

MIL-HDBK-141

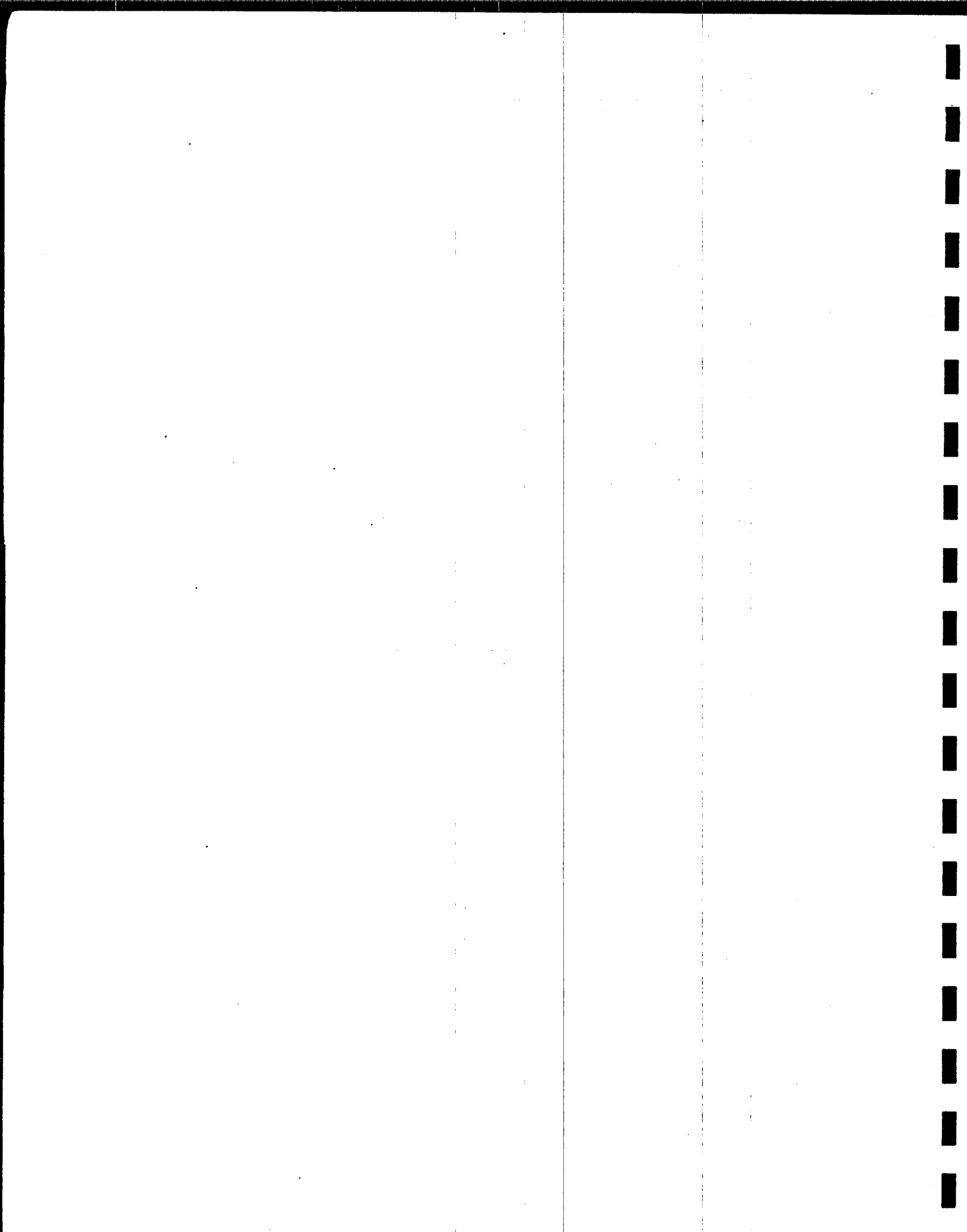
5 OCTOBER 1962

MILITARY STANDARDIZATION HANDBOOK

OPTICAL DESIGN



FSC 6650



DEFENSE SUPPLY AGENCY
WASHINGTON 25, D.C.

MIL-HDEK-141
Optical Design
5 October 1962

1. This handbook was developed by the Department of Defense with the assistance of a leading optical manufacturer. Major contributions were made by persons who, by virtue of experience in their particular fields, are recognized as qualified authorities on the subject of optical design.

2. This publication was approved on 5 October 1962 for printing and inclusion in the military standardization handbook series.

3. This document provides engineering personnel with an introduction to optical theory, and treats to an advanced level the fundamentals and principles of optical design. It is expected that wide distribution of the methods of design and computation presented in this handbook will result in a more efficient and accurate method of complying with military optical requirements.

4. To aid in maintaining the intended status of this handbook as a source of prevailing information, readers are encouraged to report any errors and suggestions for changes and additions to the Standardization Division, Defense Supply Agency, Washington 25, D. C.

This handbook contains copyright material.

CONTENTS

SECTION 1. INTRODUCTION	1-1
1.1 Scope	1-1
1.2 Definitions	1-2
1.3 Reference documents	1-5
SECTION 2. FUNDAMENTALS OF GEOMETRICAL OPTICS ¹	2-1
2.1 General	2-1
2.2 Law of refraction	2-1
2.3 Law of reflection	2-3
2.4 Total internal reflection	2-4
2.5 Index of refraction	2-4
2.6 Dispersion of light	2-5
2.7 Characteristics of optical glass	2-6
SECTION 3. CONSIDERATIONS OF PHYSICAL OPTICS ²	3-1
3.1 Introduction	3-1
3.2 Physical nature of light	3-1
3.3 Interference between waves	3-4
SECTION 4. VISUAL OPTICS ³	4-1
4.1 Introduction	4-1
4.2 Anatomy of the eye	4-1
4.3 Optical constants of the eye	4-3
4.4 Image formation and the retina	4-5
4.5 Seeing	4-10
4.6 Movement of the eyes	4-14
4.7 Binocular vision	4-16
4.8 Fatigue and ageing	4-18
SECTION 5. FUNDAMENTAL METHODS OF RAY TRACING ¹	5-1
5.1 General	5-1
5.2 Definitions and conventions	5-3
5.3 Basic ray trace procedure	5-5
5.4 Skew ray trace equations for spherical surfaces	5-5
5.5 Skew ray trace equations for aspheric surfaces	5-13
5.6 Meridional rays	5-21
5.7 Graphical ray tracing procedure	5-26
5.8 Differential ray tracing procedure	5-27
5.9 Paraxial rays	5-32
5.10 Graphical ray trace for paraxial rays	5-34
5.11 Different "orders" of optics	5-35
SECTION 6. FIRST ORDER OPTICS ¹	6-1
6.1 General	6-1
6.2 Numerical tracing of paraxial rays	6-1
6.3 Optical invariant	6-5
6.4 Linearity of paraxial ray tracing equations	6-8
6.5 Cardinal points of an optical system	6-10
6.6 Calculation of focal length from finite conjugate data	6-17

Ref 1, 2, 3 - See page vi for Author.

6.7	Systems of thin lenses in air	6-17
6.8	Optical systems involving mirrors	6-19
6.9	Differential changes in first order optics	6-23
6.10	Chromatic aberration	6-26
6.11	Entrance and exit pupils, the chief ray and vignetting	6-37

SECTION 7. SIMPLE THIN LENS. OPTICAL SYSTEMS¹

7.1	Introduction	7-1
7.2	Simple magnifier	7-1
7.3	Microscope	7-4
7.4	Telescope	7-8
7.5	Optical relay systems, Periscopes	7-8
7.6	Galilean telescope	7-11

SECTION 8. ABERRATION ANALYSIS AND THIRD ORDER THEORY¹ ..

