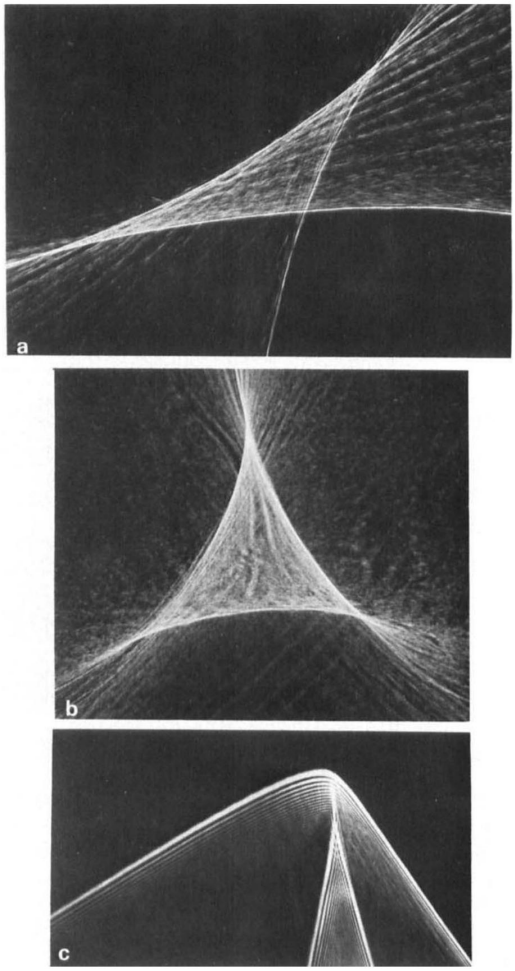



Fig. 2.10. Characteristic sections of caustics close to singular points of (a) swallowtail; (b) elliptic umbilic; (c) hyperbolic umbilic; (d) butterfly; (e) parabolic umbilic (elliptic umbilic piercing fold); (f) parabolic umbilic ("mushroom"). ((a), (b), (d) and (e) courtesy of P. N. Kesterton and D. B. White; (c) and (f) courtesy of S. Q. Hooker and J. C. Rugg.)

terms; in the geometrical optics limit these terms describe polarisation effects (Born and Wolf [1975] chap. 3), which we shall ignore.

To determine the asymptotic behaviour as $\kappa \to \infty$, the standard procedure is to express $\psi$ as

$$\psi(\mathscr{R}) = a(\mathscr{R}) \exp[\mathrm{i}\kappa\chi(\mathscr{R})], \tag{3.2}$$

where the modulus $a$ and phase $\kappa\chi$ are both real. Substitution into (3.1) gives, on taking real and imaginary parts,

$$|\nabla\chi|^2 = n^2 + \nabla^2 a/\kappa^2 a \tag{3.3}$$

and

$$\nabla \cdot \left(a^2 \nabla_\chi\right) = 0 \tag{3.4}$$

The lowest-order approximation is to neglect the term $\nabla^2 a/\kappa^2 a$ in (3.3). Then $\chi$ satisfies the Hamilton-Jacobi equation (2.4), showing that the phase of the wave is $\kappa$ times the action $\mathscr{S}$ of the underlying ray family. To solve equation (3.4) for the wave intensity $a^2$, Gauss's theorem is applied to a thin tube of rays passing through $\Re$ (Fig. 3.1), and gives

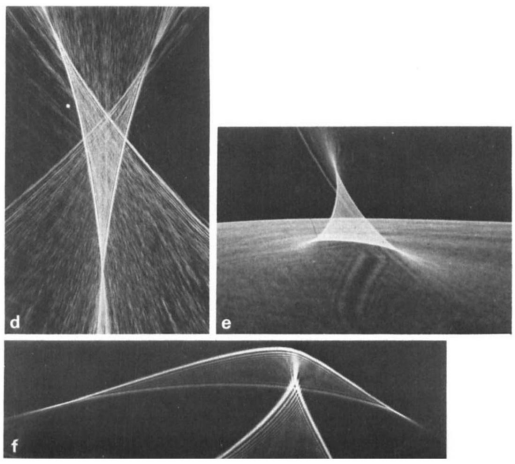



Fig. 2.10 (contd.)

$$a^2(\mathscr{R}) = a^2(\mathscr{R}_0)\,\frac{\mathrm{d}\mathscr{A}(\mathscr{R}_0)\,n(\mathscr{R}_0)}{\mathrm{d}\mathscr{A}(\mathscr{R})n(\mathscr{R})}, \tag{3.5}$$

where $\mathscr{R}_0$ is a point on the ray through $\mathscr{R}$, and $\mathrm{d}\mathscr{A}(\mathscr{R}_0)$ and $\mathrm{d}\mathscr{A}(\mathscr{R})$ are areas of wavefronts cutting the ray tube at $\mathscr{R}_0$ and $\mathscr{R}$. The simplest asymptotic formula for $\psi$ is therefore

$$\psi(\mathscr{R}) \approx a(\mathscr{R}_0)\left[\frac{n(\mathscr{R}_0)\,\mathrm{d}\mathscr{A}(\mathscr{R}_0)}{n(\mathscr{R})\mathrm{d}\mathscr{A}(\mathscr{R})}\right]^{\frac{1}{2}} \exp[\mathrm{i}\kappa\mathscr{S}(\mathscr{R})]. \tag{3.6}$$

For propagation from an initial surface $\Sigma$ on which $\psi$ is given, $\mathscr{R}_0$ should be chosen to lie on $\Sigma$; $\mathscr{S}$ is determined as explained in § 2.1, and $\mathrm{d}\mathscr{A}(\mathscr{R}_0)/\mathrm{d}\mathscr{A}(\mathscr{R})$ can be written as the Jacobian determinant between coordinates on $\Sigma$ and coordinates on the wavefront through $\mathscr{R}$ (Kline and Kay [1965]).

If the ray family has caustics between $\Sigma$ and $\mathscr{R}$, $\mathscr{S}$ is multivalued. In this case $\psi$ can be represented as a sum of interfering terms of the form (3.6), one for each ray through $\mathscr{R}$ :

$$\psi(\mathscr{R}) \approx \sum_{\mu} a_\mu(\mathscr{R}) \exp\left[\mathrm{i}\left(\kappa\mathscr{S}_\mu(\mathscr{R}) - m_\mu(\mathscr{R})\pi/2\right)\right]. \tag{3.7}$$

In this formula $a_\mu$ is the amplitude along the $\mu$ th ray, constructed according to (3.5), and $m_\mu(\mathscr{R})$ is the number of times the $\mu$ th ray has touched a caustic en route from $\Sigma$ to $\mathscr{R}$. The origin of the phase

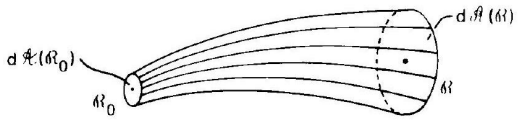



Fig. 3.1. Tube of rays through $\mathscr{R}_0$ and $\mathscr{R}$.

factor $\exp[-im_\mu\pi/2]$ will emerge in §3.2; a simple way of understanding it on the basis of (3.6) is to consider $\mathrm{d}\mathscr{A}(\mathscr{R})$ as having a simple zero when $\mathscr{R}$ lies on a typical (fold catastrophe) caustic, so that $\psi$ gains a factor $\exp[-\mathrm{i}\pi/2]$ each time $\mathscr{R}$ passes through a caustic. The systematic exploitation of the "interfering rays" formula (3.7) is the geometrical theory of diffraction of Keller [1958].

As an asymptotic approximation to $\psi$, (3.7) has a serious deficiency: it fails completely on caustics. On a caustic, the ray tube area $\mathrm{d}\mathscr{A}$ shrinks to zero, so (3.6) predicts that $\psi$ has a singularity of infinite intensity. This cannot represent a solution of (3.1), which must be a smooth function of $\mathscr{R}$ for all finite $\kappa$, however large. The error in deriving (3.7) was to neglect the term $\nabla^2 a/\kappa^2$ in (3.3), which in this approximation diverges on caustics no matter how large $\kappa$ is. This divergence is indeed a catastrophe, and justifies using the term in the context of short-wave optics.

The fact that (3.7) fails at the most important places, where the light is brightest, suggests that the remedy lies not in supplementing (3.7) by correction terms in falling powers of $\kappa$ (see e.g. Ahluwalia, Lewis and Boersma [1968]), but in replacing it by an expression whose large- $\kappa$ analytic properties near caustics are completely different. Indeed we shall see that the diffraction near a caustic depends on its catastrophe type.

## 31.4.2 3.2. Integral Representations for Short Waves

To obtain asymptotic representations that do not diverge on caustics, the principle of superposition is invoked, to express $\psi$ as an integral over a collection of contributions of the form (3.6). Each member of the collection is labelled by the values of variables $s = (s_1, s_2 \cdots)$, and has action function, denoted by $\phi(s;\mathscr{R})$, chosen to satisfy two conditions. First, $\phi$ is a single-valued function of $\mathscr{R}$ for fixed $s$. Second, the values $\mathscr{S}_\mu(\mathscr{R})$ of the actions at $\mathscr{R}$ for the ray family under study are the extreme values of $\phi(s;\mathscr{R})$ as $s$ is varied. Therefore $\phi$ is precisely the generating function of catastrophe theory introduced in § 2.1, and the rays are given by the gradient map (2.9). The integral representation for $\psi$ is then

$$\psi(\mathscr{R}) = \mathrm{e}^{-\mathrm{i}N\pi/4} \left(\frac{\kappa}{2\pi}\right)^{N/2} \int \cdots \int \mathrm{d}^{\mathrm{N}} s b(s;\mathscr{R}) \exp[\mathrm{i}\kappa\phi(s;\mathscr{R})] \tag{3.8}$$

where $N$ is the number of state variables $s$, and the prefactors have been chosen for later convenience.

Before discussing explicit realisations of this procedure, and the choice of weighting function $b$, we establish the connection between (3.8) and the simple ray approximation (3.7). This relies on the principle of stationary phase, according to which the main contributions to (3.8) for given $\mathscr{R}$ come from the stationary points, that is from those values $s^\mu(\mathscr{R})$ of the integration variables for which the gradient map (2.9) holds. The principle can be applied because we are studying the limit $\kappa \to \infty$ for which the integrand is a rapidly oscillating function of $s$, so that other than near stationary points, destructive interference occurs. If $\mathscr{R}$ is not near a caustic the stationary points are well separated and the method of stationary phase can be used in its simplest form (see e.g. appendix III of Born and Wolf [1975]). This gives a series of separate contributions with phases $\kappa\mathscr{S}_\mu$, precisely as in (3.7). Near a caustic, two or more of the stationary points lie close together and their contributions cannot be separated. The nature of this failure of simple stationary phase, and the resulting description of diffraction near caustics in terms of catastrophes will be discussed in § 3.3.

First, however, we point out that because there are many ways of choosing the function $\phi$, as described in § 2.1, there are, correspondingly, many integral representations (3.8) for $\psi$; all are equivalent as $\kappa \to \infty$. We discuss in detail the diffraction integral resulting from the choice $\phi\,(s_1, s_2; \mathscr{R})$ described in § 2.1; $\phi$ is the optical path along the ray from $\mathscr{R}_0$ on the initial surface $\Sigma$ to $\mathscr{R}$. For convenience we now take $\Sigma$ to be a wavefront. Then (3.8) has the explicit form

$$\psi(\mathscr{R}) = \frac{-\mathrm{i}\kappa}{2\pi} \int \mathrm{d}s_1 \int \mathrm{d}s_2 a\left(\mathscr{R}_0(s)\right) \left[\frac{n\left(\mathscr{R}_0(s)\right)}{n(\mathscr{R})} \det \frac{\partial^2 \phi(s;\mathscr{R})}{\partial s_i \partial X_j}\right]^{\frac{1}{2}} \times \exp[\mathrm{i}\kappa\phi(s;\mathscr{R})] \tag{3.9}$$

where the variables $X_j$ in the $2\times2$ matrix of mixed second derivatives of $\phi$ denote coordinates in the wavefront through $\mathscr{R}$ of the family of rays emanating from $\mathscr{R}_0$.

To see that (3.9) is correct, consider the case where $\mathscr{R}$ is not near a caustic and apply the method of stationary phase. The contributions $s^\mu$ correspond to the points $\mathscr{R}_0$ from which rays through $\mathscr{R}$ originate on $\Sigma$. Standard formulae (e.g. appendix III of Born and Wolf [1975]) give

$$\psi(\mathscr{R}) \approx \sum_\mu a\left(\mathscr{R}_0\left(s^\mu\right)\right) \left[\frac{n\left(\mathscr{R}_0\left(s^\mu\right)\right) \det \partial^2\phi/\partial s_i \partial X_\mathrm{j}}{n(\mathscr{R}) \det \partial^2\phi/\partial s_\mathrm{i}\partial s_\mathrm{j}}\right]^{\frac{1}{2}} \times \exp\left[\mathrm{i}\left(\kappa \mathscr{S}_\mu(\mathscr{R}) - m_\mu(\mathscr{R})\pi/2\right)\right], \tag{3.10}$$

where $m_\mu$ is the number of negative eigenvalues of the matrix $\partial^2\phi/\partial s_i \partial s_j$. The ratio of determinants is

$$\frac{\det \partial^2\phi/\partial s_i \partial X_i}{\det \partial^2\phi/\partial s_i \partial s_j} = \det \frac{\mathrm{d}s_i}{\mathrm{d}X_i} = \frac{\mathrm{d}\mathscr{A}\left(\mathscr{R}_0\right)}{\mathrm{d}\mathscr{A}(\mathscr{R})}, \tag{3.11}$$

which is precisely the Jacobian between ray tube area elements in (3.5). Prior to encountering a caustic the path length $\phi$ has a minimum, so both eigenvalues of $\partial^2\phi/\partial s_i \partial s_j$ are positive, and $m = 0$ for the (unique) path through $\mathscr{R}$. On encountering a caustic, $\mathrm{d}\,\mathscr{A}(\mathscr{R})$ passes through zero and $\phi$ has a saddle in $s_1, s_2$, so that one eigenvalue has changed sign and $m = 1$. Therefore the integer $m_\mu(\mathscr{R})$ in (3.10) is the same as that in (3.7). This completes the justification of the integral representation (3.9) by identifying its off-caustic approximation (3.10) with the simple asymptotic formulae (3.6) and (3.7).

The diffraction integral (3.9) is of Kirchhoff type, giving $\psi(\mathscr{R})$ as a superposition of wavelets from points $\mathscr{R}_0$ on $\Sigma$. Each contribution has the simple form (3.6) appropriate to the family of rays issuing from $\mathscr{R}_0$. To see this it is necessary to understand the determinant in (3.9). Let $\hat{\boldsymbol{k}}_0$ (cf (2.2)) be the unit direction vector at $\mathscr{R}_0$ of the ray from $\mathscr{R}_0$ to $\mathscr{R}$. From (2.3), (2.4) and (2.6),

$$\nabla_{\mathscr{R}_0}\phi = -n\left(\mathscr{R}_0\right)\hat{\boldsymbol{k}}_0. \tag{3.12}$$

Then the derivatives $\partial\phi/\partial s_i$ in (3.9) give the projections of $\hat{\boldsymbol{k}}_\mathrm{o}$ onto $\Sigma$. If $\theta$ is the obliquity angle between $\hat{\boldsymbol{k}}_0$ and the normal to $\Sigma$ at $\mathscr{R}_0$, the determinant in (3.9) is the Jacobian

$$\det \frac{\partial^2\phi}{\partial s_i \partial X_j} = n\left(\mathscr{R}_\mathrm{o}\right)\cos\theta \frac{\mathrm{d}\hat{\boldsymbol{k}}_\mathrm{o}}{\mathrm{d}\mathscr{A}(\mathscr{R})}, \tag{3.13}$$

showing that the amplitude is indeed of the form (3.5). Moreover, (3.13) is finite and positive under the conditions discussed in § 2.1, namely that the rays from $\mathscr{R}_0$ have no caustics en route to $\mathscr{R}$. Therefore (3.9) gives a smooth representation of $\psi$, even when the family of rays from $\Sigma$ has caustics, near which the stationary-phase approximation (3.10) breaks down.

In a uniform medium the rays are straight, $\phi$ is simply the distance from $\mathscr{R}_0$ to $\mathscr{R}$, and the Jacobian $\mathrm{d}\hat{\boldsymbol{k}}_0/\mathrm{d}\mathscr{A}(\mathscr{R})$ is the inverse square of this distance. Therefore (3.9) becomes the

familiar Kirchhoff integral (chapter 8 of Born and Wolf [1975]), apart from a slight modification to the obliquity factor, which is of no consequence here since as $\kappa \to \infty$ all contributions to $\psi$ originate near rays on $\Sigma$, for which $\theta = 0$.

We chose to describe (3.9) in detail because it is the simplest and most useful short-wave asymptotic integral representation. However, two others are frequently discussed. The first is the path integral of Feynman [1948] (for simplified accounts see Feynman and Hibbs [1965] and Berry and Mount [1972]). This is based on paths from $\Sigma$ to $\mathscr{R}$ with infinitely many segments as discussed in § 2.1 (cf Fig. 2.3). The optical path $\phi(s;\mathscr{R})$ has infinitely many variables, so that (3.8) is an infinite-dimensional "functional integral", presenting great difficulties of precise definition and evaluation, especially near caustics (Schulman [1975], Levit and Smilansky [1977], Albeverio and Höegh-Krohn [1977], Dangelmayr and Vert [1979]). The second alternative to (3.9) is due to Maslov [1965] (see also Kravtsov [1968]), and is specifically designed to apply close to caustics. The wave (3.8) is represented as a Fourier transform where the integration variables $s$ are components of wave vectors $\boldsymbol{k}$ perpendicular to the caustic being studied. Maslov's method is very ingenious, but suffers from the disadvantage of often failing when there are no caustics at all, as in the important case of a uniform medium where $\Sigma$ is a plane and $\psi$ is simply a plane wave. To our knowledge neither Feynman's nor Maslov's method has ever been employed in wave theory to derive a result not previously obtained otherwise.

### 31.4.3 3.3. The Hierarchy of Catastrophe Diffraction Integrals

Having obtained the integral representations (3.8) for $\psi$, the remaining problem is to evaluate them for large $\kappa$ in the important regions, near caustics, where the simple method of stationary phase (leading to (3.7)) breaks down. The reason for the failure lies in approximating the generating function $\phi$ near each stationary point $s^\mu$ by a constant plus quadratic terms, followed by exact evaluation of the resulting complex Gaussian integrals - in other words, the ray contributions to $\psi$ are considered separately. But when stationary points approach closer than a distance of
order $\kappa^{-\frac{1}{2}}$ the contributions are not separated by a region where destructive interference occurs, and when the stationary points coalesce the action extremum is degenerate-i.e. $\phi$ is stationary to higher than first order, and the quadratic terms in $s - s^\mu$ vanish, as well as the linear terms.

The diagnosis suggests the remedy: replace $\phi$ in (3.8) by a simpler function with the same stationary-point structure, and evaluate the resulting integral (which is more complicated than a Gaussian) exactly. If the caustic is structurally stable it must be equivalent to one of the catastrophes, in the sense explained in § 2.2, so $\phi$ can be replaced by one of the normal forms $\Phi$ which are listed in Table 1 for codimension $\leq 4$. Catastrophe theory guarantees that in doing this we are not simply analyzing one of infinitely many possible diffraction integrals, but a generic integral which will occur in many different contexts; this is what makes the effort of computation (and experiment) worthwhile.

Near the caustic, all stationary points lie close together, and the factor $b(s;\mathscr{R})$ in (3.9) can be evaluated at any one of them and taken outside the integral sign. The factor $\kappa$ in the exponent can be removed by a scaling to be discussed in §3.4. The result of this procedure is a series of standard diffraction integrals $\Psi$, one for each catastrophe, of the form

$$\Psi(C) = \frac{1}{(2\pi)^{N/2}} \int \cdots \int \mathrm{d}^N s \exp[\mathrm{i}\Phi(s;C)]. \tag{3.14}$$

Following Trinkaus and Drepper [1977], we call these integrals diffraction catastrophes.

As in § 2.2 we have denoted the catastrophe polynomials by $\Phi(s;C)$, where $s$ represents the state variables and $C$ the control parameters (which include the coordinates of the field point $\mathscr{R}$ ). The number of state variables, $N$, is the corank of the catastrophe. It is the minimum

number of state variables necessary for $\Phi$ to reproduce the stationary-point structure of $\phi$. Extra state variables (such as the infinitely many occurring in Feynman's path integrals) can be transformed by the splitting lemma of catastrophe theory (Poston and Stewart [1978]) so as to appear as isolated quadratic terms in $\Phi$, which can trivially be integrated out of (3.14). Recall from § 2.1 that in optics $N$ is at most 2.

Comparison of (2.9) and (3.14) shows that the crucial difference between the ray and wave versions of catastrophe optics lies in the way the state variables are eliminated. In ray theory, the $s$ are removed by differentiation via the gradient map (2.9). In wave theory, the $s$ are removed by integration via the diffraction functions (3.14).

The diffraction catastrophes $\Psi(C)$ give transitional approximations,
valid as $\kappa \to \infty$ and close to the caustic. Far from the caustic, the deformation of the generating function $\phi$ away from the standard polynomial $\Phi$, and the variation of the factor $b$ in (3.8), results in a loss of accuracy; in particular, the stationary-phase approximations of (3.8) and (3.14) do not coincide. To overcome this deficiency, uniform approximations, valid near to or far from the caustic, can be obtained by deforming the standard integrals. Rules for carrying out these deformations are given by Berry [1976]; more mathematically elaborate treatments (including correction terms in descending powers of $\kappa$ ) are given by Duistermati [1974] and Guillemin and Sternberg [1977]. Uniform approximations to integrals were first derived by Chester, Friedman and Ursell [1957], and their results applied to wave theory by Kravtsov [1964a,b] and Berry [1966]. In addition, Ludwig [1966] gave explicit formulae for the whole hierarchy of what we now call cuspoid diffraction catastrophes. It should be pointed out that the technique of uniform approximation is more general than catastrophe optics, since it can be applied to structurally unstable caustics, such as the glory (Nussenzveig [1969a,b], Berry [1969a,b, 1976]), whose codimension, as defined in § 2.2, is infinite (see also the end of § 3.4).

The diffraction catastrophes (3.14) constitute standard functions that stably (i.e. up to deformation) represent wave patterns near caustics. Apart from that corresponding to the fold, these functions $\Psi(C)$ cannot be expressed in closed form by means of the special functions of analysis (as tabulated in, say, Abramowitz and Stegun [1964]). The first few members of the hierarchy have been studied in detail, and we now discuss these.

**(i) The fold diffraction catastrophe**

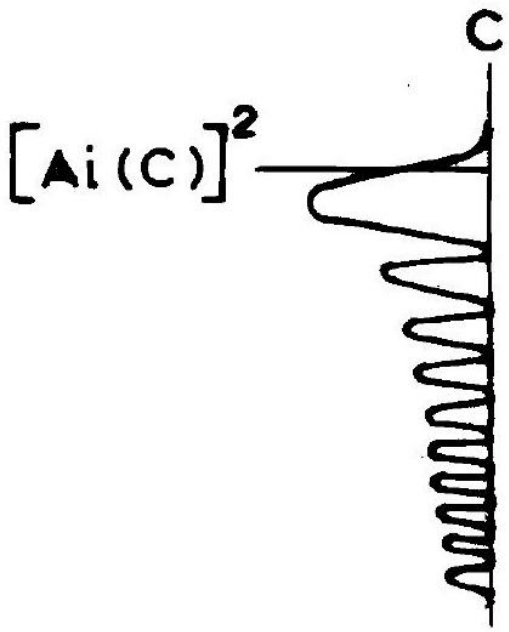
From eq. (3.14) and Table 1, this has the formula

$$\Psi(C) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\left[\mathrm{i}\left(s^3/3 + Cs\right)\right] = \sqrt{(2\pi)}\mathrm{Ai}(C) \tag{3.15}$$

where Ai denotes the Airy function (Abramowitz and Stegun [1964]), a real function of one variable C. A graph of $\mathrm{Ai}^2$ is shown in Fig. 3.2a. For $C < 0$ there are two rays (stationary points of the integrand) whose interference gives rise to oscillations in $\Psi(C)$. For $C > 0$ there is one contributing complex ray (that is, a complex solution of the mapping (2.9)), and $\Psi(C)$ falls exponentially to zero without oscillation. Fig. 3.2b

shows observed diffraction near a fold caustic. Airy [1838] introduced his function to study diffraction along the asymptote of a caustic, and in particular to establish the law of distribution of supernumerary rainbows (see Nussenzveig [1977] for a clear modern account). Larmor [1891] made a more detailed study of diffraction at a caustic surface, and obtained power laws for the variation of fringe spacing with the curvature of the caustic and with wavelength (the fringe index to be discussed in §3.4 ); he was aware of, but did not study, diffraction at a cusp - see (ii) below.

The Airy function (3.15) is of course very different from the function of the same name describing diffraction in the focal plane of a circular patch of spherical wavefront (Born and




a

Wolf [1975] p. 396).

**(ii) The cusp diffraction catastrophe**

This has the formula

$$\Psi(C_1, C_2) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\left[\mathrm{i}\left(s^4/4 + C_2 s^2/2 + C_1 s\right)\right], \tag{3.16}$$

and is a complex function of two variables. A map of $|\Psi|$ in the $C_1, C_2$ plane is shown in Fig. 3.3a, and should be compared with observed diffraction near the cusp of a caustic, shown in Fig. 3.3b. $\Psi(C_1, C_2)$ was first studied by Pearcey [1946], who produced contour plots of $|\Psi|$ and the phase of $\Psi$ in what was then a tour de force of computing. More recently, diffraction at a cusped caustic has been considered in radiowave propagation by Budden [1976] and Maslin [1976a,b], in the context of quantum scattering by Berry [1975], Connor [1976] and Trinkaus and Drepper [1977], and in underwater acoustics by Holford [1972].

The most delicate features of this rich wavefield are the wavefront dislocations (see Appendix 1 and Nye and Berry [1974]), which are the points where $|\Psi|$ vanishes and the phase of $\Psi$ has a singularity. They are formed by destructive interference between the contributing rays, and their positions are given very accurately by the method of stationary phase applied to (3.16) (Berry, Nye and Wright [1979] appendix C). Within the cusp there are three real rays, whose interference produces pairs of dislocation points arranged on a distorted hexagonal lattice. Outside the cusp there is one real ray, and, in certain regions, one complex ray decaying exponentially away from the caustic (Wright [1980]); close to the caustic the complex ray is sufficiently intense for its interference with the real ray to produce a single row of dislocations. The production of dislocations in this way (which also occurs for diffraction catastrophes of higher codimension) is the only circumstance known to us where complex rays affect the topology of a wavefield. Dislocations persist under perturbation, showing that the structural stability of diffraction catastrophes is a very strong property, extending to the finest scales.

Pearcey's function (3.16) should be distinguished from two superficially similar patterns, namely diffraction near the rotationally symmetric cusped cone produced by primary

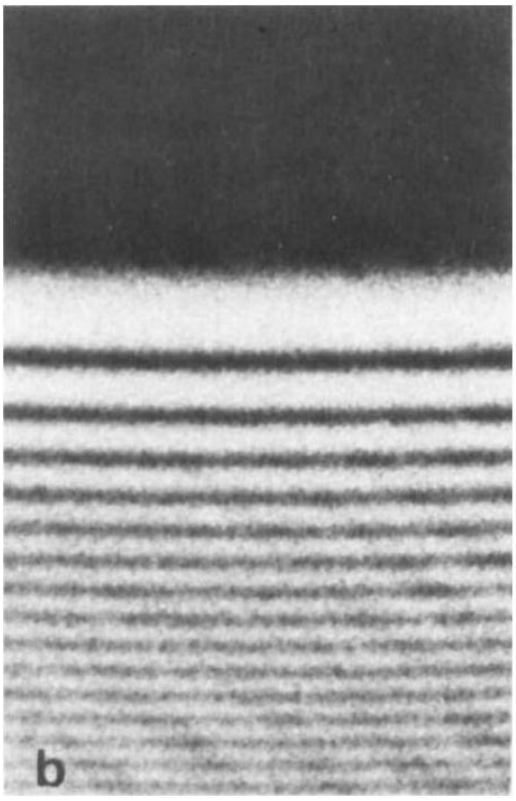



Fig. 3.2. (a) Graph of the fold diffraction catastrophe intensity (square of Airy function); (b) observation of diffraction intensity at a fold caustic.

spherical aberration (see Appendix 2), which is structurally unstable in the sense explained in § 2.2 and so not a diffraction catastrophe, and diffraction associated with coma (see (iii) below, and Appendix 2).

**(iii) The elliptic and hyperbolic umbilic diffraction catastrophes**

These are defined by the double integrals

$$\Psi(C_1, C_2, C_3) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}s_1 \int_{-\infty}^{\infty} \mathrm{d}s_2 \times \exp\left[\mathrm{i}\left(s_1^3 - 3s_1 s_2^2 - C_3\left(s_1^2 + s_2^2\right) - C_2 s_2 - C_1 s_1\right)\right] \tag{3.17}$$

for the elliptic umbilic and

$$\Psi(C_1, C_2, C_3) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}s_1 \int_{-\infty}^{\infty} \mathrm{d}s_2 \exp\left[\mathrm{i}\left(s_1^3 + s_2^3 - C_3 s_1 s_2 - C_2 s_2 - C_1 s_1\right)\right] \tag{3.18}$$

for the hyperbolic umbilic. They are complex functions of three variables,

realisable as diffraction patterns in space. Both have a mirror plane $C_3 = 0$. A brief analysis of the singular ( $C_3 = 0$ ) sections was given by Trinkaus and Drepper [1977], and Connor [1976] commented on the integrals in the context of molecular scattering. A complete theoretical and experimental study of the elliptic umbilic diffraction catastrophe has been made by Berry, Nye and Wright [1979] and a similar one is in preparation for the hyperbolic umbilic.

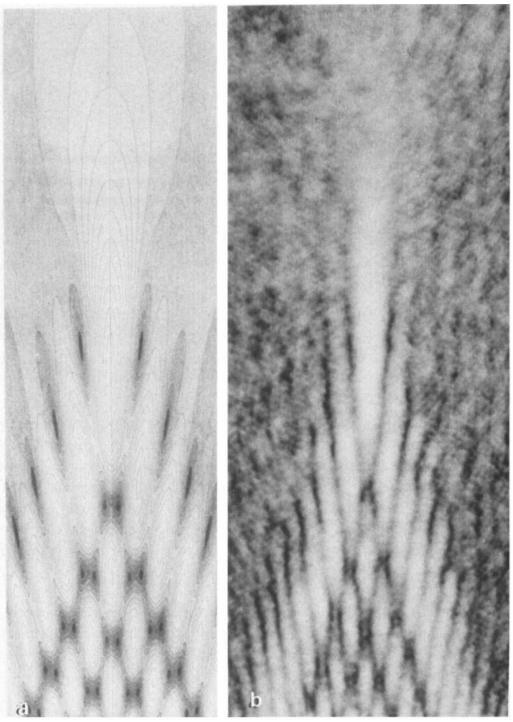


Fig. 3.3. (a) Map of the modulus of the cusp diffraction catastrophe integral (Pearcey function) (courtesy of F. J. Wright); (b) observation of diffraction intensity at a cusped caustic (courtesy of J. F. Nye).

For the elliptic umbilic (eq. (3.17)), maps of $|\Psi|$ in the $C_1, C_2$ plane, for four values of $C_3$, are shown in Figs. 3.4 (a-d). These should be compared with the observed diffraction in Figs. 3.4 (e-h), obtained by focusing a microscope on various planes in the space above a water droplet lens with a triangular perimeter, through which a parallel beam of light has been refracted (see § 5). Note the remarkable agreement between theory and experiment, down to the finest details.

In the plane of the focus of the elliptic umbilic, $C_3 = 0$ (Figs. 3.4a and e), the diffraction pattern has a bright central region, roughly triangular, surrounded by fringes receding to asymptotes $2\pi/3$ apart. Surprisingly, these do not correspond to the ribs, which lie between the asymptotes. As $C_3$ is increased, the diffraction pattern unfolds in the space bounded by the caustic surface. Formed by the interference of four waves, the pattern consists of intensity maxima and minima resembling hexagonal prisms, stacked in space like the atoms of a distorted crystal with space group $R\bar{3}m$ (Figs. 3.4b, c, f and g). The fringes parallel to the arcs of the caustic arise from the fold diffraction catastrophe: the superposition of the three sets of fringes is responsible for the array of hexagons inside the caustic. Outside the caustic, the interference of two rays produces the fringes which dominated the focal section, but they are faint compared with the fringes inside the caustic, and as $C_3$ is further increased the cusps dominate the wavefield (Figs. 3.4 d and h ).

The dislocation lines can be considered as a skeleton on which the whole pattern is built. Within the caustic, not too near the ribs, they form hexagonally puckered rings. As $C_3$ is increased and the ribs approached, the rings tilt towards one another and eventually join to form "hairpins", whose arms are tightly wound helices which pierce the planes $C_3 =$ constant in the pairs of dislocation points associated with the cusp diffraction pattern. In the focal plane $C_3 = 0, \Psi$ is real and there is an array of dislocation lines formed by the interference between the two real rays; these are the black lines in Fig. 3.4a: the associated regions of low

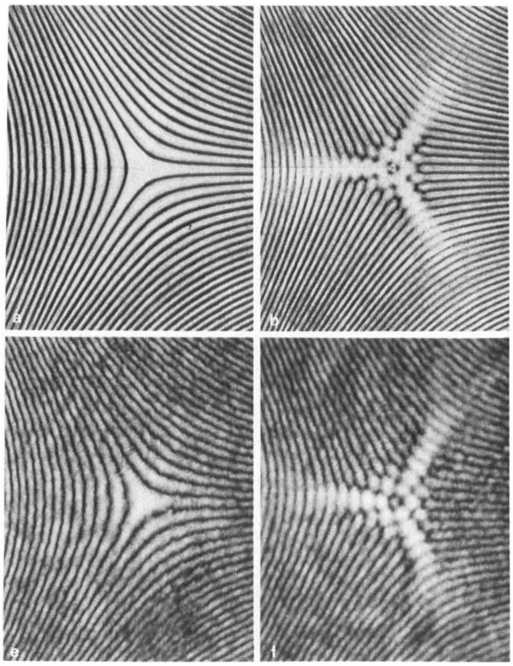


Fig. 3.4. Elliptic umbilic diffraction catastrophe. (a-d): Maps of $|\Psi|$ in the $C_{1}, C_{2}$ plane, for increasing $C_3$ (courtesy of F . J. Wright); (e-h): corresponding observed diffraction intensity (courtesy of J. F. Nye). (a) and (e) show the focal (singular) section $C_3 = 0$.}

amplitude are greatly extended in the $C_3$ direction. Outside the caustic, far from the focus, a single complex ray interferes with the two real rays. This results in an infinite sequence of helices lying close to the caustic surfaces, roughly parallel to the ribs, which do not cross the focal plane but join in opposite handed pairs with their equivalents originating from the next rib, like a pair of curly antelope horns.

For the hyperbolic umbilic (eq. (3.18)) we show maps of $|\Psi|$ and comparable observations, for the singular section $C_3 = 0$, showing the

characteristic "parallelograms", and a typical section $C_3 > 0$ (Figs. 3.5 a − d ). The singular section is familiar in classical aberration theory as coma (Born and Wolf [1975] pp. 475-479; see also Appendix 2). Note that this is non-generic in that a perturbation will yield a nonsingular section: structural stability is a property of the full threedimensional pattern.

### (iv) The swallowtail diffraction catastrophe

This is defined by

$$\Psi\left(C_1, C_2, C_3\right) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\left[\mathrm{i}\left(s^5/5 + C_3 s^3/3 + C_2 s^2 + C_1 s\right)\right]. \tag{3.19}$$

Many details are known (Wright [1977]), and a complete study is in preparation. Maps of $|\Psi|$ for the singular section $C_3 = 0$ and typical sections $C_3 < 0$ and $C_3 > 0$ are shown in Figs. 3.6 (a-c), together with observations for comparison (Figs. 3.6 (d-f)). Note that, unlike the other codimension 3 diffraction catastrophes, $C_3 = 0$ is not a mirror plane.

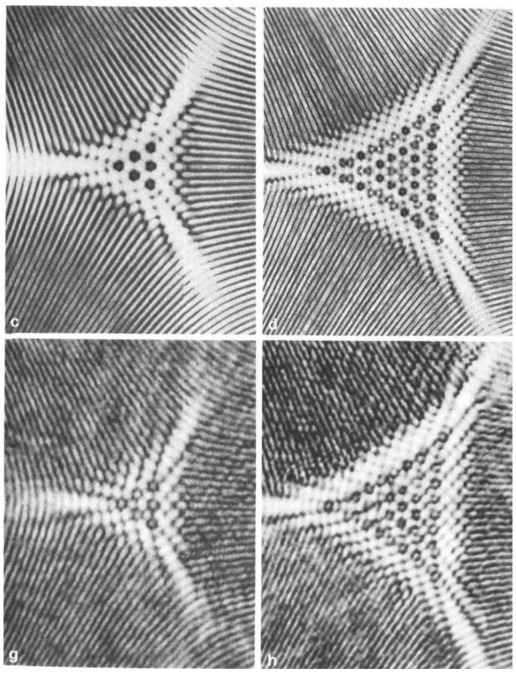


Fig. 3.4. (contd.)

**(v) Diffraction catastrophes whose codimension exceeds three**

It is clear that each diffraction catastrophe will have its characteristic architecture, whose complexity will increase very rapidly with codimension $K$. It seems unlikely that full analyses of diffraction catastrophes with $K > 3$ will be worth undertaking; the parabolic umbilic ( $K = 4$ ), for example, is a family of three-dimensional diffraction patterns. To identify catastrophes observed experimentally, however, it is valuable to characterise at least the possible two-dimensional singular sections for diffraction catastrophes with $K > 3$. Studies of caustics in light refracted by water droplet "lenses" (§5) have yielded the characteristic "anvil" and hexagons of the parabolic umbilic (Fig. 3.7), and the straight fringes of $E_6$ (Fig. 3.8) for which $K = 5$ (Nye [1979]).

## 31.4.4 3.4. Diffraction Scaling Laws

As $\kappa \to \infty$, the intensity at a caustic diverges and the scale of the associated diffraction fringes shrinks to zero. It is our purpose in this section to describe the laws governing these phenomena. The results are a series of exponents for each catastrophe. These relate the standard diffraction catastrophes $\Psi$ (eq. (3.14)) to the physical wave functions $\psi$ (eq. (3.8)). Now $\Psi$ involves standard control parameters $C$, while $\psi$ involves physical controls, independent of $\kappa$, that we call $C'$ (and which could represent, for example, coordinates of position $\mathscr{R}$ in space, the time of observation, or orientations of refracting objects). The relation between $\Psi(C)$ and $\psi(C')$ is the following scaling law, written explicitly in terms of the separate control parameters $C'_i (1 \leq j \leq K)$ :

$$\psi(C') = \kappa^{\beta} \Psi\left(\kappa^{\sigma_1} C'_j\right) \tag{3.20}$$

$\beta$ is a measure of the divergence of $|\psi|$, as $\kappa \to \infty$, at the most singular control point $C = 0$. It was introduced by Arnold [1975], who called it

the singularity index, and studied in detail by Varchenko [1976]. $\sigma_j$ are measures of the fringe spacings in the different control directions $C_j$, as $\kappa \to \infty$. These indices were introduced

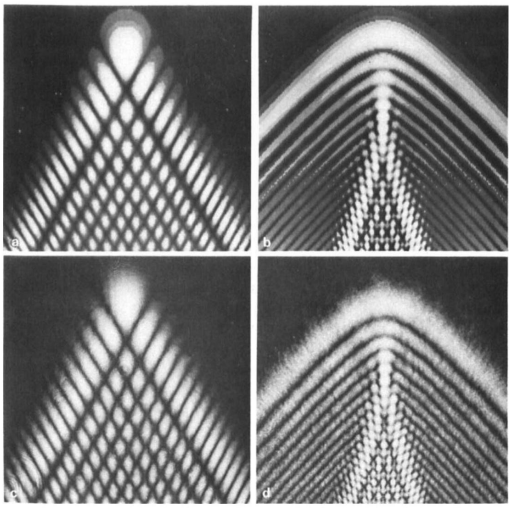


Fig. 3.5. Hyperbolic umbilic diffraction catastrophe. (a) and (b): maps of $|\Psi|$ in the $C_{1}, C_{2}$ plane for (a) the focal section $C_3 = 0$ and (b) a typical section near the focus (courtesy of *F.* J. Wright); (c) and (d): corresponding observed diffraction intensity (courtesy of J. F. Nye).}

by Berry [1977]. The sum

$$\gamma \equiv \sum_{i=1}^{K} \sigma_i \tag{3.21}$$

governs the $K$-dimensional hypervolume of the main diffraction maximum (in control space); this hypervolume scales as $\kappa^{-\gamma}$. We call $\gamma$ the
fringe index; it has been studied by Romero [1978] who shows that for each catastrophe it is invariant under diffeomorphism.

The derivation of (3.20) and computation of $\beta, \sigma_i$ and $\gamma$ are fully described by Berry [1977]. Here we give an outline and work through the simplest example in detail. Our starting point is the diffraction integral (3.8) for the wave $\psi(C')$. This involves $\kappa$ in two ways: as a factor in the exponent, multiplying the generating function $\phi$, which has dimensions of length, and as a prefactor $\kappa^{N/2}$, ensuring that far from the caustic, where the integral can be approximated by $N$-dimensional stationary phase, each ray contribution to the sum (3.7) has an amplitude independent of $\kappa$, as the geometrical-optics limit requires. Near the caustic, we can deform $\phi$ into the normal form $\Phi$ and then, denoting the physical state variables in (3.8) by $s'$, the integral becomes

$$\psi(C') = \left(\frac{\kappa}{2\pi}\right)^{N/2} \int \cdots \int \mathrm{d}^N s' \exp\left[\mathrm{i}\kappa\Phi(s'; C')\right] \tag{3.22}$$

To carry out the scaling, and hence show that (3.20) is indeed the relation between the $\kappa$-dependent (3.22) and the $\kappa$-independent (3.14), the crucial observation is that $\Phi$ consists of a germ depending on $s$ but not $C$, together with unfolding terms linear in $C$. Scaling is a two-stage process. Firstly, the $\kappa$-dependence of the germ (in (3.22)) is removed by a suitable linear $\kappa$-dependent transformation from $s'$ to $s$, giving a factor $\kappa^{\beta}$ outside the integral. Secondly, the $\kappa$-dependence of the unfolding terms is removed by suitable linear $\kappa$-dependent transformations $C'_j = \kappa^{-\sigma_i} C_j$. This procedure leads directly to the relation (3.20). Table 2 shows the resulting exponents $\beta, \sigma_j$ and $\gamma$ for the catastrophes with $K \leq 4$.

We illustrate this procedure with the fold diffraction catastrophe. In this case the integral (3.22) is

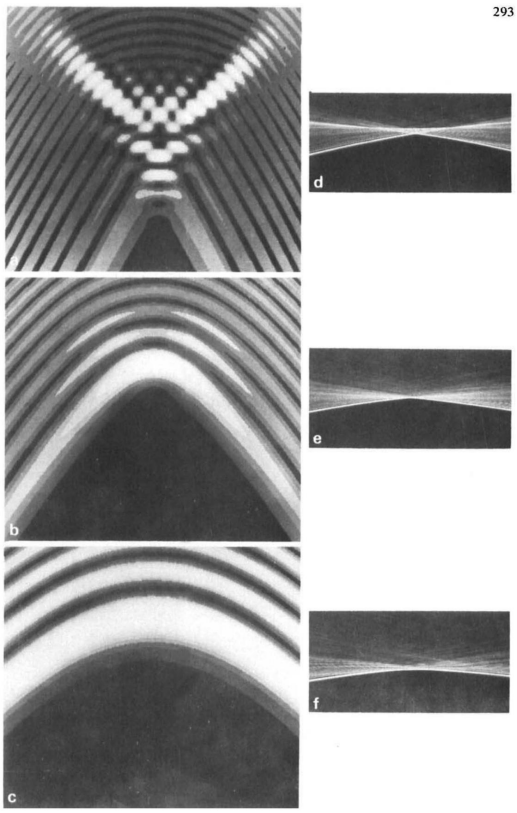



Fig. 3.6. Swallowtail diffraction catastrophe (a-c): maps of $|\Psi|$ in the $C_{1}, C_{2}$ plane for (a) a typical section $C_3 < 0$, (b) the singular section $C_3 = 0$ and (c) a typical section $C_3 > 0$ (courtesy of F. J. Wright); (d-f) observed diffraction intensity (courtesy of P. N. Kesterton and D. B. White)}

$$\psi(C') = \left(\frac{\kappa}{2\pi}\right)^{\frac{1}{2}} \int_{-\infty}^{\infty} \mathrm{d}s' \exp\left[\mathrm{i}\kappa\left(s'^3/3 + C's'\right)\right] \tag{3.23}$$

where $s'$ is proportional to distance along an initial wavefront and $C'$ is proportional to distance from the fold caustic. The scaling $\kappa s'^3 = s^3$ removes $\kappa$ from the germ, and gives

$$\psi(C') = \frac{\kappa'}{\sqrt{(2\pi)}} \int_{-\infty}^{\infty} \mathrm{d}s \exp\left[\mathrm{i}\left(s^3/3 + \kappa^{\frac{2}{3}} C's\right)\right] \tag{3.24}$$

which is precisely of the form (3.20) with $\beta = \frac{1}{6}$ and $\sigma_1 = \gamma = \frac{2}{3}$. Near a fold caustic, therefore, the light intensity $|\psi|^2$ rises to $\mathscr{O}\left(\kappa^{\frac{1}{3}}\right)$ (Airy [1838]) and the fringe spacing is $\mathcal{O}\left(\kappa^{-\frac{2}{3}}\right)$ (Larmor [1891]).

The singularity index $\beta$ increases with codimension $K$ (and faster for umbilics than cuspoids). This is reasonable, since the larger $K$ is, the greater is the number of rays coalescing to touch each point of the caustic. For waves in three-dimensional space, the greatest possible value of $\beta$ is unity and this is attained at the (structurally unstable) perfect focus of a patch of spherical wavefront.

The fringe index $\gamma$ also increases with $K$, showing that the intense region near the singularity, where all rays interfere constructively, shrinks very rapidly as $\kappa \to \infty$ for the higher catastrophes. This shrinking is anisotropic in control space, because for each catastrophe the exponents $\sigma_i$ are not all the same. Thus for the cusp the fringe spacing along the symmetry axis is $\mathscr{O}\left(\kappa^{-\frac{1}{2}}\right)$ while that across the cusp is $\mathscr{O}\left(\kappa^{-\frac{3}{4}}\right)$, explaining why the diffraction patterns

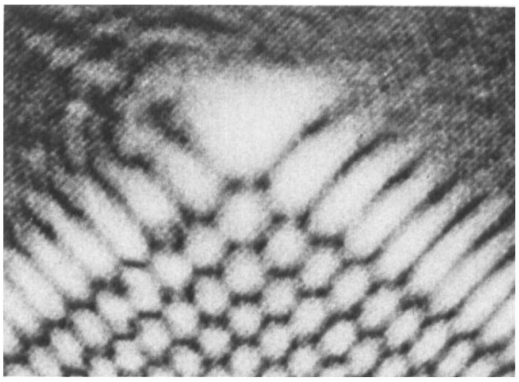


Fig. 3.7. Observed diffraction intensity for the singular section of $D_{5 $ (the parabolic umbilic) (courtesy of J. F. Nye).}

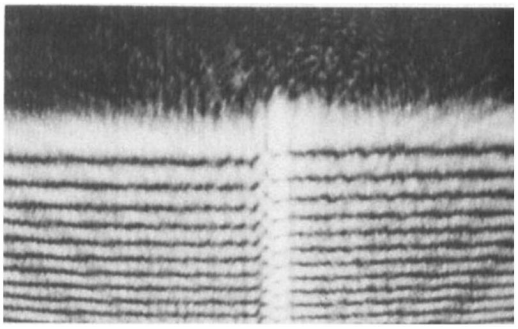


Fig. 3.8. Observed diffraction intensity for the singular section of $E_{6 $ (courtesy of **J**. **F**. Nye).}

near cusps formed by objects large in comparison with the wavelength of light are greatly elongated in the cusp direction. Similar arguments explain the stretching of the elliptic umbilic diffraction pattern along its symmetry axis as $\kappa \to \infty$ (in the experiment reported by Berry, Nye and Wright [1979], the hexagonal prismatic maxima near the axis were elongated about 50 times as compared with those in the standard diffraction function (3.17)).

The computation of the exponents for the higher catastrophes on the list of Arnol'd [1975] involves several difficulties, discussed in detail by Berry [1977] in connection with an application to random short waves which is to be described in §7.

Outside catastrophe optics, wavelength scaling laws need not take the form (3.20). There are richer possibilities, which we illustrate with a brief discussion of the glory (Bryant and Jarmie [1974], Tricker [1970],

Nussenzveig [1969a,b], Khare and Nussenzveig [1977]). This is intense backscattering of light from a spherical water droplet. Rays emerging backwards would form a (structurally unstable) focal line extending to infinity, and would give rise to an amplitude $|\psi| \propto \kappa^{\frac{1}{2}}$. Such rays only exist, however, for spheres with refractive indices between $\sqrt{2}$ and 2 , a condition which excludes water. For water, the most deflected geometricaloptics rays, which emerge at 14° to the backward direction with total amplitude $\kappa^{-\frac{1}{n}}$, creep by diffraction around the surface of the drop, suffering wavelength dependent attenuation as they do so. It is these creeping rays which form the focal line, giving rise to a final backscattered wave amplitude

$$|\psi| \propto \kappa^{\frac{1}{3}} \exp\left[-D(\kappa a)^{\frac{1}{3}}\right], \tag{3.25}$$

where $a$ is the radius of the drop and $D$ a constant (Nussenzveig [1969a,b]). Therefore the

Table 2
Exponents governing the scaling of wave amplitude and fringe spacings as $\kappa \to \infty$

| catastrophe | $\beta$ | $\sigma_{\mathrm{i}}$ | $\gamma$ |
|---|---|---|---|
| fold | 1/6 | $\sigma_1 = 2/3$ | 2/3 |
| cusp | 1/4 | $\sigma_1 = 3/4, \sigma_2 = 1/2$ | 5/4 |
| swallowtail | 3/10 | $\sigma_1 = 4/5, \sigma_2 = 3/5, \sigma_3 = 2/5$ | 9/5 |
| elliptic umbilic | 1/3 | $\sigma_1 = 2/3, \sigma_2 = 2/3, \sigma_3 = 1/3$ | 5/3 |
| hyperbolic umbilic | 1/3 | $\sigma_1 = 2/3, \sigma_2 = 2/3, \sigma_3 = 1/3$ | 5/3 |
| butterfly | 1/3 | $\sigma_1 = 5/6, \sigma_2 = 2/3, \sigma_3 = 1/2, \sigma_4 = 1/3$ | 7/3 |
| parabolic umbilic | 3/8 | $\sigma_1 = 5/8, \sigma_2 = 3/4, \sigma_3 = 1/2, \sigma_4 = 1/4$ | 17/8 |

glory is a focusing effect which disappears in the geometrical-optics limit!

## 31.5 § 4. The Geometry of Wavefronts and their Normals

### 31.5.1 4.1. Paraxial Optics of a Uniform Medium

In the most general focusing problems, we wish to consider the propagation of rays from some specified initial surface $\Sigma$. As in § 2.1 we consider monochromatic waves in isotropic media so that the rays are normal to the geometrical-optics wavefronts. Moreover we shall now consider in detail a uniform medium, where the rays are straight. This describes a whole class of problems where an initial wavefront $W$ emerges from a non-planar refracting boundary such as a water/air or glass/air interface; for convenience we identify $W$ with $\Sigma$. $W$ is specified by its deviation $f(\boldsymbol{r})$ from the plane $\boldsymbol{r} \equiv (x, y), Z = 0$, and propagates in the $Z$-direction (Fig. 4.1).

We now make the important stipulation that $f(\boldsymbol{r})$ has small slopes, so that paraxial optics is applicable, and of course we require that $f(\boldsymbol{r})$ is smooth on wavelength scales (i.e. its curvatures are small) so that geometrical optics and related short-wave approximations are valid. These restrictions imply that the rays propagate at small angles $\theta \simeq \sin\theta \simeq \tan\theta$ to some optical axis, in this case the $Z$-direction. Note that there is no restriction on the distance of a ray from the $Z$-axis, merely the angle

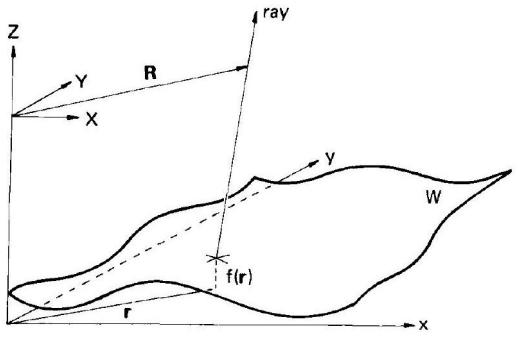



Fig. 4.1. Coordinates of wavefront and observation space.

between them-indeed in §6 we shall consider quite large extents of wavefront using this model. Paraxial optics linearizes Snell's law, with the useful consequence that $W$ may be regarded as an inversion of the refracting boundary, scaled by the ratio of the refractive indices, and we can infer the geometry of the one directly from that of the other. Henceforth we shall

consider only paraxial catastrophe optics, which is, we emphasize, an entirely self-consistent theory. There is a complete formal analogy between paraxial ray and wave optics and the nonrelativistic classical and quantum mechanics of a particle in two dimensions (Marcuse [1972], Hannay [1977]).

In the three-dimensional observation space $\mathscr{R}$ we shall generically encounter catastrophes with $K \leq 3$. But in many situations we shall need to consider not just one possible $W$ but whole families of initial wavefronts which can be explored by changing control parameters, such as rotating a water droplet lens so that gravity affects its shape (§5) or considering the time evolution of a rippling water surface (§6). In this way we can explore the unfoldings of catastrophes with $K > 3$ and even encounter their singularities. It is important to note that a singular point is interesting not just for its own sake, but more so for the geometry it imposes on its surroundings. In this way higher singularities organize lower ones; in later sections we shall talk of caustics in three space dimensions being organized by catastrophes of much higher codimension. Thus the singular points of the family of initial wavefronts assume special importance and we must analyze them in detail, to locate and determine the local configuration of the most important parts of the caustics; the global configuration is constrained by these local forms.

We denote the family of initial wavefronts by $f = f(\boldsymbol{r}; A)$, where $A$ represents control parameters influencing the form of the wavefront, and we denote spatial observation points by $\mathscr{R} \equiv (\boldsymbol{R}, Z)$ where $\boldsymbol{R} \equiv (X, Y)$ is a two-dimensional vector. Then the optical path length from $\mathbf{r}$ on $W$ to $(\boldsymbol{R}, Z)$ is $\phi$, where

$$\phi(\mathbf{r}; \boldsymbol{R}, Z, A) = \sqrt{\left[(Z - f(\mathbf{r}; A))^2 + (\boldsymbol{R} - \mathbf{r})^2\right]}. \tag{4.1}$$

For small angles (paraxiality) we can expand the square root, and the additional assumption of small curvatures means that focusing occurs for $Z \gg f$. If we write the control parameters $\boldsymbol{R}, Z$ and $A$ together as $C$, then

$$\phi = \phi(\boldsymbol{r}; C) = Z - f(\boldsymbol{r}; A) + \frac{1}{2Z}(\boldsymbol{R} - \boldsymbol{r})^2. \tag{4.2}$$

This is precisely the generating function $\phi(s; C)$ of §2.1, the state variables $s$ being coordinates $x, y$ in the $\mathbf{r}$ plane.

For rays, we require $\phi$ to be stationary with respect to $\boldsymbol{r}$, as in (2.9), thus forming the gradient map

$$\boldsymbol{R} = \boldsymbol{r} - Z\nabla_{\mathbf{r}} f(\boldsymbol{r}; A). \tag{4.3}$$

The rays focus at caustics when the Hessian (2.12) vanishes, giving

$$\mathscr{H}[\phi] = (f_{xx} - 1/Z)(f_{yy} - 1/Z) - f_{xy}^2 = 0 \tag{4.4}$$

where subscripts denote differentiation. This is quadratic in $Z$, giving the two conjugate levels at which rays from $\boldsymbol{r}$ focus. These values of $Z$ are the two principal radii of curvature of $f$.

## 31.5.2 4.2. The Curvature Landscape

In paraxial optics, rays from all points on $W$ with a given principal curvature $\mathbb{C}$ focus in the same plane $\boldsymbol{R}$, at $Z = 1/\mathscr{C}$. Thus it is advantageous to consider not the geometry of the height function $f$ of $W$, but the geometry of the curvature function. As with a height function, it is useful to visualize this as a "landscape". But it is a peculiar landscape, consisting of two surfaces which never intersect; they may join, but only at special points (see §4.3), and there

are no cliffs (infinite slopes). The curvature landscape is generally not obviously related to the height landscape. Nye [1978] considers a curvature difference landscape which is single valued; this has certain advantages in the particular case of water droplets he
considers (see § 5.2) but, as we shall shortly show, the information thereby lost is in general some of the most important to be had.

We can conveniently obtain the principal curvatures $\mathfrak{S}_1, \mathfrak{S}_2$ from (4.4) as

$$1/Z_{1,2} = \mathbb{C}_{1,2} = \frac{1}{2}\left(f_{xx} + f_{yy}\right) \pm \sqrt{\left[\frac{1}{4}\left(f_{xx} - f_{yy}\right)^2 + f_{xy}^2\right]}. \tag{4.5}$$

Note that the total curvature is just $\nabla^2 f$ : we shall use this fact when discussing water droplet optics in §5.

All points with the same principal curvature lie on a contour of the curvature landscape. These contours of curvature $\mathbb{C}$ in $r = (x, y)$ map via (4.3) to the caustics in the planes $Z =$ constant $= 1/\sqrt{C}$ (Fig. 4.2). A line element $\mathrm{d}\boldsymbol{R}$ of the caustic is perpendicular to the direction $\boldsymbol{i}$ of the principal curvature which generates it. In the notation of Fig. 4.2,

$$\mathrm{d}\boldsymbol{R} = (\mathrm{d}\boldsymbol{r} \cdot \boldsymbol{j})\,(1 - \mathfrak{C}_j/\,(\mathfrak{C}_i)\,\boldsymbol{j}. \tag{4.6}$$

Clearly something interesting happens when $\mathrm{d}\boldsymbol{r} \cdot \boldsymbol{j} = 0; \mathrm{d}\boldsymbol{r}$ and $\boldsymbol{i}$ are parallel at such a point (the two principal directions of curvature being everywhere orthogonal) and as we pass through such a tangency $\mathrm{d}\boldsymbol{R}$ changes sign and the caustic reverses direction and crosses its tangent, thus forming a cusp. We shall return to this in §4.4.

The families of caustics in $\mathscr{R}$ are completely determined by the geometry of $f$. We can consider the caustics to be organized by their most singular points, whose local unfoldings constrain the possible topologies of the three-dimensional caustic structures. It is very informative to

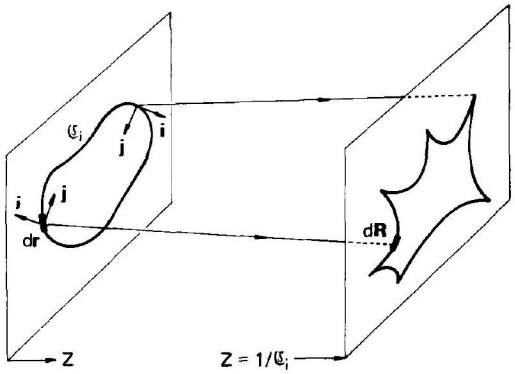



Fig. 4.2. Caustics are images of contours of principal curvature.

analyze the corresponding singular points of the curvature landscape: catastrophe theory will tell us the local behaviour of the caustics, while other differential-geometric properties will tell us about their global connectivity.

The simplest stationary points - maxima, minima and saddles - of the curvature landscape do not correspond to singular points in $\mathscr{R}$, but to places where a rib reaches an extreme value of $Z$, at which point it is tangential to a plane $\boldsymbol{R}$, and retreats again smoothly. Maxima and minima correspond to ribs convex upwards and downwards respectively: in a sequence of two-dimensional sections, these are lips (Fig. 2.9a). Saddles correspond to beak-to-beak events (Fig. 2.9b).

### 31.5.3 4.3. Umbilics

The most important places on the curvature landscape are the locally conical points where the two surfaces meet, called (for obvious reasons) umbilic points. At an umbilic point, $\mathfrak{C}_1 = \mathfrak{C}_2$, the curvature contours are singular, and the caustic has a singular point "focus". Note that umbilic points are the only possible connections between the two surfaces of the curvature landscape. In classical differential geometry, umbilic points were studied by Darboux [1896], Gullstrand [1905], Forsyth [1912] and, more recently, by Porteous [1971].

The condition for a point to be umbilic is, from (4.5),

$$f_{xx} - f_{yy} = f_{xy} = 0. \tag{4.7}$$

Umbilic points are classifiable in three different ways, discussed in detail by Berry and Hannay [1977], in terms of the coefficients of the cubic terms of a local expansion of $f$. These are third derivatives evaluated at the umbilic point, and we employ the notation

$$\alpha \equiv f_{xxx}, \quad \beta \equiv f_{xxy}, \quad \gamma \equiv f_{xyy}, \quad \delta \equiv f_{yyy}. \tag{4.8}$$

The unfolding of the point focus of the caustic is an elliptic ( $E$ ) or hyperbolic ( $H$ ) umbilic catastrophe as the contours of curvature near the umbilic point are ellipses or hyperbolae, yielding the catastrophe classification of an umbilic point:

$$\text{if} \quad C(\alpha,\beta,\gamma,\delta) \equiv 4\left(\alpha\gamma - \beta^2\right)\left(\beta\delta - \gamma^2\right) - (\alpha\delta - \beta\gamma)^2 \begin{cases} > 0 & \text{then } E \\ < 0 & \text{then } H. \end{cases} \tag{4.9}$$

These two cases are smoothly connected by the parabolic umbilic, for which the contours are parabolae; this case is non-generic in three dimensions because the parabolic umbilic has $K = 4$. The umbilic is parabolic if $C(\alpha,\beta,\gamma,\delta) = 0$, unless the other classification functions (see below) vanish, in which case it is an umbilic with $K > 4$. The transition $E \leftrightarrow H$ corresponds to a smooth change in the angle of the axis of the double cone connection between the two curvature surfaces, with respect to the propagation direction (Fig. 4.3).

The principal axes of curvature generate an orthogonal net of lines of curvature (not to be confused with contours of curvature) whose configuration around a generic umbilic point is one of the three shown in Fig. 4.4. Lemon ( $L$ ) is distinguished from star ( $S$ ) and monstar ( $M$ ) by there being one straight line through the singularity rather than three. Corresponding to this is the pattern classification:

$$\text{if} \quad P(\alpha,\beta,\gamma,\delta) \equiv 4\left[3\gamma(\alpha - 2\gamma) - (\delta - 2\beta)^2\right]\left[3\beta(\delta - 2\beta) - (\alpha - 2\gamma)^2\right]$$

$$-\left[(\delta - 2\beta)(\alpha - 2\gamma) - 9\beta\gamma\right]^2 \begin{cases} > 0 & \text{then } M \text{ or } S \\ < 0 & \text{then } L. \end{cases} \tag{4.10}$$

During a circuit of the umbilic the principal axes of curvature rotate one half-turn; co-rotation (counter-rotation) is denoted by $+\frac{1}{2}$ $\left(-\frac{1}{2}\right)$. This is the basis of the index classification:

$$\text{if } J(\alpha,\beta,\gamma,\delta) \equiv \alpha\gamma - \gamma^2 + \beta\delta - \beta^2 \begin{cases} > 0 & \text{then } +\frac{1}{2}(L \text{ or } M) \\ > 0 & \text{then } -\frac{1}{2}(S). \end{cases} \tag{4.11}$$

The threefold classification of umbilics is summarized in Fig. 4.5; this diagram illustrates the following important implications of the inequalities (4.9) and (4.11): all elliptic umbilics have index $-\frac{1}{2}$, and all umbilics with index $+\frac{1}{2}$ are hyperbolic. The classification of umbilic catastrophes with $K > 4$ depends on terms higher than cubic; we do not discuss these cases further.

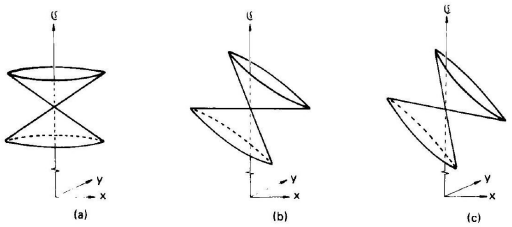



Fig. 4.3. Orientation of the conical (umbilical) connection between the two surfaces of the curvature "landscape": curvatures in the neighbourhood of (a) elliptic, (b) parabolic, (c) hyperbolic umbilic points.

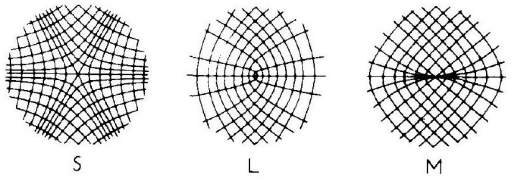



Fig. 4.4. Lines of curvature around umbilic points with star ( $S$ ), lemon ( $L$ ) and monstar ( $M$ ) configurations.

Obviously, the catastrophe classification dictates the local form of the caustics. The pattern classification serves to distinguish the rare monstars from the asymptotically indistinguishable lemons (Figs. 4.4 and 4.5), but we know of no circumstance in which this distinction affects the caustics.

The index classification constrains the configuration of caustics remote from the focus. It is especially useful in conjunction with the following conservation laws:
(i) The number of rotations of the principal axes of curvature of $f$ around any closed curve on $W$ is the algebraic sum of the indices of the enclosed umbilic points. If $A$ is varied, so that the number and positions of the umbilics vary, then provided no umbilic crosses the curve the algebraic sum of the indices is invariant.
(ii) The net index of umbilics on any unbounded surface is equal to the Euler-Poincaré characteristic of the surface (Spivak [1975]). We shall make use of this in §§ 6 and 7.

As $A$ is varied, the umbilics move along definite trajectories in the parameter space. Analyzing these, and studying the possible umbilic

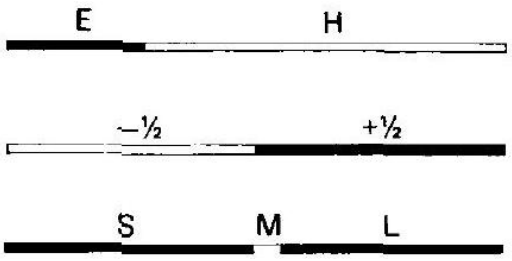



Fig. 4.5. The interrelation of the three classifications of umbilic points.

"reactions" (Nye [1979]) are valuable tools for investigating catastrophes of codimension $K > 3$. Rule (i) above shows at once that two elliptic umbilics cannot be born or annihilate, as they must both have index $-\frac{1}{2}$. The only umbilic reactions with $K \leq 4$ are:

Birth/annihilation: $H_{-\frac{1}{2}} + H_{+\frac{1}{2}} \leftrightarrow$ nothing

Transformation: $E_{-\frac{1}{2}} \leftrightarrow$ parabolic umbilic $\leftrightarrow H_{-\frac{1}{2}}$.

## 31.5.4 4.4. Cuspoids

On a plane screen $Z$ = constant, the only stable caustics are fold lines which may be interrupted by cusp points. We recall that these are the images of contours $\mathbb{C} = 1/Z$, and that a line element of caustic (Fig. 4.2) is perpendicular to that direction of principal curvature which generates it (eq. (4.6)). The points where lines of curvature touch a contour of curvature generate cusps in the focal plane. If we consider all the contours, the loci of all such points are riblines in $\boldsymbol{r}$; their images in ( $\boldsymbol{R}, Z$ ) are ribs of the caustic.

There may be points where the contours and lines of curvature not only touch but osculate (i.e. have the same curvature): these are swallowtail points. If we consider the family of wavefronts parametrized by $A$, there may also stably occur points where the contours and lines of curvature touch, osculate and also have the same rate of change of curvature: these are butterfly points. The hierarchy is shown in Fig. 4.6; we could continue similarly for cuspoids with $K > 4$. Of course, a cusp point must lie on a ribline; at a swallowtail point the ribline touches both the contour and

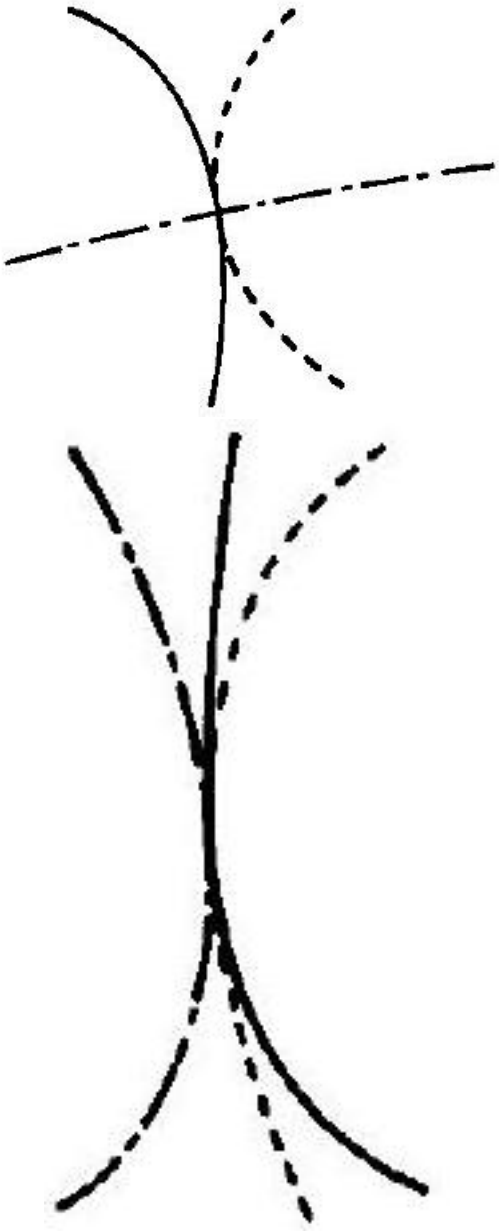




Fig. 4.6. Geometry of a wavefront in the neighbourhood of a cuspoid singularity; (a) cusp; (b) swallowtail; (c) butterfly. Contours of $\mathfrak{C}$ (full lines), lines of curvature (dashed lines), riblines (chain lines).
line of curvature, whilst at a butterfly point there are two riblines, one of which touches both the contour and line of curvature.

As with umbilics, analysis of the trajectories of the cuspoids (ribs and riblines and their analogues in higher dimensions) is extremely helpful in elucidating the geometry of higher catastrophes.

## 31.6 § 5. Liquid Droplet Lenses

### 31.6.1 5.1. The Irregular Droplet as a Lens

The distortion of an image viewed through a raindrop is familiar to all spectacle wearers, for an irregular water droplet on an inhomogeneously dirty surface is an extremely aberrant lens. Larmor [1891] wrote "on looking at a bright lamp, sufficiently distant to be treated as a luminous point, through a plate of glass covered by fine raindrops, the caustic surfaces after refraction through the drops are produced within the eye itself, and their sections by the plane of the retina appear as bright curves projected into the field of vision". Minnaert [1954] also drew attention to the optics of such droplets, but neither author saw any reason to study the forms of the seemingly infinite variety of caustics produced in this way: it is in such circumstances as these that we can use catastrophe optics and realise the power of the notions of structural stability and genericity.

Although much of the following discussion is couched in terms of geometrical optics, note that in real experiments we do not work with light of zero wavelength, and diffraction is significant: we invoke the structural stability of the diffraction catastrophes (§ 3) to relate observations to the underlying geometrical caustics.

A droplet on a flat transparent substrate illuminated normally by a plane wave is well described by the paraxial optics of §4. The form of the drop surface is constrained by the surface tension of the liquid, and the differential equation to be satisfied depends on both the size of the drop and the orientation of the substrate, the crucial condition being whether or not the internal pressure may be considered to be uniform throughout the drop. On a horizontal substrate, a water drop of more than 3 or 4 mm across is large enough for gravity to affect its shape and cause the internal pressure to be non-uniform. In §5.2 we consider the optics of small droplets, for which the effect of gravity is negligible. For a water drop of even 1 mm or so in size on a vertical surface the effect of gravity is
significant. Tilting the substrate, and hence varying the effect of gravity on the drop, provides an additional control parameter affecting the caustics. This will be considered in § 5.3.

We discuss caustics of refraction, produced by droplets of water, as this is most convenient for experiments. The discussion carries over in principle for caustics in reflected light, and for any liquid, with the proviso that the different surface tension will of course affect the size of

drop for which gravity becomes significant. We remark in passing that the caustics formed by reflection from plastic film, used to illustrate the geometry of the elementary catastrophes in § 2.3, are on a much larger scale and diffraction is much less significant. The surface tension condition for uniform internal pressure (eq. (5.2)) is obeyed approximately by the plastic film; conditions analogous to non-uniform internal pressure can be explored in a controlled way with the plastic film as part of the wall of a vessel full of liquid: the pressure gradient depends on the density of the liquid. These techniques have not proved as convenient for research as the use of water droplets.

## 31.6.2 5.2. Droplets with Uniform Internal Pressure

Observations with the naked eye such as those described by larmor [1891] are of the directional caustics at infinity-the far field. Berry [1976] considered this case in detail for small horizontal drops (infinity in this instance is anything greater than a few centimetres). The far field caustic is the image of the contour of zero curvature on the initial wavefront $W$. If $W$ is specified by its deviation $f$ from a reference plane (see §4), this is the line where the Gaussian curvature

$$\mathfrak{C}_1\mathfrak{C}_2 = f_{xx}f_{yy} - f_{xy}^2 \tag{5.1}$$

vanishes. Berry pointed out that as the far field is a two-dimensional space of directions, the only structurally stable caustics are fold lines interrupted by cusp points. He showed that if a surface-tension dominated droplet has a far field caustic (and it need not-for example, if it is shaped like the cap of an ellipsoid), the fold caustics must be concave outwards and there must be at least three cusps. This can be verified experimentally by shining a broadened laser beam vertically through a small irregular droplet on a horizontal flat glass plate and observing the far field caustic on the ceiling of a darkened room (Fig. 5.1). Such a
caustic, perhaps a metre across, with its delicate diffraction structure in the red light of a helium-neon laser, is a most beautiful demonstration of catastrophe optics.

The directional caustic at infinity is not the special asymptotic form it might appear: real and virtual far fields can be identified (with a 180° rotation) and the three-dimensional caustic structure is thus continuous from the real near field through the far field to the virtual near field. By using suitable lenses, it is possible in principle to cause any twodimensional section to be at infinity, but we reserve the term far field for the caustic at infinity when there are no lenses in the space beyond $W$.

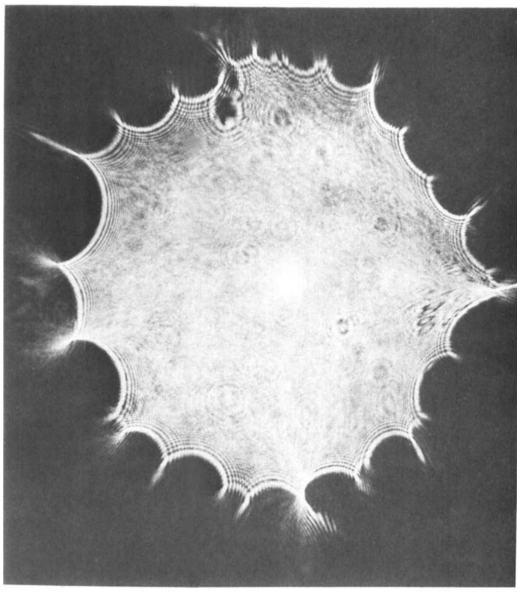


Fig. 5.1. The far field caustic from a small irregular water drop on a horizontal flat glass plate.

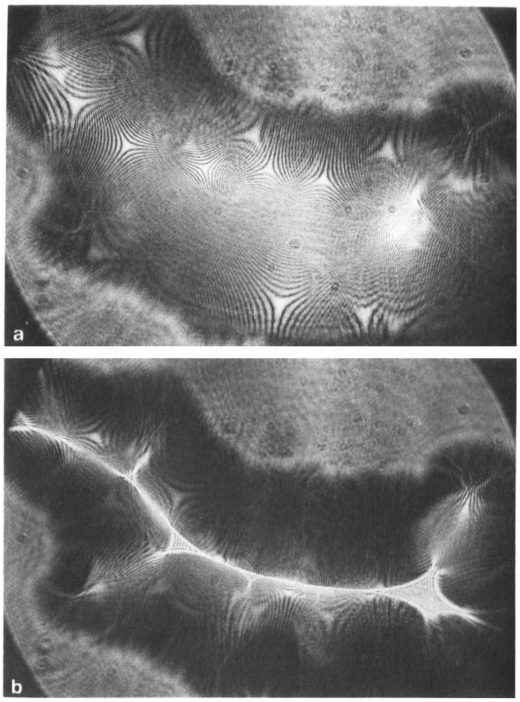



Fig. 5.2. A sequence of two-dimensional sections of the near field of a small irregular water drop on a horizontal microscope slide. In (a) the microscope is focused onto the plane of the elliptic umbilic foci; in (b) to (d) the plane of focus moves towards the drop (courtesy of J. F. Nye).

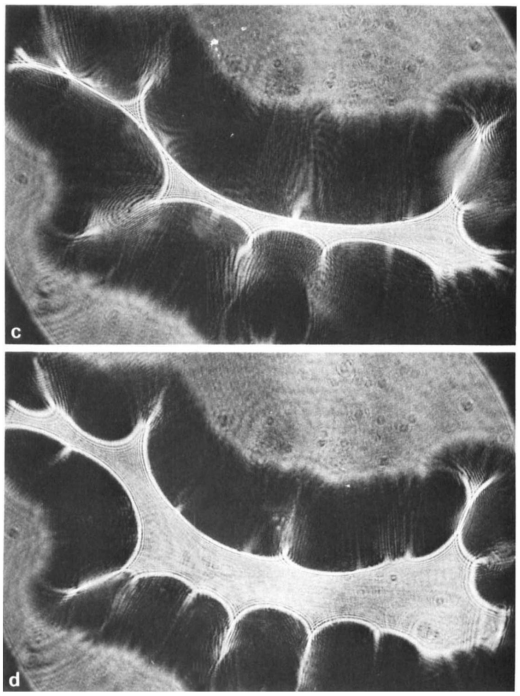



Fig. 5.2 (contd.)

Studying the spatial caustics beyond $W$ is clearly more general, and we expect to find generically catastrophes of codimension $K \leq 3$. It turns out, however, that the restriction to uniform internal pressures rules out the hyperbolic umbilic, and the only possible singularities are the fold, cusp, swallowtail and elliptic umbilic (Nye [1978]).

If the excess internal pressure $p$ is uniform throughout a drop of refractive index $n$, the surface tension $\gamma$ constrains $f$ to be a solution of

$$\nabla^2 f = (n-1)p/\gamma = \mathfrak{I} \tag{5.2}$$

where $\mathfrak{A}$ is a constant. In paraxial optics, the sum of the principal curvatures (4.5) is equal to the Laplacian, so

$$\nabla^2 f = \mathfrak{C}_1 + \mathfrak{C}_2 = \mathfrak{A} \tag{5.3}$$

Nye [1978] defines the curvature difference as

$$\mathfrak{D} \equiv |\mathfrak{S}_1 - \mathfrak{C}_2| \,. \tag{5.4}$$

As a consequence of the relationship (5.3), the contour maps of each of $\mathfrak{S}_1, \mathfrak{C}_2, \mathfrak{D}$ and the Gaussian curvature $\mathfrak{C}_1\mathfrak{C}_2$ are identical, only the labels on the individual contours being different. For convenience, Nye considered the caustic structure in terms of the single valued $\mathfrak{D}$-landscape, and showed that it has locally conical minima all at zero level, the only other stationary points being saddles; there are no maxima.

At an umbilic point on $W, \mathfrak{D} = 0, \mathfrak{C}_1 = \mathfrak{C}_2$ and the rays focus at a height $Z = 2/\mathfrak{A}$. It is important to note that all the umbilic points focus at this same level, and are all elliptic. Without the restriction on drop shape, we do not normally see all the umbilic foci at once on a two-dimensional screen; of course, all the umbilic points are simultaneously present on the wavefront.

Contours $0 < \mathfrak{D} < \mathfrak{C}_1 + \mathfrak{S}_2$ image at two conjugate levels (Nye [1978]) above and below the umbilic focusing level. For $\mathfrak{D} = \mathfrak{C}_1 + \mathfrak{C}_2$ one caustic is at infinity and the other at $1/\mathscr{A}$, and for $\mathfrak{D} > \mathfrak{C}_1 + \mathfrak{C}_2$ one caustic is virtual and its conjugate is real, lying between $1/\Re$ and $W$.

Because of their size and shape, the "focal length" of the droplets (the height of the umbilic focusing level) is small. Nye [1978] studied the near field experimentally by focusing a

microscope on successive planes above irregular horizontal droplets illuminated normally by a broadened parallel laser beam. There is indeed a unique plane of focus, containing many elliptic umbilic diffraction "stars" (Fig. 5.2a). As the plane of focus of the
microscope is altered, they unfold and interact by beak-to-beak events and swallowtails to give many-cusped figures close to the drop and in the far field (Figs. 5.2 b-d). To quote Nye, "In three dimensions, one has to picture a brilliant gothic palace: pointed arches of light, on a microscopic scale, with the tilted triangular columns of different sizes springing from a polished floor, which reflects their tracery of interactions . . .". Nye [1978] studied in detail the possible interactions between the different branches of the caustics emanating from different elliptic umbilic foci. Note that there are no lips events: these correspond to maxima of the $\mathfrak{D}$ landscape, which are precluded by the surface tension condition (5.2). To reduce the complexity and show up systematic features, Nye studied the caustics from drops with controlled perimeters by filling with water appropriately shaped holes cut in opaque tape stuck to a glass microscope slide. This technique was used with a triangular hole to study the elliptic umbilic diffraction catastrophe (Berry, Nye and Wright [1979]; see § 3) and, as we shall shortly see, is proving to be an invaluable research tool for studying the caustics from water droplets with non-uniform internal pressure.

### 31.6.3 5.3. Droplets with Non-Uniform Internal Pressure

In the caustics from an irregular droplet clinging to an inclined glass surface, all the catastrophes of codimension up to three are seen, including the hyperbolic umbilic (Nye [1979]). Patterns like Fig. 5.3 can be seen with the naked eye, close to a drop on a vertical surface, by viewing a distant point light source through a raindrop on a window or on one's spectacles. In two-dimensional plane sections of the near field of such caustics observed with a microscope (as described above), the most conspicuous new generic features are folds which are convex outwards and hyperbolic umbilic foci (Figs. 5.3 and 3.5c).

If the glass substrate is rotated in its own plane, the caustic does not rotate rigidly but continuously rearranges itself such that any fold which is convex outwards stays in the same position in the field of view. Clearly this is an effect of gravity: the drop perimeter remains fixed, but the thickest part of the drop will always be pendulous downwards. In drops not affected by gravity, folds must be concave outwards (Berry [1976]).

As the plane of focus is changed, there is a complicated sequence of interactions between the various parts of the caustic. A transition like that

between Figs. 5.4a and 5.4b is typical: it can be accomplished by changing the plane of focus, by tilting the glass substrate out of the vertical, or even dynamically "as can be verified when walking in the rain, wearing spectacles and viewing a distant light: at each pace the elliptic umbilic stars pierce the fold in striking unison" (Nye [1979]). The physically accessible "controls" are not necessarily all independent; their relation-

ship with the control parameters of catastrophe theory need not be simple, but can usually be inferred from experimental evidence.

In spite of the natural irregularities of any drop, the caustics that it produces have a remarkable degree of organization, and Nye [1979] showed that this organization can be understood by regarding higherdimensional catastrophes as organizing centres. By performing controlled experiments using the "hole in tape" method outlined above, rotating the substrate about its axes and letting gravity alter the drop shape without changing its perimeter, he obtained degenerate umbilic points on the surface corresponding to $D_5$ (the parabolic umbilic, $K = 4$ ) and $E_6(K = 5)$. These experiments have confirmed the geometry of these catastrophes (Godwin [1971], Callahan [1977]) and elucidated the characteristics of
the corresponding diffraction structure (§ 3; Figs. 3.7 and 3.8). In a randomly irregular drop,

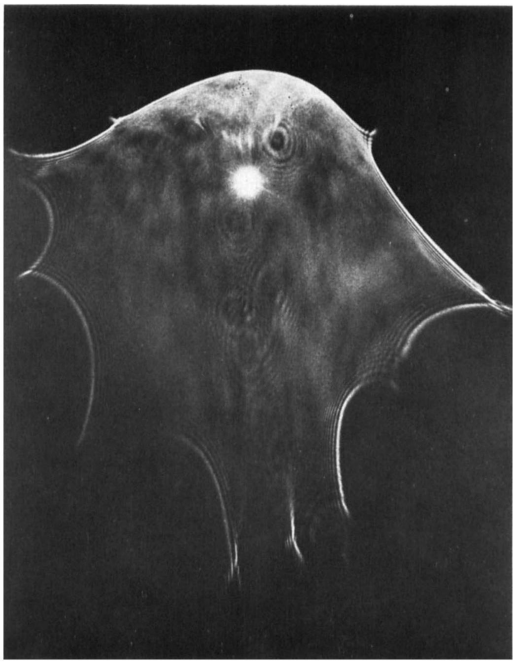


Fig. 5.3. The far field caustic from an irregular water drop on a vertical flat glass plate (courtesy of J. F. Nye).

the degeneracy is removed and the patterns can be understood in terms of unfoldings of $D_5$ and $E_6$. In addition to these catastrophes, there are persistent global features which point to there being a still higher organizing centre at work, conjectured by Nye [1979] to be $X_9(K = 8)$. Current work is confirming this (see below).

As we saw most explicitly for the umbilic classifications in §4, catastrophe theory is about the local forms of the unfoldings of singularities: typically one works in terms of expansions about umbilic (or other singular) points. Nye makes no attempt to fit solutions to boundary conditions at the edge of the drop, except for the case of a drop with a circular perimeter. This solution, due to Berry, shows the power of catastrophe optics and the value of the concept of structural stability. For a thin droplet of density $\rho$, refractive index $n$ and surface tension $\gamma$, on a substrate at an angle $\theta$ to the horizontal, with the $x$ axis as in Fig. 5.5 and the $y$ axis normal to the page, the equation governing the form of the wavefront $W$ is

$$\nabla^2 f = \frac{n-1}{\gamma}(p + g\rho x \sin\theta), \tag{5.5}$$

where $g$ is the acceleration of gravity. In dimensionless variables $f', p'$, (5.5) reduces to

$$\nabla^2 f' = p' + x \tag{5.6}$$

which, for a drop with a circular perimeter of radius $b$, has the exact solution

$$f' = \frac{1}{4}p'\left(x^2 + y^2\right) + \frac{1}{8}x\left(x^2 + y^2 - b^2\right). \tag{5.7}$$

The origin is an umbilic point, but $f$ has non-zero slope there, so the umbilic focus is off-axis. The catastrophe classification (4.9) shows that the umbilic is hyperbolic. Without gravity, the drop would be a spherical cap with a structurally unstable point focus. Gravity removes the indeterminacy and produces a structurally stable hyperbolic umbilic focus: the geometrical caustic in the focal plane can be shown to be two lines, $X = \pm\sqrt{3}Y, X \leq 0$, which is a $V$ shape

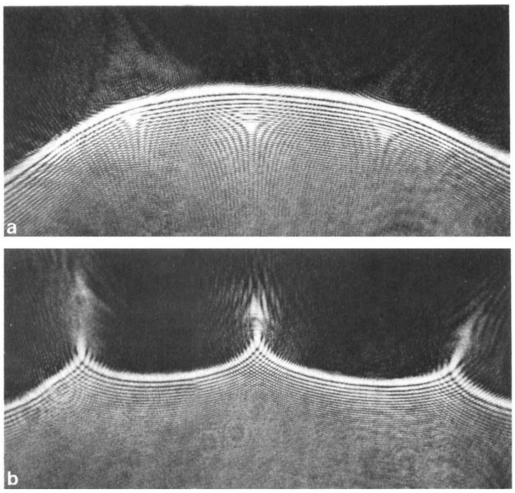



Fig. 5.4. (a, b) Near field caustics formed by the region of an irregular water drop whose shape is affected by gravity (courtesy of J. F. Nye).

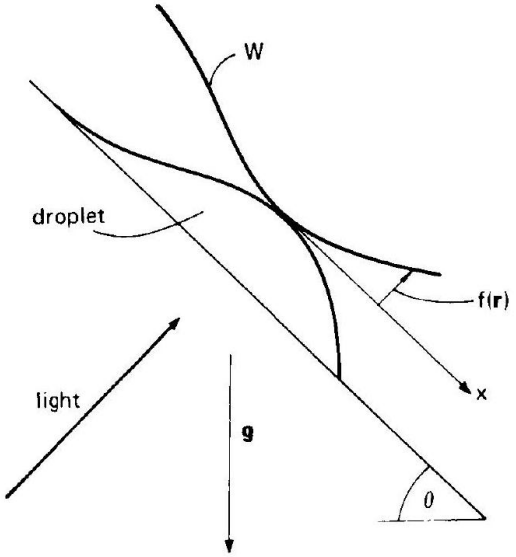



Fig. 5.5. Coordinates for a drop with non-uniform internal pressure.

with arms at 60° : see Nye [1979] for details. Fig. 3.5c was produced in this way. Structural stability guarantees that a drop in an imperfect circular or oval hole also produces a hyperbolic umbilic focus, with the arms of the V at an angle close to 60°. A good example is shown by Tanner [1978] in a study of the optics and motion of oil droplets.

The above case makes clear not only the identification of the off-axis lens aberration known as primary coma with the hyperbolic umbilic (Appendix 2) but the reason for the difficulty of avoiding it: the aberration is structurally stable!

Nye [1979] refers briefly to the "strikingly different" caustic patterns, only resolvable under high magnification, from small water droplets of such thickness that gravity affects their shape even when they are horizontal. A systematic study now in progress is showing that the central regions of such caustics can be understood in terms of the catastrophe $X_9(K = 8)$. In the next section we shall see that $X_9$ also arises in the context of the caustic networks formed by refraction and reflection by the surface of moving water or rippled (bathroom) glass.

## 31.7 § 6. Caustic Networks

### 31.7.1 6.1. Networks and Junctions

Mascrosopic caustic networks are a familiar sight as ever-changing patterns of lines of sunlight focused onto the bottom of a swimming pool or the sea bed after refraction by the wavy water surface (Minnaert [1954], Stevens [1976]). They also occur in reflection, for example on the soffit of a bridge or the overhanging side of a boat; it is sometimes possible to see both refracted and reflected caustics from the same stretch of water. Refraction of starlight by refractive-index inhomogeneities caused by atmospheric turbulence produces caustic networks, which can be seen crossing the defocused image disc of a strongly twinkling star viewed under high magnification through an astronomical telescope. More amenable to detailed study are the stationary networks of caustics formed when a plane wave is reflected or refracted by rippled glass (Berry [1976, 1978a], Upstill [1979]). Caustic networks can also occur in an electron microscope, when electrons are focused by an electrode in the form of a wire gauze (Bartz, Weissenberg and Wiskott [1956]).

These two-dimensional networks are of course sections of complicated caustic surfaces in three dimensions, which are likely to be organized by catastrophes of higher codimension, as there are many possible parameters which change the details of the caustics. It is natural to analyze these families of caustics in terms of the geometrical-optics configuration of their two-dimensional sections, using the paraxial optics of §4 and keeping in mind that in most laboratory scale experiments (and also the observation of starlight caustics) we see the characteristic diffraction structures enumerated in § 3.

Apart from the rules related to the indices of umbilic points (§4), no general theorems about caustic networks are known. It is possible, however, to construct periodic and quasi-periodic models for wavy surfaces, whose calculated caustic networks agree well with many observations, both in terms of local structure and global configurations (Upstill [1979]). More generally, it has proved instructive to consider a network as being constructed from junctions, whose details are to be understood in terms of sections of catastrophes, and which are to be connected by fold lines in the simplest possible way. We shall illustrate these considerations by analyzing two common morphologies of caustic network (§§6.2 and 6.3) for which catastrophe optics reveals previously unsuspected fine structure.

### 31.7.2 6.2. Sinuous Networks

The pattern of overlapping wavy lines shown in Fig. 6.1 is a type of caustic network often seen in sunlight focused by rippling water. The lines

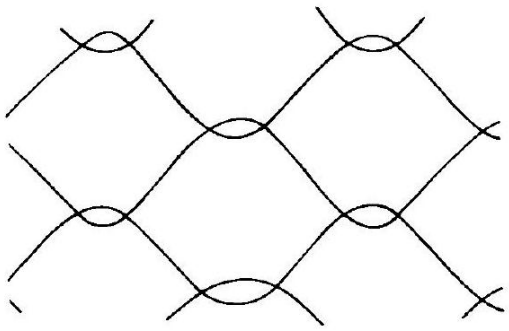


Fig. 6.1. A common morphology of caustic network formed by a wavy water surface, unresolved.

cannot all be single folds which merely cross one another, for if we attempt to label the

regions between the lines according to the number of rays contributing, which changes by $\pm 2$ across a fold, we find ambiguities, equivalent regions having different labels. This suggests that Fig. 6.1 is an imperfect resolution of a more complicated caustic network whose fine structure is blurred by the rapid motion of the water and by the halfdegree angular diameter of the sun.

The first step in resolving Fig. 6.1 is to realise that below each wave in a single train of long-crested water waves is a caustic surface (Fig. 6.2) whose two sheets meet along a cusped edge and intersect the sea bed in a line pair. This makes it likely that some (but not all) of the lines in Fig. 6.1 are line pairs. The sinuous network repeats in two directions, suggesting a refracting surface consisting of two trains of water waves. This is a special case of a model surface proposed by Upstill [1979], consisting of a superposition of $n$ sinusoids with amplitudes $A$, wave vectors $\boldsymbol{k}$ and phases $\chi$, so that

$$f = \sum_{i=1}^{n} A_i \cos\left(\boldsymbol{k}_i \cdot \boldsymbol{r} + \chi_i\right). \tag{6.1}$$

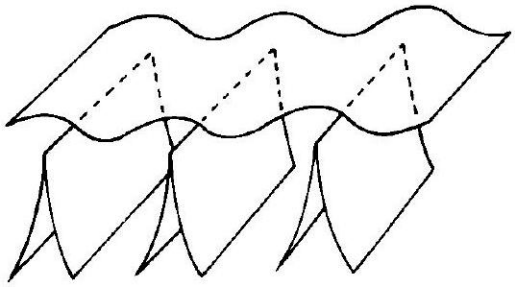


Fig. 6.2. Caustic surfaces below a single train of water waves.

For $n > 2, f$ is periodic only if the $\boldsymbol{k}_i$ are lattice vectors; but for $n = 2, f$ is periodic for any linearly independent $\boldsymbol{k}_1, \boldsymbol{k}_2$. Upstill [1979] studied in detail the caustics for the case $n = 2$, verifying his analysis by experiments using both a ripple tank and sinusoidally grooved glass slides. All the possible two-dimensional caustic sections are encountered if we let both sinusoids have unit wavelength and one sinusoid have unit amplitude, varying the amplitude $A$ of the other from 0 to 1 , the angle $\theta$ between them from 0 to $\pi/2$, and the distance $Z$ of the focal plane from 0 to $+\infty$ (this surface is invertible, so the virtual field $Z < 0$ is the same as the real field $Z > 0$ ).

Consider first the effect of increasing $A$, with $\theta$ and $Z$ fixed. Clearly if $A = 0$ there is just one sinusoid and the caustics are just parallel line pairs as in Fig. 6.2. For $0 < A \ll 1$ the line pairs start to kink (Fig. 6.3a) but, as argued above, as $A$ increases the fold lines do not simply overlap. Instead, there are swallowtail catastrophes which unfold asymmetrically (Fig. 6.3b). Finally, for $A \simeq 1$, the swallowtails link up with one another by beak-to-beak events (Fig. 6.3c) to form the network shown unresolved in Fig. 6.1; far from being composed of sinuous lines, the network is thus revealed as an array of overlapping giant lip shapes.

Now consider Fig. 6.3c as part of a focal sequence. Figs. 6.4a-c show one unit cell of such a sequence for increasing $Z$, obtained experimentally by refracting a broadened parallel laser beam through two sinusoidally grooved slides at about 60° to one another, with $A \leq 1$. Below the plane of Fig. 6.4a is a lips event, whilst between Figs. 6.4a and 6.4b two swallowtails unfold, four cusps from neighbouring unit cells enter the field of view, and there are four beak-to-beak events. The configuration of Fig. 6.4b persists for a considerable range of $Z$. Between Figs. 6.4b and 6.4c there is another lips event at right angles to the first, and beyond Fig. 6.4c the four cusps approach the fold lines which are concave towards them to give four hyperbolic umbilic catastrophes in the far field. One unit of the periodic pattern is then a large diamond shape,

each line being three coincident folds. As the far field is approached, however, neighbouring units overlap and the complete pattern becomes a confusion of caustic lines.

Finally, consider the effect of changing $\theta$. For $\theta = 0$ the surface is a single sinusoid. If $0 < \theta \ll \pi/2$ caustic patterns in planes $Z$ = constant are highly elongated; under poor resolution the typical section of Fig. 6.4b looks like two triple junctions - common features of caustic networks (cf §6.3 ). As $\theta$ increases, the caustics stretch in the $Z$ direction: whilst the

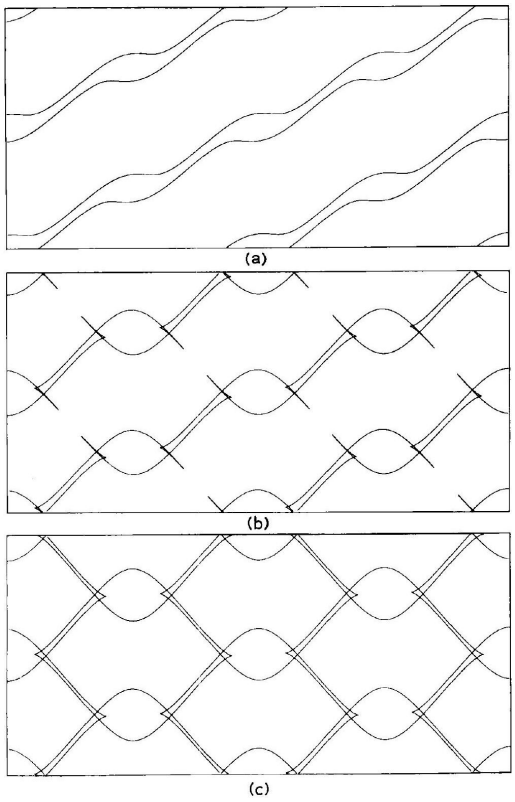



Fig. 6.3. Computed caustic sections from a surface consisting of a superposition of two sinusoids, for fixed $\theta$ and $Z$. (a) $0 < A \ll 1$, (b) $A < 1$, (c) $A = 1$.

first lips event (preceding Fig. 6.4a) and the hyperbolic umbilic foci (after Fig. 6.4c) remain fixed, the events between them occur at progressively lower $Z$. The limit $\theta = \pi/2$ is a degenerate case, for which all events occur simultaneously and there are hyperbolic umbilic singularities for a range of focusing levels, corresponding to lines of umbilics (rather than points) on $f$. It is as if there are two of the patterns from one sinusoid alone superposed orthogonally, except that within the region of intersection the caustic surface is not one but three coincident fold catastrophes; this is clearly shown experimentally by the greater intensity of that region (Fig. 6.5).

## 31.7.3 6.3. Triple Junctions

The polygonal line pattern shown in Fig. 6.6 is another common type of caustic network. Berry and Nye [1977] drew attention to the prevalence of triple junctions. This might seem unremarkable, since triple junctions are common, stable features of natural systems of lines and surfaces as diverse as mud cracks, animal markings and foams. But no catastrophe has a two-dimensional section consisting of three lines meeting at a point, and this leads to the conclusion that the caustic triple junction must be illusory, so the appearance of the network in Fig. 6.6, like the sinuous network in Fig. 6.1, is a consequence of imperfect resolution.

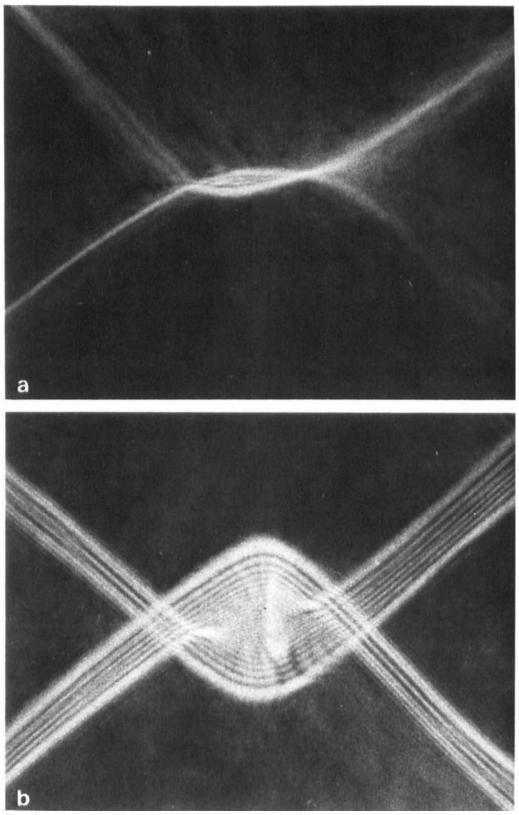



Fig. 6.4. Observation of one unit cell of caustic of refraction from two superposed sinusoidally grooved glass slides. Typical sequence of sections for fixed $\theta$ and $A; Z$ increasing from (a) to (c).

Such a junction could be produced by the meeting of three water waves, in which case each emerging line will be a line pair (cf Fig. 6.2). Berry and Nye [1977] studied the fine structure of such junctions by refracting light through a glass "lollipop", on which three smooth grooves had been impressed in the molten state, and explored the threedimensional caustic surface in a sequence of two-dimensional sections using a microscope. They found the central feature of the junction to be a section of an elliptic umbilic catastrophe, surrounded by a "foliage" of smoothly curving fold lines (Fig. 6.7). As the plane of focus approaches

the "lollipop" (Fig. 6.8), the elliptic umbilic shrinks to its singular point and unfolds again, one or more of its cusps pierce the "foliage", and the unfolded elliptic umbilic and folds interact via beak-to-beak events and swallowtails, leaving three line pairs ending in inward pointing cusps which recede from the field of view.

These triple junctions link up to form a network resembling chain mail. The existence of such networks can be confirmed experimentally by producing caustics of refraction from a type of bathroom window glass (Pilkington's "Atlantic"), as shown in Fig. 6.9. However, it is impossible for all the umbilic catastrophes produced by an undulating wavefront to be of elliptic type; this follows from the arguments outlined in §4.3, according to which all elliptic umbilics have index $-\frac{1}{2}$, whereas the surface has net index zero. Therefore there must be hyperbolic umbilic points, of index $+\frac{1}{2}$, on the surface: where do they focus? Recall that the caustics have continuity through the far field (§5.2). This implies that, provided there are no hyperbolic umbilics in the real field, the missing ones must lie in the virtual field. This possibility is exemplified by "Atlantic" glass,

whose virtual near field caustics consist of a lattice of spiky "spangles" each containing two hyperbolic umbilics (Fig. 6.10). (The existence of additional umbilics of either sort is not

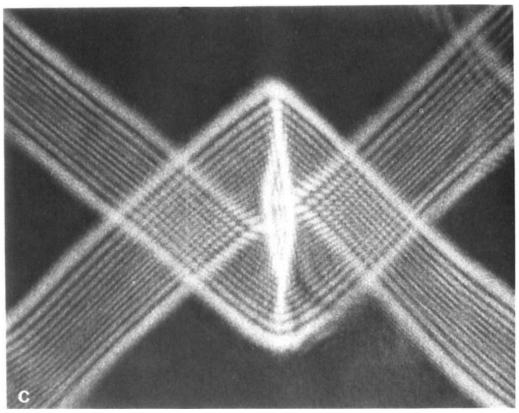



Fig. 6.4 (contd.)

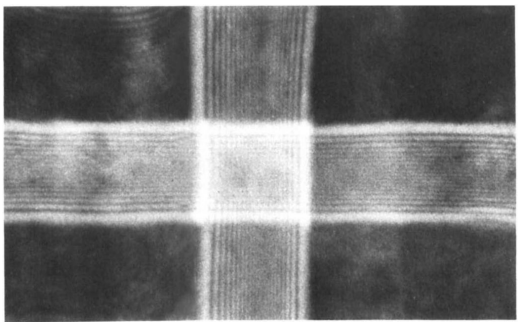


Fig. 6.5. Typical section, as Fig. 6.4 but with sinusoids at right angles.

precluded, but their net index must be zero.) Note that, in contrast to the case considered in § 6.2, real and virtual fields are different: therefore this surface is non-invertible.