8.1	Significance of ray trace data	8-1
8.2	Spot diagram	8-1
8.3	Meridional and skew fans	8-3
8.4	Use of third order theory in aberration analysis	8-3
8.5	Zero-degree image in D light	8-5
8.6	Imagery for an off-axis object point	8-9
8.7	Calculation of third order contributions	8-14
8.8	Afocal optical systems	8-15
8.9	Stop shift equations	8-16
8.10	Thin lens aberration theory	8-18

SECTION 9. METHOD OF LENS DESIGN⁴

9.1	Process of designing a lens system	9-1
9.2	Description and analysis of basic procedure	9-i
9.3	Summary of equations used in calculation of third order aberrations	9-11

SECTION 10. AN APPLICATION OF THE METHOD OF LENS DESIGN⁴ ..

10.1	Step one - selecting the lens type	10-1
10.2	Step two - first order thin lens solution	10-2
10.3	Step three - third order thin lens solution	10-13
10.4	Step four - thick lens first order and third order aberrations ..	10-17
10.5	Step five - tracing a few selected rays	10-19
10.6	Step six - readjusting third order aberrations	10-20
10.7	Evaluation of over-all performance	10-27
10.8	Summary	10-27

SECTION 11. TELESCOPE OBJECTIVES⁴

11.1	Introduction	11-1
11.2	Design procedure for a thin lens telescope objective	11-3
11.3	Design procedure for a thick lens telescope objective	11-6
11.4	Secondary spectrum of telescope objectives	11-18
11.5	Summary	11-25

SECTION 12. LENS RELAY SYSTEMS⁴

12.1	Introduction	12-1
12.2	The basic lens problem of a relay system	12-1
12.3	A visual system. Numerical example	12-1

Ref 1, 4 - See page vi for Author.

12.4	Secondary color in a relay system	12-2
12.5	Further details on design of doublets as relay lenses	12-2
12.6	Double relay systems	12-3
12.7	Summary	12-4

SECTION 13. MIRROR AND PRISM SYSTEMS ⁴

13.1	Introduction	13-1
13.2	Reflection	13-1
13.3	Location of image	13-4
13.4	Orientation of image	13-6
13.5	Image sphere	13-10
13.6	Reflection from two mirrors	13-13
13.7	Typical prism systems	13-15
13.8	Tunnel diagram	13-21
13.9	Aberrations introduced by prisms	13-23
13.10	Prism data sheets	13-25

SECTION 14. EYEPIECES ⁴

14.1	General principles	14-1
14.2	Method of description	14-1
14.3	Huygenian eyepiece	14-2
14.4	Ramsden eyepiece	14-4
14.5	Kellner eyepiece	14-6
14.6	Orthoscopic eyepiece	14-8
14.7	Symmetrical (Plössl) eyepiece	14-10
14.8	Berthele eyepiece	14-12
14.9	Erfle eyepiece	14-14
14.10	Modified Erfle eyepiece	14-16
14.11	Wild eyepiece	14-18
14.12	Summary	14-20

SECTION 15. COMPLETE TELESCOPE ⁴

15.1	Introduction	15-1
15.2	Design problem	15-1
15.3	Preliminary considerations	15-1
15.4	Design refinement	15-2
15.5	Completed design	15-3

SECTION 16. APPLICATIONS OF PHYSICAL OPTICS ²

16.1	Introduction	16-1
16.2	Fizeau interferoscope	16-3
16.3	Twyman-Green interferometer	16-5
16.4	Effect of monochromaticity on fringe contrast	16-7
16.5	Effect of pinhole size on contrast	16-8
16.6	Young's pinhole interferometer	16-9
16.7	Lloyd's interferometer	16-12
16.8	Fresnel coefficients for normal incidence	16-12
16.9	Interference with plane parallel plates and distant light sources	16-14
16.10	Interference with plane parallel plates and nearby light sources	16-16
16.11	Haidinger's interference fringes	16-17
16.12	Fizeau fringes	16-19
16.13	Newton's rings and Newton's fringes	16-19
16.14	Complex numbers	16-26
16.15	Transmittance of plane parallel plates	16-28
16.16	Reflectance from plane parallel plates	16-32
16.17	Multiple beam interference fringes from slightly inclined surfaces	16-34

16.18	Measurements with monochromatic light	16-37
16.19	Method of channeled spectra	16-41
16.20	Interpretation of measurements with channeled spectra	16-41
16.21	Huygen's principle	16-45
16.22	Fraunhofer diffraction	16-47
16.23	Fraunhofer diffraction from a rectangular aperture	16-49
16.24	Fraunhofer diffraction from circular apertures	16-50
16.25	Diffraction from spherical wavefronts	16-52
16.26	Primary diffraction integrals with objectives having circular apertures	16-54
16.27	Resolution with circular apertures	16-56
16.28	Out-of-focus aberration	16-58

SECTION 17. OPTICAL MATERIAL ⁵

17.1	Introduction	17-1
17.2	Refracting material characteristics	17-1
17.3	Refractivity and dispersion	17-3
17.4	Inclusions	17-4
17.5	Environmental characteristics	17-5
17.6	Refractive materials for specific wavelength ranges	17-5
17.7	Reflecting materials	17-8
17.8	Availability, cost, ease of working	17-10

SECTION 18. ATMOSPHERIC OPTICS ⁶

18.1	Introduction	18-1
18.2	Extinction	18-1
18.3	Extinction and visual instruments	18-4
18.4	Extinction and photographic instruments	18-5
18.5	Seeing	18-6
18.6	Thermal effects	18-8
18.7	Atmospheric contaminants	18-9
18.8	Effect of atmospheric optics on instrument design	18-10

SECTION 19. OPTICS FOR MISSILE TRACKING ⁷

19.1	Introduction	19-1
19.2	Refractive systems	19-2
19.3	Reflective systems	19-7
19.4	Catadioptric systems	19-10
19.5	Applied systems	19-14