An obvious global model surface corresponding to the meeting of three trains of water waves is eq. (6.1) with $n = 3$. We have confirmed that in the appropriate region of the ( $A, \boldsymbol{k}, \chi$ ) parameter space this model gives both the "chain mail" network and the lattice of "spangles", correctly reproducing the details of the entire observed real and virtual focal sequence. Moreover, this model describes various other experimentally observable configurations and transitions, including the transition from the networks with elliptic umbilic centred triple junctions to the sinuous networks discussed in §6.2 (Upstill and Berry, in preparation). Recall that two trains of water waves superposed at a fairly acute angle gave caustics which under poor resolution looked like triple junctions with a fine structure comprising only one cusp (cf Fig. 6.4b). Therefore three waves are not necessary for the occurrence of caustic triple junctions; it is the curvature landscape that matters, not the height landscape (§ 4.2).

## 31.7.4 6.4. Junctions Are Organized By $^{\circ}X_9$

The extended caustic networks generated by the surface (6.1) are not part of the unfolding of any single catastrophe (except perhaps one with $K = \infty$ ). We are finding, however, that all fine structures yet discovered in
the junctions of caustic networks are sections of the universal unfolding of the compact subset $^{\circ}X_9$ of the catastrophe $X_9$ (Arnol'd [1975], Zeeman [1977], Callahan [1978]), which has codimension 8. Some plane sections of $X_9$ possessing discrete rotation symmetry were calculated by Lenz and Mecking [1961] (before the development of catastrophe theory) to simulate caustics in electron optics (see also Lenz [1956]).

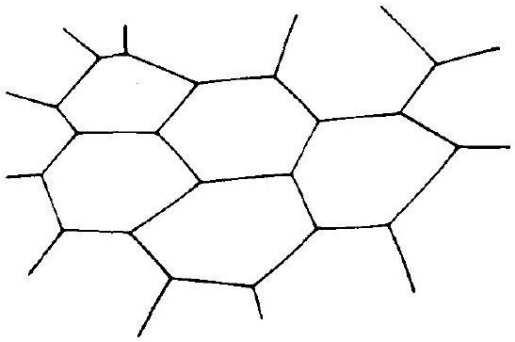


Fig. 6.6. Another common morphology of caustic network, unresolved.

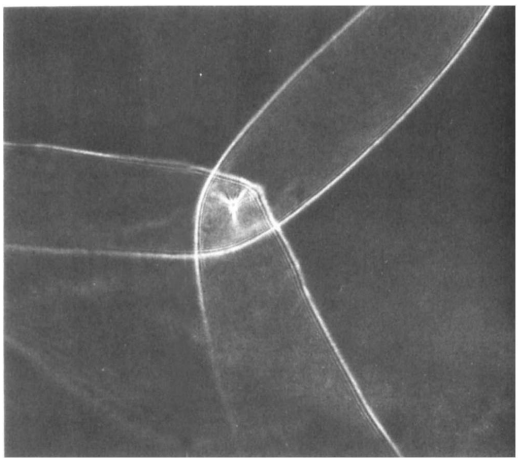


Fig. 6.7. An elliptic umbilic centred caustic triple junction resolved (courtesy of J. F. Nye).

It seems that the importance of ${}^{\circ}X_9$ in the context of caustic networks stems from the fact that it is the simplest catastrophe of corank 2 which is compact, i.e. for which the generating function $\Phi(s;C)$ has an absolute minimum, rising to $+\infty$ for any control point $C$, as $s \to \infty$ in all directions in state space. The generating function $\phi$ for an unbounded initial wavefront must also be compact because through any observation point there is an absolutely minimal ray (Poston [1980]). A non-compact $\phi$ obtained by expansion about some point can only give the local form of the caustic. By further expanding until $\phi$ is compact, we obtain the simplest possible models for significant units of caustic networks, such as junctions. Thus we expect such units to be organized (at the least) by ${}^{0}X_9$, and we are currently engaged in a comprehensive study of this. Of course, it is possible to obtain even higher compact catastrophes describing larger features of networks, such as several linked junctions.

## 31.8 § 7. Statistical Catastrophe Optics

### 31.8.1 7.1. Gaussian Random Wavefronts

The conditions under which light propagates are frequently specified not precisely but only in a statistical sense. Examples are starlight twinkling through a randomly turbulent atmosphere, and sunlight glittering on randomly rippling water. When the wave number $\kappa$ is large enough, catastrophe optics is applicable and gives a classification of caustics and diffraction catastrophes and a description of caustic networks. The additional feature of randomness immediately suggests questions of a statistical character. What is the average number of rays passing through a given point? What is the average density of umbilic foci in space? How do the wave

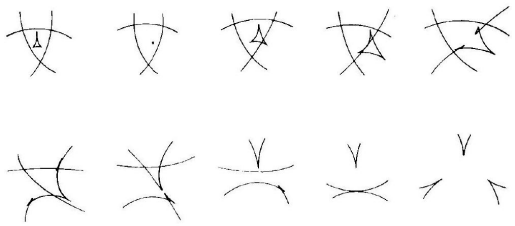


Fig. 6.8. Sequence of sections showing how the caustic junction of Fig. 6.7 is observed to evolve for decreasing $Z$.

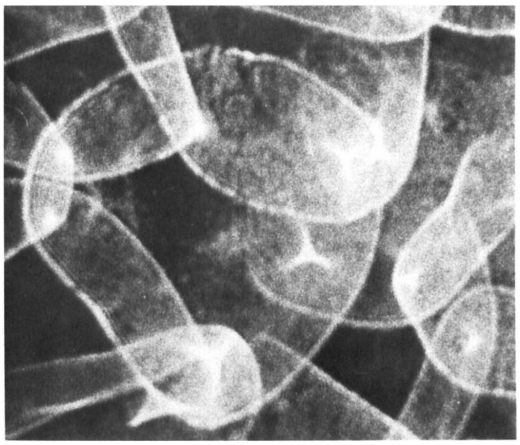


Fig. 6.9. Network of elliptic umbilic centred triple junctions in the real caustic of refraction from Pilkington's "Atlantic" glass.

intensity fluctuations depend on $\kappa$ ? The study of such problems is a promising new field of research, lying in the intersection of wave theory, topology, and probability.

It is of course necessary to choose an ensemble over which averages are to be taken; the appropriate choice depends on the physical system. In §§7.2 and 7.3 we restrict ourselves to the only case where progress has been made in elucidating the statistical geometry of rays and caustics: the paraxial optics of §4 with the height function $f$ of the initial wavefront $W$ chosen to be a Gaussian random function of $\boldsymbol{r}$. This is a superposition of infinitely many sinusoids with random phases, given by eq. (6.1) with $n \to \infty$; each choice of phases $\chi$ gives a sample $f$ in the ensemble and hence a sample family of rays and caustics. The mean value $\langle Q \rangle$ of any quantity $Q$ that depends on the choice of phases is defined as the average over all phases

$$\langle Q \rangle = \lim_{n\to\infty} \frac{1}{(2\pi)^n} \int_0^{2\pi} \mathrm{d}\chi_1 \cdots \int_0^{2\pi} \mathrm{d}\chi_n Q(\chi) \tag{7.1}$$

By standard noise theory (Rice [1944, 1945], Longuet-Higgins [1957]), $\langle f \rangle = 0, f$ and its derivatives are normally distributed, and odd and even derivatives are statistically independent (so that, for example, $\langle f f_x \rangle = 0$). The statistics of $f$ are completely determined by the autocorrelation function $\rho(\boldsymbol{r})$ or the power spectrum $E(\boldsymbol{k})$, defined by

$$\rho(\boldsymbol{r}) \equiv \langle f(\boldsymbol{r}_0) f(\boldsymbol{r}_0 + \boldsymbol{r}) \rangle \equiv \iint_{-\infty}^{\infty} \mathrm{d}\boldsymbol{k} E(k) \exp[\mathrm{i}\boldsymbol{k} \cdot \boldsymbol{r}] \tag{7.2}$$

(these functions do not depend on $\boldsymbol{r}_0$ ).

We begin in § 7.2 with the case where $W$ is corrugated, i.e. where $f$ depends only on one

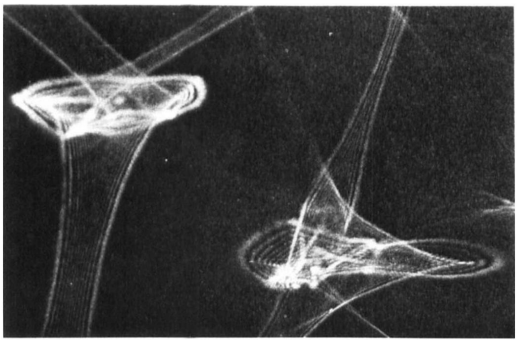


Fig. 6.10. Virtual caustic of refraction from Pilkington's "Atlantic" glass, showing hyperbolic umbilics.

coordinate $x$ so that the ray families are effectively two-dimensional; it is possible to obtain a full statistical description of the rays and caustics. The general case, where $f$ varies in two dimensions so that the ray families are three-dimensional, is considered in §7.3. Before the development of catastrophe theory, the topology of ray families was not understood. Nevertheless, Longuet-Higgins [1956, 1957, 1958, 1959, 1960a,b] pioneered techniques for evaluating the averages required to describe the statistical geometry, devoting considerable attention to the effects of the time-dependence of $W$. His methods were recently employed by Berry and Hannay [1977] to calculate the statistics of umbilic points. We confine ourselves in §§ 7.2 and 7.3 to statements of results, which were derived from the material in these papers. Unless otherwise indicated, the results are new.

In §7.4 we consider statistics that are not purely geometric: the wave intensity fluctuations. These scale with $\kappa$ according to universal laws, independent of any particular statistical ensemble or propagation model.

## 31.8.2 7.2. Statistical Geometry in Two Dimensions

If $W$ is corrugated, there is only one state variable $x$ and from eq. (4.3) the equation determining which ray(s) pass through the control point $X, Z$ (Fig. 7.1) is

$$X = x - Zf_x. \tag{7.3}$$

Stable caustics must be organised by the cuspoid catastrophes, since only these have corank one. For a typical $f$ in the ensemble, the $XZ$ plane contains fold lines, which from eq. (4.4) satisfy

$$Zf_{xx} = 1 \tag{7.4}$$

(as well as eq. (7.3)), and cusp points, which satisfy the additional condition

$$f_{xxx} = 0. \tag{7.5}$$

The autocorrelation $\rho$ (eq. (7.2)) depends only on $x$ and determines the statistical geometry via the quantities

$$F_1 \equiv \left[\left\langle \left(\frac{\mathrm{d}^i f}{\mathrm{d}x^i}\right)^2 \right\rangle\right]^{\frac{1}{2}} = \left[(-1)^i \frac{\mathrm{d}^{2i}}{\mathrm{d}x^{2i}}\rho(0)\right]^{\frac{1}{2}} \tag{7.6}$$

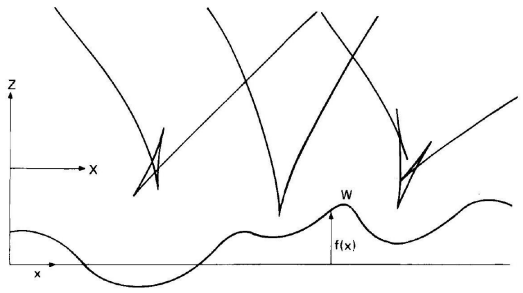



Fig. 7.1. Clusters of cusps in the caustics above a corrugated wavefront.

Consider first the most elementary average, namely the mean number of rays passing through a point at height $Z$. This is independent of the $X$ coordinate of the point, and will be denoted by $N_{\mathrm{r}}(Z)$. The rays must satisfy (7.3), so that

$$\begin{aligned} N_{\mathrm{r}}(Z) &= \left\langle \int_{-\infty}^{\infty} \mathrm{d}x \delta\left(X - x + Z f_x\right) \middle| 1 - Z f_{xx} | \right\rangle \\ &= \langle | 1 - Z f_{xx} | \rangle \end{aligned} \tag{7.7}$$

where the second equality follows from the statistical independence of $f_x$ and $f_{xx}$. Longuet-Higgins [1960a] evaluated the average over $f_{xx}$, obtaining

$$N_r(Z) = \operatorname{erf}\left[\left(Z F_2 \sqrt{2}\right)^{-1}\right] + \frac{2 Z F_2}{\sqrt{(2\pi)}} \exp\left[-\left(2 Z^2 F_2^2\right)^{-1}\right], \tag{7.8}$$

where erf denotes the error function. Thus $N_{\mathrm{r}}$ increases from 1 close to $W$ to $\sim Z$ far from $W$.

Next consider the mean linear density of fold crossings at height $Z$, denoted by $N_{\mathrm{f}}(Z)$. It is convenient to count the folds crossed in a very large distance $L$ and then transform from $X$ to $x$; bearing in mind that folds satisfy (7.4), we get

$$\begin{aligned} N_{\mathrm{f}}(Z) &= \lim_{L \to \infty} \left\langle \int_0^L \mathrm{d}x \delta\left(1 - Z f_{xx}\right) \middle| Z f_{xxx} | \right\rangle \\ &= \langle \delta\left(1 - Z f_{xx}\right) \rangle \langle | Z f_{xxx} | \rangle \end{aligned} \tag{7.9}$$

This reduces to

$$N_{\mathrm{f}}(Z) = \frac{F_3}{\pi F_2} \exp\left[-\left(2 Z^2 F_2^2\right)^{-1}\right]. \tag{7.10}$$

The asymptotic value, $F_3/\pi F_2$, is the density of inflections of $f$ from which the far field caustics originate.

Clearly, the final statistic in this hierarchy is the areal density of cusp points at height $Z$, denoted by $N_{\mathrm{c}}(Z)$. Arguments similar to those used to obtain $N_{\mathrm{f}}$, together with the cusp conditions (7.4) and (7.5), lead to

$$\begin{aligned} N_c(Z) &= \lim_{L \to \infty} \left\langle \int_0^L \mathrm{d}x \delta\left(f_{xxx}\right) \middle| f_{xxxx} \left| \delta\left(Z - 1/f_{xx}\right) \right\rangle \right. \\ &= \langle \delta\left(f_{xxx}\right) \rangle \langle \delta\left(f_{xx} - 1/Z\right) | f_{xxxx} | \rangle / Z^2 \end{aligned} \tag{7.11}$$

This gives

$$N_c(Z) = \frac{2F_3\sqrt{(\mathcal{N}^2-1)}\exp\left[-\left(2Z^2F_2^2\right)^{-1}\right]}{Z^2F_2^2(2\pi)^{\frac{3}{2}}} \times\left\{\exp\left[-\left(2Z^2F_2^2\left(\mathcal{N}^2-1\right)\right)^{-1}\right] + \frac{1}{ZF_2}\left(\frac{\pi}{2\left(\mathcal{N}^2-1\right)}\right)^{\frac{1}{2}}\operatorname{erf}\left[\left(ZF_2\sqrt{\left(\mathcal{N}^2-1\right)^{-1}}\right]\right\}\right. \tag{7.12}$$

where $\mathcal{N}$ is the quantity

$$\mathcal{N} \equiv F_4F_2/F_3^2 \tag{7.13}$$

whose significance will soon be apparent. $N_c$ vanishes near to and far from $W$, reaching a maximum for all $\mathcal{N}$ when $Z \sim F_2^{-1}$; not surprisingly, the cusps are most dense near the level where the centres of curvature of $W$ are most dense.

Globally, the fold lines consist of branches beginning and ending in the far field (Fig. 7.1). Each branch contains a cluster of an odd number of cusps, corresponding to the zeroes of $f_{xxx}$ between successive zeroes of $f_{xx}$. It is not hard to show that the mean number of cusps in a cluster is

$$\frac{2}{N_{\mathrm{f}}(\infty)}\int_0^\infty \mathrm{d}Z N_{\mathrm{c}}(Z) = \mathcal{N} \tag{7.14}$$

where $\mathcal{N}$ is also given by eq. (7.13).
One gets a very clear picture of the statistical geometry of the caustics by evaluating $\mathcal{N}$ for different functions $\rho$. Since $W$ is smooth, $\rho$ must be an even function of $x$, and so

$$\rho(x) = \sum_{n=0}^{\infty} a_n x^{2n}. \tag{7.15}$$

Then (7.13) and (7.6) give

$$\mathcal{N} = \left(\frac{28}{15}\left|a_2a_4/a_3^2\right|\right)^{\frac{1}{2}} \tag{7.16}$$

The simplest case is the (nonrandom) sinusoidal wavefront, for which $\rho = F_0^2\cos\alpha x$ and (7.16) gives $\mathcal{N} = 1$. If the correlations decay in a single-scale Gaussian manner, so that $\rho = F_0^2\exp\left[-\alpha x^2\right]$, then $\mathcal{N} = \sqrt{\frac{7}{5}} = 1.183$; if the decay is Lorentzian, so that $\rho = F_0^2/\left(1+\alpha x^2\right)$, then $\mathcal{N} = \sqrt{\frac{28}{15}} = 1.366$. Bearing in mind that the number of cusps must be odd, we see that most clusters contain just one cusp. For wavefronts with two or more scales, however, it is not difficult to construct cases where $\mathcal{N}$ is arbitrarily large.

## 31.8.3 7.3. Statistical Geometry in Three Dimensions

If $W$ is curved in two dimensions, there are two state variables $\boldsymbol{r} \equiv (x, y)$. The rays in control space $X, Y, Z$ satisfy (4.3) and the caustics satisfy (4.4) as well. For a generic $f$ the highest singularities are elliptic and hyperbolic umbilic foci, and these satisfy the additional conditions (4.7).

In the most general case, the correlations on $W$ will be anisotropic. Here, however, we restrict ourselves to the case of isotropic disorder, where $\rho$ and $E$ (eq. (7.2)) are functions only

of the lengths $r$ and $k$ of the vectors $\mathbf{r}$ and $\boldsymbol{k}$. The statistical geometry is described in terms of the moments

$$M_{\mathrm{j}} \equiv 2\pi \int_0^\infty \mathrm{d}k k^{j+1} E(k) \tag{7.17}$$

Note that $M_4 = \left\langle \left(\nabla^2 f\right)^2 \right\rangle$.
The mean number of rays passing through a point at height $Z$ is given by (cf (7.7))

$$\begin{aligned} N_{\mathrm{r}}(Z) &= \left\langle \iint_{\mathrm{w}} \mathrm{d}\boldsymbol{r} \delta(\boldsymbol{R} - \boldsymbol{r} + Z\nabla f) \left| \det \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{r}} (\boldsymbol{r} - Z\nabla f) \right| \right\rangle \\ &= \left\langle \left| 1 - Z\nabla^2 f + Z^2 \left( f_{xx} f_{yy} - f_{xy}^2 \right) \right| \right\rangle \end{aligned} \tag{7.18}$$

This average was evaluated by Longuet-Higgins [1960a] as an estimate of the number of reflected images of the sun seen from different heights above the sea. He obtained the result

$$N_{\mathrm{r}}(Z) = 1 + \frac{M_4 Z^2}{2\sqrt{3}} \exp\left[ -4/\left( 3M_4 Z^2 \right) \right]. \tag{7.19}$$

Thus $N_{\mathrm{r}}$ increases from 1 close to $W$ to $\sim Z^2$ far from $W$.
In a systematic development, the next statistic would be the average number of times a line of unit length pierces fold surfaces, or the average length of fold cutting unit area, followed by the average number of times unit area is pierced by ribs, and the spatial density of swallowtail points. However, all these cuspoid statistics are extraordinarily difficult to calculate. One result which can be derived relatively easily is that as $Z \to \infty$ the mean length of fold cutting unit area is proportional to $Z$; this explains the confusion of caustic lines in the far field of extended wavefronts, noted in § 6, and contrasts with the two-dimensional case for which the corresponding quantity $N_{\mathrm{f}}(Z)$ (eq. (7.10)) saturates as $Z \to \infty$.

Fortunately, it is not difficult to calculate the spatial density of umbilic foci at height $Z$, denoted by $N_{\mathrm{u}}(Z)$. By analogy with (7.11), use of (4.7) and (4.4) gives

$$\begin{aligned} N_{\mathrm{u}}(Z) = \lim_{\mathrm{A}\to\infty} \Bigg\langle \iint_{\mathrm{A}} \mathrm{d}\boldsymbol{r} \delta\left(f_{xx} - f_{yy}\right) \delta\left(f_{xy}\right) \\ \times \left| \det \frac{\partial\left(f_{xx} - f_{yy}, f_{xy}\right)}{\partial(x,y)} \right| \Bigg\rangle \\ = \left\langle \delta\left(f_{xx} - f_{yy}\right) \delta\left(f_{xy}\right) \delta\left(f_{xx} - 1/Z\right) \right\rangle \\ \times \left\langle \left| f_{xxx} f_{xyy} - f_{xyy}^2 + f_{xxy} - f_{xxy}^2 \right| \right\rangle / Z^2. \end{aligned} \tag{7.20}$$

This leads to the result

$$N_{\mathrm{u}}(Z) = \frac{M_6 \exp\left[ -2/\left( M_4 Z^2 \right) \right]}{\left( 2\pi M_4 \right)^{\frac{3}{2}} Z^2}. \tag{7.21}$$

$N_{\mathrm{u}}$ vanishes near to and far from $W$, and reaches a maximum at $Z = \sqrt{(2/M_4)}$, showing that umbilic foci are most dense where the centres of principal curvature of $W$ are most dense.

The density of umbilic points on $W$ is simply

$$\int_{-\infty}^{\infty} N_u(Z) \mathrm{d}Z = M_6 / \left( 4\pi M_4 \right) \tag{7.22}$$

On the average, equal numbers of these umbilics have indices $+\frac{1}{2}$ and $-\frac{1}{2}$; this follows from the second index rule in § 4.3. However, as suggested by the inequalities in § 4.3, elliptic umbilics

are less common than hyperbolic ones. By calculating an average involving the catastrophe classification function (4.9), Berry and Hannay [1977] showed that the fraction $\alpha_E$ of umbilics of elliptic type is

$$\alpha_{\mathrm{E}} = \frac{9}{50} + \frac{8}{\pi}\int_{\frac{1}{3}}^{1} \mathrm{d}u \frac{u\,(1-u^2)}{(1+u^2)^3} \arcsin\left[\left\{\frac{(1-u)^2(3u+1)}{16u^3}\right\}^{\frac{1}{2}}\right] = 0.268 \tag{7.23}$$

This result does not depend on the nature of the correlations on $W$, provided these are isotropic.

## 31.8.4 7.4. Random Diffraction Catastrophes in Twinkling Light

A random structure $S$ produces random caustics, as in the special case just discussed. Associated with these caustics are random waves $\psi$. The statistics of $\psi$ as $\kappa \to \infty$, in particular the probability distribution of the intensity $|\psi|^2$, are not well understood theoretically. Experimentally, it is convenient to measure the moments $\mathscr{I}_m$ of the fluctuating intensity, defined by

$$\mathscr{I}_m \equiv \left\langle |\psi|^{2m} \right\rangle, \tag{7.24}$$

where ( ) denotes averaging over the ensemble of $S$. If $\psi$ were a Gaussian random function of position and/or time, $\mathscr{I}_m$ would be equal to $\mathscr{I}_1 m\,!$ for all cases considered here. For large $\kappa$, however, the intensity fluctuations are much stronger than those of a Gaussian wave, even when the randomness of $S$ is Gaussian as in the model of §7.1; these strong fluctuations were measured by Jakeman, Pike and Pusey [1976] in twinkling starlight.

Conventional techniques based on perturbation theory (Uscinski [1977]) fail completely for strong fluctuations. Elaborate analysis for paraxial wave propagation from a Gaussian random wavefront (§4 and § 7.1), by Shishov [1971], Buckley [1971a,b] and Jakeman and McWhirTER [1977] has led to an understanding of $\mathscr{I}_2$ and shown that $\mathscr{I}_2 \sim \ln \kappa$ as $\kappa \to \infty$. Therefore $\mathscr{I}_2$, and a fortiori the higher moments, diverge in the geometrical-optics limit.

As has long been known, the non-Gaussian strong fluctuations in $\psi$ and associated divergences of $\mathscr{I}_m$ are due to focusing (Mercier [1962], Salpeter [1967]). When $\kappa = \infty, |\psi|$ is infinite on the caustics of any ray family in the ensemble; during a time, space or ensemble average these infinities pass across the radiation detector. It follows from conservation of energy, and can also be shown directly, that these infinities in the intensity $|\psi|^2$ are integrable, so that $\mathscr{I}_1$ is finite. But higher powers of $|\psi|^2$ have non-integrable divergences, so that $\mathscr{I}_{m\geq 2}$ diverge. Prior to catastrophe theory, the elaboration of this idea was frustrated by lack of understanding of the geometry of caustics.

In practice several factors prevent $\mathscr{I}_m$ being infinite. The source may be spatially and temporally incoherent, as with the finite size and polychromaticity of the sun, and this blurs caustic networks on the sea bed (§ 6). Most fundamental, however, is the finite value of $\kappa$, which causes the divergences of $\mathscr{I}_m$ to be softened by diffraction.

It was suggested by Berry [1977] that a natural measure of this effect is the set of critical exponents $\nu_m$, defined by

$$\nu_m \equiv \lim_{\kappa\to\infty} \mathrm{d}\left(\log \mathscr{I}_m\right) / \mathrm{d}(\log \kappa). \tag{7.25}$$

Up to logarithmic factors such as appears in $\mathscr{I}_2$, the existence of the $\nu_m$ implies the scaling laws

$$\mathscr{I}_m \sim \kappa^{\nu_m}. \tag{7.26}$$

The $\nu_m$ can be determined by catastrophe theory; we give a simplified version of the argument presented by Berry [1977].

The main result is that as $m$ increases, the exponents depend on catastrophes of ever higher codimension $K$. To see that this must be so, consider the familiar case where the $\mathscr{I}_m$ are measured by time-averaging the intensity at fixed position $\mathscr{R}$. Over long times, diffraction catastrophes of arbitrarily large $K$ will pass arbitrarily close to $\mathscr{R}$. The higher catastrophes are rare but cause large localized fluctuations in $|\psi|^2$. But large rare fluctuations are precisely what dominate high moments $\mathscr{I}_m$, which therefore depend on catastrophes of large $K$.

To calculate the $\nu_m$, a scaling argument is employed. This depends crucially on the fact that the ensemble of $S$ is parameterized by a large number of variables $C$ (e.g. the phases $\chi$ in §7.1). $\psi$ depends on $C$, which therefore constitutes a tremendous augmentation of control space, fully explored during the ensemble averaging for $\mathscr{I}_m$ which (cf eq. (7.1)) is given by

$$\Phi_m = \int \cdots \int \mathrm{d}C P(C) |\psi(C)|^{2m} \tag{7.27}$$

where $P(C)$ is the density of realizations of $S$ over the ensemble. For large $\kappa$ the integral will be dominated by those values of $C$ lying on caustics. $\mathscr{I}_m$ will be the sum of contributions $\mathscr{I}_{mj}$ from the different catastrophes in the hierarchy of § 2.2, here labelled $j$. Each contribution scales differently with $\kappa$ and can be written

$$\mathscr{I}_{mj} \sim \kappa^{\nu_{mi}}, \tag{7.28}$$

where $\nu_{mj}$ is thus the exponent governing the contribution of the $j$ th catastrophe to the $m$ th moment.

$\nu_{mj}$ is determined by two quantities describing the architecture of the $j$ th diffraction catastrophe as discussed in § 3.4, namely the singularity index, here written $\beta_j$ and the fringe index, here written $\gamma_j$. The contribution $\mathscr{I}_{mj}$ is estimated from eq. (7.27) as follows:

$$\begin{aligned} \mathscr{I}_{mj} &\sim (\max |\psi|)^{2m} \times \begin{pmatrix} \text{hypervolume of diffraction} \\ \text{maximum in control space} \end{pmatrix} \\ &\sim \kappa^{2m\beta_i} \kappa^{-\gamma_i}, \end{aligned} \tag{7.29}$$

whence (7.28) gives

$$\nu_{mj} = 2m\beta_j - \gamma_j. \tag{7.30}$$

Eqs. (7.28) and (7.30) establish how each catastrophe contributes to each moment. Summing over all catastrophes gives the $m$ th moment as

$$\mathscr{I}_m \sim \sum_j B_{mj} \kappa^{\nu_{mj}} \quad \text{as} \quad \kappa \to \infty, \tag{7.31}$$

where $B_{mj}$ are constants. The dominant term is clearly the one with the largest $\nu_{mj}$, so that from (7.26) the critical exponents are given by

$$\nu_m = \max_j \left(2m\beta_j - \gamma_j\right). \tag{7.32}$$

Eq. (7.32) is the main result. Calculation of the $\nu_m$ (a tedious exercise) shows that the maximum over $j$ does actually exist: for fixed $m, \nu_{mj}$ first increases and then decreases as catastrophes with increasing $K$ are considered. The resulting $\nu_m$ are sets of rational numbers depending only on the dimensionality of the space in which $\psi$ propagates. The reason for this is that although $K$ can be arbitrarily large, the corank (§ 2.2) cannot exceed the dimensionality of

the wavefronts. Therefore for waves in the plane only catastrophes of corank one (cuspoids) can contribute, while for waves in space only the catastrophes of corank one or two can contribute; the extra singularities make $\nu_m$ larger in the latter case. Apart from this dimension dependence, the $\nu_m$ are universal: they do not depend on the details of $S$, only on the fact that it is smooth and its randomness is described by many parameters $C$.

Table 3 shows the first few $\nu_m$, and the dominant catastrophes. The value $\nu_2 = 0$ does not mean that $\mathscr{I}_2$ is not singular as $\kappa \to \infty$; in fact, as already mentioned $\mathscr{I}_2 \sim \ln \kappa$. These results constitute testable predictions

Table 3
Critical exponents $\nu_{m} and dominant catastrophes for $2 \leq m \leq 5$

| $m$ | $\nu_m$ (waves in the plane) | Dominant catastrophe | $\nu_m$ (waves in space) | Dominant catastrophe(s) |
|---|---|---|---|---|
| 2 | 0 | $A_2$ | 0 | $A_2$ |
| 3 | 1/3 | $A_2$ | 1/3 | $A_2$ and $D_4$ |
| 4 | 3/4 | $A_3$ | 1 | $D_4$ |
| 5 | 5/4 | $A_3$ | 5/3 | $D_4$ and $E_6$ |

about the wavelength dependence of the moments $\mathscr{I}_m$ of random short waves. Usually $\mathscr{I}_m$ are measured as functions of other parameters, such as distance from a turbulent medium, or strength of turbulence, but it should be possible for $\nu_m$ to be measured directly on the basis of the definition (7.25).

For $m > 5$, serious difficulties arise in the calculation of $\nu_m$ for waves in space because of the appearance of catastrophes with modality (Arnol'd [1975]). Making plausible assumptions, Berry [1977] was able to calculate critical exponents up to $\nu_{13}$.

## 31.9 Appendix 1: Wavefronts

It is natural to define wavefronts as surfaces of constant phase. There are, however, two ways of defining phase, leading to two sorts of wavefront whose geometric properties - in particular their singularities are very different. Firstly, there are diffraction wavefronts, defined in terms of the phase $\kappa\chi$ of the exact complex scalar wave function $\psi$ (eq. (3.2)) by

$$\kappa\chi(\mathscr{R}) = A \mod 2\pi, \tag{A1.1}$$

where $A$ is constant. Secondly, there are geometrical wavefronts, defined in terms of the phases $\kappa\mathscr{S}_\mu$ of contributions to the short-wave approximation to $\psi$ (eq. (3.7)) by

$$\kappa\mathscr{S}_\mu(\mathscr{R}) = B, \tag{A1.2}$$

where $B$ is constant; geometrical wavefronts are contour surfaces of action, i.e. of solutions of the Hamilton-Jacobi equation (2.4). We shall give a brief description of these two sorts of wavefronts, emphasizing their differences.

Diffraction wavefronts, defined by given values of $A$ in eq. (A1.1), are surfaces in space (or lines in the plane), intersecting neither themselves nor wavefronts with different values of $A$. They have singularities where the function $\chi(\mathscr{R})$ is singular. Since $\psi$ is a smooth function, $\chi$ may be singular only where the amplitude $a$ in eq. (3.2) is zero. Because two equations must be satisfied ( $\operatorname{Re}\psi = \operatorname{Im}\psi = 0$ ), singularities of diffraction wavefronts generically take the form of lines in space, called wavefront dislocations by analogy with similar structures in crystals (Nye and Berry [1974]). On a dislocation line, infinitely many diffraction wavefronts join. During a

circuit around a dislocation line, the phase $\kappa\chi$ changes by $2n\pi$, where $n$ is an integer defining the dislocation strength. As explained in detail by Nye and Berry [1974], wavefront dislocations may be of edge or screw type, they may be curved and, when $\psi$ is time-dependent, they may interact in a variety of ways. Near a dislocation line, the vector field
$\nabla\chi$ has a vortex structure, a feature emphasized (in the context of quantum-mechanical waves) by Hirschfelder, Christoph and Palke [1974], Hirschfelder, Goebel and Bruch [1974], and Hirschfelder and Tang [1976a,b] (see also Dirac [1931] and Riess [1970a,b, 1976]).

As explained above, wavefront dislocations are singularities derived from the complex function $\psi$. An alternative approach, in which dislocations are studied in terms of the topography of the function $\mathrm{Re}\,\psi(\mathscr{R})$ whose stable features are analyzed using catastrophe theory, has been explored by Wright [1979].

Considered as wave morphologies (Berry [1979a]), wavefront dislocations are complementary to caustics, for the following reasons: at a dislocation the wave amplitude is zero, while on a caustic the geometricaloptics amplitude (§3.1) is infinite; observation of a dislocation requires discrimination of wave structure on the scale of $\lambda$, a circumstance in which caustics are blurred by diffraction (§ 3.3), whereas observation of a caustic takes place on scales large compared with $\lambda$, a circumstance in which phase details, and hence dislocations, are hard to discern. This complementarity is exemplified by the diffraction catastrophes described in § 3.3, all having caustics as their grossest features and (except the fold) dislocations as their finest details. However, dislocations are by no means associated only with caustics, because they occur generically in cases outside the scope of catastrophe optics where $\lambda$ is not small in comparison with other relevant scales (see, for example, Born and Wolf [1975] p. 576). In general, caustics are the singularities of ray theory, whereas dislocations are the singularities of wave theory.

We now turn to the geometrical wavefronts, defined by given values of $B$ in eq. (A1.2). These are surfaces in space, with multiple branches labelled by the index $\mu$, which corresponds to the different rays passing through each point $\mathscr{R}$ (§ 2.1). Because of this multiplicity, several geometrical wavefronts can pass through $\mathscr{R}$, as shown in Fig. 2.1a where they intersect in pairs, and in Fig. 2.1b where they intersect in threes. Also illustrated in Fig. 2.1b is the fact that a geometrical wavefront may self-intersect.

The geometrical wavefronts are singular where two or more branches $\mathscr{S}_\mu$ coincide; this occurs on the caustics of the associated ray family (§ 2.1). An individual wavefront in space generically has line singularities, but the family of wavefronts (for all $B$ ) has as its singularity the whole caustic surface. These wavefront singularities are called Legendre singularities, for mathematical reasons explained by Arnol'd [1975] and

Sewell [1977, 1978]. They form a hierarchy corresponding to the hierarchy of catastrophes. If the caustic is a cuspoid catastrophe of codimension $K$, the wavefronts have the form of sections through the singular set of a cuspoid catastrophe of codimension $K + 1$. Thus, for example, if the caustic is a fold line in the plane, the wavefronts are cusped lines (Fig. 2.1a), while if the caustic is a cusp in the plane the wavefronts are sections of a swallowtail (Fig. 2.1b and Wright [1979]). A general proof of this result for cuspoids is given by Sewell [1977]. But if the caustic is an umbilic catastrophe, the wavefronts are not sections of higher catastrophes, and their forms have to be computed ad hoc (Sewell [1978]). Pictures of sequences of wavefronts for all stable caustics that can occur in space are shown by Arnol'd [1976] and Poston and Stewart [1978] p. 280.

It is evident that geometrical and diffraction wavefronts, and their singularities, are very different in character. In monochromatic waves, the geometrical wavefronts are of little physical significance, whereas the diffraction wavefronts are related to observable features of the wave function. For shock waves, however, the geometrical wavefronts are important because they correspond to propagating discontinuities of the field; for a brief account and references see

Poston and Stewart [1978]. For the intermediate case of quasi-monochromatic pulses, the gross structure is of a geometrical wavefront but there are dislocations on fine scales (Wright [1977], Humphrey, Nye and Wright, in preparation).

## 31.10 Appendix 2: Primary Aberrations

Consider the caustic in the image space of an optical system with rotation symmetry, for a point source of monochromatic light. This caustic can be classified both in terms of catastrophe theory and in terms of the traditional aberrations. Our purpose here is to describe the relationship between these two classifications for the simplest case: primary geometrical (Seidel) aberrations. Most common is the so-called third order theory, in which $\sin\theta$ is approximated by $\theta - \theta^3/3$ ! in the ray formulae, but we will use the equivalent procedure of calculating the optical path length to fourth order in the paraxial geometrical optics of §4. Our analysis is based on unpublished notes by Dr. J. H. Hannay.

The wavefront $W$ emerging from a general optical system with rotation symmetry about the $Z$ axis will be specified, as in §4, by its deviation $f(\boldsymbol{r})$
from the plane $Z = 0$.$f$ depends parametrically on the source point $\rho$ (Fig. A2.1), the five Seidel aberration coefficients $B$ (spherical aberration), $C$ (astigmatism), $D$ (curvature), $E$ (distortion) and $F$ (coma), and on coefficients $G$ and $H$ describing the shape of the aberration-free wavefront. Rotation symmetry requires that $f$ be a function only of invariants under co-rotation of $\boldsymbol{r}$ and $\boldsymbol{\rho}$, and this gives, on choosing signs and numerical factors to agree with the usual development of aberration theory (Born and Wolf [1975] chap. 5),

$$f(\boldsymbol{r}) = \frac{1}{4}Br^4 + C(\boldsymbol{\rho}\cdot\boldsymbol{r})^2 + \frac{1}{2}D\rho^2 r^2 - E\rho^2(\boldsymbol{\rho}\cdot\boldsymbol{r}) - F(\boldsymbol{\rho}\cdot\boldsymbol{r})r^2 - G\boldsymbol{\rho}\cdot\boldsymbol{r} - Hr^2, \tag{A2.1}$$

where $r$ denotes $|\boldsymbol{r}|$ and $\boldsymbol{\rho}$ denotes $|\boldsymbol{\rho}|$. Taking coordinates $\boldsymbol{\rho} \equiv (\xi, \eta)$ and $\boldsymbol{r} \equiv (x, y)$, and choosing $\boldsymbol{\xi} = 0$ without loss of generality, we obtain

$$\begin{aligned} f(x,y) =& \frac{1}{4}B\left(x^4 + 2x^2y^2 + y^4\right) + C\eta^2y^2 + \frac{1}{2}D\eta^2\left(x^2 + y^2\right) \\ & - E\eta^3 y - F\eta\left(x^2y + y^3\right) - G\eta y - H\left(x^2 + y^2\right). \end{aligned} \tag{A2.2}$$

The most important features of $W$ are its umbilic points (§ 4.3), which occur where the conditions (4.7) are satisfied. A short calculation shows that there are two umbilics whose relative positions depend on the
quantity

$$Q = \frac{\eta^2}{B}\left(\frac{F^2}{B} - C\right), \tag{A2.3}$$

and whose coordinates are

$$\left.\begin{array}{ll} x = 0, & y = F\eta/B \pm \sqrt{Q} \\ x = \pm\sqrt{|Q|}, & \text{for} \end{array} \qquad Q > 0\right\}. \tag{A2.4}$$

Substitution of third derivatives of $f$, evaluated at these points, into the catastrophe classification (4.9) shows that the umbilics are always hyperbolic.

The wavefront (A2.2) has mirror symmetry about the lines $x = 0$ and $y = F\eta/B$, which are thus necessarily riblines (§4.4): contours of curvature must cross mirror lines normally - or be straight along them, which is clearly not possible in this case-and are therefore tangent to that principal axis of curvature which is normal to the mirror line, the other principal axis necessarily

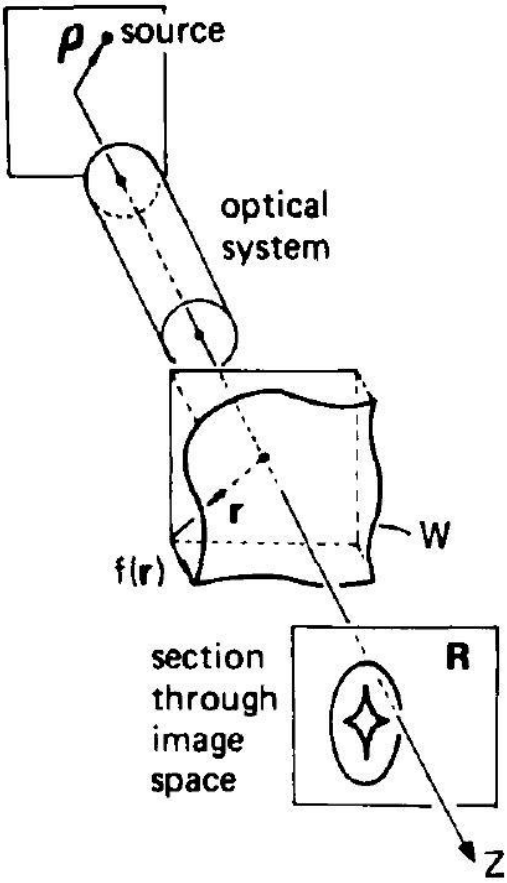



Fig. A2.1. Rotationally symmetric optical system; a source at $\rho$ produces a wavefront $W$, giving rise to a caustic in the image space ( $\boldsymbol{R , Z$ ).}

being along the mirror line. Hence it is not difficult to establish that the structurally stable form of the caustic is the sequence of sections $Z =$ constant shown in Fig. A2.2; the direction of increasing $Z$ depends on the sign of $B$ in (A2.1). The only catastrophes of codimension three are the two hyperbolic umbilic foci: there are no elliptic umbilic foci and no swallowtails. Note that in practice not all of the caustic need be seen in the image space; part of it may be virtual, and part of it may be missing due to aperture effects.

Clearly the curvature and distortion coefficients $D$ and $E$ do not affect the form of the caustic but merely its spatial position, and henceforth we ignore them. If the other aberration coefficients $B, C$ and $F$ are all zero,

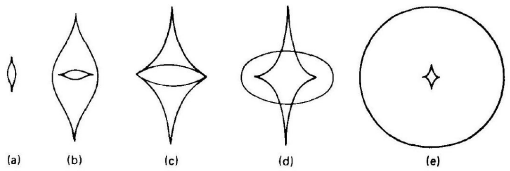



Fig. A2.2. Structurally stable caustic of primary aberration; sequence of sections $Z =$ constant showing (a) lips event, (b) conjugate lips event, (c) two hyperbolic umbilic foci, (d) and (e) unfolding into conical outer caustic and four-cusped "needle".

the caustic is a perfect point focus, which is structurally unstable. If the astigmatism coefficient $C$ alone is not zero, the caustic consists of two straight orthogonal focal lines in

different planes, which again is structurally unstable. In the generic case where $B$ and $F$ are non-zero, these focal lines become the ribs of Fig. A2.2, and the two umbilic foci come into the field from infinity.

The behaviour of the umbilics is determined principally by $Q$ (eq. (A2.3)). When $Q = 0$, we have "especially good" imaging; both umbilics and both lips events are coincident, the inner four-cusped needle in Fig. A2.2 collapses to a degenerate line caustic and the outer caustic surface is a rotationally symmetric cusped cone (Fig. A2.3). This caustic is not structurally stable, and corresponds to a catastrophe of infinite codimension. It occurs not only for $C = F = 0$, in which case the only contribution to the caustic is from the term with spherical aberration coefficient $B$, but also for $F^2 = BC$, which is the transition between the two possible orientations of the caustic of Fig. A2.2.

If $B = 0$, one umbilic point is at the origin and the other at infinity in the $\boldsymbol{r}$ plane. The generating function $\phi$ of eq. (4.2) is simply the hyperbolic umbilic, and may be put into the normal form $\Phi$ of Table 1 by the transformation $x \to x - y, y \to (x + y)/\sqrt{3}$. This is the case of primary coma, whose coefficient $F$ multiplies the germ of the hyperbolic umbilic. Note that $C$ need not be zero-the degree of astigmatism simultaneously present does not change the topology of the caustic in this case. The existence of hyperbolic umbilic foci does not, however, depend on there being a nonvanishing coma coefficient; if $F$ alone is zero, then $Q$ is nonzero and the effect of astigmatism ( $C \neq 0$ ) in addition to spherical aberration ( $B \neq 0$ ) is to break the degeneracy of the caustic in Fig. A2.3 to give that of Fig. A2.2.

The geometrical caustics will of course be clothed with diffraction, as described for the stable caustics of Fig. A2.2 by the diffraction catastrophes of § 3.3 and for the unstable caustic of Fig. A2.3 in chapter 9 of Born and Wolf [1975].

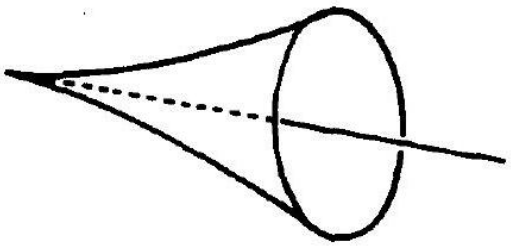


Fig. A2.3. Structurally unstable caustic of primary aberration when the two umbilics are coincident.

# Part X
# Dead, Useless, Pointless Theories in Optics

## 32 Different Ideas

### 32.1 Some Hypothetical Ideas

Думая про свет, можно придумать много идей, которые совершенно далеки от реальности.

Обсудим тут их, потому что мб потом интересно будет их развить, а мб и создать отдельную модель.

Мб даже что-то полезное от них можно получить будет, в любом случае, сперва то, что на самом деле важно.

#### 32.1.1 Different Ideas

(мб когда-то подумаю про них, но вряд ли скоро)

**а если свет неотделим от пространства?**

хз

### 32.2 Ideas from History

(see books about philosophy about it)

### 32.3 Philosophical Ideas about Light

(there were a lot of analogies and ideas, I'll write them here and discuss, how meaningful are they?)

### 32.4 Special Approach to Basic Propagation of Waves

#### 32.4.1 The corner eikonal by Landau

(that shitty theory from Landau)

**Суть углового эйконала**

(это специфика от ЛЛ2, которая нигде не нужна, потому что используется только для ввввода одной формулы, которую можно и другими путями получить. пока так думаю)

**Theory**

ИдущИй в пустоте луч света, попадая в какое-либо прозрачное материальное тело, имеет по выходе из этого тела направление, вообще говоря, отличное от первоначального.

Это изменение направления зависит, конечно, от конкретных свойств тела и от его формы.

ОКазывается, однако, возможным вывести некоторые общие законы, относящиеся К изменению направления лучей света при прохождении через проИзвольные материальные тела.

При этом предполагается только, что длЯ лучей, распространяющихся внутри рассматриваемого тела, имеет место геометрическая оптика.

Такие прозрачные тела, через которые пропускают лучи света, мы будем называть, как это принято, оптичеCKuMU cucmemanu.

В силу указанной в & 53 аналогии между распространением лучей и движением частицы, те же общие законы справедливы и для изменениЯ направления движения частиЦ, двигавшихся сначала прямолинейно в пустоте, затем проходящих через какое-нибудь электромагнитное поле и снова выходящих из этого поля в пустоту.

ДлЯ определенности мы будем, однако, ниже говорить о распространении лучей света.

Мы видели, что уравнение эйконала, определяющее распространение лучей, может быть написано (для света с определенной частотой) В виде (53.11). Ниже мы будем для удобства обозначать посредством $\psi$ эйконал $\psi_0$, деленный на постоянную величину $\omega/c$. Тогда основное уравнение геометрической оптики будет иметь вид

$$(\nabla\psi)^2 = 1$$

Каждое решение этого уравнения описывает собой определенный пучок лучей, причем направление луча, проходящего через данную точку пространства, определяется градиентом $\psi$ в этой точке.

ДлЯ наШИХ целей, однако, такое описание недостаточно, поскольку мы иЩем общие соотношения, определяющие прохождение через оПтические системы не какого-либо одного определенного пучка лучей, а соотношенИЯ, относящиеся к любым лучам.

Поэтому мы должны пользоваться эйконалом, взятым в таком виде, в котором он описывал бы все вообШе возможные лучи света, т.е. лучи, проходящие через любую пару точек в пространстве.

В обычной своей форме эйконал $\psi(\mathbf{r})$ есть фаза луча из некоторого пучка, проходящего через точку r.

Теперь же мы доЛЖНы ввести эйконал как фунКцию $\psi(\mathbf{r}, \mathbf{r}')$ коорДинат двух точек ($\mathbf{r}, \mathbf{r}'-$ радиус-векторы начальной и конечной точек луча).

Через всякую пару точек $\mathbf{r}, \mathbf{r}'$ можно провести луч, и $\psi(\mathbf{r}, \mathbf{r}')$ есть разность фаз (или, как говорят, оптическая длина пути) этого луча между точками r и $\mathrm{r}'$.

Ниже мы будем везде подразумевать под r и $\mathrm{r}'$ радиус-векторы точек на луче соответственно до и после его прохождения через оптическую chctemy.

Если В $\psi(\mathbf{r}, \mathbf{r}')$ один из радиус-векторов, скажем $\mathbf{r}'$, считать заданНым, то $\psi$ как функция от r будет описывать определенный пучОК лучей, а именно пучок лучей, проходящих через точку $\mathbf{r}'$. Тогда $\psi$ ДОлжно удовлетворять уравнению (55.1), в котором дифференцирование производится по компонентам r.

Аналогично, считая r заданным, находим еще одно уравнение длЯ $\psi(\mathbf{r}, \mathbf{r}')$, так что

$$(\nabla\psi)^2 = 1, \quad (\nabla'\psi)^2 = 1$$

Направление луча определяется градиентом его фазы.

Поскольку $\psi(\mathbf{r}, \mathbf{r}')$ есть разность фаз в точках $\mathbf{r}'$ и $\mathbf{r}$, то направление луча в точке $\mathbf{r}'$ определяется вектором $\mathbf{n}' = \partial\psi/\partial\mathbf{r}'$, а в точке $\mathbf{r}$ - вектором $\mathbf{n} = -\partial\psi/\partial\mathbf{r}$. Из (55.2) видно, что векторы $\mathbf{n}$ и $\mathbf{n}'$ единичные:

$$\mathrm{n}^2 = \mathrm{n}'^2 = 1$$

Четыре вектора $\mathrm{r}, \mathrm{r}', \mathrm{n}, \mathrm{n}'$ связаны между собой некоторым соотноШением, поскольку два из них $(\mathbf{n}, \mathbf{n}')$ являются производными по двум другим $(\mathrm{r}, \mathrm{r}')$ от некоторой функции $\psi$. Что

касается самой функции $\psi$, то она удовлетворяет дополнительным условиям - уравнениям (55.2).

ДлЯ нахождения соотношения между $\mathbf{n}, \mathbf{n}', \mathbf{r}, \mathbf{r}'$ удобно ввести вместо $\psi$ другую величину, на которую бы не налагалось никаких дополнительных условий (т.е. которая не должна была бы удовлетворять каким-либо дифференциальным уравнениям).

Это можно сделать следующим образом.

В функции $\psi$ независимыми переменными являЮтся r и r′, так что для дифференциала $d\psi$ имеем:

$$d\psi = \frac{\partial\psi}{\partial\mathbf{r}}d\mathbf{r} + \frac{\partial\psi}{\partial\mathbf{r}'}d\mathbf{r}' = -\mathbf{n}d\mathbf{r} + \mathbf{n}'d\mathbf{r}'$$

Произведем теперь преобразование Лежандра к независимым переменным $\mathbf{n}$ и $\mathbf{n}'$ вместо $\mathbf{r}$ и $\mathbf{r}'$, т.е. напиШем:

$$d\psi = -d(\mathbf{nr}) + \mathbf{r}d\mathbf{n} + d\left(\mathbf{n}'\mathbf{r}'\right) - \mathbf{r}'d\mathbf{n}'$$

откуда, вводя фунКЦИЮ

$$\chi = \mathbf{n}'\mathbf{r}' - \mathbf{nr} - \psi$$

имеем:

$$d\chi = -\mathbf{r}d\mathbf{n} + \mathbf{r}'d\mathbf{n}'$$

ФунКцию $\chi$ называют угловым эйконалом; как видно из (55.5), независимыми переменными в нем являются $\mathbf{n}$ и $\mathbf{n}'$. На $\chi$ не налагается уже никаких дополнительных условий.