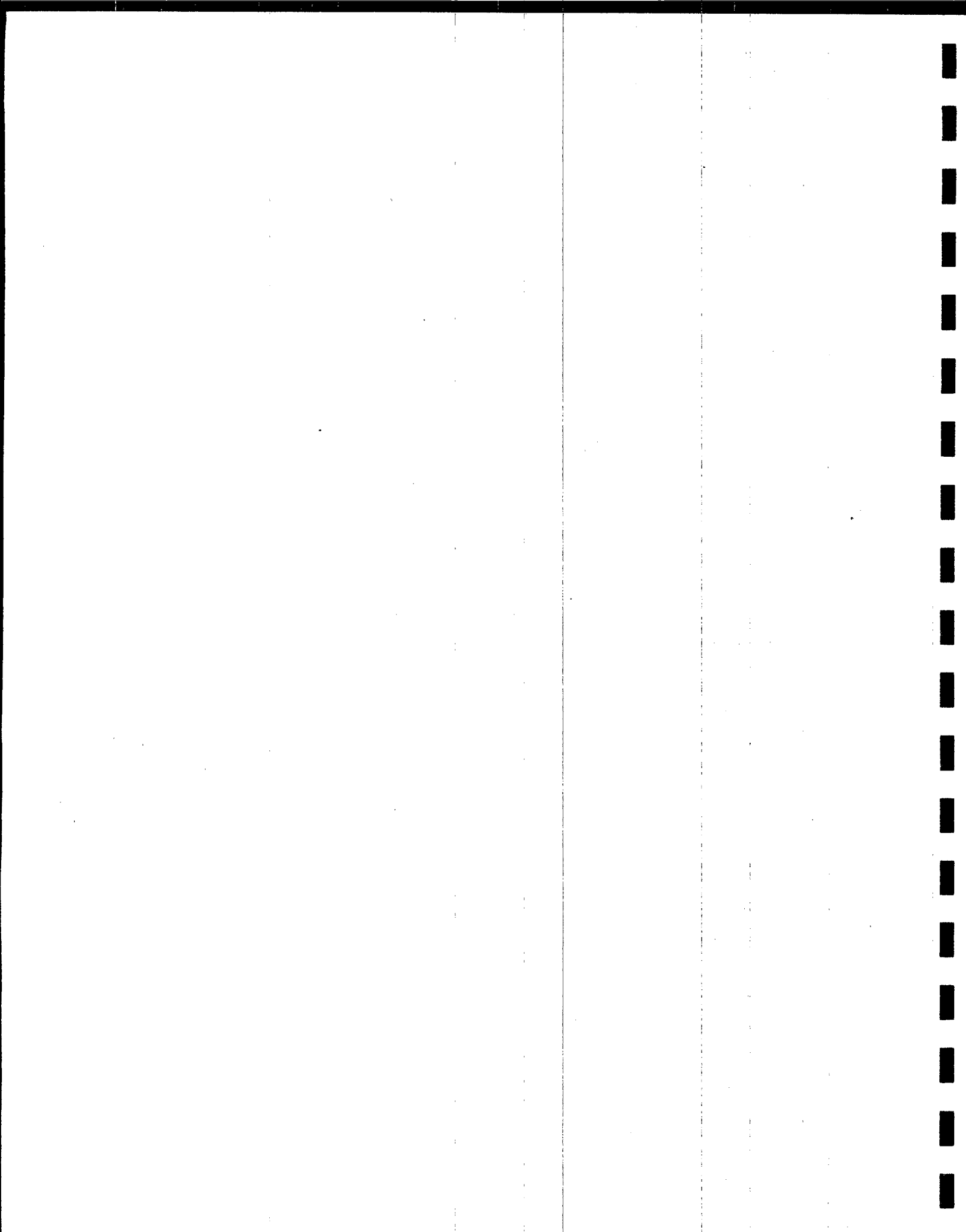
SECTION 20. APPLICATIONS OF THIN FILM COATINGS ⁸

20.1	Introduction	20-1
20.2	Manufacture of multilayer filters	20-14
20.3	Antireflection coatings	20-18
20.4	Reflectivity of multilayers with periodic structure	20-39
20.5	Long-wave pass filters	20-56
20.6	Short-wave pass filters	20-63
20.7	Beam splitters	20-64
20.8	Mirrors	20-68
20.9	Band pass filters	20-71
20.10	Fabry-Perot type filters (interference filters)	20-71
20.11	References for further study	20-91

SECTION 21. COATING OF OPTICAL SURFACES ²	21-1
21.1 Introduction	21-1
21.2 Definitions and principles	21-1
21.3 Zero reflectance from non-absorbing monolayers and substrates	21-27
21.4 Matrix methods	21-30
21.5 Quaternion methods	21-35
21.6 Monolayer coatings	21-41
21.7 Bilayer coatings	21-49
21.8 Trilayers	21-64
21.9 Quadrilayers	21-66
21.10 Quarter-wave multilayers	21-67
21.11 Materials and texts	21-77
SECTION 22. INFRARED OPTICAL DESIGN ⁵	22-1
22.1 Introduction	22-1
22.2 Infrared optical material	22-1
22.3 Environmental requirements	22-2
22.4 Operational requirements	22-2
22.5 Near infrared region	22-3
22.6 Intermediate and far infrared region	22-5
22.7 Summary and conclusion	22-12
SECTION 23. MICROSCOPE OPTICS ⁹	23-1
23.1 Introduction	23-1
23.2 Characteristics	23-1
23.3 Components of a compound microscope	23-3
23.4 Darkfield microscopy	23-11
23.5 Ultramicroscopy	23-15
23.6 Phase microscopy	23-16
23.7 Interference microscopy	23-19
23.8 Polarizing microscopy	23-23
23.9 Fluorescence microscopes	23-23
23.10 Stereoscopic microscope	23-24
23.11 Petrographic microscope	23-24
SECTION 24. DESIGN PHASE OPTICAL TESTS ²	24-1
24.1 Introduction	24-1
24.2 Calculation of Seidel aberrations	24-2
24.3 Spot diagram	24-3
24.4 Phase front calculations	24-4
SECTION 25. PRODUCTION PHASE OPTICAL TESTS ²	25-1
25.1 Introduction	25-1
25.2 Focal length	25-1
25.3 Longitudinal spherical aberration	25-3
25.4 Coma	25-4
25.5 Astigmatism and curvature field	25-4
25.6 Distortion	25-5
25.7 Auxiliary optical measurements	25-5
25.8 Optical devices, testing systems and procedures	25-8
25.9 Ronchi test	25-20
25.10 Foucault test	25-24
25.11 Star test	25-29

SECTION 26. EVALUATION PHASE OPTICAL TESTS ²	26-1
26.1 Resolving power tests	26-1
26.2 General discussion of sine-wave testing	26-10
26.3 Sine-wave testing with sine-wave targets	26-12
26.4 Sine-wave testing with square-wave targets	26-17
APPENDIX	27-1
INDEX	28-1

REF	AUTHOR
1.	Dr. Robert E. Hopkins, University of Rochester Dr. Richard Hanau, University of Kentucky
2.	Dr. Harold Osterberg, American Optical Company
3.	Dr. Oscar W. Richards, American Optical Company
4.	Dr. Robert E. Hopkins, University of Rochester
5.	Mr. A. J. Kavanagh, American Optical Company
6.	Dr. Ralph Wight, Photronics Corporation
7.	Dr. Seymour Rosin, Scanoptic, Incorporated
8.	Dr. Philip Baumeister, University of Rochester
9.	Mr. Alva Bennett, American Optical Company



1 INTRODUCTION

1.1 SCOPE

In 1952 the Ordnance Corps published ORDM 2-1, Design of Fire Control Optics. The purpose of that publication was to make available to engineering and design personnel all pertinent optical data that had been accumulated by Frankford Arsenal. In the meantime, the rapidly increasing application of optical features in the design of military systems, and the accelerated rate of over-all technical advancement in the optical field bypassed ORDM 2-1 to such an extent that, in 1958, a tri-service project was initiated to gather and present, in a single volume, up-to-date engineering information, formulas, and calculations currently applicable in the design of individual optical elements and complete optical systems. Military handbook MIL-HDBK-141 is the result of that project.

This Department of Defense handbook was developed by a leading optical manufacturer under Department of the Army Contract DA-36-038-ORD-20590. Major contributions were made by a group of recognized authorities in the field of optical design. All work was performed under the guidance of Frankford Arsenal.

Although many excellent reference works at the college and advanced-design level are available, there is a lack of transition among them from one subject to another. To provide this needed transitional feature, MIL-HDBK-141 presents as nearly as possible the full range of subjects encountered in the field of optical design, including sections covering fundamentals, principles of design, and design data.

The first seven sections serve mainly to acquaint the reader with the basic concepts of optics, and to introduce the mathematical notation employed in later sections. These initial sections require that the reader have a working knowledge of analytical geometry, differential and integral calculus, and physics.

The sections on principles of design introduce typical design considerations encountered in basic types of optical systems. Included are discussions on system aberrations and their computation and correction. The computing schemes described should enable the designer to work efficiently and accurately.

The remaining sections of the handbook apply to various commonly used components and combinations, discussions of problems and solutions in special design areas, and data on general topics related to problems of optical design and manufacture.

1.2 DEFINITIONS.

1.2.1 Symbols and Notations. The following symbols are used in this handbook. Table I contains the English alphabet notation, Table II, the Greek alphabet, and Table III, the mathematical symbols.

TABLE I

Symbol	Usage	Symbol	Usage
A	Area, points, linear dimension.	H	Magnetic vector.
\overline{A}	Aperture area.	h	Diameter of a mirror, fringe width.
a	Real number, points, mirror aperture, amplitude of a wave. Special (Geometrical optics): 3rd order chromatic aberration.	I	Angle of incidence; positive if the ray can be made coincident with the normal to the surface by rotating the ray in a clockwise direction by an angle less than 90° .
B	Points, 3rd order surface contribution for spherical aberration.	I'	Angle of refraction; positive if the ray can be made coincident with the normal to the surface by rotating the ray in a clockwise direction by an angle less than 90° .
b	Real number, coefficient of a power series. Special (Geom. optics): 3rd order chromatic aberration.	I_c	Critical angle.
C	Points. As a subscript denotes red light using hydrogen line. Special (Geom. optics): 3rd order surface contribution for astigmatism.	i	Imaginary number, paraxial angle of incidence.
c	Points, constant velocity for all electromagnetic waves in a vacuum, curvature; positive if the center of curvature is to the right of a surface.	i'	Paraxial angle of refraction.
D	Lens diameter. As a subscript denotes yellow light using sodium line. Special (Geom. optics): distortion.	$J_n(x)$	Bessel function of order n.
d	Thickness, pupil diameter in mm., as a subscript denotes yellow light using helium line. Special (Geom. optics): distance or distance along a ray (not along optical axis).	K	Absorption constant, constant of proportionality, optical constant, optical direction cosine, Ratio of the energy density at the diffraction head when objective is out-of-focus by an amount to the energy density at the diffraction head when objective is in focus.
E	Electric vector, 3rd order contribution for distortion.	k	Image surface, $2\pi/\lambda$.
F	Principal focal point. As a subscript denotes blue light of hydrogen line. Total flux radiated by a surface.	L.	Distance, optical direction cosine.
f	Focal length of a lens; positive if the first principal point is to the right of the first principal focus. A function related to the phase changes on reflection at the reflecting coated surfaces.	l	Path length through the particular medium.
f'	Focal length of a lens; positive if the second principal focal point is to the right of the second principal point.	M	Magnification ratio, optical direction cosine, unit normal vector.
		MP	Magnifying power.
		m	Lateral magnification.
		N	Number of inter-reflections, nodal point of a lens.
		n	Index of refraction, optical constant of an homogeneous isotropic film, n^{th} order of terms.
		O	Origin, object surface.

TABLE I (Cont.)

Symbol	Usage	Symbol	Usage
o	As subscript pertains to object.	v	Velocity of light in vacuum, size of field of view.
P	Object point, principal point of a lens. Special (Geom. optics): Petzval contribution.	W	Energy density or energy flux.
P'	Image point.	w	Optical half-width of the Fabry-Perot fringes.
\bar{P}	Partial dispersion ratio.	X, Y	Rectangular coordinate system of the Z plane, with subscripts they denote the position coordinate of the ray intercepts on the subscript surface.
PD	Interpupillary distance.	X_ν	Radii of the dark fringes.
Q	Incident unit ray vector, quaternion, ratio.	X_μ	Radii of the bright fringes.
Q'	Reflected unit ray vector.	x	Distance along X-coordinate.
q	Scalar coefficient.	Y	Radius of entrance pupil.
R	Radius, reflectance, resolving power in seconds of arc.	\bar{Y}	Height of chief ray.
r	Radius.	Y_ν	Admittance when electric vector is perpendicular to the plane of incidence in the ν th layer.
S	Object conjugate of a lens, surface of a lens, time-averaged Poynting vector.	\bar{y}	Object height, height of oblique paraxial ray.
S'	Image conjugate of a lens.	y_ν	Admittance when the magnetic vector is perpendicular to the plane of incidence in the ν th layer.
T	Internal transmittance, time-averaged energy transmittance, period.	Z	The abscissa of the rectangular coordinate system used. In general the axis of propagation or optical axis; with subscript, denotes a position coordinate of the ray intercept on the subscript surface. Complex number, sag.
t	Thickness measured along optical axis, Special: (Physical optics): time.	z	Distance along Z axis.
U	Angle between meridional ray and optical axis, vector.		
u	Angle between paraxial ray and optical axis, polar coordinate.		
V	Distance, optical path, vector, wavefront.		

TABLE II

Symbol	Usage
α	Absorption coefficient, angle, angular magnification, direction cosine with respect to X axis.
β	Absorption coefficient, angle, direction cosine with respect to Y axis.
γ	Constant, direction cosine with respect to Z axis.
Δ	Total phase difference, increment of change.
δ	Angle of deviation, phase difference.
ϵ	Dielectric constant.
ζ	Abscissa.
η	Ordinate.
κ	Extinction coefficient.

Symbol	Usage
θ	Angular limit of resolution, angular measurement.
λ	Wavelength.
μ	Magnetic permeability.
ν	Abbe constant, extinction coefficient, frequency of vibration, integer.
ρ	Amplitude reflectance.
σ	Electric conductivity, phase change, unit vector.
τ	Amplitude transmittance.
ϕ	Angle, phase angle, power of a thin lens.
Φ	Optical invariant.
ω	Angular velocity, angle.

TABLE III

Symbol	Usage
\pm	Plus or minus.
$=$	Equal to.
\equiv	Identity, defined as.
\approx	Nearly equal to.
\sim	Similar to, special designator when used to overline a capital letter.
\rightarrow	Approaches (from left hand side).
\leftarrow	Approaches (from right hand side).
$>$	Greater than.
$<$	Less than.
\leq	Less than or equal to.
\geq	Greater than or equal to.
$^\circ$	Degree.
\therefore	Therefore.
$()$	Parentheses; multiplication operator.

Symbol	Usage
$*$	Transverse chromatic aberration for some oblique ray displaced from the ray passing through $y_l=0$.
$\sqrt{\quad}$	Square root.
$n\sqrt{\quad}$	n^{th} root.
Σ	Summation operator.
\sum	Sigma-summation operator.
∞	Infinity.
$[\]$	Brackets; multiplication or matrix operators.
∂	Denotes partial differentiation.
\int	Integration operator.
\oint	Integration operator.
π	Pi = 3.1416. π radian = 180° .
e	Base of Napierian or natural logarithm = 2.71828.
Π	Quaternion summation operator.

1.2.2 Terms. In general, the terms used in this handbook conform to Military Standard No. 1241, Optical Terms and Definitions; where special terms are used, the definitions are given in the text. An alphabetical index is provided at the end of the volume for easy reference to these definitions.