Действительно, уравнения (55.3) представляют собой теперь лишь условия, относящиеся К независимым переменным, показывающие, что из трех компонент $n_x, n_y, n_z$ вектора $\mathbf{n}$ (и аналогично длЯ $\mathbf{n}'$) только две явлЯются независимыми.

Мы будем ниже в качестве независимых переменных пользоваться компонентами $n_y, n_z, n'_y, n'_z$, и тогДа

$$n_x = \sqrt{1 - n_y^2 - n_z^2}, \quad n'_x = \sqrt{1 - n_y'^2 - n_z'^2}$$

Подставляя эти выражения В

$$d\chi = -xdn_x - ydn_y - zdn_z + x'dn'_x + y'dn'_y + z'dn'_z$$

находим для дифференциала $d\chi$:

$$\begin{aligned} d\chi = &-\left(y - \tfrac{n_y}{n_x}x\right)dn_y - \left(z - \tfrac{n_z}{n_x}x\right)dn_z + \\ &+\left(y' - \tfrac{n'_y}{n'_x}x'\right)dn'_y + \left(z' - \tfrac{n'_z}{n'_x}x'\right)dn'_z \end{aligned}$$

Отсюда находим окончательно следующие уравнения:

$$\begin{aligned} y - \frac{n_y}{n_x}x &= -\frac{\partial\chi}{\partial n_y}, \quad & z - \frac{n_z}{n_x}x &= -\frac{\partial\chi}{\partial n_z} \\ y' - \frac{n'_y}{n'_x}x' &= \frac{\partial\chi}{\partial n'_y}, \quad & z' - \frac{n'_z}{n'_x}x' &= \frac{\partial\chi}{\partial n'_z} \end{aligned}$$

определяющие искомое общее соотношение между $\mathbf{n}, \mathbf{n}', \mathbf{r}, \mathbf{r}'$. Функция $\chi$ характеризует конкретные свойства тел, через которые проходят лучи (или свойства поля - в случае движения заряженных частиц).

При заданных значениях $\mathbf{n}, \mathbf{n}'$ каждая из двух пар уравнений (55.6) изображает собой прямую линию.

Эти прямые являются не чем иным, как лучами до и после прохождения через оптическую систему.

Таким образом, уравнения (55.6) непосредственно определяют ход лучей по обе стороны оптической системы.

### 32.4.2 Thin beams of rays by Landau

(that shitty theory from Landau)

**Теория по Ландау**

При рассмотрении прохождения пучков лучей через оптические системы особый интерес представляют пучки, все лучи которых пересекаются в одной точке (так называемые гомоцентрические пучКи).

Гомоцентрический пучок лучей после прохождения через оПтическую систему, вообще говоря, перестает быть гомоцентрическим, т.е. после прохождения через тела лучи не собираются вновь в какойнибудь одной точке.

Только в особых случаях лучи, исходящие из светящейся точки, после прохождения через оптическую систему вновь пересекаются все в одной точке - изображении светящейся точки [1)] . Можно показать (см. $ 57), что единственный случай, когда все гомОЦентрические пучки остаются после прохождения через оптическую систему строго гомоцентрическими, есть тождественное отображение, Т.е. случай, когда изображение отличается от предмета только его переносом, поворотом или зеркальным отражением как целого.

Таким образом, никакая оптическая система не может дать вПолне резкое изображение предмета (обладающего конечными размерами), за исключением только тривиального случая тождественного изображения [2)].

Возможно лишь приближенное, не вполне резкое осуществление нетождественного изображения протяженных предметов.

Наиболее важным случаем приближенного перехода гомоцентрическиХ пучков в гомоцентрические же являются достаточно тонКие пучКи (Т.

е.

пучКи с малым углом раствора), идущие вблизи определенной (длЯ данной оптической системы) линии.

Эта линия называется оптической осью системы.

Необходимо при этом отметить, что даже бесконечно узкие пучКИ лучей (в трехмерном пространстве) в общем случае не являются

гомоцентрическими; мы видели (рис.

7), что и в таком пучке различные лучи пересекаются в различных точках (это явление называется астигматизмом).

Исключение представляют те точки волновой поверхности, в которых оба ее главных радиуса кривизны равны друг другу, - вблизи такой точки малый участок поверхности можно рассматривать как сферический и соответствующий тонкий пучок лучей явлЯется гомоцентрическим.

Будем рассматривать оптические системы, обладающие аксиальной симметрией 3).

Ось симметрии такой системы является в то же времЯ ее оптической осью.

Действительно, волновая поверхность пучка

лучей, идущего вдоль этой оси, тоже имеет аксиальную симметрию; Поверхности же вращениЯ имеют в точках своего пересечениЯ с осьЮ симметрии два равных друг другу радиуса кривизны.

Поэтому тонКий пучок, идущий в этом направлении, остается гомоЦентрическим.

ДлЯ нахождения общих количественных соотнОШений, определяюЩИХ отображениЯ с помощью тонких пучКов, проходящИХ через аксиально-симметричные оптические системы, воспользуемся уравнениями (55.6), определив предварительно вид фунКциИ $\chi$ в рассматриваемОм случае.

Поскольку пучки лучей тонкие и идут вблизи оптической оси, То векторы $\mathbf{n}$ и $\mathbf{n}'$ длЯ каждого пучка направлены почти вдоль этой оси.

Если выбрать оптическую ось в качестве оси $x$, то компоненты $n_y, n_z, n'_y, n'_z$ будут малы по сравнению с единицей.

Что касается Компонент $n_x, n'_x$, то $n_x \approx 1$, а $n'_x$ может быть приближенно равным $+1$ ИлИ -1.

В первом случае лучи продолжают идти почти в прежнем направлениИ, попадая в пространство по другую сторону оптической системы, которую в этом случае называют линзой.

Во втором случае ская система называется зеркалом.

ВоспользовавшИсь малостью $n_y, n_z, n'_y, n'_z$, разложим углОВой эйконал $\chi\left(n_y, n_z, n'_y, n'_z\right)$ в ряд и ограничимся первыми членами.

В силу аксиальной симметрии всей системы $\chi$ должно быть инвариантно по отношению к поворотам системы координат вокруг оптической оси.

Отсюда видно, что членов первого порядка, пропорЦИОнальных Первым степенЯм $y$ - и $z$ - компонент векторов $\mathbf{n}$ и $\mathbf{n}'$, в разложениИ $\chi$ не моЖет быть, - эти члены не обладали бы требуемой инвариантностью.

Из членов второго порядка требуемым свойством обладают квадраТы $\mathbf{n}^2, \mathbf{n}'^2$ и скалярное произведение $\mathbf{nn}'$. Таким образом, с точнОстью До членов второго порядка угловой эйконал длЯ аксиально-симметричной оПтической системы имеет вид

$$\chi = \text{const} \ + \frac{g}{2}\left(n_y^2 + n_z^2\right) + f\left(n_y n'_y + n_z n'_z\right) + \frac{h}{2}\left(n_y'^2 + n_z'^2\right)$$

где $f, g, h$ - постоянные.

Мы будем рассматривать сейчас для определенности случай ЛИНЗы, В связи с чем Положим $n'_x \approx 1$; длЯ зеркал, как будет нИже указано, все формулы имеют аналогичный вид.

Подставляя теперь выражение (56.1) в общие уравнения (55.6), находим:

$$\begin{aligned} n_y(x-g) - fn'_y = y, \quad fn_y + n'_y\,(x'+h) = y' \\ n_z(x-g) - fn'_z = z, \quad fn_z + n'_z\,(x'+h) = z' \end{aligned}$$

Рассмотрим гомоцентрический пучок, исходящий из точки $x, y, z$; точка $x', y', z'$ пусть будет той, в которой пересекаются все лучи

пучка после прохождения через линзу.

Если бы первая и вторая пары уравнений (56.2) были независимы, то эти четыре уравнения при заДанНых $x, y, z, x', y', z'$ определили бы одну определенную систему значений $n_y, n_z, n'_y, n'_z$, т.е. всего только один из лучей, выходящИХ ИЗ точки $x, y, z$, прошел бы через точку $x', y', z'$. ДлЯ того чтобы все лучи, выходящие из $x, y, z$, прошли через $x', y', z'$, необхоДИмо, следовательно, чтобы уравнения (56.2) не были независимы, т.е. чтобы одна пара этиХ уравнений была следствием другой.

Необходимым длЯ такой зависимости условием явлЯется, очевидно, проПорцИональность Коэффициентов одной пары уравнений коэффиЦиентам другой пары.

Таким образом, должно быть

$$\frac{x-g}{f} = -\frac{f}{x'+h} = \frac{y}{y'} = \frac{z}{z'}$$

В частностИ,

$$(x-g)(x'+h) = -f^2$$

полученные уравнения определяют искомую зависимость координат точки изображения от координат предмета при отображении с помощью тонких пучков.

точки $x = g, x' = -h$ на оптической оси называются главными фокусами оптической системы.

рассмотрим пучки лучей, параллельных оптической оси.

точка испускания такого луча находится, очевидно, в бесконечности на оптической оси, т.е. $x = \infty$. из (56.3) видно, что в этом случае $x' =$ - h.

таким образом, параллельный пучок лучей после прохождения через оптическую систему пересекается в главном фокусе.

наоборот, пучок лучей, исходящий из главного фокуса, становится после прохождения через систему параллельным.

В уравнениях (56.3) координаты $x$ и $x'$ отсчитываются от одного И того же начала координат, лежащего на оптической оси.

Удобнее, однако, отсчитывать координаты предмета и изображения от разных начал координат, выбрав их соответственно в главных фокусах.

В качестве положительного направления отсчета координат выберем направления от соответствующего фокуса в сторону, направленную по ходу луча.

Обозначая новые координаты предмета и изображения большИми буквами, имеем:

$$X = x - g, \quad X' = x' + h, \quad Y = y, \quad Y' = y', \quad Z = z, \quad Z' = z'$$

Уравнения отображения (56.3) и (56.4) принимают в новых обозначениях вид

$$XX' = -f^2$$
$$\frac{Y'}{Y} = \frac{Z'}{Z} = \frac{f}{X} = -\frac{X'}{f}$$

Величину $f$ называют главньим фокусным расстоянием системы.

Отношение $Y'/Y$ называется боковым увеличением.

Что касается продольного увеличения, то, поскольку координаты не просто пропор-

ЦИОнальны друг другу, его следует писать в дифференциальном виде, сравнивая элемент длины предмета (в направлении оси) с элементом длины изображения.

Из (56.5) пишем для продольного увеличения:

$$\left|\frac{dX'}{dX}\right| = \frac{f^2}{X^2} = \left(\frac{Y'}{Y}\right)^2$$

Мы видим отсюда, что даже для бесконечно малых предметов нельзя получить геометрически подобного изображения.

Продольное увеличение никогда не равно поперечному (за исключением тривиального случая тождественного отображения).

Пучок, вышедший из точКи $X = f$ на оптической оси, пересекается ВНОВь В точке $X' = -f$ на той же оси; эти две точКи называются главньими.

Из уравнений (56.2)$\big(n_y X - f n'_y = Y, n_z X - f n'_z = Z\big)$ видно, что в этом случае ($X = f, Y = Z = 0$) имеют место равенства $n_y = n'_y, n_z = n'_z$. Таким образом, всякий луч, выходящий из главной точки, пересекает вновь оптическую ось в другой главной точке в направлении, параллельном первоначальному.

Если координаты предмета и его изображения отсчитывать от главных точек (а не от главных фокусов), то длЯ этих координат $\xi$ и $\xi'$ имеем:

$$\xi' = X' + f, \quad \xi = X - f$$

Подставляя это в (56.5), легко получаем уравнение отображения в виде

$$\frac{1}{\xi} - \frac{1}{\xi'} = -\frac{1}{f}$$

Можно поКазать, что у оптических систем с малой толщиной (например, у зеркала, узкой линзы) обе главные точки почти совпадают.

В этом случае в особенности удобно уравнение (56.8), так как В нем $\xi$ и $\xi'$ отсчитываются тогДа практически от одной и той же точки.

Если фокусное расстояние положительно, то предметы, находящиеся спереди (по ходу луча) от фокуса ($X > 0$), отображаются прямо ($Y'/Y > 0$); такие оптические системы называются собирательными.

Если же $f < 0$, то при $X > 0$ имеем $Y'/Y < 0$, т.е. предмет отображается обратным образом; такие системы называются рассеивающими.

Существует один предельный случай отображения, который не содержится в формулах (56.8), - это случай, когда все три коэффиЦиента $f, g, h$ делаются бесконечными (т.е. оПтическая система имеет бесконечное фокусное расстояние и ее главные фокусы находятся в бесконечности).

Переходя в уравнении (56.4) к пределу бесконечных $f, g, h$, находИМ:

$$x' = \frac{h}{g}x + \frac{f^2 - gh}{g}$$

Поскольку представляет интерес только тот случай, когда предмет и его изображение находятся на конечных расстоянияХ от оптической системы, то $f, g, h$ должны стремиться к бесконечности так,

чтобы отношения $h/g, (f^2 - gh)/g$ были конечными.

Обозначая их соответственно посредством $\alpha^2$ и $\beta$, имеем: $x' = \alpha^2 x + \beta$.

Для двух других координат мы имеем теперь из уравнения (56.7):

$$\frac{y'}{y} = \frac{z'}{z} = \pm\alpha$$

Наконец, отсчитывая снова координаты $x$ и $x'$ от разных начал координат, соответственно от произвольной точки на отражаемой оси и от изображения этой точки, получаем окончательно уравнения отображения в простом виде:

$$X' = \alpha^2 X, \quad Y' = \pm\alpha Y, \quad Z' = \pm\alpha Z$$

Таким образом, продольные и поперечные увеличения постоянны (но не равны друг другу).

Рассмотренный случай отображения называется телескопическим.

Все выведенные нами для линз формулы (56.5) – (56.9) в равной мере применимы и к зеркалам, и даже к оптическим системам без аксиальной симметрии, если только отображение осуществляется тонкими пучками лучей, идущими вблизи оптической оси.

При этом всегда отсчет $x$ -координат предмета и изображения должен производиться вдоль оптической оси от соответствующих точек (главных фокусов или главных точек) по направлению распространения луча.

Надо иметь в виду при этом, что у оптических систем, не обладающих аксиальной симметрией, направления оптической оси впереди и позади системы не лежат на одной прямой.

**фокусное расстояние для двух аксиально-симметричных оптических систем**

Определить фокусное расстояние для отображения с помощью двух аксиально-симметричных оптических систем с совпадающими оптическими осями.

Пусть $f_1$ и $f_2$ - фокусные расстояния обеих систем.

Для каждой системы в отдельности имеем:

$$X_1 X_1' = -f_1^2, \quad X_2 X_2' = -f_2^2$$

Поскольку изображения, даваемые первой системой, являются предметом для второй, то, если $l$ расстояние между задним главным фокусом первой системы и передним фокусом второй, имеем $X_2 = X_1' - l$; выражая $X_2'$ через $X_1$, находим:

$$X_2' = \frac{X_1 f_2^2}{f_1^2 + l X_1}$$

или

$$\left(X_1 + \frac{f_1^2}{l}\right)\left(X_2' - \frac{f_2^2}{l}\right) = -\left(\frac{f_1 f_2}{l}\right)^2$$

откуда видно, что главные фокусы составной системы находятся в точках $X_1 = -f_1^2/l, X_2' = f_2^2/l$.

Итоговое фокусное расстояние равно

$$f = -\frac{f_1 f_2}{l}$$

(для выбора знака в этом выражении надо написать соответствующее уравнение для поперечного увеличения).

В случае, если $l = 0$, фокусное расстояние $f = \infty$, т.е. составная система дает телескопическое отображение.

В этом случае имеем: $X_2' = X_1 (f_2/f_1)^2$, т.е. параметр $\alpha$ в общей формуле (56.9) равен $\alpha = f_2/f_1$.

**фокусное расстояние «магнитной линзы»**

Определим фокусное расстояние «магнитной линзы» для заряженных частиц, представляющей собой продольное однородное магнитное поле в участке длины

(тут рис 8)

При движении в магнитном доле кинетическая энергия частицы сохраняется; поэтому уравнение Гамильтона-Якоби для укороченного действия $S_0$(r) из полного $S = -\mathscr{E}t + S_0$ имеет вид

$$\left(\nabla S_0 - \frac{e}{c}\mathbf{A}\right)^2 = p^2$$

где

$$p^2 = \frac{\mathscr{E}^2}{c^2} - m^2 c^2 = \text{const}.$$

Воспользовавшись формулой для векторного потенциала однородного магнитного поля,

(19.4)

выбирая ось $x$ вдоль направления последнего и рассматривая ее как оптическую ось аксиально-симметричной оптической системы, получим уравнение Гамильтона-Якоби в виде

$$\left(\frac{\partial S_0}{\partial x}\right)^2 + \left(\frac{\partial S_0}{\partial r}\right)^2 + \frac{e^2}{4c^2} H^2 r^2 = p^2$$

где $r$ - расстояние от оси $x$, а $S_0$ - функция от $x$ и $r$.

Для тонких пучков частиц, распространяющихся вблизи оптической оси, координата $r$ мала, поэтому ищем $S_0$ в виде ряда по степеням $r$. Первые два члена этого ряда:

$$S_0 = px + \frac{1}{2}\sigma(x)r^2$$

где $\sigma(x)$ удовлетворяет уравнению

$$p\sigma'(x) + \sigma^2 + \frac{e^2}{4c^2}H^2 = 0$$

В области 1 перед линзой имеем:

$$\sigma^{(1)} = \frac{p}{x - x_1}$$

где $x_1 < 0$ - постоянная.

Это решение соответствует свободному пучку частиц, разлетающихся по прямым лучам из точки $x = x_1$ на оптической оси В области 1.

Действительно, свободному движению частицы с импульсом $p$ в направлении от точки $x = x_1$ соответствует действие

$$S_0 = p\sqrt{r^2 + (x - x_1)^2} \approx p\,(x - x_1) + \frac{pr^2}{2\,(x - x_1)}$$

Аналогично, в области 2 позади линзы пишем:

$$\sigma^{(2)} = \frac{p}{x - x_2}$$

где постоянная $x_2$ представляет собой координату изображения точки $x_1$.

В области же 3 внутри линзы решение уравнения (3):

$$\sigma^{(3)} = \frac{eH}{2c}\operatorname{ctg}\left(\frac{eH}{2cp}x + C\right)$$

где $C$ - произвольная постоянная.

Постоянные $C$ и $x_2$ (при заданном $x_1$) определяются условиями непрерывности $\sigma(x)$ при $x = 0$ и $x = l$ :

$$-\frac{p}{x_1} = \frac{eH}{2c}\operatorname{ctg} C, \quad \frac{p}{l - x_2} = \frac{eH}{2c}\operatorname{ctg}\left(\frac{eH}{2cp}l + C\right)$$

Исключая из этих равенств постоянную $C$, получим:

$$(x_1 - g)\,(x_2 + h) = -f^2$$

где

$$g = -\frac{2cp}{eH}\operatorname{ctg}\frac{eHl}{2cp}, \quad h = g + l, \quad f = \frac{2cp}{eH\sin\frac{eHl}{2cp}}$$

## 32.4.3 Wide beam mapping by Landau

(that shitty theory from Landau)

**Теория по Ландау**

Рассмотренное в предыдущем параграфе отображение предметов с помощью тонких пучков лучей является приближенным; оно тем точнее (т.е. резче), чем уже эти пучки.

Перейдем теперь к вопросу об отображении предметов пучками лучей произвольной ширины.

В противоположность отображению предметов тонкими пучками, которое можно осуществить с любой оптической системой, обладающей аксиальной симметрией, отображение широкими пучками возможно только с помощью определенным образом построенных оптических систем.

Даже с этим ограничением возможно, как уже указывалось в §56, (???) отображение далеко не всех точек пространства.

**Theory**

Дальнейшие выводы основаны на следующем существенном замечании.

Пусть все лучи, выходящие из некоторой точки $O$ и проходящие через оптическую систему, вновь пересекаются в некоторой другой точке $O'$.

Оптическая длина пути $\psi$ одинакова для всех этих лучей.

Действительно, вблизи каждой из точек $O, O'$ волновые поверхности для пересекающихся в них лучей являются сферами с центрами, соответственно в $O$ и $O'$ и в пределе, при приближении к $O$ и $O'$, вырождаются в эти точки.

Но волновые поверхности являются поверхностями постоянной фазы, и поэтому изменения фазы вдоль разных лучей между точками их пересечения двух определенных волновых поверхностей одинаковы.

Из сказанного следует, что одинаковы (для разных лучей) и полные изменения фазы между точка $O$ и $O'$. Выясним условия, необходимые для осуществления отображения Широкими пучками малого отрезка прямой; изображение представляет собой при этом тоже малый отрезок прямой.

Выберем направления этих отрезков за направления осей (назовем их $\xi$ и $\xi'$) с началами $O$ и $O'$ в каких-либо соответствующих друг другу точках предмета и изображения.

Пусть $\psi$ есть оптическая длина пути для лучей, выходящих из $O$ и приходящих в $O'$. Для лучей, выходящих из бесконечно близкой к $O$ точки с координатой $d\xi$ и сходящихся в точке изображения с координатой $d\xi'$, оптическая длина пути есть $\psi + d\psi$, где

$$d\psi = \frac{\partial \psi}{\partial \xi} d\xi + \frac{\partial \psi}{\partial \xi'} d\xi'$$

Введем «увеличение» при отображении

$$\alpha_\xi = \frac{d\xi'}{d\xi}$$

как отношение длины $d\xi'$ элемента изображения к длине отображаемого элемента $d\xi$.

В силу малости отображаемого отрезка увеличение $\alpha$ можно считать величиной, постоянной вдоль его длины.

Написав так же, как обычно, $\partial\psi/\partial\xi = -n_\xi, \partial\psi/\partial\xi' = n'_\xi$ ($n_\xi, n'_\xi$ - косинусы углов между направлениями луча и соответственно осями $\xi$ и $\xi'$), получим:

$$d\psi = \left(\alpha_\xi n'_\xi - n_\xi\right) d\xi$$

Как и для всякой пары соответствующих друг другу точек предмета и изображения, оптическая длина пути $\psi + d\psi$ должна быть одинаковой для всех лучей, выходящих из точки с координатой $d\xi$ и приходящих в точку $d\xi'$.

Отсюда получаем условие:

$$\alpha_\xi n'_\xi - n_\xi = \text{const}.$$

Это и есть искомое условие, которому должен удовлетворять ход лучей В оптической системе при отображении широкими пучками малого отрезка прямой.

Соотношение (57.1) должно выполняться для всех лучей, выходящих из точки О.

Применим теперь полученное условие к отображению с помощью аксиально-симметричной оптической системы.

Начнем с отображения отрезка прямой, лежащего на оптической оси системы (ось $x$); из соображений симметрии очевидно, что изображение будет тоже лежать на оси.

Луч, идущий вдоль оптической оси ($n_x = 1$), в силу аксиальной симметрии системы не меняет своего направления при прохождении через нее, т.е.

$n'_x = 1$. Отсюда следует, что const в (57.

1) равна в рассматриваемом случае $\alpha_x - 1$, и мы можем переписать (57.1) в виде

$$\frac{1 - n_x}{1 - n'_x} = \alpha_x$$

Обозначая посредством $\theta$ и $\theta'$ углы, образуемые лучами с оптической осью в точках предмета и изображения, имеем:

$$1 - n_x = 1 - \cos\theta = 2\sin^2\frac{\theta}{2}, \quad 1 - n'_x = 2\sin^2\frac{\theta'}{2}$$

Таким образом, получим условие отображения в виде

$$\frac{\sin\frac{\theta}{2}}{\sin\frac{\theta'}{2}} = \text{const} \ = \sqrt{\alpha_x}$$

Далее, рассмотрим отображение малого участка плоскости, перпендиКулярной к оптической оси аксиально-симметричной системы; изображение будет, очевидно, тоже перпендикулярно к этой оси.

ПрименяЯ (57.1) к любому отрезку, лежашему в отображенной плоскости, ПОЛучИМ:

$$\alpha_r \sin\theta' - \sin\theta = \text{const}$$

где $\theta$ и $\theta'$ - по-прежнему углы между лучом и оптической осью.

ДлЯ лучей, вышедшИх из точки пересечения изображаемой плоскости с оптической осью в направлении этой оси ($\theta = 0$), должно быть, в силу симметрии, и $\theta' = 0$. Поэтому const $= 0$, и мы получаем условие отображения в виде

$$\frac{\sin\theta}{\sin\theta'} = \text{const} \ = \alpha_r$$

Что касается отображения Широкими пучками трехмерных предметов, то легКо видеть, что оно невозможно даже при малом объеме тела, поскольку условия (57.2) и (57.3) несовместимы друг с другом.

# Part XI
# Appendix

## A Introduction to Optics

### A.1 Extra Motivation

Обсудим, что нас мотивирует заниматься оптикой.

#### A.1.1 Amazing Facts and Phenomena in Optics

(???)

#### A.1.2 Gaining Skills in General Physics

**Оптика играет почти ключевую роль в принципиальном понимании мира**

потому что тут и квантовые явления, и классические и чего только нет!
(потом пропишу)
и это фундаментальная вещь, которая просто необходима тому, кто понимает физику.
вот только не совсем ключевую, есть и другие предметы, но не важно.

**Многие передовые исследования в лабораториях связаны с оптикой**

То есть и многие лаборатории станут доступны. Многие исследования сразу используют оптику. И зарабатывать можно будет деньги.

Поэтому база по оптике очень вероятно, что будет использоваться, поэтому заготовки по ней очень важны.

по идее.

**Оптика глубоко связана с многими другими разделами физики**

(пойму - четко укажу как)

К оптике также приходится крайне много возвращаться, потому что для многих предметов она необходима. Просто так или иначе, занимаясь ими, упираешься в то, что хочется просто иметь сильную оптическую базу.

Поэтому собрав ее - откроются многие предметы. Так что выгодно на ней пару месяцев посидеть, а потом уже идти в прикладные предметы.

**оптические приборы в промышленности играют большую роль**

**прикладная теория поля во многом в оптикой связана**

Так что когда буду изучать его, знания оптики могут ускорить процесс в разы.

**много используется актуальных разделов математики в различных оптических задачах**

Параллельно будет выучен гармонический анализ, ...

### A.1.3 Variety of Applications in Optics

(позже раскрою, какие)

#### Любые вопросы визуализации решаются с помощью понимания прикладной оптики (!!?)

Поэтому она крайне часто и нужна в программировании, в представлении результатов. Как камеры в играх двигать, как 4d пространство видеть. Как смоделировать тор со стороны, чтобы на нем была динамика, а не на плоскости?

(Этому пока что не научился, потому что еще неделя ушла бы у меня просто на доработку основ. Но потом нужно будет вывести оптику в прикладной уровень, отдельно напишу эти методы по многим статьям или видео!!! пока развиваться еще нужно.)

### A.1.4 Motivation for Branches of Optics

(по ходу дела пойму её)

#### Мотивация к геометрической оптики

(её важно прописать, ибо не очевидно, зачем её изучать. и я понимаю. что она нужна, но пока хз, как конкретно)

##### задачи на линзы решать гораздо проще в пределе геометрической оптики

(я даже не знаю, как сделать это с помощью волновой. как????)

##### Встречается в задачах о движении света в ОТО

Просто знаю, что встречается.

##### (?) Является приближенным решением волнового уравнения

Просто знаю, что встречается.

#### Мотивация к экспериментальной оптике

(будучи у Лесовика, пропишу)

##### Можно устроиться в оптическую лабораторию и зарабатывать деньги

такие есть (на 2021 год)

##### Можно ставить эксперименты по другим разделам физики

Ибо часто оптические приборы используются в них.
(с опытом пропишу)

#### Мотивация к другим разделам оптики

##### мотивация к Фурье-оптике

вроде бы методы фурье оптики используются для распознавания сигналов
но я не уверен.

## A.2 Mindset of a Professional in Optics

### A.2.1 Ways to Make Money and Overview of Applications

(тут напишу про оптические лаборатории и компании. пока что самому нужно проверить, выучу, поработаю, пропишу, насколько это реально???)
(????)

### A.2.2 Which Methods Work Good and Which Work Bad?

#### Недостатки волновой оптики (?)

(про корпускулярные эффеткы, что еще?)
про то, что квантовые явления существуют и тут волновой оптики и не хватает нам.

### A.2.3 Unresolved Issues in Optics

(пока что не знаю их, потом подумаю.)

### A.2.4 Техническая сторона подхода к оптике

пока не понимаю, то ли все представлять волнами?
фурье ко всему пытаться применить?
пока не знаю это.

### A.2.5 Should One Study Exotic Topics in Optics?

(there are a lot of them, so I'll write an answer to this important question)

### A.2.6 Ways to Guess Some Main Ideas

незаменимая часть нормального понимания предмета.

### A.2.7 Effective Way of Studying Optics

#### Необходимые предметы для изучения оптики

##### Классическая электродинамика как часть теории поля

по сути оптика - во многом часть электродинамики.
а вот теорию поля по основам учить как раз не нужно, они не нужны.
Поэтому здесь многие вопросы - вопросы, которые должны быть проработаны в курсе теории поля. По сути мы берем часть той теории и углубляем, развиваем ее.

**квантовая физика**

потому что лазеры и всё такое тоже есть и вообще, во многом у нас именно квантовые явления происходят.

квантовая оптика также есть и также очень много можно ей заниматься.

**Дополнительные предметы для изучения оптики**

**теория каустик и волновых фронтов**

тут связка есть, но она слишком мощная и слишком отдаленная, чтобы ей заниматься изначально.

это чисто такая продвинутая оптика, для которой еще нужны заготовки по математике.

но такое есть, позже займусь, хорошо и для оптики подготовит это.

мб.

## A.3 Acknowledgements

Currently, no one except me has worked on the sections of this note (with the exception of sections taken from books).

## A.4 Literature on Optics

### A.4.1 Basic Literature

**Основная обучающая**

[3] А.В. Максимычев и др. Лабораторный практикум по общей физике: учеб. пособие. В трёх томах. Т. 2. Оптика

Прекрасная книга, лучшая для начала.

Кингсеп Локшин Ольхов Основы физики. Т. 1. Механика. Электричество и магнетизм. Колебания и волны. Волновая оптика

здесь брал основы структуры. в принципе хорошая книга для начала, никак не пропишу их вот, всё не до этого.

[1] Кириченко Оптика

пока не очень нужен, мб почитаю, добавлю детали, когда запись сделаю

**Задачники по оптике**

Овчинкин и др. Сборник задач по общей физике часть 2

Шикарный задачник, над которым еще посижу часов 30 минимум.

Корявов Методы решения задач по оптике

Большая помощь к сборнику Овчинкину, очень полезная книга о моделях оптики.

Жукарев Матвеев Петерсои Задачи повышенной сложности в курсе общей физики: Учебное пособие. 2-е изд., испр. / Под общей ред. А. Н. Матвеева. — М.: Эдиториал УРСС, 2001. — 192 с.

Шикарные задачи, обязательно прорешаю, только кайф их решать.

Чудновский А. В., Григорьев Ю. М., Муравьёв В. М., Потапов В. Ф. Теоретические задачи по физике Международная олимпиада «Туймаада» 1994-2012

Очень крутые задачи, очень интересные, тоже разберу их постепено, уже выгружено.

В.С. Булыгин, М.Г. Кремлев, Э.В. Прут Избранные задачи студенческих олимпиад МФТИ по курсу общей физики / В.С. Булыгин, М.Г. Кремлев, Э.В. Прут - Долгопрудный: Издательский Дом «Интеллект», 2015. - 312 с.

Крутой сборник задач, тоже прорешаю потом его, очень интересные и простые задачи, шикарное проведение времени на них.

О.Ф. Кабардин В.А. Орлов Международные физические олимпиады школьников

Потом разберу, наверняка по оптике тоже задач много, наверняка они очень интересные будут.

#### Основная профессиональная литература

Landsberg Optics

Great big book. I'll add a lot of ideas from it later. I haven't found any English version of it yet. However, these topics are not so popular in the modern science, so people don't want to focus on them much.

### A.4.2 Extra Literature

#### Другая обучающая

HuygensOptics on YouTube https://www.youtube.com/@HuygensOptics

Great interesting channel with many visualizations. Many details are discussed there, but one most probably will not need to know them.

#### Другая профессиональная литература

М.Борн, Э.Вольф Основы оптики

Очень крутая книга, но к сожалению пока не нужна был, может, потому что я не особо развит еще.

[2] Сивухин Оптика

Многие разделы и параграфы у него хорошо написаны, так что много из него дополнил. Может, когда-то тоже открою, досмотрю детали, но вряд ли, слишком много более актуальных задач.

#### Литература теоретической физики, содержащая оптику

[4] Ландау Л. Д., Лифшиц Е. М. Теория поля

Много для оптики заготовок принципиально важных. А также раскрыта связь с теорией поля. Однако некоторые темы раскрыты настолько плохо, что само их изложение буду брать из других мест. а последовательность выводов основана на именно этой книге.

Landau электродинамика сплошных сред, 8

глава про нелинейную оптику из нее будет

Зин Жюстен Континуальный интеграл в квантовой механике

про эйконал и геометрическую оптику есть хорошая глава. Не уверен, что она так уж нужна.

#### Математическая и о связи оптики и математике

(я так и не понял, зачем она нужна.)
Арнольд каустики и волновые фронты
мб добавлю в оптику, мб нет, тут посмотрим уже. в принципе хочется, но это отдельная заморочка.

#### Статьи по основам оптики

И. М. Франк Эйнштейн и оптика 1979 г. Декабрь Том 129, вып. 4 УФН
интересная статья, прикольные выводы. потом посмотрю, когда пропишу структуру нормально. хз зачем она, мне и не до нее

### A.4.3 Literature about Special Topics

#### По оптическим волноводам

А.Снайдер Лав Теория оптических волноводов
огромная книга с непойми чем. Мб когда-то открою, но пока не уверен в этом.

#### По квантовой оптике

[5] Андрианов Виноградов Лекции по квантовой оптике
вроде много информации, если понадобится - буду читать это

[6] Walls, D. F. Squeezed states of light
Известная статья про сжатые состояния, мб по ней что-то можно понять, я чуть смотрел. Пока общую картину нужно много дорабатывать, чтобы это было усвоено.

#### Статьи про оптику в астрономии

тут нпр гравитационное линзирование потом будет.

#### По квантовой оптике

[7] Краснок и др. Аномалии
посмотрю потом еще раз, мб что-нибудь в этом и есть годное, просто отдельно нужно разбираться, я не на таком уровне.

#### Про некоторые экспериментальные методы

Josef F. Bille High Resolution Imaging in Microscopy and Ophthalmology New Frontiers in Biomedical Optics
Может быть, на самом деле есть какой-то смысл в ОСТ, хотя выглядит этот метод очень тупо и беспеспективно. Может, когда-то почитаю, может, это лучший обзор этого метода. Вообще, много про него статей, но смысла их читать вообще не вижу.

#### По оптическим конструкциям и приборам

(пока не до этого)
[8] Курбатов, Л.Н. Оптоэлектроника видимого и инфракрасного диапазонов спектра.
В планах есть, но в реальности не дойду скорее всего никогда. Просто знаю, что такая есть книга и всё.

[9] Локшин, Г.Р. Принципы радиооптики
Может, однажды пару принципов узнаю, в то же время сейчас вообще это не нужно. . да я хз, про прикладную оптику - самая топовая книга, тем не менее, так её не открывал. успею - пару дней прикладную оптику по ней пару дней буду прописывать. нет - так и пофиг. Основы зашарить куда более актуально

Снайдер А. Лав. Дж. Теория оптических волноводов
очень задротская книга про волноводы, мб по ней многое в эту главу пропишу. уровень задротстсва зашкаливает. но посмотрим, мб польза и есть от нее, посмотрю потом.

## A.5 Overview of Optics

### A.5.1 Popular Introduction to Optics

#### Оптические подходы

Обсудим, что из себя представляет предмет наиболее кратко, выделяя самую суть.

#### появление Предмет в нашей картине мира

#### один подход

#### второй подход

#### один большой раздел

#### такой-то набор следствий

#### Наиболее используемые темы

обсудим, какие темы на самом деле самые важные тут.

#### хз

пока не дошел до применений.

#### Обзор дальнейших развитий оптики

Обсудим, какие теории далее можно развивать, поняв законы оптики.

### A.5.2 Optics in Everyday Life

а вот хз пока что, где она, но позже наберу тут структуру.
потому что по идее каждый день можно ее замечать, нужно только собрать основы.

#### дифракция в обычной жизни

хз, где ее можно увидеть???
какие там критерии ее существования кстати?
потом напишу

#### количественные оценки

Часто важно понимать, что чему равно по величине, разберемся с этим, чтобы ориентироваться.

##### величина напряженности поля в типичном свете

хз

##### типичная частота света

### A.5.3 Relations of Optics with Other Sciences

(с опытом четко впишу, очень интересно)

#### С классической теорий поля

оптика связана и с теорией поля, и с квантовой физикой и с термодинамикой.

#### С квантовой теорий поля (!)

(она тут большая, потом сформулирую. Может быть, самая и главная, пока хз)

#### С электродинамикой сплошных сред

связь большая, по сути чуть со средой взаимодействие - именно в эту главу попадаем.

#### С квантовой механикой (??!!!)

(???)
а вот тут есть четкий ответ, надеюсь, пропишу!!!
откуда фотон взялся, например, ответ тут дается.
Киселев знает это, у него это спрошу мб!!!

##### Оптика и континуальный интеграл (??)

какая-то там связь есть, но мне не зашла она.

#### С общей теории относительности

(далеко не скоро пропишу, потому что не актуально)

##### гравитационное линзирование

гравитационное линзирование такое есть.

## A.6 History of Optics

(потом напишу, если нужно будет)

## A.7 Description of the Note

### A.7.1 The Goal and Features of the Note

(напишу все это потом, под конец, не так актуально это писать и все может еще поменяться много раз.)

(Постараюсь указать методы теоретической физики из других разделов, иметь много картинок, потому что оптика - визуальная наука, а также указать многие приложения.)

### A.7.2 Designations and Constants

## A.8 Puzzles in Optics for Different Situations

Обсудим в порядке интересности вопросы по оптике.

### A.8.1 Fundamental Puzzles for Fun Discussion

В оптике есть множество сложных принципиальных вопросов, чтобы ответить на которые, нужно знать многие другие разделы физики.

Некоторые настолько могут запутать незнающего, что просто полностью перевернуть картину мира.

Ответим вкратце на них, чтобы ориентироваться.

(только когда главное будет сделано, буду это вписывать!)

#### Что такое фотон?

(пропишу потом развернутый ответ)

#### откуда появляется фотон?

какая-то там флуктуация вакуума?

что ж, разберемся....

теорию излучения собрать нормально нужно

#### можно ли поймать фотон?

#### как свет на самом деле распространяется?

ух, глубокий вопрос..

чет много шагов мне еще, чтобы на него ответить.

#### переход из геометрической оптики в волновую

Типичный вопрос по оптике, который возникает в первых же днях знакомства с ней, это как перейти от одного описания света к другому.

вот это подробно отвечу.

**когда проявляются квантовые эффекты?**

когда-то....
интересно.

**каково влияние материи и волн друг на друга**

дисперсия возникает как раз за счет материи.
обсудим все это.

**что такое поляризация света и откуда она берется?**

популярно суммирую позже.

**что находится на границе пучка света?**

я так думаю, просто поле убывает в ноль быстро, но гладко.
не дошел всё никак до этого.

**когда не работает волновое описание света?**

там лазеры и всё такое, подробно отвечу на это потом.

**откуда берутся прозрачные или отражающие поверхности**

очень сложный вопрос, внятного ответа на который я так и не услышал.

**распространение информации**

если произошло какое-то событие, другая часть пространства узнает об этом со скоростью света с помощью фотонов?
чет не знаю.

**почему две волны проходят сквозь друг друга, не взаимодействуя?**

казалось бы, должны взаимодействовать.

**сверхсветовая скорость**

можно ли так двагаться?
(тут многое сказать можно)

**свет это больше, чем электромагнитная волна?**

все ли свойства света заключены в ее электромагнитной природе?
есть ли в нем еще что-нибудь?
вроде внезапно еще квантовые свойства возникают, но тут это КЭД, так что по сути все описано
хз

**объяснение дисперсии**

почему она есть?
тут отвечу.

### A.8.2 Puzzles for Interesting Discussion

### A.8.3 Puzzles About Everyday Life

(???? what fun questions can arise about this subject in everyday life?)

### A.8.4 Selected Fun Typical Problems

(all this should also be in the problem-solution section!)

### A.8.5 Principal Puzzles

(Puzzles about really important questions, even though they are already maybe solved)

### A.8.6 Puzzles about Details

("look, this is a tricky detail, why is it so?")

### A.8.7 Selected Puzzles for Learning Subjects

Допустим, видим человека, который в принципе хотел бы понять суть, опишем, какие в какой последовательности будем задавать, чтобы в типичном случае ему было бы интереснее въезжать. В частности это могут быть младшие студенты или редко школьники.

**Головоломки для освоения школьного уровня**

**Головоломки для освоения уровня теормина**

**Головоломки для освоения других особенностей**

### A.8.8 Puzzles to Become Demotivated

(?? it is fun to know them, so I'll write later)

### A.8.9 Puzzles to Prove that a Person Doesn't Know the Subject

(тоже такие можно собрать, иногда очень и полезно это понимать, но очень редко.)

## B Some Physics for Optics

### B.0.1 Quantum Mechanics

(??? some topics that are really important. maybe path integral, if I learn how to use it???)

### B.0.2 Quantum Field Theory

(думаю, это очень важно, но пока оставил их, ибо квантовой оптикой мб и не буду заниматься.)

(!!! оставлю только то, что на самом деле нужно. структура по Пухову.)

### B.0.3 Mechanics

(пока не четко понимаю, что конкретно нужно для оптики)

**Уравнения Гамильтона-Якоби: суть**

**Уравнения Гамильтона-Якоби: примеры в оптике**

уже это видел, пойму - впишу.

# C Some Mathematics for Optics

(пока не разбираюсь, что насколько нужно.)

## C.1 Harmonic Analysis

Первый раздел математики, который важен в оптике - это гармонический анализ. Обсудим его и его применения в оптике.

### C.1.1 Fourier Transformation and Decomposition into Modes

говорят, его много где можно использовать

### C.1.2 Kotelnikov's Theorem

**Theory**

**пример**

## C.2 Integrals in Optics

### C.2.1 Integrals in Radiation Theory

(типичные интегралы, которые нужно уметь быстро брать. потом соберу)

### C.2.2 Fresnel Integrals

(????)

**спираль Корню**

## C.3 The Calculus of Variations

IT is a general feature of equations of classical physics that they can be derived from variational principles. Two early examples are Fermat's principle in optics (1657) and Maupertuis' principle in mechanics (1744). The equations of elasticity, hydrodynamics, and electrodynamics can also be represented in this way.

However, when one deals with field equations, involving as a rule four or more independent variables $x, y, z, t \ldots$, one makes little use, owing to the great complexity of partial differential equations, of the property that the solution expresses stationary values of certain integrals. The only essential advantage of the variational approach in such cases is connected with the derivation of conservation laws-e.g. for energy. The situation is quite different in problems involving one independent variable (time in mechanics, or length of a ray in geometrical optics).

Then one deals with a set of ordinary differential equations and it turns out that a study of the behaviour of the solution is greatly facilitated by a variational approach. This approach is in fact a straightforward generalization of ordinary geometrical optics in every detail. Its modern representation owes much to David Hilbert, on whose unpublished lectures, given at Göttingen in about 1903, we base the considerations of the following sections. The theory is presented here for a three-dimensional space $(x, y, z)$ only, but can easily be extended to more dimensions.

## C.3.1 1. Euler's equations as necessary conditions for an extremum

Let $F(u, v, x, y, z)$ be a given function with continuous partial derivatives up to the second order in all the five variables. Further let $C$ be any curve $x = x(z), y = y(z)$ in the $x, y, z$ space. The derivatives of $x$ and $y$ will also be assumed to be continuous up to second order. If we set

$$u = x', \quad v = y',$$

(the prime denoting differentiation with respect to $z$ ) the integral

$$I = \int_{z_1}^{z_2} F\left(x', y', x, y, z\right) dz \tag{1}$$

is a function of the curve $C$, i.e. of the two functions $x(z), y(z)$, in other words a functional. The fundamental problem of the calculus of variation is:

To determine a curve $C$ between two given points $P_1\left[x_1 = x\left(z_1\right), y_1 = y\left(z_1\right), z_1\right]$ and $P_2\left[x_2 = x\left(z_2\right), y_2 = y\left(z_2\right), z_2\right]$ for which the integral is an extremum (minimum or maximum).

The necessary conditions which such a curve $C$, called an extremal, must satisfy may be determined by the simple process of linear variation. For this purpose we choose a function $\xi(z)$ with a continuous first order derivative, which vanishes at the end points,

$$\xi\left(z_1\right) = \xi\left(z_2\right) = 0, \tag{2}$$

and form the "varied" curve $C$ ' by replacing the $x$-coordinate of the extremal by $x + \varepsilon\xi$, where $\varepsilon$ is a small parameter. Equation (1) then becomes a function of $\varepsilon$,

$$I(\varepsilon) = \int_{z_1}^{z_3} F\left(x' + \varepsilon\xi', y', x + \varepsilon\xi, y, z\right) dz \tag{3}$$

The quantity

$$(\delta I)_x = \left(\frac{\partial I}{\partial \varepsilon}\right)_{\varepsilon=0} = \int_{z_1}^{z_2} \left(\frac{\partial F}{\partial x'}\xi' + \frac{\partial F}{\partial x}\xi\right) dz \tag{4}$$

is called the first variation with respect to $x$. The vanishing of the first variation

$$(\delta I)_x = 0 \tag{5}$$

is clearly a necessary condition for an extremum.

Now if we integrate the first term in (4) by parts and use the boundary conditions (2), we obtain

$$(\delta I)_x = \int_{z_1}^{z_2} \left(F_x - \frac{d}{dz}F_{x'}\right) \xi dz \tag{6a}$$

where $F_x$ stands for $\partial F/\partial x$, etc. Similarly, replacing $y$ by $y + \varepsilon\eta$, we find that

$$(\delta I)_y = \int_{z_1}^{z_2} \left( F_y - \frac{d}{dz} F_{y'} \right) \eta dz \tag{6b}$$

Since $\xi$ and $\eta$ may be chosen arbitrarily in the interval $z_1 \leqslant z \leqslant z_2$, it follows from (6a) and (6b), that the conditions $(\delta I)_x = 0$ and $(\delta I)_y = 0$ may be expressed in the form of two differential equations, known as Euler's equations,

$$F_x - \frac{d}{dz} F_{x'} = 0 \tag{7a}$$

$$F_y - \frac{d}{dz} F_{y'} = 0 \tag{7b}$$

(7a) and (7b) are two differential equations of the second order for $x(z)$ and $y(z)$. The leading terms, i.e. those with derivatives of the highest order, will be written out fully:

$$F_{x'x'x''} + F_{x'y'} y'' + \ldots = 0 \tag{8a}$$

$$F_{y'x'x''} + F_{y'y'y''} + \ldots = 0 \tag{8b}$$

These equations can be solved for $x''$ and $y''$ provided the associated determinant does not vanish, i.e. provided that

$$F_{uu} F_{vv} - F_{uv}^2 \neq 0 \tag{9}$$

We shall assume that this condition holds throughout the five-dimensional region with which we are concerned.

The solutions of two differential equations of the second order contain four arbitrary constants of integration so that the extremals form a four-parameter family of curves ( $\infty^4$ extremals).

## C.3.2 2. Hilbert's independence integral and the Hamilton-Jacobi equation

In order to discuss the properties of these extremals it is convenient to consider another related problem. The variables $u$ and $v$ will be regarded as functions in the $x, y, z$ space

$$u = u(x, y, z), \quad v = v(x, y, z) \tag{10}$$

Then $F[u(x, y, z), v(x, y, z), x, y, z]$ as well as its partial derivatives $F_u, F_v$ are functions of $x, y, z$. Let us now choose a curve $C[x = r(z), y = y(z)]$, and form the integral

$$S = \int_{u_1}^{z_2} \{F + (x' - u) F_u + (y' - v) F_v\} dz \tag{11}$$

The problem is this: to find such functions $u, v$ which secure that $S$ is independent of the choice of the curve $C$ : $S$ is then a function of the end points $P_1$ and $P_2$ only, where $P_1$ has the coordinates $x_1 = x(z_1), y_1 = y(z_1), z_1$, and $P_2$ the coordinates $x_2 = x(z_2)$, $y = y(z_2), z_2$.$S$ will be called Hilbert's independence integral.

To determine $u$ and $v$ we first re-write (11) in the form

$$S = \int_{P_1}^{P_2} (U dx + V dy + W dz) \tag{12}$$

where

$$U = F_u, \quad \nabla = F_v, \quad W = F - uF_u - vF_v. \tag{13}$$

Now it is well known that the necessary and sufficient conditions for (12) to be independent of the curve are the vanishing of the components of the curl of the vector $\boldsymbol{A}$ whose components are $U, V, W$, i.e.

$$\frac{\partial W}{\partial y} - \frac{\partial V}{\partial z} = 0, \quad \frac{\partial U}{\partial z} - \frac{\partial W}{\partial x} = 0, \quad \frac{\partial V}{\partial x} - \frac{\partial U}{\partial y} = 0. \tag{14}$$

This is a set of three partial differential equations for $u(x, y, z)$ and $v(x, y, z)$; the equations are however not quite independent as the identity

$$\frac{\partial}{\partial x}\left(\frac{\partial W}{\partial y} - \frac{\partial V}{\partial z}\right) + \frac{\partial}{\partial y}\left(\frac{\partial U}{\partial z} - \frac{\partial W}{\partial x}\right) + \frac{\partial}{\partial z}\left(\frac{\partial V}{\partial x} - \frac{\partial U}{\partial y}\right) = 0 \tag{15}$$

holds for any $U, V$, and $W$ ( $\operatorname{div}\operatorname{curl} A = 0$ for any vector $A$ ). If the equations (14) are satisfied, then $Udx + Vdy + Wdz$ is a total differential,

$$dS = Udx + Vdy + Wdz \tag{16}$$

and $S$ is then a function of $P_1\,(x_1, y_1, z_1)$ and $P_2\,(x_2, y_2, z_2)$ alone. Writing for short $x, y, z$ in place of $x_2, y_2, z_2$ we have

$$U = \frac{\partial S}{\partial x}, \quad V = \frac{\partial S}{\partial y}, \quad W = \frac{\partial S}{\partial z} \tag{17}$$

We now take an arbitrary surface $T(x, y, z) = 0$ and at each point $P_1$ of the surface we choose the vector ( $U, V, W$ ) normal to the surface. Then by (16) $dS = 0$, and therefore

$$S(x, y, z) = S_1 \tag{18}$$

is constant on the surface. From two of the (compatible) equations (13) we can solve for $u$ and $v$ as functions on the surface,

$$u = u(U, V, x, y, z), \quad v = v(U, V, x, y, z) \tag{19}$$

If we next solve the differential equations (14) with these boundary values we obtain a special solution of our problem, namely a solution which has a constant value $S_1$ of $S$ on the selected surface $T(x, y, z) = 0$. This solution can be determined from a
single partial differential equation for the function $S$. For on substituting from (19) into the remaining equation of (13) we obtain

$$W = W(U, V, x, y, z), \tag{20}$$

which, on using (17), becomes

$$\frac{\partial S}{\partial z} = W\left(\frac{\partial S}{\partial x}, \quad \frac{\partial S}{\partial y}, \quad x, y, z\right) \tag{21}$$

this is known as the Hamilton-Jacobi equation of the problem.
A function $S(x, y, z)$ which has a given constant value on the surface $T(x, y, z) = 0$ and satisfies (21) represents a solution of the problem. The two functions $u, v$ which make the integral independent of the path can be found by solving any two of the (compatible) equations

$$F_u = \frac{\partial S}{\partial x}, \quad F_v = \frac{\partial S}{\partial y}, \quad F - uF_u - vF_v = \frac{\partial S}{\partial z}, \tag{22}$$

obtained by combining (13) and (17).

## C.3.3 3. The field of extremals

We now establish the connection between the two problems treated in Sections 1 and 2. It is this:

If $u(x, y, z)$ and $v(x, y, z)$ are two functions which make the Hilbert integral $S$ defined by (11) independent of the path, then the differential equations

$$x' = u(x, y, z), \quad y' = v(x, y, z) \tag{23}$$

have as solutions a two-parameter set ( $\infty^2$ ) of extremals, namely those which satisfy the condition that they are "transversal" to the surfaces $S(x, y, z) = S_1$. By "transversality" we mean here that the relation

$$U dx + V dy + W dz = 0 \tag{24}$$

is satisfied: it states that the vector ( $U, V, W$ ), defined by ( 13 ) in terms of $u$ and $v$, is orthogonal to any element $dx, dy, dz$ of the surface.

Consider a region in the $x, y, z$ space and associate with each point of this region a vector ( $u, v$ ), which is continuous and has continuous first order partial derivatives. The set of these vectors defined over the given region is called a field. In the present context one speaks of a field of extremals and $u(x, y, z), v(x, y, z)$ are then called slope functions of the field.

The following converse of the theorem just enunciated also holds: If a field of $\infty^2$ extremals is constructed in such a way that it is transversal to a given surface $T(x, y, z) = 0$, and if $u$ and $v$ are its slope functions, defined by (23), then the Hilbert integral $S$, given by (11), is independent of the path.

Before proving these theorems we note the following corollary:
Let the curve $C$ in (11) be an extremal of the field; the Hilbert integral (11) then reduces to the variational integral $I = \int_{z_1}^{z_2} F dz$. Hence the value of this integral taken between each pair of "corresponding" points on the surfaces $S(x, y, z) = S_1$ and $S(x, y, z) = S_2$ (i.e. between points on the same extremal transversal to $S_1$ and $S_2$ ) is the same for all such pairs of points (see Fig. 1). The surfaces $S(x, y, z) =$ constant and
the $\infty^2$ transversals may be regarded as generalizations of wave-fronts and rays of geometrical optics.

To prove the first theorem we consider a fixed curve $C$ which satisfies (23) and is transversal to the surface $S(x, y, z) = S_1$, and apply a linear variation to it, i.e. we replace $x$ by $x + a\xi$ and $y$ by $y + b\eta$ where $a$ and $b$ are small parameters and $\xi$ and $\eta$ are arbitrary but fixed differentiable functions of $z$, which vanish for $z_1$ and $z_2$. Now since, by hypothesis, $S$ is independent of the path,

$$\left(\frac{\partial S}{\partial a}\right)_0 = 0, \quad \left(\frac{\partial S}{\partial b}\right)_0 = 0, \tag{25}$$

where the suffix 0 indicates that we substitute $a = b = 0$ after differentiation.

Differentiation of $F[u(x, y, z), v(x, y, z), x, y, z]$ gives

$$\frac{\partial F}{\partial a} = (F_u u_x + F_v v_x + F_x)\, \xi. \tag{26}$$

From (11) we obtain, with the help of (26),

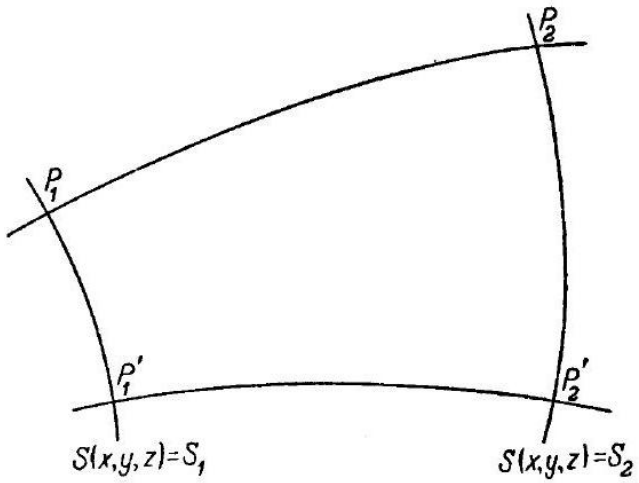



Fig. 1. Illustrating a generalization of the concept of wave-fronts and rays of geometrical optics. The variational integral (1) has a constant value for all extremals such as $P_{1 P_{2}$, $P_1'P_2'$, . which are transversal to the surfaces $S_1$ and $S_2$.}

$$\left(\frac{\partial S}{\partial a}\right)_0 = \int_{z_1}^{z_2} \{F_u u_x + F_v v_x + F_x)\,\xi + (\xi' - u_x\xi)\,F_u - v_x\xi F_v + (x' - u)\,\frac{\partial F_u}{\partial a} + (y' - v)\,\frac{\partial F_v}{\partial a}\Big\}_0 dz \tag{27}$$

The terms in the second line vanish, since the curve is assumed to satisfy (23). In the first line several terms cancel and we obtain

$$\left(\frac{\partial S}{\partial a}\right)_0 = \int_{z_1}^{z_2} (F_x\xi + F_u\xi')\,dz \tag{28}$$

or, on integrating by parts,

$$\left(\frac{\partial S}{\partial a}\right)_0 = \int_{z_1}^{z_2} \left(F_x - \frac{d}{dz}F_u\right)\xi dz, \tag{28a}$$

and similarly

$$\left(\frac{\partial S}{\partial b}\right)_0 = \int_{z_1}^{z_2} \left(F_y - \frac{d}{dz}F_v\right)\eta dz. \tag{28b}$$

The right-hand side of equations (28a) and (28b) are the first variations of $I$ [cf. (6a) and (6b)], and (25) shows that they vanish. Hence the curve $C$ satisfies Euler's equations, i.e. it is an extremal, and the theorem is proved.

To establish the converse theorem we construct a field $f_1$ of $\infty^2$ extremals transversal to a given surface $T(x, y, z) = 0$, and compare it with another field $f_2$ of $\infty^2$ extremals. The latter field is constructed in the following way. We solve the Hamilton-Jacobi equation (21) with the boundary condition that $S(x, y, z)$ shall be constant on $T(x, y, z) = 0$. If $u$ and $v$ are determined from (22), the solution can be represented by the integral (11). Then, according to the theorem just established, the equations (23) define a field $f_2$ of extremals which are transversal to $T = 0$. These two fields $f_1$ and $f_2$ must, however, be identical, since they satisfy the same differential equations and the same boundary conditions on $T = 0$. Hence, for the given field $f_1$, the integral $S$ is independent of the path.

## C.3.4 4. Determination of all extremals from the solution of the Hamilton-Jacobi equation

So far we have only considered an $\infty^1$ set of solutions $S(x, y, z) =$ constant of the Hamilton-Jacobi equation, which correspond to an $\infty^2$ set of transversal extremals. To obtain all $\infty^4$ extremals we must consider a larger set of solutions $S$, namely $\infty^4$; these can be obtained by rotating the surface $T = 0$ round a point, and taking, in each case, $S(x, y, z) = S_1$. Assume that we have found the "complete" solution $S(x, y, z, \alpha, \beta)$ of (21), involving two parameters $\alpha$ and $\beta$. The function $S$, corresponding to any pair of values of $\alpha$ and $\beta$ can be represented by an integral of the form (11) which is independent of the path by making an appropriate choice of the two functions $u$ and $v : u = u(x, y, z, \alpha, \beta), v = v(x, y, z, \alpha, \beta)$. Hence not only $S$ but also $\partial S/\partial\alpha$ and $\partial S/\partial\beta$ are independent of the path. Now since $F = F[u(x, y, z, \alpha, \beta), v(x, y, z, \alpha, \beta), x, y, z]$

$$\frac{\partial F}{\partial \alpha} = F_u u_\alpha + F_v v_\alpha \tag{29}$$

and we obtain from (11)

$$\left.\begin{aligned} S_\alpha &= \int_{z_1}^{z_2} \left\{(x' - u)\,\tfrac{\partial F_u}{\partial \alpha} + (y' - v)\,\tfrac{\partial F_v}{\partial \alpha}\right\} dz \\ S_\beta &= \int_{z_1}^{z_2} \left\{(x' - u)\,\tfrac{\partial F_u}{\partial \beta} + (y' - v)\,\tfrac{\partial F_v}{\partial \beta}\right\} dz \end{aligned}\right\} \tag{30}$$

where

$$\left.\begin{aligned} \tfrac{\partial F_u}{\partial \alpha} &= F_{uu} u_\alpha + F_{uv} v_\alpha, \\ \tfrac{\partial F_v}{\partial \alpha} &= F_{vu} u_\alpha + F_{vv} v_\alpha, \end{aligned}\right\} \tag{31}$$

with similar expressions for the other two derivatives.
Since the integrals (30) are independent of the path, it follows that the expressions

$$\left.\begin{aligned} dS_\alpha &= \left[(x' - u)\,\tfrac{\partial F_u}{\partial \alpha} + (y' - v)\,\tfrac{\partial F_v}{\partial \alpha}\right] dz, \\ dS_\beta &= \left[(x' - u)\,\tfrac{\partial F_u}{\partial \beta} + (y' - v)\,\tfrac{\partial F_v}{\partial \beta}\right] dz, \end{aligned}\right\} \tag{32}$$

are total differentials; $S_\alpha$ and $S_\beta$ are thus functions of the end points $P_1$ and $P_2$ only, so that surfaces defined by constant values $S_\alpha$ and $S_\beta$ exist, say

$$\frac{\partial S(x, y, z, \alpha, \beta)}{\partial \alpha} = A, \quad \frac{\partial S(x, y, z, \alpha, \beta)}{\partial \beta} = B, \tag{33}$$

$A, B$ being constants. These equations must therefore represent the solutions of the differential equations

$$x' = u(x, y, z, \alpha, \beta), \quad y' = v(x, y, z, \alpha, \beta). \tag{34}$$

For on the surfaces defined by the equations (33), $dS_\alpha = 0, dS_\beta = 0$, and by (32) this implies (34), provided that the associated determinant does not vanish. Now, according to equation (31) the determinant may be written in the form

$$\begin{vmatrix} \frac{\partial F_u}{\partial \alpha} & \frac{\partial F_v}{\partial \alpha} \\ \frac{\partial F_u}{\partial \beta} & \frac{\partial F_v}{\partial \beta} \end{vmatrix} = \begin{vmatrix} F_{uu} & F_{uv} \\ F_{vu} & F_{vv} \end{vmatrix} \begin{vmatrix} u_\alpha & u_\beta \\ v_\alpha & v_\beta \end{vmatrix}. \tag{35}$$

The first factor on the right-hand side is the expression (9), which was assumed to differ from zero. The second term can only vanish at points where the equations (34) have no solutions for $\alpha$ and $\beta$, in other words at points where the field does not cover the space uniquely. If we exclude such cases, (34) represents $\infty^2$ sets of differential equations, each with $\infty^2$ solutions; all

these together must therefore be identical with the totality of $\infty^4$ extremals. We have proved that the solutions

$$x = x(z, \alpha, \beta, A, B), \quad y = y(z, \alpha, \beta, A, B) \tag{36}$$

of (33) are just this total set of $\infty^4$ extremals; hence the whole set of $\infty^4$ extremals can be obtained from a complete solution $S(x, y, z, \alpha, \beta)$ of the Hamilton-Jacobi equation by differentiation and elimination only, according to (33) and (36).

## C.3.5 5. Hamilton's canonical equations

Each Euler's equation (7) is a differential equation of the second order. It is often convenient to replace these two second-order equations by four differential equations of the first order. This can be done in many ways. The most symmetrical way gives the so-called Hamilton canonical equations, obtained as follows:

The equations (13) are regarded as a Legendre transformation (see p. 135), which replace the variables $u, v$ by $U, V$ (retaining $x, y, z$ ), and the function $F(u, v, x, y, z)$ by $W(U, V, x, y, z)$. We can write the last equation in (13) as

$$W = F - uU - vV \tag{37}$$

so that

$$dW = dF - udU - vdV - Udu - Vdv$$

Now

$$\begin{aligned} dF &= F_u du + F_v dv + F_x dx + F_y dy + F_z dz \\ &= Udu + Vdv + F_x dx + F_y dy + F_z dz \end{aligned}$$

and therefore

$$dW = -udU - vdV + F_x dx + F_y dy + F_z dz.$$

Since $W$ is to be regarded as a function of $U, V, x, y, z$, it follows that

$$W_U = -u, \quad W_V = -v, \quad W_x = F_x, \quad W_y = F_y, \quad W_z = F_z \tag{38}$$

If we now consider a curve $x = x(z), y = y(z)$ which satisfies the equations

$$x' = u(x, y, z), \quad y' = v(x, y, z) \tag{39}$$

these two equations, together with Euler's equations (7), may be written in the form

$$\left.\begin{aligned} x' = -W_U, \quad y' = -W_V, \\ U' = W_x, \quad V' = W_y \end{aligned}\right\} \tag{40}$$

Equations (40) are four differential equations of the first order for $x, y, U, V$ as functions of $z$, and are called Hamilton's canonical equations. They may be regarded as the Euler equations of the variational integral expressed in terms of the function $W(U, V, x, y, z)$. If we substitute from (37) and (39) into (1), the integral goes over into

$$I = \int_{z_1}^{z_2} \{W(U, V, x, y, z) + x'U + y'V\}\, dz \tag{41}$$

If $U, V, x, y$ are regarded here as four unknown functions of $z$ and the Euler equation for each of them is formed, then (40) is immediately obtained.

## C.3.6 6. The special case when the independent variable does not appear explicitly in the integrand

The case when $F$ does not explicitly depend on $z$ deserves special consideration.
We have in general for $F(x', y', x, y, z)$ that

$$\frac{dF}{dz} = F_{x'}x'' + F_{y'}y'' + F_x x' + F_y y' + F_z$$

Assume now that $F_z = 0$, and substitute for $F_x, F_y$ from the Euler equations (7). This gives

$$\frac{dF}{dz} = F_{x'}x'' + F_{y'}y'' + x'\frac{d}{dz}F_{x'} + y'\frac{d}{dz}F_{y'} = \frac{d}{dz}\left(x'F_{x'} + y'F_{y'}\right)$$

Hence

$$\frac{d}{dz}\left(F - x'F_x - y'F_{y'}\right) = 0$$

so that

$$F - x'F_{x'} - y'F_{y'} = \text{ constant.} \tag{42}$$

This expression, which is just the quantity $W$ on the extremal, is independent of $z$; $W$ is a constant of integration. The same result can be seen directly from the canonical equations (40), for if $F$ does not explicitly depend on $z, W$ is also independent of $z, W_z = 0$, and

$$\frac{dW}{dz} = W_U U' + W_V V' + W_x x' + W_y y'$$

this expression vanishes by virtue of (40).
In this case one may reduce the variational problem from three to two dimensions. Consider $y$ as a function of $x$; then $y' = (dy/dx)x'$, and $F(x', y', x, y) = F(x', dy/dx, x, y)$. In the same way $F_{x'}, F_{y'}$, and also $F - x'F_{x'} - y'F_{y'}$ can be regarded as functions of $x', dy/dx, x, y$. Now the equation

$$F - x'F_x - y'F_{y'} = W \tag{43}$$

can be solved with respect to $x'$ :

$$\left.\begin{array}{ll} x' & = \Phi\left(\frac{dy}{dx}, x, y, W\right), \\ y' & = \frac{dy}{dx}\Phi\left(\frac{dy}{dx}, x, y, W\right). \end{array}\right\} \tag{44}$$

Let us consider all curves for which $W$ has some given value; then the integral (1) can be replaced by the difference of two integrals, or

$$J = \int_{z_1}^{z_2}(F - W)dz = \int_{z_1}^{z_2}\left(x'F_{x'} + y'F_{y'}\right)dz = \int_{x_1}^{x_2}\left(F_{x'} + \frac{dy}{dx}F_{y'}\right)dx$$

If now the abbreviation

$$F_{x'} + \frac{dy}{dx}F_{y'} = f \tag{45}$$

is used, and $x', y'$ are eliminated with the help of (44), $f$ becomes a function of $dy/dx, x, y, W$, and

$$J = \int_{x_1}^{x_3} f\left(\frac{dy}{dx}, x, y, W\right)dx \tag{46}$$

Thus we have, for each value of $W$, a variational integral of one dimension less. [This reduction corresponds in mechanics to the transition from Hamilton's principle to Maupertuis' principle: see (88) below.] If $y(x)$ is found from the Euler equation corresponding to (46), the complete extremal is obtained on integrating (44).

## C.3.7 7. Discontinuities

It may happen that the function $F(u, v, x, y, z)$ is not everywhere continuous. The most important case (which frequently occurs in optics) is that in which $F$ has a finite discontinuity along a surface $\sigma(x, y, z) = 0$ for all values of $u, v$.

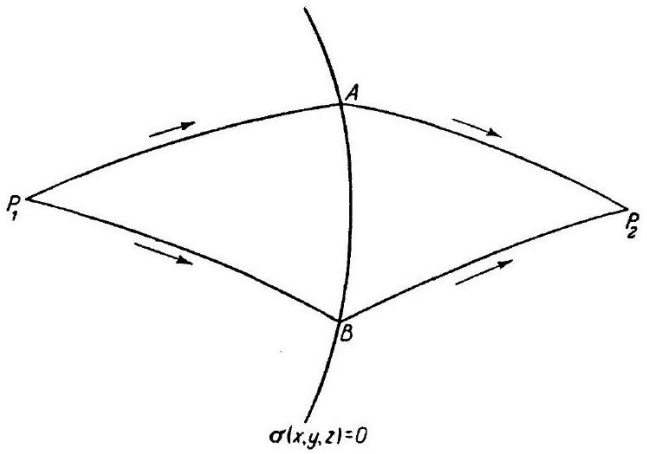



Fig. 2. Illustrating the variational analogue of the optical law of refraction.

It is obvious that outside the surface the extremals are again the solutions of the Euler equations (7); but if an extremal crosses the surface, it will have a discontinuity of direction (refraction). Let us distinguish the space to the left and right of the surface by the indices 1 and 2 (Fig. 2).

In order to find the "law of refraction" we have to establish the condition which ensures that the Hilbert integral $S$, defined by (11), extended from a point $P_1$ on the left of the surface to a point $P_2$ on the right of it, is independent of the path. Consider two paths $P_1AP_2$ and $P_1BP_2, A$ and $B$ being points on the surface. Then (with obvious notation) we demand that $S\,(P_1AP_2) = S\,(P_1BP_2)$ or

$$S_1\,(P_1A) + S_2\,(AP_2) = S_1\,(P_1B) + S_2\,(BP_2)\,. \tag{47}$$

Now on the closed path $P_1ABP_1$ on the left of the surface,

$$S_1\,(P_1A) + S_1(AB) + S_1\,(BP_1) = 0 \tag{48}$$

and on the closed path on the right

$$S_2\,(P_2B) + S_2(BA) + S_2\,(AP_2) = 0 \tag{49}$$

If we add equations (48) and (49), and use (47) and the relation $S(XY) = -S(YX)$, it follows that

$$S_1(AB) = S_2(AB) \tag{50}$$

The integral (11) taken along any path on the surface $\sigma = 0$ has, therefore, the same value, whether one takes as $u, v$ the values $u_1, v_1$ on the left, or the values $u_2, v_2$ on the right. Thus

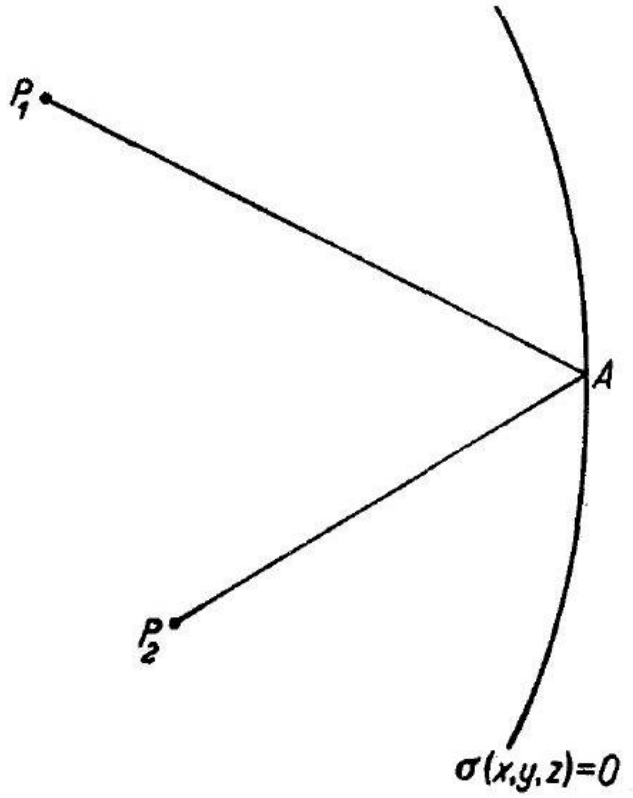



Fig. 3. Illustrating the variational analogue of the optical law of reflection.

the integrands must be equal in the two cases and the law of refraction is equivalent to the assertion that the expression

$$F + (x' - u)\, F_u + (y' - v)\, F_v \tag{51}$$

is continuous on the surface $\sigma = 0$. According to (13) this condition may be expressed in the form

$$(Ux' + Vy' + W)_1 = (Ux' + Vy' + W)_2\,, \tag{52}$$

where $x', y'$ are the derivatives of $x(z), y(z)$ for any curve on the surface. This can also be expressed by saying that the vector $U_2 - U_1, V_2 - V_1, W_2 - W_1$ is normal to the discontinuity surface:

$$(U_2 - U_1)\, dx + iV_2 - V_1)\, dy + (W_2 - W_1)\, dz = 0. \tag{53}$$

Very similar to the problem of refracted extremals is that of reflected extremals. One has to connect two points $P_1$ and $P_2$, situated in a region where $F$ is a continuous function of $x, y, z$, by a curve $P_1AP_2$ which has a discontinuity in direction at a point $A$ on a given surface $\sigma(x, y, z) = 0$, the points $P_1$ and $P_2$ being on the same side of $\sigma$ (Fig. 3).

It is obvious that the parts $P_1A$ and $AP_2$ must be extremals, and it follows, by a consideration like that leading to the law of refraction [(53)], that the condition for the validity of the independence theorem for an incident field (suffix 1) and a reflected field (suffix 2) is the law of reflection,

$$(U_1 + U_2)\, dx + (V_1 + V_2)\, dy + (W_1 + W_2)\, dz = 0 \tag{54}$$

The independence theorem also holds for any fields with a finite number of refracting or reflecting discontinuities. It will be shown in the next section that in all these cases, whether the extremal which connects $P_1$ and $P_2$ is continuous or has (directional) discontinuities, it will give rise to a minimum of the integral (1), provided that the function $F$ satisfies certain simple conditions along this curve.

## C.3.8 8. Weierstrass' and Legendre's conditions (sufficiency conditions for an extremum)

So far no distinction has been made between maxima and minima; the extremals considered (smooth or with a "kink") may even correspond to stationary cases which are not true extrema. We shall now derive conditions necessary for a real minimum.

Let $\bar{x}(z), \bar{y}(z)$ be a fixed extremal $\bar{C}$ embedded in a field $u(x, y, z), v(x, y, z)$, and let

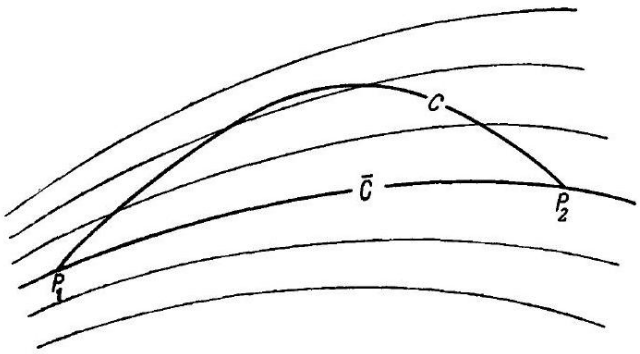



Fig. 4. Illustrating the definition of the Weierstrass $\mathscr{E}$-function.

$x(z), y(z)$ be any neighbouring curve $C$ also completely embedded in the field and with the same end points $P_1$ and $P_2$ as $\bar{C}$ (Fig. 4). The extremum will be a real minimum if

$$\int_C F\,(x', y', x, y, z)\,dz - \int_{\bar{C}} F\,(\bar{x}', \bar{y}', \bar{x}, \bar{y}, z)\,dz > 0 \tag{55}$$

According to Sections 2 and 3 we may replace the second integral by one extended not over $\bar{C}$ but over $C$, namely by

$$\int_C \{F(u, v, x, y, z) + (x' - u)\,F_u + (y' - v)\,F_v\}\,dz$$

this integral is independent of the path and reduces to $\int_{\bar{C}} F dz$, if the path is taken to coincide with $\bar{C}$. Hence (55) becomes

$$\int_C F dz - \int_{\bar{C}} F dz = \int_C \mathscr{E}\,(x', y', u, v, x, y, z)\,dz > 0 \tag{56}$$

where

$$\begin{aligned}\mathscr{E}\,(x', y', u, v, x, y, z) =& F\,(x', y', x, y, z) - F(u, v, x, y, z)\\ &- (x' - u)\,F_u - (y' - v)\,F_v,\end{aligned} \tag{57}$$

and the arguments in $F_u, F_v$ are the same as in $F(u, v, x, y, z)$. The function defined by (57) is called the $\mathscr{E}$-function (or the excess function) of Weierstrass; the arguments $x, y, z, x', y'$ refer

to a point on the curve $C$ and to its direction, while $u, v$ refer to the direction of the extremal of the field, which passes through the point $x, y, z$.

It is seen that $\mathscr{E}$ vanishes on any portion of $C$ which coincides with a field extremal. Now we can choose the field as the set of all $\infty^2$ extremals going through $P_1$. Then we construct a special curve $C$ which is such that between $P_1$ and a point $A$ the
curve coincides with a field extremal, from $A$ to a point $B$ on the given extremal it is a straight line, and from $B$ to $P_2$ it coincides with the given extremal (Fig. 5).

Then $\mathscr{E}$ vanishes on the parts $P_1A$ and $BP_2$, and there remains

$$\int_A^B \mathscr{E} dz > 0.$$

By letting $A$ approach $B$, it is seen that this inequality is only possible if

$$\mathscr{E}(x', y', \bar{x}', \bar{y}', \bar{x}, \bar{y}, \bar{z}) > 0; \tag{58}$$

here $\bar{x}, \bar{y}, \bar{z}$ refer to a typical point ( $B$ ) on the given extremal $\bar{C}$ and $x', y'$ refer to the direction $AB$ which is quite arbitrary. The formula (58)

Fig. 5. Derivation of Weierstrass' condition for a strong minimum. is Weierstrass' condition for a strong minimum; it is certainly a necessary condition. But on the assumption that the function $F$ is continuous in all its five arguments (hence $\mathscr{E}$ is continuous in its seven arguments), it follows that if (58) is satisfied for all points on the given extremal and for arbitrary directions $x', y'$, then the inequality (56) must hold for any neighbouring curve $C$ of arbitrary directions in a certain region surrounding $C$. Hence the condition (58) is also sufficient for a strong minimum. This minimum is, of course, only relative, for there may be several extremals having the property of giving a minimum value to the integral (1) in comparison with all neighbouring curves; which of them gives the absolute minimum cannot be decided in this way.

The inequality (58) has a simple geometrical interpretation. For a fixed point $\bar{x}, \bar{y}, \bar{z}$ in space, $F$ is a function of $\bar{x}', \bar{y}'$ alone; this function $F(\bar{x}', \bar{y}')$ can be represented by a surface in the three-dimensional $\bar{x}', \bar{y}', F$-space. (In Fig. 6 the twodimensional cross-section $\bar{x}'F$ is drawn.) Then

$$\mathscr{E}(x', y', \bar{x}', \bar{y}') = F(x', y') - [F(\bar{x}', \bar{y}') + (x' - \bar{x}') F_{\bar{x}'} + (y' - \bar{y}') F_{\bar{y}'}] \tag{59}$$

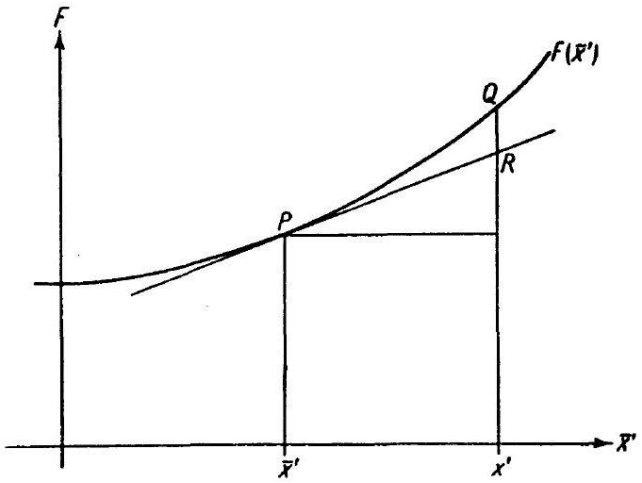



Fig. 6. A geometrical interpretation of Weimerstrass' condition for a strong minimum.

is obviously the distance $QR$ along the ordinate in $x', y'$, between the point $Q$ on the surface $F = F(\bar{x}', \bar{y}')$ and the point $R$ (see Fig. 6) where the tangential plane at $P(\bar{x}', \bar{y}')$ intersects

this ordinate. Hence $\mathscr{E}(x', y', \bar{x}', \bar{y}') > 0$ if the surface $F$ is above the tangential plane at $P$. If this holds for all $x', y'$ there is a strong minimum.

If, however, (58) holds only for small intervals of $\xi = x' - \bar{x}', \eta = y' - \bar{y}'$, there is a weak minimum; in this case we may expand $\mathscr{E}$ in powers of $\xi, \eta$ and obtain (if again we omit the arguments $x, y, z$ )

$$\mathscr{E}(x', y', \bar{x}', \bar{y}') = F(x', y') - F(\bar{x}', \bar{y}') - \xi F_{x'} - \eta F_{\bar{v}'} + \frac{1}{2}\left[F_{x'x'}\xi^2 + 2F_{x'\bar{y}'}\xi\eta + F_{\bar{y}'\bar{y}'}\eta^2\right] + \ldots$$

For small values $\xi$ and $\eta$ the quadratic terms are decisive and must evidently be positive for a minimum. Thus we obtain Legendre's condition (necessary and sufficient) for a weak minimum*:

$$F_{\tilde{x}'\tilde{x}'} > 0, \quad F_{\tilde{x}'\tilde{x}'}F_{\tilde{y}'\tilde{y}'} - F_{\tilde{x}'\tilde{y}'}{}^2 > 0. \tag{60}$$

## C.3.9 9. Minimum of the variational integral when one end point is constrained to a surface

The $\mathscr{E}$-function provides a simple solution of the problem of finding the minimum of the variational integral (1) with respect to all curves that have one end point $P_1$ in common and the other constrained to a given surface $\sigma(x, y, z) = 0$.

The curve must obviously be one of the $\infty^2$ extremals through $P_1$, the problem is "which of them?" Now amongst these $\infty^2$ extremals there is just one † that is transversal to the surface $\sigma = 0$ and it is easily seen that this represents the solution of

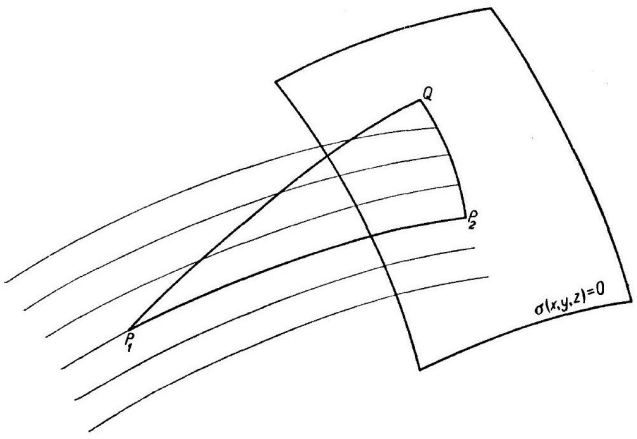



Fig. 7. Determination of the minimum of the variational integral with respect to all curves which have one end point fixed and the other constrained to a surface.

the problem. To show this let $\boldsymbol{P}_2$ be the point in which this extremal intersects the surface $\sigma = 0$, and surround it by the field of all extremals that are transversal to the surface. Let $P_1Q$ be any extremal through $P_1$ and $Q$ its point of intersection with the surface (Fig. 7). Then the Hilbert integral $S(P_2, Q)$ vanishes. Hence the integral

0

- This condition has an immediate generalization when one deals not with two but with a larger number of variables ( $n$ say). In order to have a minimum, the quadratic form in the $n$ variables must be positive definite; this implies that the associated determinant and its principal minors must all be positive.
  † It is assumed that $P_1$ is near enough to the surface, so that the case of several such extremals is

$S$ taken over the path $P_1QP_2$ is equal to the variational integral $I(P_1, Q)$. The difference $I(P_1, Q) - I(P_1, P_2)$ can now be expressed in terms of the $\mathscr{E}$-function just in the same way as before, and if (58) is satisfied, this difference is positive as long as $Q$ differs from $P_1$.

## C.3.10 10. Jacobi's criterion for a minimum

If an extremal can be embedded in a field and the Legendre condition is satisfied for all points of it between $P_1$ and $P_2$, then the integral $I$, defined by (1), is certainly a (weak) minimum. It remains to find a criterion for the existence of the field.

Let all $\infty^2$ extremals that pass through $P_1$ be given by

$$x = x(z, \alpha, \beta), \quad y = y(z, \alpha, \beta), \tag{61}$$

and let the given extremal $C$ be characterized by the values $\alpha = 0, \beta = 0$ :

$$x = x(z, 0, 0), \quad y = y(z, 0, 0). \tag{62}$$

The curves (61) form a field as long as there is one such curve through a given point $P(x, y)$ arbitrarily close to $C$, i.e. as long as the equations (61) have unique solutions for $\alpha, \beta$ as functions of $x, y$; the condition for this is that

$$\Delta = \begin{vmatrix} x_\alpha & x_\beta \\ y_\alpha & y_\beta \end{vmatrix} \neq 0 \tag{63}$$

This is Jacobi's criterion for a minimum.

The determinant $\Delta$ is a function of $z$ along the given extremal (62). The first point $\bar{P}$ where $\Delta = 0$ is called the conjugate point of $P_1$; for every interval $P_1P_2$ such that $P_2$ lies between $P_1$ and $\bar{P}$, there is a real minimum.

In $\bar{P}$ the given extremal is intersected by a neighbouring extremal (infinitesimally close to it); it is a point of the envelope of the set (61). Hence the limit of the field is determined by the envelope of the set of extremals (61). In optics these envelopes are the caustic surfaces.

## C.3.11 11. Example I: Optics

The general theory will now be illustrated by a few examples. The first concerns the shortest line in ordinary geometry and the line of shortest optical length in geometrical optics.

Euclidean geometry is based on the theorem of Pythagoras according to which the line element $ds$ is related to its projections $dx, dy, dz$ on to the axes of a rectangular system by means of the relation

$$(ds)^2 = (dx)^2 + (dy)^2 + (dz)^2 \tag{64}$$

The shortest lines between two points are then given by the minima of the variational integral

$$s = \int_{P_1}^{P_2} ds = \int_{z_1}^{z_2} \sqrt{x'^2 + y'^2 + 1}\, dz \tag{65}$$

Geometrical optics can be based on a generalization of this integral, namely on Fermat's principle of the shortest optical path (see § 3.3.2),

---

excluded.

$$\int_{P_1}^{P_2} nds = \int_{z_1}^{z_2} n(x,y,z)\sqrt{x'^2+y'^2+1}dz \tag{66}$$

where $n(x,y,z)$ is the index of refraction. We shall treat only the optical case, as (65) is the special case of (66) for $n=1$.

We now have

$$F(x',y',x,y,z) = n(x,y,z)\sqrt{x'^2+y'^2+1}. \tag{67}$$

Since $ds = \sqrt{x'^2+y'^2+1}dz$,

$$\left.\begin{aligned} U &= F_{x'} = \frac{nx'}{\sqrt{x'^2+y'^2+1}} = n\frac{dx}{ds} = ns_{x'} \\ V &= F_{y'} = \frac{ny'}{\sqrt{x'^2+y'^2+1}} = n\frac{dy}{ds} = ns_y \\ W &= F - F_{x'}x' - F_{y'}y' = \frac{n}{\sqrt{x'^2+y'^2+1}} = n\frac{dz}{ds} = ns_z \end{aligned}\right\} \tag{68}$$

where $s_x, s_y$, and $s_z$ are the components of the unit vector $s$ tangential to the curve $x = x(z), y = y(z)$. The law of refraction on a discontinuity surface of $n(x,y,z)$ may according to (53) be expressed in the form

$$(n_2 s_2 - n_1 s_1)\cdot dl = 0, \tag{69a}$$

where $dl(dx,dy,dz)$ is any line element of the surface. This equation implies that $s_1, s_2$ and the surface normal are coplanar, and that the angles $\theta_1$ and $\theta_2$ which $s_1$ and $s_2$ make with the surface normal are related by

$$n_2 \sin\theta_2 = n_1 \sin\theta_1, \tag{69b}$$

in agreement with the law of refraction (§ 3.2 (19)).
The Euler equations (7) associated with (66) are

$$\frac{\partial n}{\partial x}\sqrt{x'^2+y'^2+1} - \frac{d}{dz}\frac{nx'}{\sqrt{x'^2+y'^2+1}} = 0$$
$$\frac{\partial n}{\partial y}\sqrt{x'^2+y'^2+1} - \frac{d}{dz}\frac{ny'}{\sqrt{x'^2+y'^2+1}} = 0$$

or

$$\frac{d}{ds}\left(n\frac{dx}{ds}\right) = \frac{\partial n}{\partial x}, \quad \frac{d}{ds}\left(n\frac{dy}{ds}\right) = \frac{\partial n}{\partial y}. \tag{70a}$$

The corresponding equation in $z$, namely

$$\frac{d}{ds}\left(n\frac{dz}{ds}\right) = \frac{\partial n}{\partial z}, \tag{70b}$$

is an identity, as it follows from

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 + \left(\frac{dz}{ds}\right)^2 = 1. \tag{71}$$

To derive (70b) one may proceed as follows: First one differentiates (71) with respect to $s$. Next one multiplies (71) by $dn/ds$, the differentiated equation by $n$ and one adds the two equations. Finally, one uses (70a).* The three scalar differential equations (70) are in agreement with the vector equation § 3.2(2) for the light rays.

Since $U, V$, and $W$ now represent the components of the ray vector [cf. §4.1(4)] it is seen from (12), that the $S$-function belonging to the geometrical optics field is just the Hamilton point characteristic function [cf. §4.1 (1)]. Moreover, by following the procedure leading to (21) one finds that the Hamilton-Jacobi equation of the present variational problem is the eikonal equation.

To investigate the Legendre condition (60) one has to determine the derivatives $F_{x'x'}$, etc. One has, in the present case,

$$F_{x'x'} = n\frac{1+y'^2}{(1+x'^2+y'^2)^{3/2}},$$
$$F_{y'y'} = n\frac{1+x'^2}{(1+x'^2+y'^2)^{3/2}},$$
$$F_{x'y'} = -n\frac{x'y'}{(1+x'^2+y'^2)^{3/2}},$$

so that

$$F_{x'x'}.F_{y'y'} - F_{x'y'} = \frac{n^2}{(1+x'^2+y'^2)^2} > 0. \tag{72}$$

Every extremal therefore gives a weak minimum (p. 732) if the Jacobi condition (63) is satisfied. But as the function $F$, for given $x, y, z$, i.e. for a given $n$, is convex downwards for all values of $x', y'$, it follows from the geometrical interpretation of Weierstrass' condition that the minimum is strong.

It remains to consider Jacobi's criterion. For $n = \text{const.}$, i.e. in ordinary Euclidean geometry, the extremals are obviously straight lines; since a bundle of straight lines through a point $P_1$ never has an envelope, each straight line gives a strong minimum of the distance between any two points of it. In geometrical optics on the other hand, where $n$ in general depends on $x, y, z$ (continuously or discontinuously), the bundle of rays from $P_1$ give rise to envelopes (caustic surfaces). To determine the nature of the extremum these surfaces have to be examined separately in every particular case.

## C.3.12 12. Example II: Mechanics of material points

As a second example we consider the mechanics of systems of material points. Here the independent variable is the time $t$, and the unknown functions are the Lagrangian coordinates $q_\alpha (\alpha = 1, 2, \ldots n)$ and their derivatives the velocities $u_\alpha = \dot{q}_\alpha$.

The variational problem is given by Hamilton's principle

$$\int_{t_1}^{t_2} L\,(u_1, u_2, \ldots q_1, q_2, \ldots t)\,dt = \text{ extremum} \tag{73}$$

where $L$ is the Lagrangian. In ordinary non-relativistic mechanics one has $L = T - \Phi$, where $T$ is the kinetic energy, a quadratic form of the $u_\alpha$, and $\Phi$ the potential energy; but (73) holds also in more general cases, when a magnetic force is acting and when the relativistic variation of mass is taken into account.

---

0

- Equation (70b) can be obtained in a more symmetrical way, which, however, introduces a superfluous Euler equation. One regards $x, y$, and $z$ as functions of a parameter $\lambda$; one then has three Euler equations connected by an identity and the parameter $\lambda$ is finally identified with $s$.

The function $F(u, v, x, y, z)$ is here replaced by $L(u, q, t)$; hence [cf. (13)] $U, V$ now correspond to the momenta

$$p_\alpha = \frac{\partial L}{\partial u_\alpha} \tag{74}$$

and $W$ to $-H$, where $H$ is the Hamiltonian

$$H = \sum_\alpha u_\alpha \frac{\partial L}{\partial u_\alpha} - L = \sum_\alpha u_\alpha p_\alpha - L. \tag{75}$$

If $L = T - \Phi$, (75) becomes

$$H = \sum_\alpha u_\alpha \frac{\partial L}{\partial u_\imath} - T + \Phi.$$

From Euler's theorem on homogeneous functions* it follows that

$$2T = \Sigma u_\alpha \frac{\partial T}{\partial u_\alpha} \left( = \Sigma u_\alpha \frac{\partial L}{\partial u_\alpha} \right),$$

and $H$ reduces to the total energy

$$H = T + \Phi. \tag{76}$$

The Euler equations (7) become Lagrange's equation of motion

$$\frac{d}{dt}\left(\frac{\partial L}{\partial u_\mathrm{a}}\right) - \frac{\partial L}{\partial q_\mathrm{x}} = 0; \tag{77}$$

and the canonical equations (40) now are

$$\frac{dq_\alpha}{dt} = \frac{\partial H}{\partial p_\alpha}, \quad \frac{dp_\alpha}{dt} = -\frac{\partial H}{\partial q_\alpha}, \tag{78}$$

where $H$ is to be regarded as a function of the $p_\alpha, q_\alpha$ and $t$. If $H$ is independent of the time $t$, formula (42) expresses the law of conservation of energy

$$H = \sum_\alpha u_\alpha p_\alpha - L = \text{ constant } = E. \tag{79}$$

In this case the Hamilton-Jacobi differential equation (21) becomes

$$\frac{\partial S}{\partial t} + H\left(\frac{\partial S}{\partial q_\alpha}, q_\alpha\right) = 0. \tag{80}$$

This gives on integration,

$$S = -Et + S_1(q_\alpha) \tag{81}$$

where, because of (80), $S_1$ satisfies the equation

$$H\left(\frac{\partial S_1}{\partial q_\alpha}, q_\alpha\right) = E. \tag{82}$$

0

- See, for example, R. Courant, Differential and Integral Calculus, Vol. II (Glasgow, Blackie and Son, 1936), p. 109.

From the solution of the Hamilton-Jacobi equation one obtains the momenta in accordance with (17):

$$p_\alpha = \frac{\partial S}{\partial q_\alpha} = \frac{\partial S_1}{\partial q_\alpha} \tag{83}$$

It follows that the line integral

$$\int_{P_1}^{P_2} \sum_\alpha p_\alpha dq_\alpha \tag{84}$$

is independent of the path joining $P_1$ and $P_2$ and is, therefore, zero when taken over a closed path in a simply connected region,

$$\oint p_\alpha dq_\alpha = 0 \tag{85}$$

If the functions $p_\alpha$ are multivalued it may happen that the integral over a closed path is not zero, but is a multiple of a certain constant period. This result is a generalization of the optical invariant of Lagrange (§3.3.1) and is one of Poincare's invariants. It may also be expressed in the form

$$\frac{\partial p_\alpha}{\partial q_\beta} - \frac{\partial p_\beta}{\partial q_\alpha} = 0 \tag{86}$$

If $H$ is independent of time, $t$ may be eliminated from the minimum principle according to the general procedure expressed by (45) and (46). One then has

$$J = \int_{t_1}^{t_2} (L+E)dt = \int_{t_1}^{t_2} \sum_\alpha \dot{q}_\alpha p_\alpha dt = \int_{P_1}^{P_2} \sum_\alpha p_\alpha dq_\alpha = \text{ extremum.} \tag{87}$$

This is Maupertuis' principle of least action, generalized for arbitrary $L$; it has to be understood in the following way: the equation (79) allows the elimination of the time derivatives $u_\alpha = \dot{q}_\alpha$, expressing them in terms of purely geometrical derivatives, say $dq_\alpha/dq_1 (\alpha = 2, \ldots n)$, using $q_1$ as an independent variable. The equation (87) represents a purely geometrical principle describing the orbits, not the motions. The latter are then to be found from (78).

If $L = T - \Phi$, then $E = T + \Phi$ and one has Maupertuis' original expression

$$J = 2\int_{t_1}^{t_2} T dt \tag{88}$$

which is to be understood in the same way. An example for this reduction of the problem of the motion to the problem of the orbit by the transition from Hamilton's to Maupertuis' principle is the treatment of electron optics given in Section 2 of Appendix II.

The Legendre condition can only be investigated for a given $L$. If $L = T - \Phi$ and $T$ is a quadratic form of the $u_\alpha$, the condition is obviously equivalent to the postulate that $T$ is positive definite. Then Weierstrass' condition is also satisfied and one has a strong minimum as long as Jacobr's condition holds. The latter leads to the investigation of dynamical foci and caustics, but is of little importance in practice.

The Legendre condition for the relativistic electron with the Lagrangian (cf. Appendix II),

$$L = -mc^2\sqrt{1 - \left(\frac{v}{c}\right)^2} + e\left(\frac{v}{c} \cdot A - \phi\right) \tag{89}$$

leads to the quadratic form

$$\frac{m}{\left[1-\left(\frac{v}{c}\right)^2\right]^{3/2}}\left\{\rho^2\left(1-\frac{v^2}{c^2}\right)+\left(\frac{\rho\cdot v}{c}\right)^2\right\} \tag{90}$$

in the components of the vector $\rho(\xi,\eta,\zeta)$ and is, therefore, always satisfied.

## C.4 Light Optics, Electron Optics and Wave Mechanics

In 1831 William Rowan Hamilton discovered the analogy between the trajectory of material particles in potential fields and the path of light rays in media with continuously variable refractive index. By virtue of its great mathematical beauty the "Hamiltonian Analogy" survived in the textbooks of dynamics for almost a hundred years, but did not inspire any practical applications until 1925 when H. Busch first explained the focusing effect of electric and magnetic fields on electron beams in optical terms. Almost at the same time E. Schrödinger took the Hamiltonian Analogy a step further by passing from geometrical optics to wave optics of particles with his wave equation, in which he incorporated the wavelength of particles, first conceived by Louis de Broglie in 1923.

Practical electron optics developed rapidly from 1928 onwards. By this time the Hamiltonian Analogy was widely known and inspired the invention of electronoptical counterparts of light-optical instruments, such as the electron microscope. Though the mathematical analogy is general, the two techniques are not exactly parallel. Some electron optical instruments such as cathode-ray tubes and systems with curved optic axes have no important counterparts in light optics. In the available space only those problems of electron optics will be considered whose light optical analogues were developed at length in the previous chapters of this work, so that the results can be transferred almost in toto, with few modifications. It may be noted that this applies in particular to the most recondite chapter of electron optics: the wave theory of lens aberrations.

### C.4.1 1. The Hamiltonian analogy in elementary form

We will show first that the determination of the trajectory of a charged particle can be reduced to an optical problem by the introduction of a suitable refractive index, variable from point to point.

Consider a particle of charge $e$ and of mass $m$, to which we will for the sake of simplicity refer as an electron, moving in a steady electrostatic potential field $\phi(x,y,z)$. By Newton's law of motion

$$\frac{dp}{dt}=eE=-e\,\mathrm{grad}\,\phi, \tag{1}$$

where $\boldsymbol{p}$ is the momentum vector. This law is valid at all velocities $\mathbf{v}$ if Newton's definition of the momentum $\boldsymbol{p}=\boldsymbol{m}\mathbf{v}$ is replaced by Einstein's

$$p=\frac{mv}{\sqrt{1-\beta^2}},\quad \left(\beta=\frac{v}{c}\right), \tag{2}$$

$c$ being the vacuum velocity of light.

It is convenient to split the equation of motion (1) into two equations, of which one is the equation of the trajectory, while the other specifies the "time table" according to which the electron moves along the trajectory. For this purpose we write $v=vs$, $p=ps$, $s$ being the unit vector in the direction of motion. Then

$$\frac{dp}{dt} = \frac{dp}{dt}s + p\frac{ds}{dt} = \frac{dp}{dt}s + p\frac{ds}{ds}\frac{ds}{dt} = \frac{dp}{dt}s + pv\frac{ds}{ds}.$$

Now it is well known from differential geometry that $ds/ds$ is a vector in the direction of the unit principal normal $\nu$, whose absolute value is equal to the curvature $1/\rho$ of the trajectory. Hence

$$\frac{dp}{dt} = \frac{dp}{dt}s + \frac{pv}{\rho}\nu.$$

From this relation and from (1) it follows that the instantaneous centre of curvature is in the plane which passes through the tangent $s$ and the electric vector $E = -\operatorname{grad}\phi$. Resolving grad $\phi$ in the two directions it follows that

$$\frac{dp}{dt}s + \frac{pv}{\rho}\nu = -e[(s \cdot \operatorname{grad}\phi)s + (\nu \cdot \operatorname{grad}\phi)\nu]. \tag{3}$$

Equating the first terms on the two sides we obtain a scalar equation, which can be called the "time table," because ultimately it leads to the position of the electron on the trajectory as function of time. After multiplying it by $v = ds/dt$, it can be integrated and gives

$$\frac{mc^2}{\sqrt{1-\beta^2}} = -e\phi + \text{ constant }. \tag{4}$$

This is Einstein's energy integral. For slow-moving particles ( $\beta \ll 1$ ), it goes over into Newton's integral, $\frac{1}{2}mv^2 + e\phi =$ constant.

We now restrict the problem, for convenience, by considering only electrons with the same integration constant, i.e. with the same total energy. This is the case if all electrons have started at a certain potential surface $\phi_0$ with zero velocity. In many practical problems this surface can be identified with the cathode. Writing

$$V = \phi - \phi_0$$

i.e. measuring the potential $V$ from this level, the energy integral now appears in the form

$$mc^2\left[\frac{1}{\sqrt{1-\beta^2}} - 1\right] = -eV \tag{4a}$$

Combining this with (2) we can express both in the form of a useful duuble equation

$$\left(1 - \frac{eV}{mc^2}\right)^2 = 1 + \left(\frac{p}{mc}\right)^2 = \frac{1}{1-\beta^2}. \tag{5}$$

Thus the scalar value $p$ of the momentum of these particles is determined as a function of the position $x, y, z$ by

$$mc^2\left[\sqrt{1 + \left(\frac{p}{mc}\right)^2} - 1\right] = -eV(x, y, z). \tag{6}$$

Consider now the second part of (3), i.e. the component at right angles to the direction of motion:

$$\frac{pv}{\rho} = -e[\nu \cdot \operatorname{grad} V]. \tag{7}$$

Expressing $v$ and $V$ in terms of $p$ by means of (2) and (6) we obtain the simple law

$$\frac{1}{\rho} = \frac{\nu \cdot \operatorname{grad} p}{p} = \nu \cdot \operatorname{grad}(\log p). \tag{8}$$

Equation (8) is identical with equation (14) of § 3.2 for the curvature of rays in a medium of refractive index $n$, that is proportional to $p$, and we thus obtain a formal analogy between the paths of electrons and of light.

It must be emphasized that the absolute value $p$ of the momentum is a function of position alone only for electrons of a fixed total energy; for electrons of a different energy it is another function, given by (6). Thus the refractive index is dependent on the electron energy. This too has its analogue in light optics, in the dependence of the refractive index on the colour of light. It will be shown later that this is a very appropriate analogy, as in both cases the refractive index turns out to be a function of the wavelength.

For slow electrons $p$ is proportional to the velocity, and this in turn is proportional to $\sqrt{V}$. The relativistic equations which we have used have the advantage that they clearly indicate that it is the momentum, and not the velocity which is the characteristic quantity. Moreover these results immediately suggest a generalization for the case of a general static field, electric and magnetic. It is well known in relativity theory that in the presence of a magnetic field one must replace the mechanical momentum $p$, which now we will write $p_m$, by the "total" momentum

$$p_{\text{tot}} = p_m + eA, \tag{9}$$

where $A$ is the vector potential.* This suggests that the refractive index, which in the electrostatic case was the component of $\boldsymbol{p_m}$ in the direction of the motion, will have to be replaced in the general case by the component of $p_{\text{tot}}$ in the same direction. This guess proves to be correct, but it will now be preferable to place electron optics in electromagnetic fields on a more solid and general foundation.