1.3 REFERENCE DOCUMENTS.

1.3.1 The following government publications are used in direct reference or provide related information valuable in the general field of optical design:

JAN-G-174 Optical Glass
MIL-STD-12 Abbreviations for Use on Drawings
MIL-STD-34 General Requirements for the Preparation of Drawings for Optical Elements and Optical Systems
MIL-STD-106 Mathematical Symbols
MIL-STD-150 Photographic Lenses
MIL-STD-1241 Optical Terms and Definitions

1.3.2 The following commercial publications are used in direct reference or provide related information valuable in the general field of optical design:

Ballard, S.S., McCarthy, K.A., Wolfe, W.L., State-of-the-Art Report: Optical Materials for Infrared Instrumentation, (Report No. 2389-11-S: I.R.I.A, Univ. of Michigan, 1959).
Bennett, A.H., Jupnik, H., Osterberg, H. and Richards, O.W., Phase Microscopy, (John Wiley and Sons, 1951).
Born and Wolf, Principles of Optics, (Pergamon Press, 1959).
Committee on Colorimetry, The Science of Color, (Thomas Crowell Co., 1954).
Conrady, Applied Optics and Optical Design, Parts 1 and 2 (Dover Publications Inc., 1960).
Drude, Theory of Optics, (Dover Publications Inc., 1960).
Hardy and Perrin, The Principles of Optics, (McGraw - Hill, 1932).
Holland, L., Vacuum Deposition of Thin Films, (John Wiley and Sons, 1956).
International Lighting Vocabulary Vol. I (CIE. - 1.1. - 1957).
Jacobs, Fundamentals of Optical Engineering, (McGraw-Hill, 1943).
Jenkins and White, Fundamentals of Optics, (McGraw-Hill, 1957).
Johnson, B.K., Optics and Optical Instruments, (Dover Publications Inc., 1960).
Journal, Optical Society of America.
Linfoot, Recent Advances in Optics, (Oxford, 1955).
Martin, Technical Optics, (Pitman, 1948).
National Bureau of Standards, Circular No. 526, Optical Image Evaluation, (1954).
Optical Industry Directory, (Optical Publishing Co., 1961).
Sawyer, Experimental Spectroscopy, (Prentice-Hall, 1951).
Searle, Experimental Optics, (Cambridge Univ. Press, 1926).
Sears, F.W., Optics, (Addison-Wesley Press, Inc., 1949).
Strong, Concepts of Classical Optics, (Freeman, 1958).
Strong, Procedures in Experimental Physics, (Prentice-Hall, 1953).
Taylor, The Adjustment and Testing of Telescope Objectives, (Grubb, Parsons and Co., 1946).
Twyman, Prism and Lens Making, (Hilger, 1957).
Wagner, Experimental Optics, (John Wiley and Sons, 1929).

2 FUNDAMENTALS OF GEOMETRICAL OPTICS¹

2.1 GENERAL

2.1.1 Geometrical optics. The term geometrical optics is applied to that branch of physics which deals with the propagation of light in terms of rays. These rays are considered as straight lines in homogeneous media. Geometrical optics, however, does not include some of the wave aspects of light propagation and hence does not take into account interference or diffraction effects. It is the starting point of the design of all optical systems; often it is the end point. It offers a means of progressing from graphical representations to numerical methods of analysis, and of arriving at solutions which in most cases are sufficiently accurate. One purpose of this text is to describe the laws and principles of geometrical optics and to show their application to the design of optical elements and systems.

2.1.2 Wave surfaces and rays. A basic problem in the design of optical systems is the calculation of wave surfaces as they progress through the various optical media. In geometrical optics this calculation is approximated by considering a relatively small number of rays, and then tracing these rays through the system. The actual passage of the rays is computed using analytic geometry procedures and two simple laws, the law of reflection and the law of refraction.

2.1.3 Direction of rays. The rays are perpendicular to the wave surfaces if the radiation is passing through a medium which is optically isotropic. The position of a wave surface (often called a wavefront) with respect to a point source may be determined at any time by the following procedure. From the point source equal optical path lengths are laid off along the rays. The surface that passes through these end points and is normal to the rays is a wavefront. (The optical path length corresponding to a physical path length is the product of the physical path length and the index of refraction.) In birefringent material the ray directions are not necessarily normal to the wave surfaces. The path of a ray of light traveling in a homogeneous medium is a straight line. When the ray is incident upon a surface separating two optically different media, it is reflected and refracted. This usually results in an abrupt change in the direction of the ray.

2.1.4 Angles of incidence, reflection, and refraction. If a normal is erected to the surface separating two media at the point where the ray is incident, the angles which the normal makes with the incident, refracted, and reflected rays are termed, respectively, the angles of incidence, refraction, and reflection. The laws of refraction and reflection, which state the relations existing between these angles, are two of the fundamental laws upon which optical design is based. The third law, mentioned above, states that a ray in a homogeneous medium travels in a straight line.

2.2 THE LAW OF REFRACTION

2.2.1 Diagram for refraction. Figure 2.1 shows a ray of light refracted at an interface between two different homogeneous materials characterized by n_0 and n_1 , which are the respective indices of refraction of the materials. The interface is shown as a straight line representing the intersection of a plane surface with the plane of the paper. This is a special case of the general situation in which the interface is a curved surface. In addition to the refracted ray, shown in Figure 2.1, in general there will also be a reflected ray. This has been omitted in the figure only for the purpose of clarification. For most cases where refraction is the aim, the reflected rays account for less than 10% of the incident energy. Section 21.2 will discuss the calculation of the reflected energy.

2.2.2 Sign convention. The following sign convention will be used for the angles of incidence, refraction, and reflection. If the ray must be rotated clockwise through the acute angle to bring it into coincidence with the normal to the surface, the angle is called positive. The angles I and I' in Figure 2.1 are both positive.

2.2.3 Statement of the law of refraction. The law of refraction is stated in two parts:

(1) The incident ray, the refracted ray, and the normal to the surface all lie in a single plane.

(2) The sines of the angles of incidence and refraction are related by the equation

$$n_0 \sin I = n_1 \sin I' . \quad (1)$$

2.2.4 Vector form of the law of refraction.

2.2.4.1 In solving many three dimensional refraction problems it is convenient to express the law of refraction in vector form. This is accomplished by describing the incident ray direction by a vector of unit

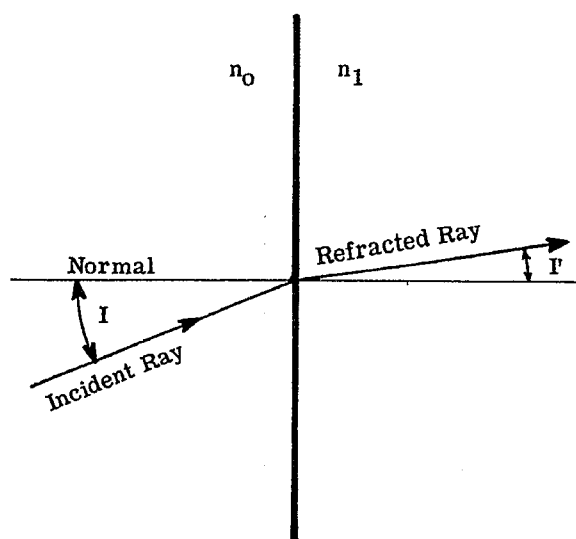


Figure 2.1 - Illustration of refraction.

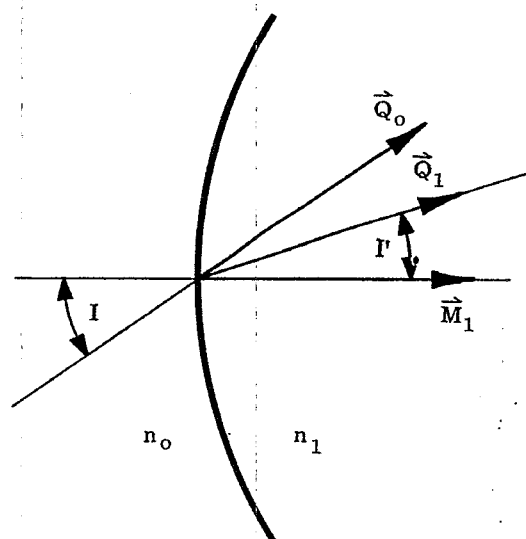


Figure 2.2 - Unit vectors for ray directions.

length \vec{Q}_0 , the refracted ray by a unit vector \vec{Q}_1 , and the normal by a unit vector \vec{M}_1 . Figure 2.2 shows the relationship between these unit vectors and the ray directions. The vector \vec{M}_1 lies along the normal in the direction incident medium to refractive medium.

2.2.4.2 The vector product (cross product) of the two vectors \vec{Q}_0 and \vec{M}_1 is a vector of magnitude

$$\vec{Q}_0 \times \vec{M}_1 = |\vec{Q}_0| |\vec{M}_1| \sin I = \sin I$$

because the angle between these vectors is I and they are each of unit length. The vector whose magnitude is $\sin I$ is perpendicular to the plane containing angle I (the plane of Figure 2.2), and directed perpendicularly into the plane of the paper. Similarly, $\vec{Q}_1 \times \vec{M}_1 = \sin I'$, and this is a vector parallel to $\vec{Q}_0 \times \vec{M}_1$, because the refracted ray lies in the plane determined by the normal and the incident ray.

2.2.4.3 We have established the parallelism of the two vectors whose magnitudes are $\sin I$ and $\sin I'$. By Equation (1) their magnitudes are in the ratio of the indices. Hence

$$\frac{\sin I}{\sin I'} = \frac{\vec{Q}_0 \times \vec{M}_1}{\vec{Q}_1 \times \vec{M}_1} = \frac{n_1}{n_0},$$

and the vector form of the law of refraction may be written as

$$n_0 (\vec{Q}_0 \times \vec{M}_1) = n_1 (\vec{Q}_1 \times \vec{M}_1). \quad (2)$$

Equation (2) indicates, as all vector equations do, that the vector given by the left hand side equals in magnitude and direction the vector given by the right hand side.

2.2.4.4 Equation (2) can be written in another form by absorbing the scalar quantities n_0 and n_1 . Replacing the two vectors $n_0 \vec{Q}_0$ and $n_1 \vec{Q}_1$ by \vec{S}_0 and \vec{S}_1 , respectively, we have

$$\vec{S}_0 \times \vec{M}_1 = \vec{S}_1 \times \vec{M}_1,$$

and

$$(\vec{S}_1 - \vec{S}_0) \times \vec{M}_1 = 0.$$

since neither \vec{M}_1 nor $(\vec{S}_1 - \vec{S}_0)$ is zero, these two vectors must be parallel or anti-parallel. Therefore we can define a quantity Γ (sometimes called the astigmatic constant) by writing

$$\vec{S}_1 - \vec{S}_0 = \Gamma \vec{M}_1. \quad (3)$$

2.2.4.5 Having found the direction of $(\vec{S}_1 - \vec{S}_0)$, we now want to determine its magnitude, Γ . From the definitions of \vec{S}_0 and \vec{S}_1 , and because \vec{Q}_0 and \vec{Q}_1 are unit length, \vec{S}_0 and \vec{S}_1 are two vectors of length n_0 and n_1 , in the directions of the incident and refracted rays respectively. The difference, $\vec{S}_1 - \vec{S}_0$, between these vectors is indicated in Figure 2.3. The length of $\vec{S}_1 - \vec{S}_0$ is the difference between the projections of \vec{S}_1 and \vec{S}_0 on \vec{M}_1 . For the case illustrated, $n_1 > n_0$ and therefore $\cos I' > \cos I$. Hence, since Γ is a positive number for Figure 2.3,

$$\Gamma = n_1 \cos I' - n_0 \cos I = -n_0 \cos I + n_1 \left[\left(\frac{n_0}{n_1} \cos I \right)^2 + \left(\frac{n_0}{n_1} \right)^2 + 1 \right]^{1/2}. \quad (4)$$

Equations (3) and (4) are used in the derivation of the skew ray formulae included in Section 5.

2.3 THE LAW OF REFLECTION

2.3.1 Diagram for reflection. Figure 2.4 shows a ray reflected from a surface. Just as in Figure 2.1, the interface is shown as a straight line, although in general it is a curve. Generally, there will also be a refracted ray which is more or less absorbed as it traverses the medium to the right of the interface. For clarity, only the incident and reflected rays are shown. The calculation of the refracted energy is discussed in Section 21.2.

2.3.2 Statement of the law of reflection. The law of reflection is also stated in two parts:

- (1) The incident ray, the reflected ray, and the normal to the surface all lie in the same plane.
- (2) The angle of incidence is numerically equal to the angle of reflection.

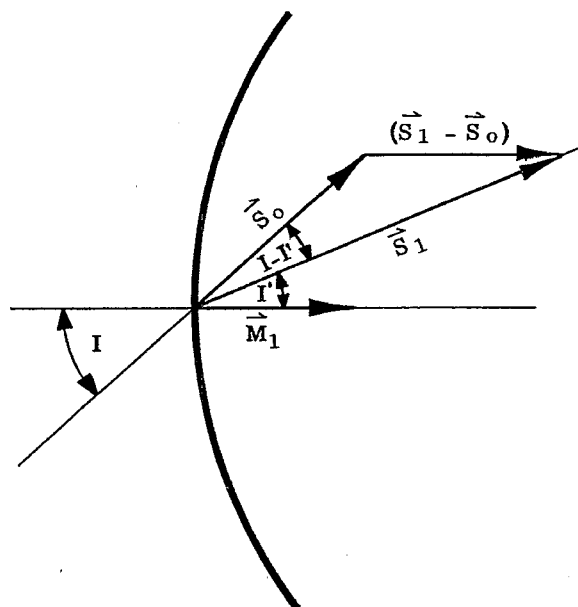


Figure 2.3 - Relation between \vec{S}_0 , \vec{S}_1 , and their difference.

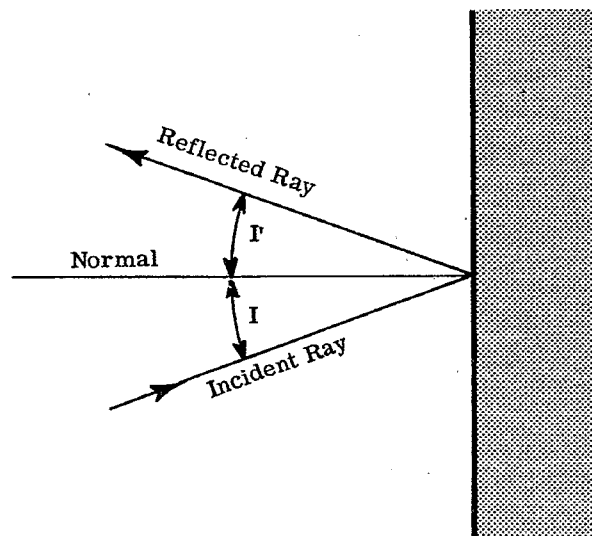


Figure 2.4 - Illustration of reflection.

Note that if I' is labelled as shown in Figure 2.4, then I' is negative while I is positive according to the sign convention. The law of reflection then is

$$I = -I' . \quad (5)$$

2.3.3 Unification of the laws of reflection and refraction. A very convenient way to unify the laws of reflection and refraction is to use the single equation (1) for the law of refraction and to say that in the case of reflection

$$n_1 = -n_0 . \quad (6)$$

With this convention, Equation (1) leads directly to Equation (5). This convention will be used later to provide a completely unified treatment of reflection and refraction problems.

2.4 TOTAL INTERNAL REFLECTION

2.4.1 The critical angle. An inspection of Equation (1) shows that if $n_1 < n_0$, and I' is 90° , the angle of incidence then would be given by

$$\sin I_C = \frac{n_1}{n_0} , \quad (7)$$

where I_C is called the critical angle. If the angle of incidence exceeds the critical angle, the reflected ray has associated with it all the incident energy, as though the interface were a perfect mirror. This effect is used to an advantage in the design of prism systems to obtain reflectivity with very little loss of energy. (See Section 13).