## C.4.2 2. The Hamiltonian analogy in variational form

The laws of geometrical optics may be derived from Fermat's principle (§ 3.3.2), according to which the path of light between two points $P_1$ and $P_2$ makes the optical length a minimum,

$$\int_{P_1}^{P_2} n ds = \min \tag{10}$$

It may be recalled that this strong formulation of Fermat's principle is valid only if the two end points are sufficiently close together, that is to say if there is no image of either $P_1$ or $P_2$ between these points on the ray connecting them. If $P_2$ is the image of $P_1$, equation (10) does not determine a ray but an infinitesimal pencil of rays connecting the two points, each with the same optical length. If $P_1$ and $P_2$ are drawn farther apart, so that an image appears between them, the weak formulation

$$\delta \int_{P_1}^{P_2} n ds = 0 \tag{10a}$$

again determines a ray, but this represents not a maximum, as is sometimes erroneously stated, but a stationary value of the integral, which is neither a minimum, nor a maximum.

Some consequences of Fermat's principle were discussed in Section 11 of Appendix I, on the calculus of variations. In Section 12 of Appendix I it was shown that the motion of a system of material points can be described in a similar variational form,

0

by Hamilton's principle, which for the special case of a single material point is expressed by the condition

$$\int_{P_1,t_1}^{P_2,t_2} L(\dot{x},\dot{y},\dot{z},x,y,z)dt = \text{m}^{\text{i}}\text{a} \tag{11}$$

This formulation is valid in any system of coordinates, but for illustration Cartesian coordinates, $x, y, z$ will suffice. $\dot{x}, \dot{y}, \dot{z}$ are the velocity components, $L$ is the Lagrangian function. Departure and arrival are specified in four dimensional space-time, and are assumed to be fixed in the variational process. It was shown in Appendix I how the equations of motion in the Lagrangian and in the Hamiltonian form can be deduced as consequences of the principle (11) [App. I, eqs. (77), (78)].

For a relativistic electron, with charge $e$ and rest mass $m$, the Lagrangian function is

$$L = -mc^2\sqrt{1-\beta^2} - e\left(\phi - \frac{1}{c}\dot{r}A\right). \tag{12}$$

Here $\dot{\boldsymbol{r}} = \mathbf{v}$ is the velocity vector, $\phi$ is the electrostatic potential, and $A$ the magnetic or vector potential.* The Lagrangian (12) can be verified by substituting it into the Lagrangian equations of motion (Appendix I (77)) and making use of the electromagnetic relations § 2.1 (7) and § 2.1 (5),

$$E = -\frac{1}{c}\dot{A} - \operatorname{grad}\phi, \quad B = \operatorname{curl} A$$

The Lagrangian equations then appear in the form

$$\frac{dp}{dt} = e\left(E + \frac{1}{c}\vee\wedge B\right) \tag{13}$$

which is the Newton-Lorentz form of the equations of motion.

Hamilton's principle expressed by (11) is too general for the purpose of electron optics. It contains the time, which is without interest if the fields are stationary. Moreover it represents $\infty^5$ extremals; $\infty^4$ because at given total energy the points $P_1, P_2$ can be freely chosen on any two given surfaces, and a further infinity because the total energy has been left undetermined. In order to reduce the cumbersome dimensionality similar steps are taken in electron optics as in light optics.

First we fix the energy constant, that is to say we restrict the discussion to monochromatic light or monoenergetic electrons. As shown in Appendix I (87) this allows a reduction by one in the dimensionality of the problem. Hamiton's principle (11) is now replaced by the principle of least action,

$$J = \int_{P_1,t_1}^{P_2,t_2} (L+E)dt = \min \tag{14}$$

where $E$ is the total energy

$$E = \Sigma\dot{x}p_x - L.$$

In (14) the time $t$ appears only in a formal way, because on substituting for $E$ one obtains

$$J = \int_{P_1,t_1}^{P_2,t_2} \Sigma p_x\dot{x}dt = \int_{P_1}^{P_2} \Sigma p_x dx = \int_{P_1}^{P_2} p\cdot dr = \min \tag{15}$$

• It is remarkable that this fundamental result was discovered by K. Schwarzschild (Königl. Gess. Wiss. Göttingen, Math. Phys. Kl., 3 (1903), 126), three years before the advent of relativity.

In the second and third expression we have dropped $t_1, t_2$, because we are dealing with time-invariant fields only (otherwise the total energy would not remain constant), and in such fields the starting time $t_1$ is of no consequence, while the transit time $t_2 - t_1$ is fully determined by the path and by the energy constant.

Thus the principle of least action (14) is a complete analogue of Fermat's principle (10). The study of electron motion is reduced to an optical problem, if one defines the electron-optical refractive index as the component of the momentum in the direction of the trajectory. For a purely electrostatic field this result is seen to be equivalent to that found from more elementary considerations in the previous section. In this case the momentum is purely mechanical, and parallel to the trajectory, its value being given by (2):

$$p_m = \frac{mv}{\sqrt{1-\beta^2}}$$

In the presence of a magnetic field we must use the general definition of momentum components (Appendix I, (74) ), as derivatives of the Lagrangian with respect to the velocity components. For a single particle with the Lagrangian (12) these are

$$p_x = \frac{\partial L}{\partial \dot{x}} = \frac{m\dot{x}}{\sqrt{1-\beta^2}} + \frac{e}{c}A_x$$

etc. In vector form

$$p = \frac{mv}{\sqrt{1-\beta^2}} + \frac{e}{c}A \tag{16}$$

Thus, apart from an arbitrary constant factor, the general electron-optical refractive index is

$$n = \frac{mv}{\sqrt{1-\beta^2}} + \frac{e}{c}\,\text{A.s} \tag{17}$$

where $A.s$ is the component of the vector potential in the direction of motion. Again it is understood that this must be interpreted as a function of position for electrons of a given total energy.

There appears to exist an important difference between the general case and the special case of purely electric fields. In an electric field the refractive index is proportional to the mechanical momentum, which is a measurable physical quantity. In the general expression (17) the second term is a component of the vector potential, which is not a physical quantity but a function, such that its curl is equal to the magnetic induction $\boldsymbol{B}$. This makes it clear that the general electron-optical refractive index itself is not a physical quantity, but a Lagrangian function.* But this is equally true of the special refractive index in a purely electric field. In both cases we could add to it the component, in the direction of motion, of the gradient of any arbitrary function of position, without altering any of the physical consequences.

There exists a more important difference between the special and the general case, which is best explained if now we follow further the simplifying procedure of optics, and reduce the $\infty^4$ manifoldness of trajectory fields to $\infty^2$. We do this by selecting

0

- The Lagrangian (12) without the relativistic corrections was found by K. Schwarzschild, loc. cit., in 1903.

0

- This has been particularly emphasized by W. Ehrenberg and R. E. Siday, Proc. Phys. Soc., B, 62 (1949), 8, who have also pointed out the interesting fact that unless the magnetic field vanishes everywhere it is not possible to normalize $\boldsymbol{A}$ by a gauge transformation in such a way that it vanishes with the magnetic

those trajectories which have started on some surface $\mathscr{S}(x, y, z) = \mathscr{S}_0$ at right angles to this surface. In geometrical optics this surface can be considered as a wavefront, and it can then be shown that the pencil of trajectories will be everywhere at right angles to a family of surfaces $\mathscr{S}(x, y, z) =$ constant; this is the theorem of Malus and Dupin (§ 3.3.3).

A certain property of "transversality" exists also in the general case, as shown in Sections 2 and 3 of Appendix I, but its meaning is less simple than in the MalusDupin theorem. What remains at right angles to a family of surfaces $\mathscr{S} =$ const. is not the unit vector $s$ in the direction of the trajectories, but the momentum $p$. As the vector potential is defined only up to a gauge transformation, there exists an infinity of families of transversal surfaces, but in the presence of a magnetic field these cannot be made transversal to the trajectories by any gauge-normalizing procedure.

In the language of geometry, two-dimensional pencils of curves which are at right angles to a family of surfaces form a "normal congruence," otherwise they form a "skew congruence" (cf. §3.2.3). In light optics and in electrostatic electron optics the trajectories can be ordered in normal congruences, and the surfaces transversal to these are identified with the "wave-fronts." In magnetic fields the two-dimensional pencils of trajectories, usually called "beams," form skew congruences, and the concept of transversal wave-fronts cannot be applied to them. This is a rather essential difference between electron optics and light optics.

## C.4.3 3. Wave mechanics of free electrons

The dual nature of light was first conjectured by Einstein in 1905. Light propagates as if it were an electromagnetic wave, but it interacts with matter as if its energy were concentrated in photons, each with an energy quantum. Soon afterwards Einstein's conjecture was brilliantly verified by observations on photoelectrical and photochemical processes.

The dual conception of material particles is due to Louis de Broglie, who showed in 1923 that if there is a wavelength to be associated, in a relativistically invariant way, with a particle having a mechanical momentum $p_m$, this could only be of the form

$$\lambda = \frac{h}{p_m}, \tag{18}$$

where $h$ is a universal constant with the dimension of an action, which de Broglie identified with Planck's constant.

Soon afterwards Heisenberg, Born, and Jordan developed, quite independently of de Broglie, the first complete mathematical formulation of quantum mechanics, but their methods are less convenient for the discussion of free particles than Schrödinger's wave mechanics, which will now be briefly outlined.

Schrödinger combined de Broglie's ideas with Hamilton's, and was led to the problem of a wave-description of particle motion, which stands in the same relation to the dynamics of mass points as wave optics is to geometrical optics. His historical approach, which as we now know is correct only to a certain point, involved of course a certain amount of guessing because, though geometrical optics is logically contained in wave optics, the converse is not true.

Assume that there exists a wave field, whose intensity indicates the density of electrons in a similar way as the intensity of the electromagnetic field indicates the photon density. Moreover, assume that this is a scalar field, whose amplitude is
represented by some scalar function $\Psi(x, y, z, t)$; in order to account for its presumed wave-like properties, assume that it satisfies a wave equation

field.

$$\nabla^2\Psi = \frac{1}{u^2}\ddot{\Psi}$$

$u$ being the wave velocity-in general a function of position. This, of course, is a highly restrictive hypothesis, because the ordinary wave equation with constant velocity of propogation can be generalized in many different ways, of which this is only one, and the simplest.

Substituting for $\Psi$ a "monochromatic" wave

$$\Psi(x, y, z, t) = \psi(x, y, z)e^{-i\omega t}$$

we obtain a time-independent equation

$$\begin{aligned}\nabla^2\psi &= -\left(\frac{\omega}{u}\right)^2\psi \\ &= -\left(\frac{2\pi}{\lambda}\right)^2\psi \end{aligned} \tag{19}$$

in which only the wavelength $\lambda$ occurs, but not the wave velocity. Assume now that $\lambda$ is identical with the de Broglie wavelength, if one substitutes for the momentum $p_m$ the value which a particle would possess at the point $x, y, z$ according to classical mechanics. This can be calculated from (6). For simplicity consider a slow electron in an electrostatic field. In this case de Broglie's relation (18) combined with (6) gives

$$\lambda = \frac{h}{\sqrt{2m(E - e\phi)}}. \tag{20}$$

Substituting this into (19) we obtain

$$\nabla^2\psi + \frac{8\pi^2 m}{h^2}(E - e\phi)\psi = 0 \tag{21}$$

This is Schrödinger's wave equation for a free particle in a scalar potential field. As it is a time-independent equation, we can interpret it as describing the stationary, e.g. periodic motion of a particle in a field of force. But we can equally apply it to the stationary beams with which we deal in electron optics, in which many particles appear, one after the other, but all under identical conditions. In either case, it appears reasonable to assume, in accordance with Born's statistical interpretation, that the absolute square of the amplitude $|\psi|^2 = \psi\psi^\star$ is proportional to the particle density at the point $x, y, z$, measured over long times; or, what is the same in this case, that it is proportional to the probability that a particle will be observed at this spot at any instant.

The first verification of (21) was carried out by Schrödinger, who showed that it accounted for the atomic spectra if one assumed that the electron was bound in a field of force which was the same as in the old atom model of Rutherford and Bohr. More relevant for the present purposes was its verification in the case of free electrons, in the discovery of electron diffraction by Davisson and Germer, and independently by G. P. Thomson in 1927-1928.

For slow electrons with a kinetic energy equivalent to $eV$ electron volts, de Broglie's relation (18) gives a wavelength of approximately

$$\lambda = \sqrt{\frac{150}{V}} \text{ Angstrom units.}$$

Thus the wavelength of electrons which can be easily handled in laboratory experiments is of the order of fractions of an Angstrom, of the same order of magnitude as X-ray wavelengths.

Hence the wave behaviour of free electrons can be most easily demonstrated in experiments similar to X-ray diffraction in crystal lattices.

The terminology of X-ray analysis, which was also applied to electrons, is somewhat different from that of light optics. What is called X-ray or electron diffraction is really the interference of coherent secondary wavelets emitted by the more or less regularly arranged atoms in a lattice. Diffraction of electrons in the optical sense, by relatively large material obstacles, whose atomic structure does not come into play, results in extremely small diffraction angles, and was first observed by H. Boersch, in 1940 in an electron microscope.*

## C.4.4 4. The application of optical principles to electron optics

This elementary and incomplete sketch of wave mechanics is sufficient for almost all practical purposes of electron optics. It is not even necessary to make use of the extension of wave mechanics to magnetic fields, because in electron optics it is never necessary to go beyond Kirchhoff's approximation in diffraction problems. The assumptions on which this approximation is based are fully justified in electron optics. With the exception of the fields inside and in the immediate neighbourhood of atomic nuclei, there exist no electric fields which are so abrupt that they change appreciably within a wavelength of the electrons used in electron optics. This is a fortiori true for magnetic fields, and the specifically magnetic diffraction effects, which can be inferred from the Klein-Gordon or Dirac wave equation, are far too small to be detected in experiments with free electrons. Thus, with a certain caution which will be explained below, one can safely apply Kirchioff's diffraction theory (§ 8.3) which, in its somewhat extended form, may be summarized as follows:
I. Construct the wave-front from the source to the obstacle or object by measuring equal optical lengths along the rays as calculated by the rules of geometrical optics.
II. If the obstacle is a dark screen, consider the free part of the wavefront as undisturbed, in phase and in amplitude, and the rest as cut off. In the case of a partially transmitting object, trace the rays through it by the laws of geometrical optics, with the changes in phase and in intensity as determined by its complex refractive index. This is justified in all practical cases because the objects which one considers, e.g. in a microscope, have such small depth that diffraction effects inside them can be safely neglected.
III. In order to calculate the diffraction effects at some remote point, for instance in an optical image, calculate, by the rules of geometrical optics, the transmission between this point and each element of the wave-front emerging from the object or obstacle, and sum the complex amplitudes, taking into account the inclination factors.

Kirchinoff's method, as outlined, can be transferred without modification to electron optics if only electric fields are present, including the strong microscopic

fields due to the atomic structure of solid matter, which are predominantly electric. If however magnetic fields are present, the wave-fronts can no longer be determined by plotting equal numbers of de Broglie wavelengths along the trajectories. As has been pointed out at the end of Section 3 of this Appendix, the wave-fronts in magnetic fields are no longer orthogonal to the trajectories, but to the lines of total momentum or " $p$-lines," which have no simple geometric-optical interpretation. In principle one could determine the wave-fronts by starting from some given wavefront, and plotting optical lengths along the $p$-lines. Practically this method is almost useless, because it gives the phases, but not the amplitudes. In practice one is able to avoid this complication, by the artifice of imagining the magnetic lenses replaced

0

- H. Boersch, Naturwissenschaften, 28 (1940), 709. Boersch's diffraction photographs are reproduced in W. Glaser, Grundlagen der Elektronenoptik (Wien, Springer, 1952), p. 548.

by equivalent electric lenses, which produce the same image field, apart from a rotation of the image as a whole, and a certain rotational distortion of the image. By ignoring these specifically magnetic effects in the propagation process, and adding them only at the end, one can safely apply Kirchhoff's method, and obtain results which are sufficient for all practical purposes.

Though electron optics has the same general mathematical foundation as light optics, its practical structure is very different. Practical light optics has developed on the technical basis of grinding and polishing suitably shaped surfaces on transparent or reflecting solid media. In electron optics there is only one medium: the electromagnetic field. This has its limitations when it comes to the correction of aberrations, especially as space charges and space currents cannot be used for practical reasons. It was discovered early in the development of electron optics that it was impossible to construct diverging lenses and systems free from spherical or from chromatic aberrations, so long as one considered rotationally symmetrical fields only. This led to the development of systems without rotational symmetry, in which these limitations do not hold, and also to systems with curved optic axes, which are almost unknown in light optics. Hence the knowledge of optical instruments and design principles is useful in electron optics only in so far as it may pose problems and encourage inventions; the realizations are bound to be widely divergent. On the other hand optical thinking, based on the fundamental analogy of the two fields is likely to be as useful in the future as it was in the first quarter-century of electron optics.

## C.5 Asymptotic Approximations to Integrals

The purpose of this Appendix is to provide the mathematical background to methods of some generality, referred to in the main text, for obtaining asymptotic approximations to certain types of integral that frequently occur in optics.

### C.5.1 1. The method of steepest descent

THIS is a method* for obtaining, for large values of $k$, asymptotic approximations to complex integrals of the type

$$\int g(z)e^{kf(z)}dz \tag{1}$$

where $g(z)$ and $f(z)$ are independent of $k$.

For present purposes $k$ is to be considered real and positive, and is assumed so in the later discussion. In fact, however, the results to be given are in general also true for complex values of $k$, and it is desirable to make a few prefatory remarks about asymptotic expansions of a function of a complex variable, $\zeta$, say.

First, a definition of what is meant by an asymptotic expansion, due to Porncaré, † may be stated as follows: if

$$F(\zeta) = \sum_{m=0}^{n} \frac{a_m}{\zeta^m} + R_n(\zeta) \tag{2}$$

where, for all $n, \zeta^n R_n(\zeta) \to 0$ as $\zeta \to \infty$ for $\arg\zeta$ within a given interval, $a_0, a_1, \ldots a_n$ being constants, then one writes

$$F(\zeta) \sim a_0 + \frac{a_1}{\zeta} + \frac{a_2}{\zeta^2} + \ldots \tag{3}$$

and the right-hand side of (3) is called an asymptotic expansion of $F(\zeta)$ for the given range of $\arg\zeta$.

If $F(\zeta)$ is the quotient of two functions, say $G(\zeta)$ and $H(\zeta)$, one can write

$$G(\zeta) \sim H(\zeta)\left(a_0 + \frac{a_1}{\zeta} + \frac{a_2}{\zeta^2} + \ldots\right) \tag{4}$$

In fact it should be stressed that further on in the discussion the expansions considered are in the form of the right-hand side of (4), $H(\zeta)$ being $\exp(a\zeta)$, where $a$ is some constant.

Some of the main properties of asymptotic expansions are now briefly set out. ‡ If the series in (3) terminates, or converges for sufficiently large $|\zeta|$, it is asymptotic; but it often fails to converge for any value of $|\zeta|$. In general, for a given $F(\zeta)$ a particular expansion holds only for a specified range of $\arg\zeta$; if it holds for all $\arg\zeta$

then it converges. Again, for a given $F(\zeta)$ in the appropriate range of $\arg\zeta$ the asymptotic expansion is unique, in the sense that the coefficients in (3) are unique; on the other hand the asymptotic expansion of any one function also belongs to an infinity of other functions, that of $F(\zeta) + e^{-\zeta}$, for example, being the same as that of $F(\zeta)$ for $-\frac{1}{2}\pi < \arg\zeta < \frac{1}{2}\pi$. The asymptotic expansion of the product of two functions is obtained by multiplying out their respective asymptotic expansions. Finally (3) may be integrated term by term to yield unconditionally the asymptotic expansion of the integral of $F(\zeta)$; it also may be differentiated term by term to yield, if it exists, the asymptotic expansion of the differential of $F(\zeta)$.

In general, for a given (sufficiently large) value of $|\zeta|$, the moduli of the terms of (3) start by decreasing successively to a minimum and then subsequently increase. Roughly speaking, if the expansion is summed to any term prior to the smallest, the error is of the order of the first omitted term.* Evidently, the larger $|\zeta|$ is, the greater is the available accuracy. In physical applications it is often sufficient to use the first term only; to borrow an illustration from electromagnetic theory, the radiation field of a finite source distribution is the first term of the asymptotic expansion of the complete field in inverse powers of distance from the source.

The basis of the method for obtaining the asymptotic expansion of (1) in inverse powers of $k$ is to relate it to integrals of the form †

$$\int_0^\infty h(\mu)e^{-k\mu^2}d\mu \tag{5}$$

For the asymptotic development of (5) can be easily derived; it results from expanding $h(\mu)$ as an ascending power series in $\mu$ and carrying out the integration term by term.

This last remark is, in fact, the substance of Watson's lemma, ‡ which may be stated in the following form: Let

$$h(\mu) = \frac{1}{\mu^\alpha}\sum_{s=0}^{\infty} c_s \mu^{\beta s} \tag{6}$$

with radius of convergence $\rho$, where $\beta$ is real and positive and the real part of $\alpha$ is less than 1 . Let a real number $d$ exist such that $\mu^\alpha e^{-d\mu^2}h(\mu)$ is bounded for all real values of $\mu$ greater than $\rho$. Then (with $\Gamma$ denoting the gamma function)

$$\frac{1}{2k^{\frac{1}{4}(1-\alpha)}}\left\{c_0\Gamma\left(\tfrac{-\alpha+1}{2}\right) + c_1\Gamma\left(\tfrac{\beta-\alpha+1}{2}\right)\frac{1}{k^{\frac{1}{2}\beta}} + c_2\Gamma\left(\tfrac{2\beta-\alpha+1}{2}\right)\frac{1}{k^{\beta}} + c_3\Gamma\left(\tfrac{3\beta-\alpha+1}{2}\right)\frac{1}{k^{\frac{3}{2}\beta}} + \ldots\right\} \tag{7}$$

0

- Essentially due to P. Debye, Math. Ann., 67 (1909), 535.
  † H. Ponncarí, Acta Math., 8 (1886), 295.
  ‡ See, for example, H. and B. S. Jeffreys, Methods of Mathematical Physics (Cambridge University Press, 1946), Chapter 17; or A. Erdfucyi, Asymptotic Expansions (New York, Dover Publications, 1956).

is the asymptotic expansion of (5).
In cases of physical interest it is likely that $\alpha$ will be zero and $\beta$ unity. Often too, as already suggested, the first term of (7) provides an adequate approximation.

In order to be able to change the variable of integration so that (1) is represented by one or more integrals of the type (5) it is necessary that the path of integration should be made up of sections along each of which the imaginary part of $f(z)$ is constant and the real part of $f(z)$ decreases monotonically to $-\infty$. If this is not the case, the first step is to distort the path appropriately. The path distortion is, of course, governed by the fundamental rules of integration in the complex plane; here it will merely be shown how it may be possible to close a path, by means of sections having the required properties, with the supposition that the evaluation of any remaining contour integral is to be attempted by standard techniques.

Some quite general remarks can be made about paths along which the imaginary part of $f(z)$ is constant, irrespective of the nature of the particular function. These will suffice to indicate how two points in the complex plane are joined in the desired manner, though the method may perhaps fail if $f(z)$ has singularities. In the present connection an especially important role is played by the points at which

$$\frac{df}{dz} = 0. \tag{8}$$

These are called saddle-points, because they are the points where the real and imaginary parts of $f(z)$ are stationary with respect to position in the complex plane, without being absolute maxima or minima.*

Now if

$$f(z) = u(x, y) + iv(x, y) \tag{9}$$

it is easy to show from the Cauchy-Riemann relations that along any path $v(x, y) =$ constant the rate of change of $u(x, y)$ can only vanish at a saddle-point. In other words, on a path $v(x, y) =$ constant which does not traverse a saddle-point, $u(x, y)$ is strictly monotonic throughout, and on a path $v(x, y) =$ constant which goes through one or more saddle-points $u(x, y)$ is strictly monotonic between adjacent saddle-points and from the terminal saddle-points to the respective ends of the path. In fact, from any point, the direction in which $u(x, y)$ decreases most rapidly is one along which $v(x, y)$ is constant, and in this sense such paths are paths of steepest descent. Provided $f(z)$ is single-valued, or is made so by the use of branch-cuts, it follows that, starting from any point $(\ x_0, y_0\ )$, a path $v(x, y) = v\,(x_0, y_0)$ can be chosen which continues always in the direction of decreasing $u(x, y)$ and terminates either at infinity or at a singularity. It is readily seen that, apart from singularities, the only points at which a path $v(x, y) =$ constant can go in more than one direction are where $df/dz = 0$; thus, once the path $v(x, y) = v\,(x_0, y_0)$ has been started correctly, with $u(x, y)$ decreasing, no question of an alternative route arises unless a saddle-point is encountered, in which case at least one direction of decreasing $u(x, y)$ will be available.

Now suppose that the end-points of the given path of integration in (1) are $A$ and $B$, and that it is possible to find paths of steepest descent from $A$ and $B$, respectively, to infinity.

0

- For a method of estimating the error see H. and B. S. Jeffreys, loc. cit., Chapter 17.
† The choice of $\mu^2$ in the exponential of the integrand rather than the first (or some other) power of $\mu$ is of no great significance; an even power is convenient because it oftens happens that the path of integration in (1) begins and ends at infinity in such a way that (1) can be transformed to (5) with the lower limit replaced by $-\infty$.
‡ G. N. Watson, Theory of Bessel Functions (Cambridge University Press, 2nd ed., 1948), Chapter 8. See also H. and B. S. Jeffreys, loc. cit., Chapter 17.

If both these paths terminate at infinity in the same section of convergence of the integral, then the procedure is complete; but if the terminations at infinity are in different sections of convergence, then one must be joined to the other by a path $v(x, y) =$ constant along which the rate of change of $u(x, y)$ changes sign only once (at a saddle-point), or if necessary by several such paths via intermediary

regions at infinity. An asymptotic expansion, in the extended sense of (4), can be obtained for each of the different paths involved, but the true asymptotic expansion of the original integral is simply that for the path on which $u(x, y)$ attains its greatest value.* Similar remarks apply when the paths of steepest descent terminate in singularities.

It should be noted that there are cases outside the precise scope of the method just outlined which may nevertheless be treated by a slight modification. As a preliminary it is remarked that replacement of the upper limit in (5) by any positive number (independent of $k$ ) leaves the asymptotic expansion (7) unaltered. Thus the case can be treated involving a path of steepest descent along which $u(x, y)$ tends down to a finite value instead of $-\infty$. Again, it is possible to use a path $v(x, y) =$ constant traversing more than one point at which the rate of change of $u(x, y)$ changes sign.

Perhaps the commonest situation where the method of steepest descent is applicable is that in which the path of integration $v(x, y) =$ constant runs from a saddlepoint to infinity with $u(x, y)$ decreasing monotonically all the way. † A well-known formula for the first term of the asymptotic expansion is then readily derived. For suppose the saddle-point is at $z_0$, and change the variable of integration in (1) by the transformation

$$f(z) = f(z_0) - \mu^2 \tag{10}$$

Then (1) becomes

$$-2e^{kf(z_0)} \int_0^\infty \frac{g(z)}{f'(z)} \mu e^{-k\mu^2} d\mu \tag{11}$$

To get the first term in the asymptotic expansion of (11) the value of $\mu g(z)/f'(z)$ at $\mu = 0\,(z = z_0)$ is required. Provided $f''(z_0)$ is not zero and $g(z_0)$ is not infinite this is easily seen to be $g(z_0)\,/\sqrt{\{-2f''(z_0)\}}$, where the sign of the square root must be determined by an examination of each particular case. The required approximation is therefore

$$\sqrt{\left\{-\frac{\pi}{2f''(z_0)}\right\}}\, g(z_0)\, \frac{e^{kf(z_0)}}{\sqrt{k}} \tag{12}$$

A few remarks should be added regarding the cases in which the approximation (12) needs supplementing or is invalid. The full asymptotic expansion is, of course, obtained by expanding $\mu g(z)/f'(z)$ as a power series in $\mu$ and carrying out the integration in (11) term by term. As indicated in (7), the power of $k$ appearing in the first term of the asymptotic expansion is determined by the power of $\mu$ with which the series expansion of $\mu g(z)/f'(z)$ begins. If the first term of this series expansion is $A\mu^{-p}$, where $p$ must be less than unity for the integral to converge, then the first term of the asymptotic expansion is

$$-A\Gamma\left(\frac{-p+1}{2}\right) \quad \frac{e^{kf(z_0)}}{k^{\sharp(1-p)}}. \tag{13}$$

0

- The real and imaginary parts of an analytic function have no absolute maxima or minima in the complex plane.

0

Thus if $g(z_0)$ is infinite or $f''(z_0)$ is zero, (12) is invalid and must be replaced by (13). In these cases the factor multiplying $\exp\{kf(z_0)\}$ tends to zero less rapidly than $k^{-1/2}$, as $k$ tends to infinity. If $g(z_0) = 0$ or $f''(z_0)$ is infinite (12) is not incorrect, but then merely states that the factor multiplying $\exp\{kf(z_0)\}$ tends to zero more rapidly than $k^{-1/2}$, as $k$ tends to infinity.

If the path of integration is one of steepest descent running to infinity as before but starting from a point which is not a saddle-point, it is easy to see that the nonexponential part of the first term of the asymptotic expansion is, in general, proportional to $k^{-1}$, in contrast to the factor $k^{-1/2}$ of (12). Again, though, if $g(z)/f'(z)$ is singular or zero at the end point, the power of $k$ is different, and depends on the order of the singularity or zero.

What has been said so far covers fairly comprehensively the results yielded by the method of steepest descent for asymptotic expansions in the strict mathematical sense, in which $k$ is allowed to become indefinitely great, and other parameters are supposed to have prescribed values independent of $k$. But it has been shown that the form of the asymptotic approximation depends on particular conditions, for example, whether the path of steepest descent starts at a saddle-point or not. That is, it may depend on parameters other than $k$ in the sense that it changes abruptly when these parameters take certain critical values.* Thus for any given value of $k$, no matter how large, if the values of other parameters are sufficiently close to those for which the asymptotic form would be different, the formulae mentioned above clearly fail to provide good numerical approximations. Hence there is a practical requirement for further expressions which will give smooth transitions from one asymptotic form to another. These are naturally more elaborate functions of $k$ than (13), but it is worth mentioning three cases which can be satisfactorily treated. In the brief outline which follows it is tacitly assumed in each case that the only complexity is the specific one under discussion.

First, suppose that $f''(z_0)$ approaches zero. Then (12) is only a good approximation for increasingly large values of $k$, and an expression is required to effect the transition from (12) to the different form which is appropriate when $f''(z_0) = 0$. Since $f''(z_0)$ is nearly zero there must be a second saddle-point, $z_1$ say, near $z_0$. Then the required expression can be obtained by making the transformation

$$f(z) = \alpha - \beta\mu + \frac{1}{3}\mu^3$$

where

$$\alpha = \frac{1}{2}\left[f(z_1) + f(z_0)\right], \quad \frac{2}{3}\beta^{\frac{3}{2}} = \frac{1}{2}\left[f(z_1) - f(z_0)\right].$$

This transformation gives $z$ as a regular function of $\mu$ in the vicinity of $z_0$ and $z_1$, and leads to an asymptotic development expressed in terms of the Arry integral and its first derivative with argument $k\beta$.†

Next, suppose that the expansion of $\mu g(z)/f'(z)$ as a power series in $\mu$ has a radius of convergence which approaches zero because $g(z)$ has a simple pole which gets close to the saddle-point. Again, (12) is only a good approximation for increasingly large values of $k$, and an expression is required to effect the transition from (12) to

- It is, of course, possible that the greatest value of $u(x, y)$ will be attained on more than one of the paths.
† Strictly the commonest case is that in which the path begins and ends at infinity with $u(x, y)$ running monotonically from $-\infty$ up to a maximum at a saddle-point, and then monotonically from the maximum down to $-\infty$. Formula (12) holds in this case if multiplied by 2, but for the subsequent discussion it is more convenient to take a path starting at the saddle-point. Because the case in which it is arranged that the path of integration starts from or traverses a saddlepoint is so common, the method of steepest descent is sometimes called the saddle-point method.

0

the different form which is appropriate when the pole is at the saddle-point. This case has been treated by various authors.* The idea is to write the non-exponential part of the integrand as the sum of two terms, one containing the pole only, the other being regular. The latter can then be handled in the usual way and the former yields a Fresnel or error integral, with, in general, a complex argument.

Finally, suppose that the starting point of a path of steepest descent is not a saddlepoint, but approaches very close to one. In this case it is clear that the error integral can again be used to link the two asymptotic forms. †

## C.5.2 2. The method of stationary phase

This is an alternative, if very similar, method to that of steepest descent. Though perhaps less general and less immediately convincing analytically, it often has the advantage of closer contact with the physical problem. The integral to be considered is more appropriately written

$$\int g(z)e^{ikf(z)}dz \tag{14}$$

than in the form (1), and in practice the exponential commonly represents a travelling wave.

Keeping to the notation of (9), the paths of integration used are $v(x, y) =$ constant; but in (14) this means that the amplitude part of the exponential is constant along the path, whilst the phase part varies most rapidly, a reversal of the situation in the method of steepest descent. It can still be established that the only significant contribution to the integral arises from portions of the path in the vicinity of saddlepoints or end-points, but the physical interpretation of the mechanism by which this comes about is now in terms of "phase interference" (cf. §8.3, p. 386) rather than amplitude decay.

The method of stationary phase was first introduced explicitly by Lord Kelvin. ‡ A rigorous mathematical treatment which would justify the statements made above was subsequently given by Watson§; this treatment is based on the fact that, if $0 < m < 1, a$ is a positive constant and $F(x)$ has limited total fluctuation in the range $x \geqslant 0$, then

$$k^m \int_0^a x^{m-1}F(x)e^{ikx}dx \to F(0)\Gamma(m)e^{\frac{1}{2}imn}$$

as $k \to \infty$. ‖

Watson's discussion, however, is rather restricted in scope. In particular, it does not seem capable of producing the complete asymptotic expansion. This is given in some detail by Focke $q$ for the case in which $f(z)$ is a real function and the path of integration is confined to the real axis. Focke makes use of a neutralizing function, a method suggested earlier by van der Corput.**

---

- This is essentially a matter of non-uniform convergence. To take a trivial illustration, as $k \to \infty, 1/(1 + ka) \sim 1/ka$ for $a \neq 0$, and $\sim 1$ for $a = 0$.
  † The role of the Arry integral in uniform approximations of this type has long been recognized (see, for example, H. and B. S. Jeffreys, loc. cit., Chapter 17), but only relatively recently has a satisfactory method, as outlined, been fully explored. (See C. Chester, B. Friedman, and F. Ursell, Proc. Camb. Phil. Soc., 53 (1957), 599.)

0

- W. Paulr, Phys. Rev., 54 (1938), 925. H. Ott, Ann. d. Physik (5), 43 (1943), 393. P. C. Clemmow, Quart. J. Mech. Appl. Maths., 3 (1950), 241; Proc. Roy. Soc., A, 205 (1951), 286. B. L. van der Waerden, Appl. Sci. Res., Hague, B, 2 (1952), 33.
  † P. C. Clemmow, Quart. J. Mech. Appl. Maths., 3 (1950), 241.
  ‡ W. Thomson, Proc. Roy. Soc., A, 42 (1887), 80.
  § G. N. Watson, Proc. Camb. Phil. Soc., 19 (1918), 49.

The deductions from the method of stationary phase follow much the same pattern as those from the method of steepest descent. For example, in the case when the path of integration in (14) starts from a saddle-point at $z_0$ and runs to infinity along $v(x, y) =$ constant without encountering another saddle-point, the approximation corresponding to (12) is

$$\sqrt{\left\{-\frac{\pi}{2f''(z_0)}\right\}}\, g(z_0)\, e^{-i\pi}\, \frac{e^{ikf(z_0)}}{\sqrt{k}} \tag{15}$$

But one aspect in which there is some distinction between the methods should be noted. With a steepest descent path which starts at a saddle-point and does not go to infinity, the contribution of the point at the end of the path to the strict asymptotic expansion is zero in comparison with the saddle-point contribution by virtue of the extra exponential factor it contains. On the other hand, with a stationary phase path of the same type the contribution of the point at the end of the path is, in general, of the order of that of the saddle-point merely divided by $k^{1/2}$; it is excluded, therefore, from the asymptotic approximation only if the first term of the asymptotic expansion is alone retained.

To sum up, the methods of steepest descent and stationary phase, stripped of their mathematical expression, depend on choosing a path of integration in such a way that the integrand, by virtue of its exponential factor, contributes negligibly to the integral except in the vicinity of certain critical points, these being either saddlepoints or end-points of the path of integration.

## C.5.3 3. Double integrals

It was shown in § 8.3 and §9.1 that the problem of a field diffracted by an aperture leads to double integrals of the form

$$\iint g(x, y) e^{ikf(x,y)} dxdy \tag{16}$$

where $g(x, y)$ and $f(x, y)$ are independent of $k$, and the domain of integration is determined by the aperture. Clearly (16) is analogous to (14), and approximations for large values of $k$ are likewise provided by asymptotic expansions.

The theory of asymptotic expansions of double integrals is naturally more complicated than that of single integrals. The techniques of integration in the complex plane are only readily applicable to single integrals, so for double integrals it seems as though the attack must be based on an approach somewhat different from that developed in the preceding pages of this Appendix.

The case in which $f(x, y)$ is a real function is comprehensively discussed by Focke,* who uses the concept of a neutralizing function, in the paper already referred to in connection with the application of this idea to single integrals. The analysis again shows that contributions to the asymptotic expansion come only from regions in the vicinity of certain critical points, and that different types of critical point give rise to different powers of $k$ in the leading terms of their respective contributions. †

There are three types of critical point. The leading terms of their respective

‖ A result given by T. J. I'A. Bromwich, An Introduction to the Theory of Infinite Series (London, Macmillan, 1908), p. 447.

T J. Focke, Ber. Sächs. Ges. (Akad.) Wiss., 101 (1954), Heft 3.

** J. G. van der Corput, Indag. Math., 10 (1948), 201.

0

- J. Focke (loc. cit.). See also D. S. Jones and M. Kline, J. Math. Phys., 37 (1958), 1; N. Chako, C.R.Acad.Sci. Paris, 247 (1958), 436, 580, 637; M. Kline and I. W. Kay, Electromagnetic Theory and Geometrical Optics (New York, Interscience Publishers, 1965), Chapt. XII.

contributions to the asymptotic expansion are now briefly discussed, excluding cases in which a critical point is of more than one type.

A critical point of the first kind is a point within the domain of integration at which

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y} = 0. \tag{17}$$

Then near the critical point, $( x_0, y_0 )$ say,

$$f(x,y) = f(x_0,y_0) + \frac{1}{2}\alpha (x-x_0)^2 + \frac{1}{2}\beta (y-y_0)^2 + \gamma (x-x_0)(y-y_0) + \dots \tag{18}$$

where $\alpha = \partial^2 f/\partial x^2, \beta = \partial^2 f/\partial y^2, \gamma = \partial^2 f/\partial x \partial y$, the partial derivatives all being evaluated at $( x_0, y_0 )$. Now choose certain new variables of integration $\xi, \eta$ which are such that

$$f(x,y) = f(x_0,y_0) + \frac{1}{2}\alpha\xi^2 + \frac{1}{2}\beta\eta^2 + \gamma\xi\eta. \tag{19}$$

Then the required asymptotic approximation to (16) is

$$g(x_0,y_0)\, e^{ikf(x_0,y_0)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{1}{2}ik\left(\alpha\xi^2+\beta\eta^2+2\gamma\xi\eta\right)} d\xi d\eta = \frac{2\pi i\sigma}{\sqrt{|\alpha\beta-\gamma^2|}} g(x_0,y_0)\, \frac{e^{ikf(x_0,y_0)}}{k} \tag{20}$$

where the positive square root* is taken and

$$\sigma = \begin{cases} +1 \text{ for } \alpha\beta > \gamma^2, \alpha > 0, \\ -1 \text{ for } \alpha\beta > \gamma^2, \alpha < 0, \\ -i \text{ for } \alpha\beta < \gamma^2. \end{cases} \tag{21}$$

The expression (20) is the analogue for double integrals of the asymptotic approximation (15) for single integrals.

Critical points of the second kind are points on the curve bounding the domain of integration at which $\partial f/\partial s = 0$, where $ds$ is an element of arc of the bounding curve. The power of $k$ in the non-exponential part of the leading term of the corresponding contribution to the asymptotic expansion is $k^{-3/2}$, in contrast to the factor $k^{-1}$ of (20).

Finally, critical points of the third kind are corner points on the curve bounding the domain of integration, that is, points at which the slope of the curve is discontinuous. In this case the corresponding factor is $k^{-2}$.

- The term under the square root sign has a simple geometrical interpretation. Consider the surface $z = f(x,y)$. Let $R_1$ and $R_2$ be the principal radii of curvature and $K = 1/R_1R_2$ the Gaussian curvature at a typical point of the surface. Then [see, for example, G. Salmon, Analytic Geometry of Three Dimensions, Vol. I, revised by R. A. P. Rogers (London, Longmans, Green & Co., 5th ed., 1912), p. 411],

$$K = \frac{f_{xxx}f_{yy} - {f_{xy}}^2}{(1 + {f_x}^2 + {f_y}^2)^2}$$

where $f_x = \partial f/\partial x, f_{xx} = \partial^2 f/\partial x^2$, etc. At a critical point of the first kind $f_x = f_y = 0, f_{xx} = \alpha$, etc., and this expression reduces to

$$K = \alpha\beta - \gamma^2$$

† Critical points for double integrals were briefly mentioned by J. G. van der Corput, Indag. Math., 10 (1948), 201; and their qualitative significance for optics was pointed out by N. G. vAN Kampen, Physica, 14 (1949), 575.

## C.6 The Dirac Delta Function

The purpose of this Appendix is to summarize the main properties of the delta function,* which we found useful for representing point sources, point charges, etc. This function which is used extensively in quantum mechanics as well as in classical applied mathematics, may be defined by the equations

$$\left.\begin{array}{rl} \delta(x) = 0, & \text{when } x \neq 0, \\ \int_{-\infty}^{+\infty} \delta(x)dx = 1. & \text{(a)} \end{array}\right\} \tag{1}$$

Evidently $\delta(x)$ is not a function in the ordinary mathematical sense, † since if a function is zero everywhere except at one point, and the integral of this function exists, the value of the integral is necessarily also equal to zero. It is more appropriate to regard $\delta(x)$ as a quantity with a certain symbolic meaning.

Consider a set of functions $\delta(x, \mu)$ which, with increasing $\mu$, differ appreciably from zero only over a smaller and smaller $x$-interval around the origin and which are such that for all values of $\mu$

$$\int_{-\infty}^{+\infty} \delta(x, \mu)dx = 1 \tag{2}$$

Examples are the functions [see Fig. 8]

$$\delta(x, \mu) = \frac{\mu}{\sqrt{\pi}} e^{-\mu^2 x^2} \tag{3}$$

It is tempting to try to interpret the Dirac delta function as a limit of such a set, for $\mu \to \infty$, but it must be noted that the limit of $\delta(x, \mu)$ need not exist for all $x$. However

$$\mathrm{Lim}_{\mu\to\infty} \int_{-\infty}^{+\infty} \delta(x, \mu)dx \tag{4}$$

does exist and is equal to unity. We interpret any operation involving $\delta(x)$ as implying that this operation is to be performed with a function $\delta(x, \mu)$ of a suitable chosen set such as (3), and that the limit $\mu \to \infty$ is taken at the end of the calculation. With this interpretation (1b) evidently holds. The exact choice of the functions $\delta(x, \mu)$ is not important, provided that their oscillations (if any) near the origin are not too violent.

An important property of the Dirac delta function is the so-called sifting property, expressed by the relation

$$\int_{-\infty}^{+\infty} f(x)\delta(x - a)dx = f(a) \tag{5}$$

0

- Also known as the impulse function. It was brought into prominence by P. Dirac, The Principles of Quantum Mechanics (Oxford, Clarendon Press, 1930), but was known to mathematicians and physicists much earlier, chiefly through the writings of O. Heaviside. Cf. B. van der Pol and H. Bremmer, Operational Calculus based on the two-sided Laplace Integral (Cambridge University Press, 1950), pp. 62-66.
† The theory of the delta function may be made mathematically rigorous by using the notion of distributions, as developed by L. Schwartz in his Théorie des distributions (Paris, Hermann et Cie., Vol. I (1950), Vol. II (1951)). A simplified version of Schwartz' theory was developed by G. Temple, Proc. Roy. Soc., A, 228 (1955), 175, and is described fully in M. J. Lighthill, An Introduction to Fourier Analysis and Generalised Functions (Cambridge University Press, 1958).

Here $f(x)$ is any continuous function of $x$. The validity of (5) is seen immediately if $\delta(x-a)$ is replaced by $\delta(x-a,\mu)$ and the behaviour of the integral is examined for large values of $\mu$. Evidently when $\mu$ is large

$$\int_{-\infty}^{+\infty} f(x)\delta(x-a,\mu)dx \tag{6}$$

depends essentially on the values of $f(x)$ in the immediate neighbourhood of the point $x=a$ only, and the error of replacing $f(x)$ by $f(a)$ may be made negligible by taking $\mu$ sufficiently large. Using (lb), equation (5) then follows. This result implies that the

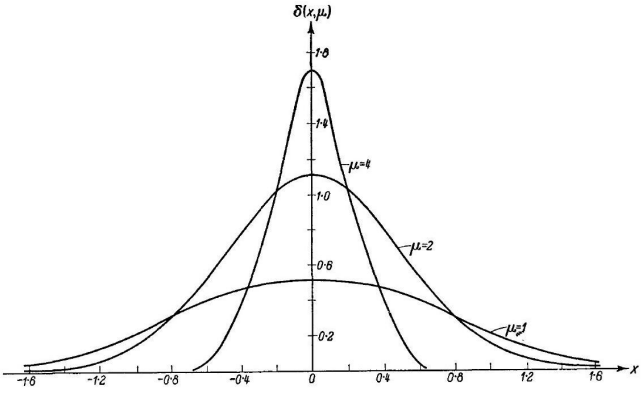



Fig. 8. Illustrating the significance of the Dirad delta function. The functions

$$\delta(x,\mu)=\frac{\mu}{\sqrt{\pi}}e^{-\mu^2x^2} \text{ for } \mu=1,2,4.$$

The total area under each curve is unity.
process of multiplying a continuous function by $\delta(x-a)$ and integrating over all values of $x$ is equivalent to the process of substituting $a$ for the argument of the function. Actually, for this result to hold, the range of integration need not be taken from $-\infty$ to $+\infty$. It is only necessary that the domain of integration contains the point $x=a$ in its interior. The result is also written symbolically as

$$f(x)\delta(x-a)=f(a)\delta(x-a), \tag{7}$$

the meaning of such a relation being that the two sides give the same results when used as factors in an integral. In particular, with $f(x)=x, a=0$, (7) gives

$$x\delta(x)=0. \tag{8}$$

With a similar interpretation the following relations may easily be verified:

$$\delta(-x)=\delta(x), \tag{9}$$

$$\delta(ax)=\frac{1}{|a|}\delta(x), \tag{10}$$

$$\delta\left(x^2-a^2\right)=\frac{1}{2|a|}\{\delta(x-a)+\delta(x+a)\}, \tag{11}$$

$$\int_{-\infty}^{+\infty}\delta(a-x)\delta(x-b)dx=\delta(a-b). \tag{12}$$

To verify the relation (10) for example, we compare the integrals of $f(x)\delta(ax)$ and of $f(x)\frac{1}{|a|}\delta(ax)$. We have

$$\int_{-\infty}^{+\infty} f(x)\delta(ax)dx = \pm\int_{-\infty}^{+\infty} f\left(\frac{y}{a}\right)\delta(y)\frac{1}{a}dy = \frac{1}{|a|}f(0)$$

where the upper or lower sign is taken in front of the second integral according as $a \gtrless 0$. We also have

$$\int_{-\infty}^{+\infty} f(x)\frac{1}{|a|}\delta(x)dx = \frac{1}{|a|}f(0)$$

because of (5). The integrals are seen to be equal and this is the meaning of (10). Similarly (12) implies that if the two sides are multiplied by a continuous function of $a$ or $b$ and are integrated over all values of $a$ or $b$ respectively, an identity is obtained.

Let us next consider what interpretation may be given to the derivatives of the $\delta$-function. We have, using the "approximation functions" $\delta(x,\mu)$ and integrating by parts

$$\int_{-\infty}^{+\infty} f(x)\delta'(x,\mu)dx = f(\infty)\delta(\infty,\mu) - f(-\infty)\delta(-\infty,\mu) - \int_{-\infty}^{+\infty} f'(x)\delta(x,\mu)dx$$

Proceeding to the limit as $\mu \to \infty$, the first two terms on the right disappear and we obtain

$$\int_{-\infty}^{+\infty} f(x)\delta'(x)dx = -f'(0) \tag{13}$$

Repeating this process we find that

$$\int_{-\infty}^{+\infty} f(x)\delta^{(n)}(x)dx = (-1)^n f^{(n)}(0) \tag{14}$$

The following relations may easily be verified:

$$\delta'(-x) = -\delta'(x), \tag{15}$$

$$x\delta'(x) = -\delta(x). \tag{16}$$

It is often convenient (see, for example Appendix VI) to express the Dirac delta function in terms of the Heaviside unit function (called also the step function) $U(x)$, this function being defined as

$$\left.\begin{aligned} U(x) &= 0, \quad \text{when } x < 0, \\ &= 1, \quad \text{when } x > 0. \end{aligned}\right\} \tag{17}$$

If as before, prime denotes the derivative with respect to the argument $x$, we obtain formally, on integrating by parts (with $x_1 > 0, x_2 > 0$ ),

$$\begin{aligned} \int_{-x_1}^{x_2} f(x)U'(x)dx &= [f(x)U(x)]_{-x_1}^{x_2} - \int_{-x_1}^{x_2} f'(x)U(x)dx \\ &= f(x_2) - \int_0^{x_2} f'(x)dx \\ &= f(x_2) - f(x_2) + f(0) \\ &= f(0) \end{aligned}$$

If we set $x = y - a, f(x) = f(y - a) = F(y)$ and proceed to the limit $x_1 \to \infty$, $x_2 \to \infty$, this becomes

$$\int_{-\infty}^{+\infty} F(y)U'(y - a)dy = F(a)$$

so that $U'$ has the sifting property. In particular, with $F \equiv 1, a = 0$ this relation becomes

$$\int_{-\infty}^{+\infty} U'(y)dy = 1$$

and shows that $U'$ satisfies a relation of the form (lb). Moreover $U'(x) = 0$ when $x \neq 0$. Hence we may identify the derivative of the unit function with the delta function:

$$\delta(x) = \frac{d}{dx}U(x) \tag{18}$$

The delta function may also be introduced with the help of the Fourier integral theorem

$$f(a) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dk \int_{-\infty}^{+\infty} f(x)e^{-ik(x-a)}dx \tag{19}$$

If we set

$$K(x - a, \mu) = \frac{1}{2\pi}\int_{-\mu}^{\mu} e^{-ik(x-a)}dk = \frac{\sin \mu(x - a)}{\pi(x - a)} \tag{20}$$

and invert the order of integration, then (19) may be formally written as

$$f(a) = \int_{-\infty}^{+\infty} f(x)K(x - a)dx \tag{21}$$

where $K(x - a)$ is regarded as the limit of $K(x - a, \mu)$ when $\mu \to \infty$. Strictly, this limit does not exist in the ordinary sense when $x - a \neq 0$,* but (21) has a similar symbolic meaning as the integrals discussed before, i.e. it should be interpreted as meaning that

$$f(a) = \lim_{\mu\to\infty} \int_{-\infty}^{+\infty} f(x)K(x - a, \mu)dx \tag{22}$$

Thus $K$ has the sifting property. If we set $f(x) = 1$ in (21) we see that the integral of $K(x)$ taken over all $x$ is equal to unity. We thus have another representation of the Dirac delta function, viz.

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-ikx}dk \tag{23}$$

i.e. $\delta(x)$ may be regarded as the Fourier transform of unity. There is a reciprocal relation, which follows from (21) on setting $f(x) = e^{ikx}, a = 0$,

$$\int_{-\infty}^{+\infty} \delta(x)e^{ikx}dx = 1 \tag{24}$$

0

- The limit exists and has the value zero if interpreted in the sense of a Cesaro limit [cf. B. van der Pol and H. Bremmer, loc. cit., pp. 100-104].