2.4.2 Table of critical angles and indices. Table 2.1 lists the critical angle* corresponding to various indices of refraction. These data are useful in the design of prism systems, where it is necessary to be sure that the prism totally reflects all the desired rays.

n	I_C (radians)	n	I_C (radians)	n	I_C (radians)
1.50	0.729728	1.57	0.690526	1.64	0.655753
1.51	0.723820	1.58	0.685308	1.65	0.651099
1.52	0.718020	1.59	0.680177	1.66	0.646517
1.53	0.712324	1.60	0.675132	1.67	0.642005
1.54	0.706730	1.61	0.670168	1.68	0.637562
1.55	0.701234	1.62	0.665286	1.69	0.633186
1.56	0.695834	1.63	0.660481	1.70	0.628875

Table 2.1 - Table of critical angles (n vs I_C).

2.5 INDEX OF REFRACTION

2.5.1 Absolute index of refraction. It is appropriate at this time to discuss the meaning of index of refraction, referred to as n . The absolute refractive index of a material is defined as the ratio of the velocity of light in a vacuum to that in the material,

$$n_0 = \frac{v_{vac}}{v_o} . \quad (8)$$

2.5.2 Relative index of refraction. In practice the absolute index of refraction is never directly measured. Instead the velocity in the material is compared to the velocity in air. From this comparison the relative index of refraction can be determined. The relative index of one material with respect to another is equal to the ratio of the absolute indices. For example, the relative index of a substance with respect to air is

$$(n_o)_{rel} = \frac{n_o}{n_{air}} = \frac{v_{vac}/v_o}{v_{vac}/v_{air}} = \frac{v_{air}}{v_o} .$$

* As indicated here the angle is expressed in radians. In the future, if an angle is given in radians, the word "radian" will be omitted; if the angle is given in degrees, the degree sign ($^\circ$) will be used.

Equation (1), which is the basic equation applying to a ray as it traverses a boundary, can be applied without knowing the absolute indices n_0 and n_1 . Only the relative index, n_1/n_0 , is needed. Hence all refraction problems involve only a ratio of two indices and it is not necessary to know the absolute index of optical materials. Therefore, unless specifically stated, the indices of refraction of optical materials relative to air are used, and it is these relative indices which are measured. (See Section 25.7.3). In problems involving vacuum the absolute index of refraction of air must be used to calculate the absolute index of the material.

2.5.3 Table of refractive indices. The index of refraction of several optical materials is shown in Table 2.2. Except for silicon, where the index applies to the infrared, the indices are for the visible spectrum. Detailed refractive index data on optical glasses are available in catalogs from glass manufacturers. (See paragraph 2.7.9). Materials other than glass are available and are used for optical elements. Refractive index and other data on these materials are discussed in Section 17. It should be noted that the indices given in Table 2.2, as well as in other references, are not only functions of wavelength, which is discussed in Section 2.6, but are also functions of temperature and pressure. The pressure dependence becomes of major importance in the case of gases; sometimes a particular gas at relatively high pressure is used to enclose part or all of an optical system.

Material	n
Vacuum	1.
Air	1.0003
Water	1.33
Fused quartz	1.46
Borosilicate crown glass	1.51
Ordinary crown glass	1.52
Canada balsam	1.53
Light flint	1.57
Dense barium crown	1.62
Extra dense flint	1.72
Silicon (in the infrared)	3.4

Table 2.2 - Refractive indices of various materials.

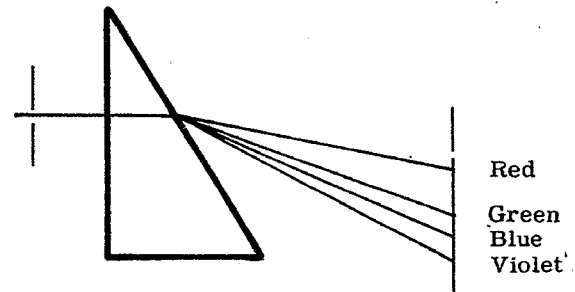


Figure 2.5 - Beam of white light passing through a dispersing prism.

2.6 DISPERSION OF LIGHT

2.6.1 General. It was shown by Newton that white light is not to be considered as a fundamental type, but is rather a composite mixture which can be separated into a range of colors, - that is a spectrum -, by passage through a prism as shown schematically in Figure 2.5. According to the wave theory of light, each color corresponds to a definite frequency of vibration or, when the light is traveling in a vacuum, to a definite wavelength (λ). The shorter waves correspond to the violet end of the spectrum; the longer, to the red. Further investigation has shown that the radiation spectrum extends to longer wavelengths beyond the red, the infrared (IR) region, and to shorter waves beyond the violet, the ultraviolet (UV) region.

2.6.2 Variation of index with wavelength.

2.6.2.1 Since index is inversely proportional to the velocity of light in a given medium, and since this velocity is not constant for all colors, the index is a function of the color of the light. The color may be specified either by stating the frequency or wavelength in vacuum; hence, the index may be considered a function of either frequency or wavelength. Which functional dependence is used depends on the specific problem involved. In geometrical optics, since spectrum lines are used to measure indices, and since these lines are indicated by wavelength (instead of frequency), it is customary to use the functional dependence on wavelength.

2.6.2.2 For a given refracting medium, the absolute refractive index takes on a different value for each wavelength. In all practical cases it is higher for short wavelengths, and lower for long ones. Thus in Figure 2.5 a ray of composite light is incident normally on the first surface. Since the angle of incidence on this surface is zero, the angle of refraction is also zero and the ray is undeviated. At the second surface, however, the light is deviated, the blue ray being bent more than the red. This unequal refraction

is called dispersion. The variation of index with wavelength, for most optical materials in the wavelength region where they are used, is such that the index decreases as the wavelength increases. The index varies approximately linearly with $1/\lambda^2$ where λ is the wavelength of the radiation.

2.6.3 Fraunhofer lines. In optical design work, the indices of refraction of the media to be used must be known in the wavelength region in which the device is to be used. (Methods of measuring index will be discussed in Section 24.6). Within the region, the choice of wavelengths at which measurements are made depends partly on convenience in measurement, and partly on custom. The section on glass characteristics applies generally to the visible region. Similar considerations apply to the ultraviolet and infrared regions, but the use of specific wavelengths for reference in those regions is not yet so well established. The range of visible wavelengths runs from about 0.380μ to about 0.740μ . (See Section 4.5). Within this region several reference wavelengths are used which, for historical reasons, are known as Fraunhofer lines, and are customarily denoted by letters assigned to them in a system originated by Joseph von Fraunhofer in his studies of the solar spectrum. In Table 2.3 are given the wavelengths of light of some of the Fraunhofer lines, and the elements from which the lines result. Also included are two additional lines, one in the near infrared, the other in the near ultraviolet, which are being used as standard wavelengths for index measurements.

Color of light	Line	Wavelength, Microns	Element
Infrared		1.0140	Hg
Red	A'	0.7665	K
Red	C	0.6563	H
Yellow	D	0.5893	Na
Yellow	d	0.5876	He
Green	e	0.5461	Hg
Light Blue	F	0.4861	H
Blue	g	0.4358	Hg
Dark Blue	G'	0.4340	H
Violet	h	0.4047	Hg
Ultraviolet		0.3650	Hg

Table 2.3-Fraunhofer and other standard lines.