So far we have considered a space of one dimension only, but the definition may easily be extended to spaces of several dimensions. In particular, consider a space of three dimensions. Then the function

$$\delta(x,y,z) = \delta(x)\delta(y)\delta(z) \tag{25}$$

often also denoted by $\delta(r)$, where $r$ is the vector with components $x, y, z$, evidently satisfies relations analogous to (1), viz.

$$\left.\begin{aligned} \delta(x,y,z) &= 0, \text{ when } x \neq 0, y \neq 0, z \neq 0 \\ \iint_{-\infty}^{+\infty} \int \delta(x,y,z)dxdydz &= 1 \end{aligned}\right\} \tag{26}$$

The sifting property is now expressed by the relation

$$\iiint_{-\infty}^{+\infty} \int_{f(x,y,z)\delta(x-a,y-b,z-c)dxdydz=f(a,b,c)} \tag{27}$$

and $\delta(x,y,z)$ satisfies the Fourier reciprocity relations

$$\delta(x,y,z) = \frac{1}{(2\pi)^3} \iiint_{-\infty}^{+\infty} e^{-i(k_x x+k_y y+k_z z)} dk_x dk_y dk_z \tag{28}$$

$$\iint_{-\infty}^{+\infty} \int \delta(x,y,z) e^{i(k_x x+k_y y+k_z z)} dxdydz = 1 \tag{29}$$

## C.7 A Mathematical Lemma used in the Rigorous Derivation of the Lorentz-Lorenz Law (§ 2.4.2)

In this Appendix we shall establish a lemma used in § 2.4, according to which

$$\text{curl curl} \int_\sigma^\Sigma Q(r')\, G(R) dV' \to \int_\sigma^\Sigma \text{curl curl}\, Q(r')\, G(R) dV' + \frac{8\pi}{3} Q(r) \tag{1}$$

as $a \to 0$. Here $Q(r)$ is an arbitrary vector function of position and $G(R) = e^{ikR}/R$. The integrals are taken throughout the volume bounded externally by a surface $\Sigma$ and internally by a sphere $\sigma$ of radius $a$, which is centered at a point $P$ specified by the position vector $r(x,y,z)$. $R$ denotes the distance $|r - r'|$, where the vector $r'(x',y',z')$ specifies the position of a typical volume element $dV'$.

Let $A$ be an arbitrary vector function of position. The components of curl curl $A$ are

$$(\text{curl curl}\, A)_x = \frac{\partial^2 A_y}{\partial y \partial x} + \frac{\partial^2 A_z}{\partial z \partial x} - \frac{\partial^2 A_x}{\partial y^2} - \frac{\partial^2 A_x}{\partial z^2}$$

etc., so that

$$\left(\text{curl curl} \int_\sigma^\Sigma Q G dV'\right)_x = \frac{\partial^2}{\partial y \partial x} \int_\sigma^\Sigma Q_y G dV' + \frac{\partial^2}{\partial z \partial x} \int_\sigma^\Sigma Q_z G dV' - \left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) \int_\sigma^\Sigma Q_x G dV' \tag{2}$$

Now we have for any differentiable scalar function $F(r, r')$,

$$\frac{\partial}{\partial x} \int_\sigma^\Sigma F dV' = \int_\sigma^\Sigma \frac{\partial F}{\partial x} dV' + \text{Lim}_{\delta x \to 0} \frac{1}{\delta x} \left[\int_{\sigma'}^\Sigma F dV' - \int_\sigma^\Sigma F dV'\right] \tag{3}$$

where $\sigma'$ denotes a small sphere of radius $a$, centred on the point $T(x+\delta x, y, z)$. To evaluate the limit in (3), we note that the difference of the two integrals represents contributions from the two regions shown shaded in Fig. 9. The volume element may be expressed in the form $\delta V' = -\delta S'.\delta x.\rho_x$, where $\delta S'$ is the surface element and $\rho_x$ is the $x$-component of the unit radial vector $\rho$ pointing away from the point $P$. Hence

$$\mathrm{Lim}_{\delta x\to 0}\frac{1}{\delta x}\left[\int_{\sigma'}^{\Sigma} F dV' - \int_{\sigma}^{\Sigma} F dV'\right] = -\int_{\sigma} F\rho_x dS' \tag{4}$$

From (3) and (4) we have, if we sot $F = Q_j(r)G(R)$ where $Q_j (j = x, y \text{ or } z)$ is any of the Cartesian components of $Q$,

$$\frac{\partial}{\partial x}\int_{\sigma}^{\Sigma}(Q_j G)\, dV' = \int_{\sigma}^{\Sigma}\frac{\partial}{\partial x}(Q_j G)\, dV' - \int_{\sigma}(Q_j G)\,\rho_x dS' \tag{5}$$

Next we consider the partial derivatives of the second order. We differentiate (5) with respect to $x$ and use (5) again. This gives

$$\frac{\partial^2}{\partial x^2}\int_{\sigma}^{\Sigma}(Q_j G)\, dV' = \int_{\sigma}^{\Sigma}\frac{\partial^2}{\partial x^2}(Q_j G)\, dV' - \int_{\sigma}\frac{\partial}{\partial x}(Q_j G)\,\rho_x dS' - \frac{\partial}{\partial x}\int_{\sigma}(Q_j G)\,\rho_x dS' \tag{6}$$

Since
and

$$\left.\begin{array}{ll} \frac{\partial G}{\partial x} & = \frac{dG}{dR}\frac{\partial R}{\partial x} = -\rho_x\frac{d}{dR}\frac{e^{ikR}}{R} = \rho_x\left(\frac{1}{R} - ik\right)\frac{e^{ikR}}{R}, \\ dS' & = a^2 d\Omega, \end{array}\right\} \tag{7}$$

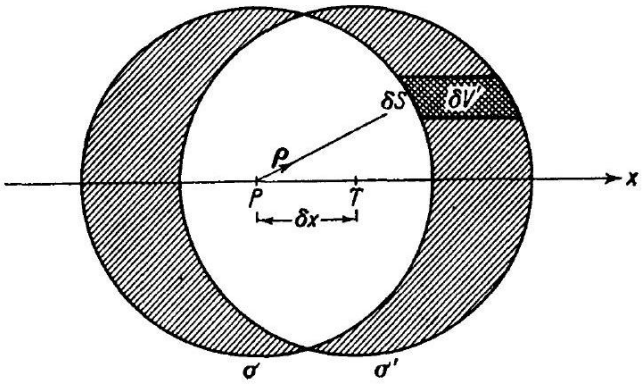



Fig. 9. Evaluation of

$$\mathrm{Lim}_{\delta x\to 0}\frac{1}{\delta x}\left[\int_{\sigma'}^{\Sigma} F dV' - \int_{\sigma}^{\Sigma} F dV'\right].$$

The centre of $\sigma$ is at $P(x, y, z)$, the centre of $\sigma'$ is at $T(x + \delta x, y, z)$.
where $d\Omega$ is an element of the solid angle, it follows that

$$\int_{\sigma}\frac{\partial}{\partial x}(Q_j G)\,\rho_x dS' = \int_{\sigma}\rho_x Q_j\frac{\partial G}{\partial x}dS' = \int_{\Omega}\rho_x^2 Q_j e^{ika}(1 - iak)d\Omega \to \frac{4\pi}{3}Q_j(r) \quad \text{as} \quad a \to 0, \tag{8}$$

$\Omega$ denoting the surface of the unit sphere. The last integral in (6) tends to zero with $a$, so that we have, as $a \to 0$,

$$\frac{\partial^2}{\partial x^2}\int_\sigma^\Sigma (Q_j G)\, dV' \to \int_\sigma^\Sigma \frac{\partial^2}{\partial x^2}(Q_j G)\, dV' - \frac{4\pi}{3}Q_j(r). \tag{9}$$

The mixed second order partial derivatives may be evaluated in the same way, and we have, for example,

$$\frac{\partial^2}{\partial y \partial x}\int_\sigma^\Sigma (Q_j G)\, dV' \to \int_\sigma^\Sigma \frac{\partial^2}{\partial y \partial x}(Q_j G)\, dV'. \tag{10}$$

The term $-\frac{4\pi}{3}Q_j(r)$ is now absent, since the integral corresponding to (8) is

$$\int_\sigma \frac{\partial}{\partial x}(Q_j G)\,\rho_y dS' = \int_\Omega \rho_x \rho_y Q_j (1 - iak) d\Omega$$

and this tends to zero with $a$.

On substituting from equations of the form (9) and (10) into (2) one finds that, as $a \to 0$,

$$\left[\text{curlcurl}\int_\sigma^\Sigma (QG)dV'\right]_x \to \left[\int_\sigma^\Sigma \text{curl}\,\text{curl}(QG)dV'\right]_x + \frac{8\pi}{3}Q_x \tag{11}$$

Similar expressions are obtained for the $y$ and $z$ components. Combining these three expressions into vector form, the formula (1) follows.

## C.8 Propagation of Discontinuities in an Electromagnetic Field (§3.1.1)

It was mentioned in § 3.1.1 that the eikonal equation of geometrical optics is identical with an equation which describes rigorously the propagation of discontinuities in an electromagnetic field. More generally, the four equations § 3.1 (11a)-(14a) governing the behaviour of the electromagnetic field associated with the geometrical light rays may be shown to be identical with equations which connect the field vectors on a moving discontinuity surface. It is the purpose of this Appendix to demonstrate this mathematical equivalence.

### C.8.1 1. Relations connecting discontinuous changes in field vectors

In § 1.1.3 we considered discontinuities in field vectors which arise from abrupt changes in the material parameters $\varepsilon$ and $\mu$, for example at a surface of a lens. Discontinuous fields may also arise from entirely different reasons, namely because a source suddenly begins to radiate. The field then spreads into the space surrounding the source and with increasing time fills a larger and larger region. On the boundary of this region the field has a discontinuity, the field vectors being in general finite inside this region and zero outside it. We shall first establish certain general relations which hold on any surface at which the field is discontinuous. For simplicity we assume that at any instant of time $t > 0$ there is only one such surface; the extension to several discontinuity surfaces (which may arise, for example, from reflections at obstacles present in the medium) is straightforward.

Let $F(x, y, z, t) = 0$ be any surface at which at least one of the field vectors is discontinuous. If this surface is fixed in space, $F$ is, of course, independent of $t$. The points on either side of the surface may be distinguished by the inequalities $F < 0$ and $F > 0$ respectively. Let $E$ be the electric vector and let

$$\left.\begin{aligned} E(x,y,z,t) &= E^{(1)}(x,y,z,t), \qquad \text{when } F(x,y,z,t)<0,\\ &= E^{(2)}(x,y,z,t), \qquad \text{when } F(x,y,z,t)>0. \end{aligned}\right\} \tag{1}$$

$E$ may then be written in the form

$$E = E^{(1)}U(-F) + E^{(2)}U(F), \tag{2}$$

where $U$ is the Heaviside unit function [cf. eq. (17), Appendix IV].
Using the representation (2), we derive expressions for the quantities curl $E$, div $E, \partial E/\partial t$, etc., which enter Maxwell's equations. To differentiate a sum or a product that contains a discontinuous factor, we apply the ordinary rules of differentiation and use relation (18) of Appendix IV,

$$\frac{d}{dx}U(x) = \delta(x), \tag{3}$$

where $\delta$ is the Dirac delta function. Thus from (2) we have, for example,

$$\begin{aligned} \operatorname{curl} E = U\left(-F'\right)\operatorname{curl} E^{(1)} + U\left(F'\right)\operatorname{curl} E^{(2)} \\ + \left[\operatorname{grad} U\left(-F'\right)\right] \wedge E^{(1)} + \left[\operatorname{grad} U\left(F'\right)\right] \wedge E^{(2)}. \end{aligned} \tag{4}$$

Now

$$\operatorname{grad} U(-F) = -\operatorname{grad} U(F) = -\frac{dU(F)}{dF}\operatorname{grad} F = -\delta\left(F'\right)\operatorname{grad} F, \tag{5}$$

and from (4) and (5) it follows that

$$\operatorname{curl} E = U(-F)\operatorname{curl} E^{(1)} + U(F)\operatorname{curl} E^{(2)} + \delta(F)\operatorname{grad} F \wedge (\Delta E), \tag{6}$$

where

$$\Delta E = E^{(2)} - E^{(1)} \tag{7}$$

In a similar way we find

$$\operatorname{div} \boldsymbol{E} = U\left(-F'\right)\operatorname{div} \boldsymbol{E}^{(1)} + U\left(F'\right)\operatorname{div} \boldsymbol{E}^{(2)} + \delta\left(F'\right)\operatorname{grad} \boldsymbol{F}.\Delta \boldsymbol{E}, \tag{8}$$

$$\frac{\partial \boldsymbol{E}}{\partial t} = U(-F)\frac{\partial E^{(1)}}{\partial t} + U\left(F'\right)\frac{\partial E^{(2)}}{\partial t} + \delta\left(F'\right)\frac{\partial F}{\partial t}\Delta \boldsymbol{E}. \tag{9}$$

Currents and charges may be represented in a similar form as the field vectors, but if the material parameters $\varepsilon$ and $\mu$ are discontinuous at the surface $F = 0$, there may be additional terms representing the current charge density $\hat{j}$ and the surface charge density $\hat{\rho}$. Their contribution is given by equations (17a) and (18a) of § 1.1, and we have in all

$$\boldsymbol{j} = \boldsymbol{j}^{(1)}U(-F) + \boldsymbol{j}^{(2)}U(F) + \hat{\boldsymbol{j}}|\operatorname{grad} F|\delta(F), \tag{10}$$

$$\rho = \rho^{(1)}U(-F) + \rho^{(2)}U(F) + \hat{\rho}|\operatorname{grad} F|\delta(F). \tag{11}$$

We now substitute from equations (6), (8), (9), (10), (11) and from similar expressions involving the other field vectors into Maxwell's equations § 1.1 (1)-(4). The terms with superscript (1) cancel out, and so do the terms with superscript (2), since the fields on either side of

the discontinuity surface satisfy separately Maxwell's equations. The remaining terms give the following relations connecting the discontinuous changes in the field vectors

$$\operatorname{grad} F \wedge \Delta \boldsymbol{H} - \frac{1}{c}\frac{\partial F}{\partial t}\Delta \boldsymbol{D} = \frac{4\pi}{c}\hat{j}|\operatorname{grad} F|, \tag{12}$$

$$\operatorname{grad} F \wedge \Delta \boldsymbol{E} + \frac{1}{c}\frac{\partial F}{\partial t}\Delta \boldsymbol{B} = 0 \tag{13}$$

$$\operatorname{grad} F \cdot \Delta \boldsymbol{D} = 4\pi\hat{\rho}|\operatorname{grad} F| \tag{14}$$

$$\operatorname{grad} F \cdot \Delta \boldsymbol{B} = 0 \tag{15}$$

It is of interest to note that these equations may be formally obtained from Maxwell's equations by replacing the vectors $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{D}, \boldsymbol{B}$ by the differences $\Delta\boldsymbol{E}, \Delta\boldsymbol{H}, \Delta\boldsymbol{D}, \Delta\boldsymbol{B}$; $j$ and $\rho$ by $\hat{j}$ and $\rho$; and the differential operators $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}, \frac{\partial}{\partial t}$ by the multiplicative operators $\frac{1}{|\operatorname{grad} F|}\frac{\partial F}{\partial x}, \frac{1}{|\operatorname{grad} F|}\frac{\partial F}{\partial y}, \frac{1}{|\operatorname{grad} F|}\frac{\partial F}{\partial z}, \frac{1}{|\operatorname{grad} F|}\frac{\partial F}{\partial t}$.

Let $n_{12}$ be the unit vector normal to the discontinuity surface and pointing from the region $F < 0$ (suffix 1), into the region $F > 0$ (suffix 2),

$$n_{12} = \frac{\operatorname{grad} F}{|\operatorname{grad} F|} \tag{16}$$

We also introduce the speed $v$ with which the discontinuity surface advances. To a small displacement $\delta\boldsymbol{r}(\delta x, \delta y, \delta z)$ from a point on the discontinuity surface $F(x, y, z, t) = 0$ to a point on a neighbouring discontinuity surface, there corresponds a change $\delta t$ in time, such that

$$\operatorname{grad} F.\delta r + \frac{\partial F}{\partial t}\delta t = 0 \tag{17}$$

In particular, for a displacement along the normal, $\delta\boldsymbol{r} = \delta s\boldsymbol{n}_{12}$, so that the speed $v$ is given by

$$v = \frac{ds}{dt} = -\frac{1}{|\operatorname{grad} F|}\frac{\partial F}{\partial t} \tag{18}$$

It follows that the relations (12)-(15) may also be written in the form*

$$n_{12} \wedge \Delta H + \frac{v}{c}\Delta D = \frac{4\pi}{c}\hat{j} \tag{12a}$$

$$n_{12} \wedge \Delta E - \frac{v}{c}\Delta B = 0 \tag{13a}$$

$$n_{12} \cdot \Delta D = 4\pi\hat{\rho} \tag{14a}$$

$$n_{12} \cdot \Delta B = 0 \tag{15a}$$

If the discontinuities in the field vectors arise from abrupt changes in the material parameters $\varepsilon$ and $\mu$ on a surface $F(x, y, z) = 0$ which is fixed in space, then $v = 0$ and (12a)-(15a) reduce to equations (25), (23), (19), and (15) of § 1.1.

## C.8.2 2. The field on a moving discontinuity surface

Consider a moving discontinuity surface which arises from the presence of a source which suddenly begins to radiate. We represent the surface in the form

$$F(x, y, z, t) \equiv \mathscr{S}(x, y, z) - ct = 0 \tag{19}$$

where $c$ is the vacuum velocity of light. The field vectors on this discontinuity surface will be denoted by small letters,

$$e(x,y,z) = E\left[x,y,z,\frac{1}{c}\mathscr{S}(x,y,z)\right] \tag{20}$$

with similar expressions for the other field vectors. Now in the region outside the moving surface (say $F(x,y,z,t) > 0$ ) there is no field, so that according to (1) and

(7) $\Delta E = -E^{(1)} = -e$, etc. The material equations § 1.1 (10), (11) give $d = \varepsilon e$, $b = \mu h$, and if we also set $\hat{j} = 0, \hat{\rho} = 0$, the relations (12)-(15) give

$$\begin{aligned} \operatorname{grad}\mathscr{S} \wedge \boldsymbol{h} + \varepsilon \boldsymbol{e} &= 0, \\ \operatorname{grad}\mathscr{S} \wedge \boldsymbol{e} - \mu \boldsymbol{h} &= 0, \\ \operatorname{grad}\mathscr{S} \cdot \boldsymbol{e} &= 0, \\ \operatorname{grad}\mathscr{S} \cdot \boldsymbol{h} &= 0. \end{aligned} \tag{24}$$

These equations are formally identical with the basic equations [(11a)-(14a), §3.1] of geometrical optics. Hence the field vectors on a moving discontinuity surface obey rigorously the same equations as the field vectors associated with the geometrical optics approximation to a time-harmonic field, the moving discontinuity surface corresponding to the geometrical wave-fronts.*

Evidently the moving discontinuity surface must obey the eikonal equation

$$\left(\frac{\partial\mathscr{S}}{\partial x}\right)^2 + \left(\frac{\partial\mathscr{S}}{\partial y}\right)^2 + \left(\frac{\partial\mathscr{S}}{\partial z}\right)^2 = n^2 \tag{25}$$

where $n^2 = \varepsilon\mu$. This equation follows as before. It is the consistency condition between the equations (21) and (22) and follows from them on eliminating $h$ or $e$ and using (23) or (24). According to (18), (19), and (25), the discontinuity surface is propagated with the velocity $v = c/n$.

## C.9 The Circle Polynomials of Zernike (§ 9.2.1)

In this Appendix, the circle polynomials discussed briefly in § 9.2.1 will be considered more fully. These polynomials were introduced and first investigated by Zernike,* in his important paper on the knife edge test and the phase contrast method, and were studied further by

0

- These equations are a generalization of a set of discontinuity relations which appear to have been known already to O. Heaviside. See also A. E. H. Love, Proc. Lond. Math. Soc., 1 (1904), 56; H. Bateman, Partial Differential Equations of Mathematical Physics (Cambridge University Press, 1932), p. 196; T. Levi-Civita, Caractéristiques des Systèmes Différentiels et Propagation des Ondes (Paris, Librairie Félix Alcan, 1932), § 10; R. K. Lunebura, Mathematical Theory of Optics, mimeographed lecture notes, Brown University, 1944; printed version published by University of California Press, Berkeley and Los Angeles, 1964, § 6 and §7; M. Kline, Comm. Pure and Appl. Math., 4 (1951), 239; A. Rubinowicz, Acta Phys. Polonica, 14 (1955), 209; M. Kline and I. W. Kay, Electromagnetic Theory and Geometrical Optics (New York, Interscience Publishers, 1965), pp. 37-51. The method of derivation used here is due to H. Bremmer, Comm. Pure and Appl. Math., 4 (1951), 419.

0

- It may also be shown that $e$ and $h$ obey the same transport equations [eq. (41) and (42), §3.1] as the complex amplitude vectors of the geometrical optics field. This result was first established by R. K. Luneburg, loc. cit., 1944, p. 46; loc. cit., 1964, p. 44. See also E. T. Copson, Comm. Pure and Appl. Math., 4 (1951), 427; M. Kline and I. W. Kay, loc.cit., p. 162.

him and Brinkman, † and by Nujbokr. ‡ They were later derived from the requirement of orthogonality and invariance alone by Bhatia and Wolf§; we shall follow substantially their treatment.

## C.9.1 1. Some general considerations

It is not difficult to show that there exists an infinity of complete sets of polynomials in two real variables $x, y$ which are orthogonal for the interior of the unit circle, i.e. which satisfy the orthogonality condition

$$\iint_{x^2+y^3\leqslant 1} V^{\star}_{(\alpha)}(x,y)V_{(\beta)}(x,y)dxdy = A_{\alpha\beta}\delta_{\alpha\beta} \tag{1}$$

Here $V_{(\alpha)}$ and $V_{(\beta)}$ denote two typical polynomials of the set, the asterisk denotes complex conjugate, $\delta$ is the Kronecker symbol, and $A_{\alpha\beta}$ are normalization constants, to be chosen later. The circle polynomials of Zernike are distinguished from the other sets by certain simple invariance properties which can best be explained from group theoretical considerations. It is however possiple to avoid the abstract formalism of group theory with the help of a kind of normalization. One considers first of all such sets as are "invariant in form" with respect to rotations of axes about the origin. By such invariance we mean that when any rotation

$$\left.\begin{aligned} x' &= x\cos\phi + y\sin\phi, \\ y' &= -x\sin\phi + y\cos\phi, \end{aligned}\right\} \tag{2}$$

is applied, each polynomial $V(x,y)$ is transformed into a polynomial of the same form, i.e. $V$ satisfies the following relation under the transformation (2):

$$V(x,y) = G(\phi)V\left(x',y'\right) \tag{3}$$

where $G(\phi)$ is a continuous function with period $2\pi$ of the angle of rotation $\phi$, and $G(0) = 1$.

Now the application of two successive rotations through angles $\phi_1$ and $\phi_2$ is equivalent to a single rotation through an angle $\phi_1 + \phi_2$. Hence it follows from (3) that $G$ must satisfy the functional equation

$$G\left(\phi_1\right)G\left(\phi_2\right) = G\left(\phi_1+\phi_2\right). \tag{4}$$

The general solution with the period $2\pi$ of this equation is well known, and is*

$$G(\phi) = e^{il\phi}. \tag{5}$$

Here $l$ is any integer, positive, negative, or zero. On substituting from (5) into (3), setting $x' = \rho, y' = 0$, and using (2) it follows that $V$ must be of the form

$$V(\rho\cos\phi, \rho\sin\phi) = R(\rho)e^{il\phi}, \tag{6}$$

where $R(\rho) = V(\rho, 0)$ is a function of $\rho$ alone. Next we expand $e^{il\phi}$ in powers of $\cos\phi$ and $\sin\phi$. Suppose that $V$ is a polynomial of degree $n$ in the variables $x = \rho\cos\phi$, $y = \rho\sin\phi$; it then follows from (6) that $R(\rho)$ is a polynomial in $\rho$ of degree $n$ and contains no power of $\rho$ of degree

0

- F. Zernike, Physica, 1 (1934), 689.
  † F. Zernike and H. C. Brinkman, Verh. Akad. Wet. Amst. (Proc. Sec. Sci.), 38 (1935), 11.
  ‡ B. R. A. Nijbole, Thesis (University of Groningen, 1942).
  § A. B. Bhatia and E. Wolf, Proc. Camb. Phil. Soc., 50 (1954), 40.

lower than $|l|$. Moreover $R(\rho)$ is evidently an even or an odd polynomial according as $l$ is even or odd. The set of the Zernike circle polynomials is distinguished from all other such sets by the property that it contains a polynomial for each pair of the permissible values of $n$ (degree) and $l$ (angular dependence), i.e. for integral values of $n$ and $l$, such that $n \geqslant 0, l \geqslant 0, n \geqslant |l|$, and $n - |l|$ is even. We denote a typical polynomial of this set by

$$V_n^l(\rho\cos\phi, \rho\sin\phi) = R_n^l(\rho)e^{il\phi}. \tag{7}$$

It follows from (1) and (7) that the radial polynomials $R_n^l(\rho)$ satisfy the relation

$$\int_0^1 R_n^l(\rho)R_{n'}^l(\rho)\rho d\rho = a_n^l\delta_{nn'} \tag{8}$$

where

$$a_n = \frac{A_n^l}{2\pi}. \tag{9}$$

For any given value $l$, the lower index $n$ can only take the values $|l|, |l|+2, |l|+4, \ldots$ The corresponding sequence $R_{|l|}^l(\rho), R_{|l|+2}^l(\rho), R_{|l|+4}^l(\rho)\ldots$ may be obtained by orthogonalizing the powers

$$\rho^{|l|}, \quad \rho^{|l|+2}, \quad \rho^{|l|+4}, \ldots \tag{10}$$

with the weighting factor $\rho$ over the interval $0 \leqslant \rho \leqslant 1$. Moreover, since only the absolute values of $l$ occur in (10),

$$R_n^{-l}(\rho) = \beta_n^l R_n^l(\rho), \tag{11}$$

where $\beta_n^l$ is a constant depending only on the normalization of the two polynomials $R_n^{-l}$ and $R_n^l$. In particular we may normalize in such a way that $\beta_n^l = 1$ for all $l$ and $n$, and then

$$V_n^{\pm m}(\rho\cos\phi, \rho\sin\phi) = R_n^m(\rho)e^{\pm im\phi}, \tag{12}$$

where $m = |l|$ is a non-negative integer.
The set of the circle polynomials contains $\frac{1}{2}(n+1)(n+2)$ linearly independent polynomials of degree $\leqslant n$. Hence every monomial $x^iy^j (i \geqslant 0, j \geqslant 0$ integers) and, consequently every polynomial in $x, y$ may be expressed as a linear combination of a finite number of the circle polynomials $V_n^l$. By Weierstrass' approximation theorem † it then follows that the set is complete.

## C.9.2 2. Explicit expressions for the radial polynomials $R_n^{\pm m}(\rho)$

Since $R_n^{\pm m}(\rho)$ is a polynomial in $\rho$ of degree $n$ and contains no power of $\rho$ lower than $m$ and is an even or odd polynomial according as $n$ is even or odd, it follows that $R$ may be expressed in the form

$$R_n^{\pm m}(\rho) = t^{\frac{m}{2}}Q_{\frac{n-m}{2}}(t), \tag{13}$$

0

- See, for example, M. Born, Natural Philosophy of Cause and Chance (Oxford, Clarendon Press, 1949, p. 153; Dover Publications, New York, 1964), p. 153.
  † See, for example, R. Courant and D. Hilbert, Methods of Mathematical Physics, Vol. 1 (New York, Interscience Publishers, 1st English edition, 1953), p. 65.

where $t = \rho^2$ and $\frac{Q_{n-m}}{2}(t)$ is a polynomial in $t$ of degree $\frac{1}{2}(n-m)$. According to (8), the $Q$ polynomials must satisfy the relations
where

$$\left.\begin{array}{l} \frac{1}{2}\int_0^1 t^m Q_k(t) Q_{k'}(t) dt = a_n^{\pm m}\delta_{kk'}, \\ k = \frac{1}{2}(n-m), \quad k' = \frac{1}{2}(n'-m). \end{array}\right\} \tag{14}$$

It follows that the polynomials $Q_0(t), Q_1(t), \ldots Q_k(t) \ldots$ may be obtained by orthogonalizing the sequence of natural powers

$$1, t, t^2, \ldots\ldots t^k, \ldots \tag{15}$$

with the weighting factor $w(t) = t^m$ over the range $0 \leqslant t \leqslant 1$. Now the well-known Jacobi (or hypergeometric) polynomials*

$$G_k(p,q,t) = \frac{(q-1)!}{(q+k-1)!} t^{1-q}(1-t)^{q-p}\frac{d^k}{dt^k}\left[t^{q-1+k}(1-t)^{p-q+k}\right] \tag{16}$$

$$= \frac{k!(q-1)!}{(p+k-1)!}\sum_{s=0}^{k}(-1)^s \frac{(p+k+s-1)!}{(k-s)!s!(q+s-1)!}t^s \tag{17}$$

( $k \geqslant 0, q > 0, p-q > -1$ ) may be defined as functions obtained by orthogonalizing (15) with the more general weighting function

$$w(t) = t^{q-1}(1-t)^{p-q}$$

over the range $0 \leqslant t \leqslant 1$. Their orthogonality and normalization properties are given by †

$$\int_0^1 t^{q-1}(1-t)^{p-q} G_k(p,q,t) G_{k'}(p,q,t) dt = b_k(p,q)\delta_{kk'} \tag{18}$$

where

$$b_k(p,q) = \frac{k![(q-1)!]^2[p-q+k]!}{[q-1+k]![p-1+k]![p+2k]}. \tag{19}$$

(With this choice of $b_k, G_k(p,q,0) = 1$ for all $k$.) On comparing (18) and (14) it follows that ‡

$$Q_k(t) = \sqrt{\frac{2a_n^{\pm m}}{b_k(m+1,m+1)}} G_k(m+1, m+1, t). \tag{20}$$

From (13) and (20) we obtain the following expression for the radial polynomials in terms of Jacobi polynomials:

$$R_n^{\pm m}(\rho) = \sqrt{\frac{2a_n^{\pm m}}{b_k(m+1,m+1)}}\rho^m G_k\left(m+1, m+1, \rho^2\right) \quad \left[k = \frac{1}{2}(n-m)\right] \tag{21}$$

0

- R. Courant and D. Hilbert, loc. cit., Vol. I, p. 90.
  † E. Kemble, The Fundamental Principles of Quantum Mechanics (New York, McGraw-Hill, 1937), p. 594.
  ‡ The sign of the square root on the right hand side of eq. (20) is determined from eq. (26) below.

Following Zernike, we choose the normalization so that for all $n$ and $m$

$$R_n^{\pm m}(1) = 1 \tag{22}$$

Then from (21) and (22)

$$\sqrt{\frac{b_k(m+1,m+1)}{2a_n^{\pm m}}} = G_k(m+1,m+1,1) \tag{23}$$

The value of $G_k(m+1,m+1,1)$ can be obtained from the generating function for the Jacobi polynomials.* We have

$$\frac{\left[z-1+\sqrt{1-2z\,(1-2\rho^2)+z^2}\right]^m}{(2z\rho^2)^m\sqrt{1-2z\,(1-2\rho^2)+z^2}} = \sum_{s=0}^{\infty}\binom{m+s}{s}G_s\left(m+1,m+1,\rho^2\right)z^s \tag{24}$$

For $\rho = 1$, the left-hand side reduces to $(1+z)^{-1}$ and we obtain on expanding it in a power series and on comparing it with the right-hand side:

$$G_s(m+1,m+1,1) = \frac{(-1)^s}{\binom{m+s}{s}} \tag{25}$$

From (25) and (23) it follows that

$$\sqrt{\frac{2a_n^{\pm m}}{b_k(m+1,m+1)}} = (-1)^{\frac{n-m}{2}}\binom{\frac{1}{2}(n+m)}{\frac{1}{2}(n-m)} \tag{26}$$

and using (16), (17), and (26) we finally obtain from (21) the following expressions for the radial polynomials:

$$R_n^{\pm m}(\rho) = \frac{1}{\left(\frac{n-m}{2}\right)!\rho^m}\left\{\frac{d}{d\,(\rho^2)}\right\}^{\frac{n-m}{2}}\left\{\left(\rho^2\right)^{\frac{n+m}{2}}\left(\rho^2-1\right)^{\frac{n-m}{2}}\right\} \tag{27}$$

$$= \sum_{s=0}^{1(n-m)}(-1)^s\frac{(n-s)!}{s!\left(\frac{n+m}{2}-s\right)!\left(\frac{n-m}{2}-s\right)!}\rho^{n-2s} \tag{28}$$

Explicit expressions for the first few polynomials are given in Table XXI on p. 465.
The normalization constant $a_n^{\pm m}$ is obtained from (26) and (19):

$$a_n^{\pm m} = \frac{1}{2n+2} \tag{29}$$

To obtain the generating function for the radial polynomials, we write $s$ in place of $k = (n-m)/2$ and $m+2s$ in place of $n$ in (21) and (26) and substitute into (24). We then obtain

$$\frac{\left[1+z-\sqrt{1+2z\,(1-2\rho^2)+z^2}\right]^m}{(2z\rho)^m\sqrt{1+2z\,(1-2\rho^2)+z^2}} = \sum_{s=0}^{\infty}z^s R_{m+2s}^{\pm m}(\rho). \tag{30}$$

Finally we shall evaluate the integral

0

- Cf. R. Courant and D. Hilbert, loc. cit., Vol. I, p. 91.

$$\int_0^1 R_n^m(\rho) J_m(v\rho)\rho d\rho$$

which, as we saw in Chapter 9, plays an important part in the Zernike-Nijboer diffraction theory of aberrations. We substitute for $R_n^m(\rho)$ the expression (27) and for the Bessel function $J_m$ its series expansion.* The resulting expression may be written in the form

$$\begin{aligned}&\int_0^1 R_n^m(\rho) J_m(v\rho)\rho d\rho \\ &\quad = \frac{1}{2\left(\frac{n-m}{2}\right)!}\sum_{s=0}^{\infty}\frac{(-1)^s}{s!(s+m)!}\left(\frac{v}{2}\right)^{m+2s} f\left(s, \frac{n-m}{2}, \frac{n+m}{2}, \frac{n-m}{2}\right),\end{aligned} \tag{31}$$

where

$$f(s,p,q,r) = \int_0^1 u^s \left(\frac{d}{du}\right)^p \{u^q(u-1)^r\}\, du \tag{32}$$

$p, q, r, s$ being non-negative integers. Integrating (32) by parts it follows that

$$f(s,p,q,r) = \left\{u^s\left(\frac{d}{du}\right)^{p-1}[u^q(u-1)^r]\right\}_0^1 - s\int_0^1 u^{s-1}\left(\frac{d}{du}\right)^{p-1}\{u^q(u-1)^r\}\, du. \tag{33}$$

Now if $r \geqslant p$ and $s+q-p \geqslant 0$, the first term on the right vanishes, so that

$$f(s,p,q,r) = -sf(s-1,p-1,q,r). \tag{34}$$

We consider separately the case $s \geqslant p$ and $s < p$.
When $s \geqslant p$, we have, on applying (34) $p$ times,

$$\begin{aligned}f(s,p,q,r) &= (-1)^p s(s-1)(s-2)\ldots(s-p+1)f(s-p,0,q,r) \\ &= \frac{(-1)^{p+r}s!}{(s-p)!}\int_0^1 u^{s+q-p}(1-u)^r du\end{aligned} \tag{35}$$

The integral in (35) is the Euler integral of the first kind (beta function) and its value is †$(s+q-p)!r!/(s+q+r-p+1)!$ Hence, for $s \geqslant p$,

$$f(s,p,q,r) = (-1)^{p+r}\frac{s!(s+q-p)!r!}{(s-p)!(s+q+r-p+1)!}. \tag{36}$$

Next consider the case $s < p$. Applying (34) $s$ times it follows that

$$\begin{aligned}f(s,p,q,r) &= (-1)^s s(s-1)\ldots f(0,p-s,q,r) \\ &= (-1)^s s!\left\{\left(\frac{d}{du}\right)^{p-s-1}[u^q(u-1)^r]\right\}_0^1 \\ &= 0\end{aligned} \tag{37}$$

0

- R. Courant and D. Hilbert, loc. cit., Vol. I, p. 484.
† R. Courant and D. Hlbert, loc, cit., Vol. I, p. 483.

We now substitute from (36) and (37) into (31) and introduce a new variable $l$ such that $s = \frac{1}{2}(n-m) + l$. We then obtain

$$\int_0^1 R_n^m(\rho) J_m(v\rho)\rho d\rho = \frac{(-1)^{\frac{3(n-m)}{2}}}{v} \sum_{l=0}^{\infty} \frac{(-1)^l}{l!(n+l+1)!} \left(\frac{v}{2}\right)^{n+2l+1} \tag{38}$$

The series on the right will be recognized as the expansion of $J_{n+1}(v)$. Since $n-m$ is even the factor $(-1)^{\frac{3(n-m)}{2}}$ may be replaced by $(-1)^{\frac{n-m}{2}}$, and we finally obtain

$$\int_0^1 R_n^m(\rho) J_m(v\rho)\rho d\rho = (-1)^{\frac{n-m}{2}} \frac{J_{n+1}(v)}{v} \tag{39}$$

## C.10 Proof of an Inequality (§ 10.7.3)

Let $f(\tau)$ and $g(\tau)$ be any two functions, generally complex, of the real variable $\tau$ and let $\lambda$ be a real parameter. Then

$$\int_{-\infty}^{+\infty} |f + \lambda g^\star|^2 \, d\tau = \int_{-\infty}^{+\infty} (f + \lambda g^\star)(f^\star + \lambda g) \, d\tau \geqslant 0, \tag{1}$$

or

$$\int_{-\infty}^{+\infty} f f^\star d\tau + \lambda \int_{-\infty}^{+\infty} (fg + f^\star g^\star) \, d\tau + \lambda^2 \int_{-\infty}^{+\infty} g g^\star d\tau \geqslant 0 \tag{2}$$

The minimum of this quadratic expression in $\lambda$ is obtained by differentiating:

$$\int_{-\infty}^{+\infty} (fg + f^\star g^\star) \, d\tau + 2\lambda \int_{-\infty}^{+\infty} g g^\star d\tau = 0 \tag{3}$$

The root $\lambda = \lambda_{\min}$ of this expression is

$$\lambda_{\min} = -\frac{1}{2} \frac{\int_{-\infty}^{+\infty} (fg + f^\star g^\star) \, d\tau}{\int_{-\infty}^{+\infty} g g^\star d\tau} \tag{4}$$

If this value is substituted into (2) we obtain

$$4 \left( \int_{-\infty}^{+\infty} f f^\star d\tau \right) \left( \int_{-\infty}^{+\infty} g g^\star d\tau \right) \geqslant \left( \int_{-\infty}^{+\infty} (fg + f^\star g^\star) \, d\tau \right)^2 \tag{5}$$

Let

$$f = \tau\psi(\tau), \quad g = \frac{d\psi^\star(\tau)}{d\tau} \tag{6}$$

Then

$$fg + f^\star g^\star = \tau \left( \psi \frac{d\psi^\star}{d\tau} + \psi^\star \frac{d\psi}{d\tau} \right) = \tau \frac{d}{d\tau} (\psi\psi^\star), \tag{7}$$

and (5) becomes, if we integrate by parts on the right and assume that* $\tau\psi\psi^\star \to 0$ as $\tau \to \pm\infty$,

$$4 \left( \int_{-\infty}^{+\infty} \tau^2 \psi\psi^\star d\tau \right) \left( \int_{-\infty}^{+\infty} \frac{d\psi}{d\tau} \frac{d\psi^\star}{d\tau} d\tau \right) \geqslant \left( \int_{-\infty}^{+\infty} \psi\psi^\star d\tau \right)^2 \tag{8}$$

This is the required inequality.
The equals sign in (8) can only hold if it holds in (1); this is only possible if $f \equiv -\lambda g^\star$, or, using (6), if

$$\frac{d\psi}{d\tau} = -\frac{1}{\lambda}\tau\psi. \tag{9}$$

The general solution of this differential equation is

$$\psi(\tau) = Ae^{-\frac{\tau^2}{2\lambda}} \tag{10}$$

where $A$ is a constant. Only solutions with $\lambda \geqslant 0$ apply, since otherwise $\psi(\tau)$ would not vanish at infinity. Hence (8) becomes an equality if and only if $\psi$ is a Gaussian function.

## C.11 Evaluation of T'wo Integrals (§ 12.2.2)*

In this Appendix we shall evaluate the two integrals of § 12.2 (8), (9)

$$\mathscr{J}_1 = \frac{1}{4\pi}\iiint\left\{\int_{V_1} e^{i(px_1+qy_1)}\frac{\partial^2}{\partial z_1^2}\left(\frac{e^{i\omega R/c}}{R}\right)\right\}dx_1dy_1dz_1 \tag{1}$$

$$\mathscr{J}_2 = \frac{1}{4\pi}\frac{\omega^2}{c^2}\iiint\left\{e_{V_1}\int(px_1+qy_1)\left(\frac{e^{i\omega R/c}}{R}\right)\right\}dx_1dy_1dz_1 \tag{2}$$

where

$$R = +\sqrt{x_1^2+y_1^2+z_1^2} \tag{3}$$

and $\omega^2 > c^2p^2$. We are interested in two cases, namely:
(a) $0 < y < d$, where $y$ and $d$ are constants and the volume $V_1$ of integration is the slab $-\infty < x_1 < \infty, -y \leqslant y_1 \leqslant d-y, -\infty < z_1 < \infty$, except for a small sphere of vanishingly small radius around the origin $x_1 = y_1 = z_1 = 0$.
(b) $y > d$ or $y < 0$, the volume $V_1$ of integration now being the full region $-\infty < x_1 < \infty, -y \leqslant y_1 \leqslant d-y, -\infty < z_1 < \infty$.

To evaluate $\mathscr{J}_1$ we apply Gauss's theorem, in the form

$$\iiint_{V_1}\operatorname{div} G dV_1 = \iint_{S_1} G.ndS_1, \tag{4}$$

where $\boldsymbol{G}$ is an arbitrary vector function of position and $\boldsymbol{n}\,(n_x, n_y, n_z)$ is the unit outward normal to the surface $S_1$ bounding the volume $V_1$. We take

$$G_{x_1} = G_{y_1} = 0, \quad G_{z_1} = \frac{1}{4\pi}\frac{\partial}{\partial z_1}\left[\frac{e^{i(px_1+qy_1+\omega R/c)}}{R}\right] \tag{5}$$

and obtain from (1) and (4)

$$\mathscr{J}_1 = \frac{1}{4\pi}\iint_{S_1} n_{z_1}\frac{\partial}{\partial z_1}\left(\frac{e^{i(px_1+qy_1+\omega R/c)}}{R}\right)dS_1. \tag{6}$$

0

- This condition is in fact satisfied whenever the integrals on the left in (8) converge [cf. H. Weyl, The Theory of Groups and Quantum Mechanics, translated from German (London, Methuen, 1931; also, New York, Dover Publications, Inc.), pp. 393-394].

Since $n_{z_1}$ is zero at each face ( $y_1 = -y$ and $y_1 = d - y$ ) of the slab, the integral $\mathscr{J}_1$ is zero in cases $(b)$†. In case (a), we must also include the contribution from the small sphere $\sigma_1$ around the origin. If $a$ is the radius of the sphere, we have $n_z = -z_1/a$ and the contribution is

$$\frac{1}{4\pi}\iint_{\sigma_1}\left(\frac{z_1}{a}\right)^2 e^{i(px_1+qy_1+\omega a/c)}\left(\frac{1}{a^2}-\frac{i\omega}{ac}\right)d\sigma_1 \to \frac{1}{3} \quad \text{as } a \to 0. \tag{7}$$

Hence we have in all

$$\left.\begin{aligned} \mathscr{J}_1 = \tfrac{1}{3} \quad &\text{when} \quad 0 < y < d, \\ = 0 \quad &\text{when} \quad y < 0 \text{ or } y > d. \end{aligned}\right\} \tag{8}$$

Before calculating $\mathscr{J}_2$ we note that because the integrand contains only a singularity of order $1/R$, integration throughout the vanishingly small sphere around the origin will not give a contribution. Hence in case (a), just as in case (b) we may now integrate throughout the full volume bounded by the slab:

$$\mathscr{J}_2 = \frac{\omega^2}{4\pi c^2}\int_{-y}^{d-y} e^{iqy_1}\mathscr{L}(y_1)\,dy_1, \tag{9}$$

where

$$\mathscr{L}(y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{e^{i(px_1+\omega R/c)}}{R}dx_1dz_1. \tag{10}$$

In (10) we take new variables $\rho$ and $\chi$ defined by

$$px_1 + \frac{\omega}{c}R = \rho\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2} \tag{11a}$$

$$z_1 = \sqrt{\rho^2 - y_1^2}\sin\chi \tag{11b}$$

in each case the positive square root being taken. On the $x_1z_1$-plane, the curves of constant $\rho$ are ellipses, with $\chi$ as the eccentric angle. For $\rho = |y_1|$ the ellipse degenerates into a point. Thus the whole $x_1z_1$-plane is covered if $\rho$ and $\chi$ run through the values $|y_1| \leqslant \rho \leqslant \infty, 0 \leqslant \chi \leqslant 2\pi$. Hence

$$\mathscr{L}(y_1) = \int_0^{2\pi}\int_{|y_1|}^{\infty}\frac{e^{i\rho\sqrt{\frac{\omega^2}{c^2}-p^2}}}{R}\frac{\partial(x_1,z_1)}{\partial(\rho,\chi)}d\rho d\chi \tag{12}$$

where

$$\frac{\partial(x_1,z_1)}{\partial(\rho,\chi)} = \frac{\partial x_1}{\partial\rho}\frac{\partial z_1}{\partial\chi} - \frac{\partial x_1}{\partial\chi}\frac{\partial z_1}{\partial\rho} \tag{13}$$

is the Jacobian of the transformation. From (11a) and (3)

$$\frac{\partial x_1}{\partial\rho} = \frac{\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}}{p + \frac{\omega}{c}\frac{x_1}{R}}, \quad \frac{\partial x_1}{\partial\chi} = 0, \tag{14a}$$

0

- After C. G. Darwin, Trans. Camb. Phil. Soc., 23 (1924), §§ 6 and 8.
† Strictly, the surface $S_1$ should be closed, so that we should also consider the contributions from the remote edges. These contributions may, however, be rejected on physical grounds, since they would require infinitely long time to reach the point where the effect is being considered.

and from (11b)

$$\frac{\partial z_1}{\partial \rho} = \frac{\rho}{\sqrt{\rho^2 - {y_1}^2}} \sin\chi, \quad \frac{\partial z_1}{\partial \chi} = \sqrt{\rho^2 - {y_1}^2}\cos\chi = \sqrt{\rho^2 - {y_1}^2 - {z_1}^2}. \tag{14b}$$

Now we also have the identity

$$\left(pR + \frac{\omega}{c}x_1\right)^2 - \left(px_1 + \frac{\omega}{c}R\right)^2 = \left[\left(\frac{\omega}{c}\right)^2 - p^2\right]\left[x_1^2 - R^2\right],$$

or using (3) and (11a),

$$\left(pR + \frac{\omega}{c}x_1\right) = \sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}\cdot\sqrt{\rho^2 - y_1^2 - z_1^2}. \tag{15}$$

From (13), (14), and (15) it follows that

$$\frac{\partial\left(x_1, z_1\right)}{\partial(\rho, \chi)} = R. \tag{16}$$

We substitute from (16) into (12). The integration with respect to $\chi$ gives immediately the value $2\pi$. The integration with respect to $\rho$ is also straightforward and we obtain

$$\mathscr{L}(y) = \frac{2\pi i e^{i|y_1|}\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}}{\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}}, \tag{17}$$

where, with the same physical justification as before, we have rejected an oscillatory contribution from infinity.

We now substitute from (17) into (9) and evaluate the resulting integral. This gives:

$$\left.\begin{aligned} \mathscr{J}_2 &= \left(\tfrac{\omega}{c}\right)^2 \frac{e^{-igy}}{2g\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}}\left\{e^{igd} - 1\right\} \\ &= \left(\tfrac{\omega}{c}\right)^2 \frac{e^{-ihy}}{2h\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}}\left\{e^{ihd} - 1\right\} \qquad \text{when } y > d, \\ &= \left(\tfrac{\omega}{c}\right)^2 \left\{\frac{1}{gh} - \frac{e^{-igy}}{2g\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}} + \frac{e^{-ih(y-d)}}{2h\sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}}\right) \quad \text{when } 0 < y < d, \end{aligned}\right\} \tag{18}$$

where

$$g = q - \sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}, \quad h = q + \sqrt{\left(\frac{\omega}{c}\right)^2 - p^2}. \tag{19}$$

## C.12 Other, Special Mathematical Methods

то, что к хз к чему относится

### C.12.1 Compressed Measurements by Tao (?)

(????)

# D Programming and Modeling

(?????)

# Index

дисперсия, 1189
закон дисперсии, 1190
групповая скорость, 1190
плазменная частота, 1192
дифракция
Гюйгенса-Френеля принцип, 174
пятно Пуассона, 187
закон Релея, 1180
кристаллы, 1195
двулучепреломление, 1197
оптическая ось кристалла, 1195
главное сечение, 1196
кристаллические пластинки, 1198
необыкновенная волна, 1196
положительные и отрицательные, 1195
линейная поляризация, 102
поляризация
Плоскость поляризации, 102
закон Малюса, 103

саморепродукция, 188
угол Брюстера, 103
формула Релея, 1191
четырехмерный волновой вектор, 114

# E References

**References from Me**

# References


[1] А., Кириченко Н.: *Принципы оптики*, 2016.

[2] Сивухин, Д. В.: *Общий курс физики. Том 4 Оптика.* Мосвка.: Физматлит, 2002.

[3] А.В. Максимычев, Д.А. Александров, Н.С. Берюлёва и др.: *Лабораторный практикум по общей физики Том 2 Оптика*, 2014.

[4] Д., Ландау Л. and М., Лифшиц Е.: *Теория поля.* Москва; Физматлит, 2006.

[5] Андрианов, ЕС, Виноградов, АП, and Пухов, АА: *Лекции по квантовой оптике.*

[6] Walls, D. F.: *Squeezed states of light*, 1983. https://arxiv.org/abs/1907.06506.

[7] Krasnok, Alex, Baranov, Denis, Li, Huanan, Miri, Mohammad Ali, Monticone, Francesco, and Alu, Andrea: *Anomalies in light scattering*, 2019. https://arxiv.org/abs/1907.06506.

[8] Курбатов, Л.Н.: *Оптоэлектроника видимого и инфракрасного диапазонов спектра.* Изд-во МФТИ М., 2013.

[9] Локшин, Г.Р.: *Принципы радиооптики.*

[10] В., Киселев В.: *Лекции по теории поля*, 2017.


### E.0.1 References from Theoretical Articles

**References from Catastrophe Optics by Berry Upstil**


Abramowitz, M. and I. A. Stegun, eds., 1964, Handbook of Mathematical Functions (Dover, New York).
Ahluwalia, D. S., R. M. Lewis and J. Boersma, 1968, SIAM J. Appl. Math. 16, 783-807.
Airy, G. B., 1838, Trans. Camb. Phil. Soc. 6, 379-403.
Albeverio, S. and R. Höegh-Krohn, 1977, Inventiones Math. 40, 59-106.
Arnol'd, V. I., 1975, Russ. Math. Survs. 30, 1-75.
Arnol'd, V. I., 1976, Commun. Pure Appl. Math. 29, 557-582 [Errata 30, 1977, 823].
Bartz, G., G. Weissenberg and D. Wiskott, 1956, Radex Rdsch., pp. 163-165.
Berry, M. V., 1966, Proc. Phys. Soc. 89, 479-490.
Berry, M. V., 1969a, Science Progress (Oxford) 57, 43-64.
Berry, M. V., 1969b, J. Phys. B2, 381-392.
Berry, M. V., 1975, J. Phys. A8, 566-584.
Berry, M. V., 1976, Adv. Phys. 25, 1-26.
Berry, M. V., 1977, J. Phys. A10, 2061-2081.
Berry, M. V., 1978a, La Recherche 9, 760-768.
Berry, M. V., 1978b, Regular and Irregular Motion, in: Topics in non-linear dynamics, ed. S. Jorna, Am. Inst. Phys. Conf. Proc. 46, 16-120.
Berry, M. V., 1979a, Some geometric aspects of wave motion, in: Geometry of the Laplace operator, Proc. Amer. Math. Soc. Conf., Honolulu 1979, to be published.
Berry, M. V., 1979b, J. Phys. A12, 781-797.

Berry, M. V. and J. H. Hannay, 1977, J. Phys. A10, 1809-1821.
Berry, M. V. and K. E. Mount, 1972, Repts. Prog. Phys. 35, 315-397.
Berry, M. V. and J. F. Nye, 1977, Nature 267, 34-36.
Berry, M. V., J. F. Nye and F. J. Wright, 1979, Phil. Trans. R. Soc. Lond. A291, 453-484.
Born, M. and E. Wolf, 1975, Principles of Optics, 5th ed. (Pergamon Press, Oxford).
Bryant, H. C. and N. Jarmie, 1974, Sci. Am. 231, 60-71.
Buckley, R., 1971a, Aust. J. Phys. 24, 351-371.
Buckley, R., 1971b, Aust. J. Phys. 24, 373-396.
Budden, K. G., 1976, Proc. R. Soc. Lond. A350, 143-164.
Callahan, J., 1977, Geometry of $E_6$ and anorexia, Preprint, Mathematics Institute, University of Warwick, U.K.
Callahan, J., 1978, Special bifurcations of the double cusp, Preprint, Mathematics Institute, University of Warwick, U.K.
Chester, C., B. Friedman and F. Ursell, 1957, Proc. Camb. Phil. Soc. 53, 599-611.
Connor, J. N. L., 1976, Molec. Phys. 31, 33-55.
Dangelmayr, G. and W. Veit, 1979, Ann. Phys. New York 118, 108-138.
Darboux, G., 1896, Leçons sur la théorie générale des surfaces, Vol. 4 (Gauthier-Villars, Paris) Note VII [Reprinted: 1972 (Chelsea, New York)].
Descartes, R., 1637, Les Météores, Discours VIII [also in: 1902, Oeuvres de Descartes, Vol. VI, eds. C. Adam and P. Tannery (Léopold Cerf, Paris)].
Dirac, P. A. M., 1931, Proc. R. Soc. Lond. A133, 60-72.
Duistermaat, J. J., 1974, Commun. Pure Appl. Math. 27, 207-281.
Feynman, R. P., 1948, Rev. Mod. Phys. 20, 367-387.
Feynman, R. P. and A. R. Hibbs, 1965, Quantum mechanics and path integrals (McGrawHill, New York).
Forsyth, A. R., 1912, Differential geometry (Cambridge University Press, Cambridge) chap. 4.
Glaser, W., 1956, in: Handbuch der Physik 33, ed. S. Flügge (Springer, Berlin) pp. 273284.
Godwin, A. N., 1971, Publ. Math. IHES 40, 117-138.
Golubitsky, M., 1978, SIAM Review 20, 352-387.
Green, H. S. and E. Wolf, 1953, Proc. Phys. Soc. A66, 1129-1137.
Gulllemin, V. and S. Sternberg, 1977, Geometric asymptotics, American Mathematical Society Survey No. 14 (Providence, USA).
Gullstrand, A., 1905, Acta. Math. 29, 59-100.
Hannay, J. H., 1977, Paraxial optics and statistical problems of wave propagation, Ph.D. Thesis, Magdalene College, University of Cambridge, U.K.
Hirschfelder, J. O., A. C. Christoph and W. E. Palke, 1974, J. Chem. Phys. 61, 5435-5455.
Hirschfelder, J. O., C. J. Goebel and L. W. Bruch, 1974, J. Chem. Phys. 61, 5456-5459.
Hirschfelder, J. O. and K. T. Tang, 1976a, J. Chem. Phys. 64, 760-785.
Hirschfelder, J. O. and K. T. Tang, 1976b, J. Chem. Phys. 65, 470-486.
Holford, R. L., 1972, Modifications to ray theory near cusped caustics, Bell Telephone Labs. Report.
Humphreys, W. J., 1940, Physics of the air, 3rd ed. (McGraw-Hill, New York).
Jakeman, E. and J. G. McWhirter, 1977, J. Phys. A10, 1599-1643.
Jakeman, E., E. R. Pike and P. N. Pusey, 1976, Nature 263, 215-217.
Keller, J. B., 1958, Proc. Symp. Appl. Math. 8, 27-52.
Khare, V. and H. M. Nussenzveig, 1974, Phys. Rev. Lett. 33, 976-980.
Khare, V. and H. M. Nussenzveig, 1977, Phys. Rev. Lett. 38, 1279-1282.
Kline, M. and I. W. Kay, 1965, Electromagnetic theory and geometrical optics (Interscience, New York).
Kravtsov, Yu. A., 1964a, Radiofizika 7, 664-673.

Kravtsov, Yu. A., 1964b, Radiofizika 7, 1049-1056.
Kravtsov, Yu, A., 1968, Soviet Phys. Acoust. 14, 1-17.
Lanczos, C., 1966, The variational principles of mechanics, 3rd ed. (University of Toronto Press, Toronto).
Larmor, J., 1891, Proc. Camb. Phil. Soc. 7, 131-137.
Lenz, F., 1956, in: Proc. 3rd Int. Conf. on electron microscopy, London 1954 (Royal Microscopical Society, London) pp. 83-85.
Lenz, F. and H. Mecking, 1961, in: Proc. Eur. Reg. Conf. on electron microscopy, Delft 1960, Vol. I, eds. A. L. Houwink and B. J. Spit (Nederlandse Vereniging voor Electronenmicroscopie, Delft) pp. 41-43.
Levit, S. and U. Smilansky, 1977, Ann. Phys. New York 108, 165-197.
Liesegang, S., 1953, Optik 10, 5-14.
Longuet-Higgins, M. S., 1956, Proc. Camb. Phil. Soc. 52, 234-245.
Longuet-Higgins, M. S., 1957, Phil. Trans. R. Soc. Lond. A249, 321-387.
Longuet-Higgins, M. S., 1958, Proc. Camb. Phil. Soc. 54, 439-453.
Longuet-Higgins, M. S., 1959, Proc. Camb. Phil. Soc. 55, 91-100.
Longuet-Higgins, M. S., 1960a, J. Opt. Soc. Am. 50, 838-856.
Longuet-Higgins, M. S., 1960b, Proc. Camb. Phil. Soc. 56, 27-40.
Ludwig, D., 1966, Commun. Pure Appl. Math. 19, 215-250.
Mandelbrot, B. B., 1977, Fractals (W. H. Freeman, San Francisco).
Marcuse, D., 1972, Light transmission optics (Van Nostrand, New York).
Maslin, N. M., 1976a, J. Atmos. Terr. Phys. 38, 239-250.
Maslin, N. M., 1976b, J. Atmos. Terr. Phys. 38, 1251-1263.
Mercier, R. P., 1962, Proc. Camb. Phil. Soc. 58, 382-400.
Minnaert, M., 1954, The nature of light and colour in the open air (Dover, New York).
Maslov, V. P., 1965 (in Russian) [French translation: 1972, Théorie des perturbations et des méthodes asymptotiques (Dunod, Paris)].
McDonald, J. E., 1963, Am. J. Phys. 31, 282-284.
Nussenzveig, H. M., 1969a, J. Math. Phys. 10, 82-124.
Nussenzveig, H. M., 1969b, J. Math. Phys. 10, 125-176.
Nussenzveig, H. M., 1977, Sci. Am. 236, 116-127.
Nye, J. F., 1978, Proc. R. Soc. Lond. A361, 21-41.
Nye, J. F., 1979, Phil. Trans. R. Soc. Lond. A292, 25-44.
Nye, J. F. and M. V. Berry, 1974, Proc. R. Soc. Lond. A336, 165-190.
Pearcey, T., 1946, Phil. Mag. 37, 311-317.
Porteous, I. R., 1971, J. Diff. Geom. 5, 543-564.
Poston, T., 1980, Catastrophe theory in physics, Physics Reports, to be published.
Poston, T. and I. N. Stewart, 1978, Catastrophe theory and its applications (Pitman, London).
Rice, S. O., 1944, Bell Syst. Tech. J. 23, 282-332.
Rice, S. O., 1945, Bell Syst. Tech. J. 24, 46-156.
Riess, J., 1970a, Ann. Phys. New York 57, 301-321.
Riess, J., 1970b, Phys. Rev. D2, 647-653.
Riess, J., 1976, Phys. Rev. B13, 3862-3869.
Romero, M. del C., 1978, On the Berry index, M.Sc. Thesis, Mathematics Institute, University of Warwick, U.K.
Salpeter, E. E., 1967, Astrophys. J. 147, 433-448.
Scheffels, W., M. Hahn and F. Lenz, 1953, Optik 10, 455-458.
Schleich, F., D. Hoffmeister, H. Koops and F. Lenz, 1968, Optik 27, 305-316.
Schulman, L. S., 1975, Caustics and multivaluedness, in: Functional integration and its applications, ed. A. M. Arthurs (Clarendon Press, Oxford) pp. 144-156.

Sewell, M. J., 1977, Math. Proc. Camb. Phil. Soc. 82, 147-163.
Sewell, M. J., 1978, Math. Proc. Camb. Phil. Soc. 83, 273-288.
Shishov, V. I., 1971, Izv. Vuz. Radiofiz. 14, 85-92.
Spivak, M., 1975, A comprehensive introduction to differential geometry, Vol. 3 (Publish or Perish, Boston) Chaps. 4 and 6.
Stevens, P. S., 1976, Patterns in nature (Penguin, Harmondsworth, U.K.).
Tanner, L. H., 1978, Opt. Laser Technol. 10, 125-128.
Thom, R., 1972, Stabilité structurelle et morphogénèse (Benjamin, Reading, Mass.) [English translation, updated: 1975, Structural stability and morphogenesis (Benjamin, Reading, Mass.)].
Tricker, R. A. R., 1970, Introduction to meteorological optics (Mills and Boon, London).
Trinkaus, H. and F. Drepper, 1977, J. Phys, A10, L11-L16.
Upstill, C., 1979, Proc. R. Soc. Lond. A365, 95-104.
Uscinski, B. J., 1977, The elements of wave propagation in random media (McGraw-Hill, New York).
Varchenko, A. N., 1976, Funct. Anal. Appl. 10, 13-38.
Wolf, E., 1959, Proc. Phys. Soc. 74, 269-280.
Wolf, E., 1960, Some aspects of a rigorous scalar treatment of electromagnetic wave propagation, in: Electromagnetic wave propagation (Academic Press, New York) pp. 119-125.
Wright, F. J., 1977, Wavefield singularities, Ph.D. Thesis, H. H. Wills Physics Laboratory, University of Bristol. U.K.

## E.0.2 References from Articles about Applications

**References from "Optical coherence tomography today: speed, contrast, and multimodality"**

1. D. Huang et al., "Optical coherence tomography," Science 254(5035), 1178-1181 (1991).

2. B. Bouma and G. Tearney, Handbook of Optical Coherence Tomography, Taylor & Francis, USA (2001).

3. J. Schuman, C. Puliafito, and J. G. Fujimoto, Everyday OCT: A Handbook for Clinicians and Technicians, SLACK Incorporated, USA (2007).

4. W. Drexler and J. G. Fujimoto, Optical Coherence Tomography: Technology and Applications, Springer Publishing, Berlin, Heidelberg (2008).

5. A. F. Fercher and E. Roth, "Ophthalmic laser interferometer," Proc. SPIE 658, 48-51 (1986).

6. A. F. Fercher, K. Mengedoht, and W. Werner, "Eye length measurement by interferometer with partially coherent light," Opt. Lett. 13(3), 186188 (1988).

7. J. G. Fujimoto et al., "Femtosecond optical ranging in biological systems," Opt. Lett. 11(3), 150-152 (1986).

8. M. Kemerink et al., "Characteristics of a first-generation x-ray system," Radiology 259(2), 534-539 (2011).

9. J. G. Fujimoto et al., "Optical biopsy and imaging using optical coherence tomography," Nat. Med. 1(9), 970-972 (1995).

10. B. E. Bouma et al., "Self-phase-modulated Kerr-lens mode-locked Cr: forsterite laser source for optical coherence tomography," Opt. Lett. 21(22), 1839-1841 (1996).

11. B. E. Bouma et al., "Optical coherence tomographic imaging using a mode locked Cr/-sup 4+/:Forsterite laser source," presented at Proc. of Ultrafast Phenomena 1997 OSA Technical Digest Series Vol. 8, 28 May-1 June 1996, San Diego, CA, Opt. Soc. America, Washington, DC (1996).

12. W. Drexler et al., "In vivo ultrahigh-resolution optical coherence tomography," Opt. Lett. 24(17), 1221-1223 (1999).

13. W. Drexler et al., "Ultrahigh-resolution ophthalmic optical coherence tomography," Nat. Med. 7(4), 502-507 (2001).

14. W. Drexler, "Ultrahigh-resolution optical coherence tomography," J. Biomed. Opt. 9(1), 47-74 (2004).

15. A. F. Fercher et al., "Measurement of intraocular distances by backscattering spectral interferometry," Opt. Commun. 117(1-2), 43-48 (1995).

16. M. Wojtkowski et al., "In vivo human retinal imaging by Fourier domain optical coherence tomography," J. Biomed. Opt. 7(3), 457-463 (2002).

17. M. A. Choma et al., "Sensitivity advantage of swept source and Fourier domain optical coherence tomography," Opt. Express 11(18), 21832189 (2003).

18. R. Leitgeb, C. K. Hitzenberger, and A. F. Fercher, "Performance of Fourier domain vs. time domain optical coherence tomography," Opt. Express 11(8), 889-894 (2003).

19. B. Golubovic et al., "Optical frequency-domain reflectometry using rapid wavelength tuning of a Cr4+:forsterite laser," Opt. Lett. 22(22), 1704-1706 (1997).

20. S. Yun et al., "High-speed optical frequency-domain imaging," Opt. Express 11(22), 2953-2963 (2003).

21. R. Huber, M. Wojtkowski, and J. G. Fujimoto, "Fourier domain mode locking (FDML): a new laser operating regime and applications for optical coherence tomography," Opt. Express 14(8), 3225-3237 (2006).

22. T. Klein et al., "Megahertz OCT for ultrawide-field retinal imaging with a 1050 nm Fourier domain mode-locked laser," Opt. Express 19(4), 3044-3062 (2011).

23. T. Klein et al., "Multi-MHz retinal OCT," Biomed. Opt. Express 4(10), 1890-1908 (2013).

24. M. R. Hee et al., "Polarization-sensitive low-coherence reflectometer for birefringence characterization and ranging," J. Opt. Soc. Am. B 9(6), 903-908 (1992).

25. J. F. de Boer et al., "Two-dimensional birefringence imaging in biological tissue by polarization-sensitive optical coherence tomography," Opt. Lett. 22(12), 934-936 (1997).

26. C. Hitzenberger et al., "Measurement and imaging of birefringence and optic axis orientation by phase resolved polarization sensitive optical coherence tomography," Opt. Express 9(13), 780-790 (2001).

27. J. F. de Boer and T. E. Milner, "Review of polarization sensitive optical coherence tomography and Stokes vector determination," J. Biomed. Opt. 7(3), 359-371 (2002).

28. Y. Yasuno et al., "Birefringence imaging of human skin by polarizationsensitive spectral interferometric optical coherence tomography," Opt. Lett. 27(20), 1803-1805 (2002).

29. C. K. Hitzenberger and A. F. Fercher, "Differential phase contrast in optical coherence tomography," Opt. Lett. 24(9), 622-624 (1999).

30. A. Wax, S. Bali, and J. E. Thomas, "Time-resolved phase-space distributions for light backscattered from it disordered medium," Phys. Rev. Lett. 85(1), 66-69 (2000).

31. C. H. Yang et al., "Interferometric phase-dispersion microscopy," Opt. Lett. 25(20), 1526-1528 (2000).

32. C. K. Hitzenberger et al., "Differential phase measurements in lowcoherence interferometry without 2 pi ambiguity," Opt. Lett. 26(23), 1864-1866 (2001).

33. C. Yang et al., "Phase-dispersion optical tomography," Opt. Lett. 26(10), 686-688 (2001).

34. J. M. Schmitt, "OCT elastography: imaging microscopic deformation and strain of tissue," Opt. Express 3(6), 199-211 (1998).

35. R. C. Chan et al., "OCT-based arterial elastography: robust estimation exploiting tissue biomechanics," Opt. Express 12(19), 4558-4572 (2004).

36. A. S. Khalil et al., "Tissue elasticity estimation with optical coherence elastography: toward mechanical characterization of in vivo soft tissue," Ann. Biomed. Eng. 33(11), 1631-1639 (2005).

37. S. J. Kirkpatrick, R. K. Wang, and D. D. Duncan, "OCT-based elastography for large and small deformations," Opt. Express 14(24), 1158511597 (2006).

38. A. Nahas et al., "3D static elastography at the micrometer scale using full field OCT," Biomed. Opt. Express 4(10), 2138-2149 (2013).

39. A. Nahas et al., "From supersonic shear wave imaging to full-field optical coherence shear wave elastography," J. Biomed. Opt. 18(12), 121514 (2013).

40. M. D. Kulkarni and J. A. Izatt, "Spectroscopic optical coherence tomography," in Conf. Proc. of Lasers and Electro Optics Society Annual Meeting, pp. 59-60, CLEO, OSA, USA (1996).

41. U. Morgner et al., "Spectroscopic optical coherence tomography," Opt. Lett. 25(2), 111-113 (2000).

42. D. J. Faber et al., "Light absorption of (oxy-)hemoglobin assessed by spectroscopic optical coherence tomography," Opt. Lett. 28(16), 14361438 (2003).

43. D. C. Adler et al., "Optical coherence tomography contrast enhancement using spectroscopic analysis with spectral autocorrelation," Opt. Express 12(22), 5487-5501 (2004).

44. A. Wax and J. E. Thomas, "Measurement of smoothed Wigner phasespace distributions for small-angle scattering in a turbid medium," J. Opt. Soc. Am. A Opt. Image Sci. Vis. 15(7), 1896-1908 (1998).

45. A. Wax et al., "Measurement of angular distributions by use of lowcoherence interferometry for light-scattering spectroscopy," Opt. Lett. 26(6), 322-324 (2001).

46. A. Wax, C. H. Yang, and J. A. Izatt, "Fourier-domain low-coherence interferometry for light-scattering spectroscopy," Opt. Lett. 28(14), 1230-1232 (2003).

47. T. M. Lee et al., "Engineered microsphere contrast agents for optical coherence tomography," Opt. Lett. 28(17), 1546-1548 (2003).

48. D. L. Marks and S. A. Boppart, "Nonlinear interferometric vibrational imaging," Phys. Rev. Lett. 92(12), 123905 (2004).

49. C. Vinegoni et al., "Nonlinear optical contrast enhancement for optical coherence tomography," Opt. Express 12(2), 331-341 (2004).

50. C. Xu et al., "Separation of absorption and scattering profiles in spectroscopic optical coherence tomography using a least-squares algorithm," Opt. Express 12(20), 4790-4803 (2004).

51. S. A. Boppart et al., "Optical probes and techniques for molecular contrast enhancement in coherence imaging," J. Biomed. Opt. 10(4), 041208 (2005).

52. J. S. Bredfeldt et al., "Molecularly sensitive optical coherence tomography," Opt. Lett. 30(5), 495-497 (2005).

53. R. L. Shelton et al., "Optical coherence tomography for advanced screening in the primary care office," J. Biophotonics (2013).

54. A. G. Podoleanu, "Optical coherence tomography," J. Microsc. 247(3), 209-219 (2012).

55. W. Jung and S. A. Boppart, "Optical coherence tomography for rapid tissue screening and directed histological sectioning," Anal. Cell Pathol. (Amst) 35(3), 129-143 (2012).

56. G. Lamouche et al., "Review of tissue simulating phantoms with controllable optical, mechanical and structural properties for use in optical coherence tomography," Biomed. Opt. Express 3(6), 1381-1398 (2012).

57. M. Wojtkowski, "High-speed optical coherence tomography: basics and applications," Appl. Opt. 49(16), D30-D61 (2010).

58. A. M. Zysk et al., "Optical coherence tomography: a review of clinical development from bench to bedside," J. Biomed. Opt. 12(5), 051403 (2007).

59. M. Wojtkowski, B. Kaluzny, and R. J. Zawadzki, "New directions in ophthalmic optical coherence tomography," Optom. Vis. Sci. 89(5), 524-542 (2012).

60. J. L. Ramos, Y. Li, and D. Huang, "Clinical and research applications of anterior segment optical coherence tomography-a review," Clin. Experiment. Ophthalmol. 37(1), 81-89 (2009).

61. W. Drexler and J. G. Fujimoto, "State-of-the-art retinal optical coherence tomography," Prog. Retin. Eye Res. 27(1), 45-88 (2008).

62. T. Yonetsu et al., "Optical coherence tomography-15 years in cardiology," Circ. J. 77(8), 1933-1940 (2013).

63. M. J. Suter et al., "Intravascular optical imaging technology for investigating the coronary artery," JACC Cardiovasc. Imaging 4(9), 10221039 (2011).

64. H. C. Lowe et al., "Intracoronary optical diagnostics current status, limitations, and potential," JACC Cardiovasc. Interv. 4(12), 1257-1270 (2011).

65. A. Tanaka, G. J. Tearney, and B. E. Bouma, "Challenges on the frontier of intracoronary imaging: atherosclerotic plaque macrophage measurement by optical coherence tomography," J. Biomed. Opt. 15(1), 011104 (2010).

66. H. G. Bezerra et al., "Intracoronary optical coherence tomography: a comprehensive review clinical and research applications," JACC Cardiovasc. Interv. 2(11), 1035-1046 (2009).

67. E. Sattler, R. Kastle, and J. Welzel, "Optical coherence tomography in dermatology," J. Biomed. Opt. 18(6), 061224 (2013).

68. C. A. Banzhaf et al., "In vivo imaging of Sarcoptes scabiei infestation using optical coherence tomography," Case Rep. Dermatol. 5(2), 156162 (2013).

69. M. A. Boone et al., "High-definition optical coherence tomography imaging of melanocytic lesions: a pilot study," Arch. Dermatol. Res. 306(1), 11-26 (2014).

70. M. A. Boone et al., "Imaging actinic keratosis by high-definition optical coherence tomography. Histomorphologic correlation: a pilot study," Exp. Dermatol. 22(2), 93-97 (2013).

71. L. P. Hariri et al., "Seeing beyond the bronchoscope to increase the diagnostic yield of bronchoscopic biopsy," Am. J. Respir. Crit. Care Med. 187(2), 125-129 (2013).

72. B. J. Vakoc et al., "Cancer imaging by optical coherence tomography: preclinical progress and clinical potential," Nat. Rev. Cancer 12(5), 363-368 (2012).

73. X. Liang, V. Crecea, and S. A. Boppart, "Dynamic optical coherence elastography: a review," J. Innov. Opt. Health Sci. 3(4), 221-233 (2010).

74. A. Ahmad et al., "Real-time computed optical interferometric tomography," Nat. Photonics 7(6), 444-448 (2013).

75. R. Kafieh, H. Rabbani, and S. Kermani, "A review of algorithms for segmentation of optical coherence tomography from retina," J. Med. Signals Sens. 3(1), 45-60 (2013).

76. R. Huber et al., "Fourier domain mode locking at 1050 nm for ultra-high-speed optical coherence tomography of the human retina at 236,000 axial scans per second," Opt. Lett. 32(14), 2049-2051 (2007).

77. B. R. Biedermann et al., "Recent developments in Fourier domain mode locked lasers for optical coherence tomography: imaging at 1310 nm vs. 1550 nm wavelength," J. Biophotonics 2(6-7), 357-363 (2009).

78. W. Wieser et al., "Multi-megahertz OCT: high quality 3D imaging at 20 million A-scans and 4.5 GVoxels per second," Opt. Express 18(14), 14685-14704 (2010).

79. W. Drexler, "Ultrahigh resolution optical coherence tomography," J. Biomed. Opt. 9(1), 47-74 (2004).

80. E. Beaurepaire et al., "Full-field optical coherence microscopy," Opt. Lett. 23(4), 244-246 (1998).

81. A. C. Boccara et al., "Detection analogique multicanal," Frenchpatent FR90.092255 (29 June 1990).

82. A. Dubois et al., "High-resolution full-field optical coherence tomography with a Linnik microscope," Appl. Opt. 41(4), 805-812 (2002).

83. K. Grieve et al., "Ocular tissue imaging using ultrahigh-resolution, fullfield optical coherence tomography," Invest. Ophthalmol. Vis. Sci. 45(11), 4126-4131 (2004).

84. W. Y. Oh et al., "Ultrahigh-speed optical frequency domain imaging and application to laser ablation monitoring," Appl. Phys. Lett. 88(10), 103902 (2006).

85. A. Dubois, G. Moneron, and C. Boccara, "Thermal-light full-field optical coherence tomography in the 1.2 mu m wavelength region," Opt. Commun. 266(2), 738-743 (2006).

86. M. Jain et al., "Full-field optical coherence tomography for the analysis of fresh unstained human lobectomy specimens," J. Pathol. Inform. 4, 26 (2013).

87. M. Laubscher et al., "Video-rate three-dimensional optical coherence tomography," Opt. Express 10(9), 429-435 (2002).

88. W. Y. Oh et al., "Spectrally-modulated full-field optical coherence microscopy for ultrahigh-resolution endoscopic imaging," Opt. Express 14(19), 8675-8684 (2006).

89. H. D. Ford et al., "Comparative signal-to-noise analysis of fibre-optic based optical coherence tomography systems," J. Mod. Opt. 52(14), 1965-1979 (2005).

90. A. Latrive and A. C. Boccara, "In vivo and in situ cellular imaging fullfield optical coherence tomography with a rigid endoscopic probe," Biomed. Opt. Express 2(10), 2897-2904 (2011).

91. I. Zeylikovich, A. Gilerson, and R. R. Alfano, "Nonmechanical gratinggenerated scanning coherence microscopy," Opt. Lett. 23(23), 17971799 (1998).

92. Y. Watanabe, K. Yamada, and M. Sato, "Three-dimensional imaging by ultrahigh-speed axial-lateral parallel time domain optical coherence tomography," Opt. Express 14(12), 5201-5209 (2006).

93. M. Choma et al., "Sensitivity advantage of swept source and Fourier domain optical coherence tomography," Opt. Express 11(18), 21832189 (2003).

94. A. F. Zuluaga and R. Richards-Kortum, "Spatially resolved spectral interferometry for determination of subsurface structure," Opt. Lett. 24(8), 519-521 (1999).

95. B. Grajciar et al., "Parallel Fourier domain optical coherence tomography for in vivo measurement of the human eye," Opt. Express 13(4), 1131-1137 (2005).

96. Y. Nakamura et al., "High-speed three-dimensional human retinal imaging by line-field spectral domain optical coherence tomography," Opt. Express 15(12), 7103-7116 (2007).

97. Y. Yasuno et al., "Three-dimensional line-field Fourier domain optical coherence tomography for in vivo dermatological investigation," J. Biomed. Opt. 11(1), 014014 (2006).

98. Y. Zhang et al., "Adaptive optics parallel spectral domain optical coherence tomography for imaging the living retina," Opt. Express 13(12), 4792-4811 (2005).

99. S. Witte et al., "Single-shot two-dimensional full-range optical coherence tomography achieved by dispersion control," Opt. Express 17(14), 11335-11349 (2009).

100. B. Hofer et al., "Dispersion encoded full range frequency domain optical coherence tomography," Opt. Express 17(1), 7-24 (2009).

101. S. W. Lee and B. M. Kim, "Line-field optical coherence tomography using frequency-sweeping source," IEEE J. Sel. Topics Quantum Electron. 14(1), 50-55 (2008).

102. T. Bonin et al., "In vivo Fourier-domain full-field OCT of the human retina with 1.5 million A-lines/s," Opt. Lett. 35(20), 3432-3434 (2010).

103. M. K. Kim, "Tomographic three-dimensional imaging of a biological specimen using wavelength-scanning digital interference holography," Opt. Express 7(9), 305-310 (2000).

104. D. Hillmann et al., "Efficient holoscopy image reconstruction," Opt. Express 20(19), 21247-21263 (2012).

105. D. Hillmann et al., "Holoscopy-holographic optical coherence tomography," Opt. Lett. 36(13), 2390-2392 (2011).

106. B. Povazay et al., "Full-field time-encoded frequency-domain optical coherence tomography," Opt. Express 14(17), 7661-7669 (2006).

107. D. L. Marks et al., "Inverse scattering for frequency-scanned full-field optical coherence tomography," J. Opt. Soc. Am. A Opt. Image Sci. Vis. 24(4), 1034-1041 (2007).

108. S. G. Adie et al., "Computational adaptive optics for broadband optical interferometric tomography of biological tissue," Proc. Natl. Acad. Sci. U S A 109(19), 7175-7180 (2012).

109. A. Kumar, W. Drexler, and R. A. Leitgeb, "Subaperture correlation based digital adaptive optics for full field optical coherence tomography," Opt. Express 21(9), 10850-10866 (2013).

110. A. H. Dhalla, J. V. Migacz, and J. A. Izatt, "Crosstalk rejection in parallel optical coherence tomography using spatially incoherent illumination with partially coherent sources," Opt. Lett. 35(13), 2305-2307 (2010).

111. P. Krížek, I. Raška, and G. M. Hagen, "Flexible structured illumination microscope with a programmable illumination array," Optics Express 20(22), 24585-24599 (2012).

112. Y. Zhao et al., "Phase-resolved optical coherence tomography and optical Doppler tomography for imaging blood flow in human skin with fast scanning speed and high velocity sensitivity," Opt. Lett. 25(2), 114-116 (2000).

113. V. Yang et al., "High speed, wide velocity dynamic range Doppler optical coherence tomography (Part I): System design, signal processing, and performance," Opt. Express 11(7), 794-809 (2003).

114. M. A. Choma et al., "Sensitivity advantage of swept source and Fourier domain optical coherence tomography," Opt. Express 11(18), 21832189 (2003).

115. S. Makita et al., "Optical coherence angiography," Opt. Express 14(17), 7821-7840 (2006).

116. B. H. Park et al., "Real-time fiber-based multi-functional spectraldomain optical coherence tomography at 1.3$\mu$ m," Opt. Express 13(11), 3931-3944 (2005).

117. S. H. Yun et al., "Motion artifacts in optical coherence tomography with frequency-domain ranging," Opt. Express 12(13), 2977-2998 (2004).

118. A. H. Bachmann et al., "Resonant Doppler flow imaging and optical vivisection of retinal blood vessels," Opt. Express 15(2), 408-422 (2007).

119. Y. K. Tao, A. M. Davis, and J. A. Izatt, "Single-pass volumetric bidirectional blood flow imaging spectral domain optical coherence tomography using a modified Hilbert transform," Opt. Express 16(16), 12350-12361 (2008).

120. L. An and R. K. Wang, "In vivo volumetric imaging of vascular perfusion within human retina and choroids with optical microangiography," Opt. Express 16(15), 11438-11452 (2008).

121. C. Kolbitsch, T. Schmoll, and R. A. Leitgeb, "Histogram-based filtering for quantitative 3D retinal angiography," J. Biophotonics 2(6-7), 416-425 (2009).

122. A. S. G. Singh, T. Schmoll, and R. A. Leitgeb, "Segmentation of Doppler optical coherence tomography signatures using a support-vector machine," Biomed. Opt. Express 2(5), 1328-1339 (2011).

123. E. Dittrich et al., "Detection of capillary vessels in optical coherence tomography based on a probabilistic kernel," in Proc. 13th Annual Meeting, Medical Image Understanding and Analysis, A. H. Bhalerao and N. M. Rajpoot, Eds., pp. 37-41, University of Warwick, Warwick (2009).

124. T. Schmoll et al., "Imaging of the parafoveal capillary network and its integrity analysis using fractal dimension," Biomed. Opt. Express 2(5), 1159-1168 (2011).

125. V. Kaji et al., "Automated three-dimensional choroidal vessel segmentation of 3D 1060 nm OCT retinal data," Biomed. Opt. Express 4(1), 134-150 (2013).

126. F. Jaillon et al., "Enhanced imaging of choroidal vasculature by highpenetration and dual-velocity optical coherence angiography," Biomed. Opt. Express 2(5), 1147-1158 (2011).

127. S. Zotter et al., "Visualization of microvasculature by dual-beam phase-resolved Doppler optical coherence tomography," Opt. Express 19(2), 1217-1227 (2011).

128. F. Jaillon, S. Makita, and Y. Yasuno, "Variable velocity range imaging of the choroid with dual-beam optical coherence angiography," Opt. Express 20(1), 385-396 (2012).

129. T. Schmoll, C. Kolbitsch, and R. A. Leitgeb, "Ultra-high-speed volumetric tomography of human retinal blood flow," Opt. Express 17(5), 4166-4176 (2009).

130. W. Choi et al., "Choriocapillaris and choroidal microvasculature imaging with ultrahigh speed OCT angiography," PLoS ONE 8(12), e81499 (2013).

131. M. Szkulmowski et al., "Flow velocity estimation using joint spectral and time domain optical coherence tomography," Opt. Express 16(9), 6008-6025 (2008).

132. P. Meemon and J. P. Rolland, "Swept-source based, single-shot, multidetectable velocity range Doppler optical coherence tomography," Biomed. Opt. Express 1(3), 955-966 (2010).

133. C. Blatter et al., "Dove prism based rotating dual beam bidirectional Doppler OCT," 4(7), 1188-1203 (2013).

134. I. Grulkowski et al., "Scanning protocols dedicated to smart velocity ranging in spectral OCT," Opt. Express 17(26), 23736-23754 (2009).

135. B. Braaf et al., "Angiography of the retina and the choroid with phaseresolved OCT using interval-optimized backstitched B-scans," Opt. Express 20(18), 20516-20534 (2012).

136. M. S. Mahmud et al., "Review of speckle and phase variance optical coherence tomography to visualize microvascular networks," J. Biomed. Opt. 18(5), 050901 (2013).

137. D. Y. Kim et al., "In vivo volumetric imaging of human retinal circulation with phase-variance optical coherence tomography," Biomed. Opt. Express 2(6), 1504-1513 (2011).

138. B. J. Vakoc et al., "Three-dimensional microscopy of the tumor microenvironment in vivo using optical frequency domain imaging," Nat. Med. 15(10), 1219-1223 (2009).

139. A. Mariampillai et al., "Speckle variance detection of microvasculature using swept-source optical coherence tomography," Opt. Lett. 33(13), 1530-1532 (2008).

140. J. Barton and S. Stromski, "Flow measurement without phase information in optical coherence tomography images," Opt. Express 13(14), 5234-5239 (2005).

141. L. An, J. Qin, and R. K. Wang, "Ultrahigh sensitive optical microangiography for in vivo imaging of microcirculations within human skin tissue beds," Opt. Express 18(8), 8220-8228 (2010).

142. C. Blatter et al., "Ultrahigh-speed non-invasive widefield angiography," J. Biomed. Opt. 17(7), 070505 (2012).

143. C. Blatter et al., "Extended focus high-speed swept source OCT with self-reconstructive illumination," Opt. Express 19(13), 12141-12155 (2011).

144. B. Povazay et al., "Enhanced visualization of choroidal vessels using ultrahigh resolution ophthalmic OCT at 1050 nm ," Opt. Express 11(17), 1980-1986 (2003).

145. Y. Yasuno et al., "In vivo high-contrast imaging of deep posterior eye by 1-um swept source optical coherence tomography andscattering optical coherence angiography," Opt. Express 15(10), 6121-6139 (2007).

146. A. Mariampillai et al., "Optimized speckle variance OCT imaging of microvasculature," Opt. Lett. 35(8), 1257-1259 (2010).

147. C. Blatter et al., "In situ structural and microangiographic assessment of human skin lesions with high-speed OCT," Biomed. Opt. Express 3(10), 2636-2646 (2012).

148. S. Yazdanfar, M. Kulkarni, and J. Izatt, "High resolution imaging of in vivo cardiac dynamics using color Doppler optical coherence tomography," Opt. Express 1(13), 424-431 (1997).

149. R. A. Leitgeb et al., "Real-time assessment of retinal blood flow with ultrafast acquisition by color Doppler Fourier domain optical coherence tomography," Opt. Express 11(23), 3116-3121 (2003).

150. G. Liu et al., "A comparison of Doppler optical coherence tomography methods," Biomed. Opt. Express 3(10), 2669-2680 (2012).

151. J. Lee et al., "Dynamic light scattering optical coherence tomography," Opt. Express 20(20), 22262-22277 (2012).

152. M. H. Bernstein and M. J. Hollenberg, "Fine structure of the choriocapillaris and retinal capillaries," Invest. Ophthalmol. Vis. Sci. 4(6), 1016-1025 (1965).

153. S. H. Hardarson et al., "Automatic retinal oximetry," Invest. Ophthalmol. Vis. Sci. 47(11), 5011-5016 (2006).

154. F. C. Delori, "Noninvasive technique for oximetry of blood in retinal vessels," Appl. Opt. 27(6), 1113-1125 (1988).

155. M. C. Skala et al., "Combined hyperspectral and spectral domain optical coherence tomography microscope for noninvasive hemodynamic imaging," Opt. Lett. 34(3), 289-291 (2009).

156. L. V. Wang, "Multiscale photoacoustic microscopy and computed tomography," Nat. Photonics 3(9), 503-509 (2009).

157. H. F. Zhang et al., "Functional photoacoustic microscopy for highresolution and noninvasive in vivo imaging," Nat. Biotechnol. 24(7), 848-851 (2006).

158. K. K. C. Lee et al., "Real-time speckle variance swept-source optical coherence tomography using a graphics processing unit," Biomed. Opt. Express 3(7), 1557-1564 (2012).

159. T. M. Lee et al., "Engineered microsphere contrast agents for optical coherence tomography," Opt. Lett. 28(17), 1546-1548 (2003).

160. W. Denk, J. H. Strickler, and W. W. Webb, "Two-photon laser scanning fluorescence microscopy," Science 248(4951), 73-76 (1990).

161. F. S. Pavone and P. J. Campagnola, Second Harmonic Generation Imaging, p. 479, CRC Press, Boca Ratopn, FL (2013).

162. J. Squier et al., "Third harmonic generation microscopy," Opt. Express 3(9), 315-324 (1998).

163. M. D. Duncan, J. Reintjes, and T. J. Manuccia, "Scanning coherent anti-Stokes Raman microscope," Opt. Lett. 7(8), 350-352 (1982).

164. C. V. Raman and K. S. Krishnan, "A new type of secondary radiation," Nature 121, 501-502 (1928).

165. A. C. T. Ko et al., "Ex vivo detection and characterization of early dental caries by optical coherence tomography and Raman spectroscopy," J. Biomed. Opt. 10(3), 031118 (2005).

166. C. A. Patil et al., "Combined Raman spectroscopy and optical coherence tomography device for tissue characterization," Opt. Lett. 33(10), 1135-1137 (2008).

167. C. A. Patil et al., "A clinical instrument for combined Raman spectros-copy-optical coherence tomography of skin cancers," Lasers Surg. Med. 43(2), 143-151 (2011).

168. J. W. Evans et al., "Optical coherence tomography and Raman spectroscopy of the ex-vivo retina," J. Biophotonics 2(6-7), 398-406 (2009).

169. P. C. Ashok et al., "Multi-modal approach using Raman spectroscopy and optical coherence tomography for the discrimination of colonic adenocarcinoma from normal colon," Biomed. Opt. Express 4(10), 2179-2186 (2013).

170. K. M. Khan et al., "Depth-sensitive Raman spectroscopy combined with optical coherence tomography for layered tissue analysis," J. Biophotonics 7(1-2), 77-85 (2014).

171. A. Zumbusch, G. R. Holtom, and X. S. Xie, "Three-dimensional vibrational imaging by coherent anti-Stokes Raman scattering," Phys. Rev. Lett. 82, 4142-4145 (1999).

172. E. O. Potma and X. S. Xie, "CARS microscopy for biology and medicine," Opt. Photon. News 15(11), 40-45 (2004).

173. D. L. Marks et al., "Interferometric differentiation between resonant coherent anti-Stokes Raman scattering and nonresonant four-wavemixing processes," Appl. Phys. Lett. 85(23), 5787-5789 (2004).

174. M. Göppert-Mayer, "Über Elementarakte mit zwei Quantensprüngen," Annalen der Physik 401(3), 273-294 (1931).

175. W. Kaiser and C. G. Garrett, "Two-photon excitation in CaF2: Eu2+," Phys. Rev. Lett. 7(6), 229-231 (1961).

176. W. R. Zipfel et al., "Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation," Proc. Natl. Acad. Sci. U S A 100(12), 7075-7080 (2003).

177. I. Freund, M. Deutsch, and A. Sprecher, "Connective tissue polarity. Optical second-harmonic microscopy, crossed-beam summation, and small-angle scattering in rat-tail tendon," Biophys. J. 50(4), 693712 (1986).

178. Y. Jiang et al., "Second-harmonic optical coherence tomography," Opt. Lett. 29(10), 1090-1092 (2004).

179. B. E. Applegate et al., "Polarization-resolved second-harmonic-generation optical coherence tomography in collagen," Opt. Lett. 29(19), 2252-2254 (2004).

180. P. Fratzl, Collagen: Structure and Mechanics, p. 524, Springer, New York (2008).

181. P. Jeannesson, "Impact of carbamylation and glycation of collagen type I on migration of HT1080 human fibrosarcoma cells," Int. J. Oncol. 40(6), 1797-1804 (2012).

182. M. J. Farrar et al., "In vivo imaging of myelin in the vertebrate central nervous system using third harmonic generation microscopy," Biophys. J. 100(5), 1362-1371 (2011).

183. E. Beaurepaire et al., "Combined scanning optical coherence and two-photon-excited fluorescence microscopy," Opt. Lett. 24(14), 969-971 (1999).

184. R. V. Kuranov et al., "Combined application of optical methods to increase the information content of optical coherent tomography in diagnostics of neoplastic processes," Quantum Electron. 32(11), 993 (2002).

185. Y. T. Pan et al., "Enhancing early bladder cancer detection with fluo-rescence-guided endoscopic optical coherence tomography," Opt. Lett. 28(24), 2485-2487 (2003).

186. J. Liu et al., "In vivo bladder imaging with microelectromechanical-systems-based endoscopic spectral domain optical coherence tomography," J. Biomed. Opt. 12(3), 034009 (2007).

187. A. T. Yeh et al., "Imaging wound healing using optical coherence tomography and multiphoton microscopy in an in vitro skin-equivalent tissue model," J. Biomed. Opt. 9(2), 248-253 (2004).

188. C. Vinegoni et al., "Integrated structural and functional optical imaging combining spectral-domain optical coherence and multiphoton microscopy," Appl. Phys. Lett. 88(5), 053901 (2006).

189. J. A. Jo et al., "In vivo simultaneous morphological and biochemical optical imaging of oral epithelial cancer," IEEE Trans. Biomed. Eng. 57(10), 2596-2599 (2010).

190. G. M. Dobre, A. G. Podoleanu, and R. B. Rosen, "Simultaneous optical coherence tomography-indocyanine green dye fluorescence imaging system for investigations of the eye's fundus," Opt. Lett. 30(1), 58-60 (2005).

191. R. B. Rosen et al., "Simultaneous OCT/SLO/ICG imaging," Invest. Ophthalmol. Vis. Sci. 50(2), 851-860 (2008).

192. Q. Wu, B. E. Applegate, and A. T. Yeh, "Cornea microstructure and mechanical responses measured with nonlinear optical and optical coherence microscopy using sub-10-fs pulses," Biomed. Opt. Express 2(5), 1135-1146 (2011).

193. S. P. Chong et al., "Tri-modal microscopy with multiphoton and optical coherence microscopy/tomography for multi-scale and multi-contrast imaging," Biomed. Opt. Express 4(9), 1584-1594 (2013).

194. C. Dai, X. Liu, and S. Jiao, "Simultaneous optical coherence tomography and autofluorescence microscopy with a single light source," J. Biomed. Opt. 17(8), 080502 (2012).

195. S. Tang et al., "Combined multiphoton microscopy and optical coherence tomography using a 12-fs broadband source," J. Biomed. Opt. 11(2), 020502 (2006).

196. S. Yazdanfar et al., "Multifunctional imaging of endogenous contrast by simultaneous nonlinear and optical coherence microscopy of thick tissues," Microsc. Res. Techn. 70(7), 628-633 (2007).

197. A. Bradu et al., "Dual optical coherence tomography/fluorescence microscopy for monitoring of Drosophila melanogaster larval heart," J. Biophotonics 2(6-7), 380-388 (2009).

198. B. W. Graf et al., "Dual-spectrum laser source based on fiber continuum generation for integrated optical coherence and multiphoton microscopy," J. Biomed. Opt. 14(3), 034019 (2009).

199. B. Jeong et al., "Combined two-photon microscopy and optical coherence tomography using individually optimized sources," Opt. Express 19(14), 13089-13096 (2011).

200. K. König et al., "Clinical optical coherence tomography combined with multiphoton tomography of patients with skin diseases," J. Biophotonics 2(6-7), 389-397 (2009).

201. K. König, "Hybrid multiphoton multimodal tomography of in vivo human skin," IntraVital 1(1), 11-26 (2012).

202. A. Alex et al., "Three-dimensional multiphoton/optical coherence tomography for diagnostic applications in dermatology," *J.* Biophotonics 6(4), 352-362 (2013).

203. Y. Zhao et al., "Integrated multimodal optical microscopy for structural and functional imaging of engineered and natural skin," *J.* Biophotonics 5(5-6), 437-448 (2012).

204. B. W. Graf et al., "In vivo multimodal microscopy for detecting bone-marrow-derived cell contribution to skin regeneration," J. Biophotonics 7(1-2), 96-102 (2014).

205. A. R. Tumlinson et al., "Miniature endoscope for simultaneous optical coherence tomography and laser-induced fluorescence measurement," Appl. Opt. 43(1), 113-121 (2004).

206. J. B. McNally et al., "Task-based imaging of colon cancer in the Apc(Min/+) mouse model," Appl. Opt. 45(13), 3049-3062 (2006).

207. L. P. Hariri et al., "Endoscopic optical coherence tomography and laser-induced fluorescence spectroscopy in a murine colon cancer model," Lasers Surg. Med. 38(4), 305-313 (2006).

208. L. P. Hariri et al., "Serial endoscopy in azoxymethane treated mice using ultra-high resolution optical coherence tomography," Cancer Biol. Ther. 6(11), 1753-1762 (2007).

209. S. Yuan et al., "Combining optical coherence tomography with fluorescence molecular imaging: towards simultaneous morphology and molecular imaging," Phys. Med. Biol. 55(1), 191-206 (2010).

210. J. Xi et al., "Integrated multimodal endomicroscopy platform for simultaneous en face optical coherence and two-photon fluorescence imaging," Opt. Lett. 37(3), 362-364 (2012).

211. M. Wojtkowski et al., "In vivo human retinal imaging by Fourier domain optical coherence tomography," J. Biomed. Opt. 7(3), 457463 (2002).

212. I.-K. Jang et al., "Visualization of coronary atherosclerotic plaques in patients using optical coherence tomography: comparison with intravascular ultrasound," J. Am. Coll. Cardiol. 39(4), 604-609 (2002).

213. W. Wieser et al., "Extended coherence length megahertz FDML and its application for anterior segment imaging," Biomed. Opt. Express 3(10), 2647-2657 (2012).

214. J. Dunkers, M. Cicerone, and N. Washburn, "Collinear optical coherence and confocal fluorescence microscopies for tissue engineering," Opt. Express 11(23), 3074-3079 (2003).

215. X. J. Wang, T. E. Milner, and J. S. Nelson, "Characterization of fluid flow velocity by optical Doppler tomography," Opt. Lett. 20(11), 1337-1339 (1995).

216. C. E. Saxer et al., "High-speed fiber based polarization-sensitive optical coherence tomography of in vivo human skin," Opt. Lett. 25(18), 1355-1357 (2000).

217. Y. N. Billeh, M. Liu, and T. Buma, "Spectroscopic photoacoustic microscopy using a photonic crystal fiber supercontinuum source," Opt. Express 18(18), 18519-18524 (2010).

218. C. Zhang et al., "Reflection-mode submicron-resolution in vivo photoacoustic microscopy," J. Biomed. Opt. 17(2), 020501 (2012).

219. J. Laufer et al., "In vivo preclinical photoacoustic imaging of tumor vasculature development and therapy," J. Biomed. Opt. 17(5), 056016(2012).

220. S. Hu, K. Maslov, and L. V. Wang, "Second-generation optical-resolution photoacoustic microscopy with improved sensitivity and speed," Opt. Lett. 36(7), 1134-1136 (2011).

221. M. Liu, "A study of spectral domain optical coherence tomography and photoacoustic microscopy for biometric and biomedical applications," M.S. Thesis, University of Delaware, Newark, Delaware (2011).

222. M. Liu et al., "In vivo three dimensional dual wavelength photoacoustic tomography imaging of the far red fluorescent protein E2-Crimson expressed in adult zebrafish," Biomed. Opt. Express 4(10), 1846-1855 (2013).

223. M. Liu et al., "In vivo spectroscopic photoacoustic tomography imaging of a far red fluorescent protein expressed in the exocrine panreas of adult zebrafish," Proc. SPIE 8943, 894341 (2014).

224. J. A. Viator et al., "Clinical testing of a photoacoustic probe for port wine stain depth determination," Lasers Surg. Med. 30(2), 141-148 (2002).

225. G. J. Tearney et al., "Scanning single-mode fiber optic catheter-endoscope for optical coherence tomography," Opt. Lett. 21(7), 543-545 (1996).

226. J.-M. Yang et al., "Photoacoustic endoscopy," Opt. Lett. 34(10), 15911593 (2009).

227. J.-M. Yang et al., "A 2.5 − mm diameter probe for photoacoustic and ultrasonic endoscopy," Opt. Express 20(21), 23944-23953 (2012).

228. Y. Yang et al., "Integrated optical coherence tomography, ultrasound and photoacoustic imaging for ovarian tissue characterization," Biomed. Opt. Express 2(9), 2551-2561 (2011).

229. L. Li et al., "Three-dimensional combined photoacoustic and optical coherence microscopy for in vivo microcirculation studies," Opt. Express 17(19), 16450-16455 (2009).

230. S. Jiao et al., "Simultaneous multimodal imaging with integrated photoacoustic microscopy and optical coherence tomography," Opt. Lett. 34(19), 2961-2963 (2009).

231. C. Lee et al., "Combined photoacoustic and optical coherence tomography using a single near-infrared supercontinuum laser source," Appl. Opt. 52(9), 1824-1828 (2013).

232. X. Zhang, H. F. Zhang, and S. Jiao, "Optical coherence photoacoustic microscopy: accomplishing optical coherence tomography and photoacoustic microscopy with a single light source," J. Biomed. Opt. 17(3), 030502 (2012).

233. L. Xi et al., "Miniature probe combining optical-resolution photoacoustic microscopy and optical coherence tomography for in vivo microcirculation study," Appl. Opt. 52(9), 1928-1931 (2013).

234. T. Liu et al., "Combined photoacoustic microscopy and optical coherence tomography can measure metabolic rate of oxygen," Biomed. Opt. Express 2(5), 1359-1365 (2011).

235. X. Cai et al., "Investigation of neovascularization in three-dimensional porous scaffolds in vivo by a combination of multiscale photoacoustic microscopy and optical coherence tomography," Tissue Eng. Part C Methods 19(3), 196-204 (2013).

236. V. Tsytsarev et al., "Photoacoustic and optical coherence tomography of epilepsy with high temporal and spatial resolution and dual optical contrasts," J. Neurosci. Methods 216(2), 142-145 (2013).

237. S. Jiao et al., "Photoacoustic ophthalmoscopy for in vivo retinal imaging," Opt. Express 18(4), 3967-3972 (2010).

238. B. Cox et al., "Quantitative spectroscopic photoacoustic imaging: a review," J. Biomed. Opt. 17(6), 061202 (2012).

239. L. Wang et al., "Fast voice-coil scanning optical-resolution photoacoustic microscopy," Opt. Lett. 36(2), 139-141 (2011).

240. E. Z. Zhang et al., "Multimodal photoacoustic and optical coherence tomography scanner using an all optical detection scheme for 3D morphological skin imaging," Biomed. Opt. Express 2(8), 2202-2215 (2011).

241. B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," J. Biomed. Opt. 15(2), 021314 (2010).

242. W. Shi et al., "In vivo near-realtime volumetric optical-resolution photoacoustic microscopy using a high-repetition-rate nanosecond fiber-laser," Opt. Express 19(18), 17143-17150 (2011).

243. P. Shao et al., "Mosaic acquisition and processing for optical-resolution photoacoustic microscopy," J. Biomed. Opt. 17(8), 080503 (2012).

244. W. Trasischker et al., "In vitro and in vivo three-dimensional velocity vector measurement by three-beam spectral-domain Doppler optical coherence tomography," J. Biomed. Opt. 18(11), 116010 (2013).

245. M. Liu and T. Buma, "Biometric mapping of fingertip eccrine glands with optical coherence tomography," IEEE Photonics Technol. Lett. 22(22), 1677-1679 (2010).