2.7 CHARACTERISTICS OF OPTICAL GLASS

2.7.1 Reference indices. In designing chromatically corrected systems, it is necessary to make provision for the variation of the index of refraction with wavelength. This will be expanded in later sections, but for now it is important to be aware of the terms and quantities which are usually sufficient to describe the properties of an optical medium in the visible spectrum. In this and in the following paragraphs, reference should be made to specification MIL-G-174, Optical Glass, to become acquainted with approved standard requirements for the military. It is impractical to treat simply the infinite number of indices corresponding to all the wavelengths in white light. Common practice is to select a convenient wavelength near the middle of the eye's sensitive range, using one which can be easily and accurately reproduced. The refractive index of the material at this wavelength is then used as a basic reference both in design and in material designation. The material's refractive index for yellow light corresponding to the mean wavelength of the two sodium D lines is usually used in the United States and is designated n_D . European practice is to use n_d , the index corresponding to the yellow helium line. Similarly, the terms n_F and n_C are the indices of refraction for the F and C lines of hydrogen and provide reference indices in the blue and red regions.

2.7.2 Abbe constant. A commonly used expression for identifying chromatic properties is the Abbe constant, which is defined as

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

The symbol V , rather than the Greek ν is frequently used; however ν will be used in this text. The Abbe constant is named for its inventor, the German scientist Ernst Abbe. It is often called the nu value or the vee number. The numerator, $n_D - 1$, is called the refractivity for the sodium D lines.

2.7.3 Partial dispersion. The difference between any two indices for a given substance, corresponding to two different wavelengths, is called the partial dispersion. Hence $n_D - n_C$ is the partial dispersion for the D and C lines. The particular partial dispersion, $n_F - n_C$, is called the mean dispersion because it covers approximately the visual range of wavelengths. Use is sometimes made of a partial dispersion ratio, for example, $(n_D - n_C) / (n_F - n_C)$.

2.7.4 Glass type number. It has become common practice to identify a glass by the type number, which is a six-digit number. The first three digits of the type number are the first three rounded digits of the refractivity, $(n_D - 1)$, and the last three digits of the type number are the first three rounded digits of the ν -value of the glass. A glass with $n_D = 1.51250$ and $\nu = 60.5$ would have a type number of 513605.

2.7.5 Staining. In addition to the quantities involving refractive indices, which have been mentioned above, additional optical characteristics must be considered in optical design. One of these, surface staining, obviously affects the transmittance; such staining is accelerated by the presence of acidic atmospheres, for example caused by carbon dioxide or perspiration. Staining can be measured quantitatively by the time required to form a film one quarter of a wavelength thick when the sample is immersed in nitric acid under controlled conditions of concentration and temperature.

2.7.6 Dimming. A characteristic somewhat related to staining is surface dimming, which occurs when the polished sample is exposed to moist air. It can be measured quantitatively by exposing the sample to a 100% relative humidity atmosphere at a given temperature for a specified time, and classifying the appearance of the surface.

2.7.7 Bubbles. All glasses contain some bubbles, or inclusions, varying in size and number according to the glass type. A glass sample is classified according to the number of bubbles in a specified volume of material. If a bubble is less than 0.02 mm in diameter (or some other standard value), it is not counted as it is considered invisible.

2.7.8 Table of optical glass characteristics. Table 2.4 lists the quantities described above in identifying glass. The glass type number is given in both the extreme left and extreme right hand columns. The second column at the left gives the ν -number. There follow eleven columns giving the refractive index for the corresponding wavelengths. The next column gives the mean dispersion. There follow six columns listing two numbers for each glass type. The one in large type is a partial dispersion, the other a partial dispersion ratio. The specific gravity is listed in the next column; as the metal parts of optical instruments become more and more fabricated of light alloys, the glass weight becomes an important factor and must be considered in overall optical design. The next column gives the staining time in hours, and adjacent to it is listed the stain test class. In the next column is given the dimming test class number, running from 1 (not visibly dimmed) to 5 (dimming interfering with clear vision). The bubble code is given in the next to the last column; the code runs from 1 (few bubbles) to 4 (many bubbles). The letter P following a glass type indicates that this type is available in a form which makes it resistant to gamma rays and X-rays. The term fine annealed indicates that permanent strain on cooling has been virtually eliminated.

2.7.9 Availability of glass tables. Designers, or interested students should obtain from glass manufacturers the latest catalog information. Some suggested sources are: (1) in the United States, Bausch and Lomb, Rochester, New York; Corning Glass Works, Corning, New York; Eastman Kodak Co., Rochester, New York; Hayward Glass Co., Whittier, California; Pittsburgh Plate Glass Co., Pittsburgh, Pennsylvania; and (2) abroad, Chance-Pilkington Optical Works, St. Asaph, England; Tozai Boeki Kaisha, Ltd., No. 13, 4-Chome, Shiba-Tamuracho, Minatoku, Tokyo, Japan; Minex, P.W.O. Works, Jelenia Góra, Poland; Ohara Optical Glass Manufacturing Co., Sagami-hara, Kanagawa, Japan; Parra-Mantois, Le Vésinet, France; Schott Glass Works, Mainz, West Germany; Schott Glass Works, Jena, East Germany. Catalogs of Russian manufacturers are published by Gosudarstvennoe Isdatelstvo, Moscow, USSR. Additional U.S. companies and representatives of foreign companies are listed in the Optical Industry Directory (See page 1-5).

CHARACTERISTICS — OPTICAL GLASS

Indices Given are for "Fine Annealed" Glass

TYPE	V	10140	Potassium 765.5	D _K	Hydrogen 656.3	n _D	Helium 587.6	n _H	Hydrogen 486.1	n _F	Mercury 435.8	n _{H'}	Hydrogen 434.1	n _H	365.0	TYPE
BOROSILICATE CROWN																
498670	67.0	1.49316	1.49577	1.49808	1.49984	1.50320	1.50717	1.51048	1.50732	1.51655	1.52039	1.51675	1.52039	1.51048	*	498670
506596	59.6	1.50058	1.50347	1.50609	1.50811	1.51196	1.51656	1.52039	1.51675	1.52039	1.52424	1.52039	1.52424	1.52039	2	506596
511635	62.5	1.50578	1.50860	1.51107	1.51300	1.51685	1.52096	1.52454	1.52114	1.52454	1.52839	1.52454	1.52839	1.52454	1	511635
517645	64.5	1.51179	1.51461	1.51707	1.51899	1.52262	1.52690	1.53043	1.52709	1.53043	1.53428	1.53043	1.53428	1.53043	1	517645
517645P	64.5P	1.51179	1.51461	1.51707	1.51899	1.52262	1.52690	1.53043	1.52709	1.53043	1.53428	1.53043	1.53428	1.53043	1	517645P
CROWN																
513605	60.5	1.50708	1.50999	1.51258	1.51459	1.51846	1.52304	1.52685	1.52324	1.52685	1.53070	1.52685	1.53070	1.52685	1	513605
518596	58.6	1.51242	1.51544	1.51807	1.52015	1.52413	1.52886	1.53279	1.52906	1.53279	1.53664	1.53279	1.53664	1.53279	3	518596
523586	58.6	1.51729	1.52036	1.52307	1.52520	1.52929	1.53415	1.53819	1.53435	1.53819	1.54204	1.53819	1.54204	1.53819	4	523586
524595	58.5	1.51838	1.52140	1.52408	1.52618	1.53021	1.53500	1.53888	1.53520	1.53888	1.54269	1.53888	1.54269	1.53888	1	524595
LIGHT BARIUM CROWN																
541599	59.9	1.53522	1.53833	1.54109	1.54323	1.54736	1.55226	1.55633	1.55246	1.55633	1.56048	1.55633	1.56048	1.55633	1	541599
541599P	59.9P	1.53522	1.53833	1.54109	1.54323	1.54736	1.55226	1.55633	1.55246	1.55633	1.56048	1.55633	1.56048	1.55633	1	541599P
573568	56.8	1.56614	1.56954	1.57259	1.57498	1.57962	1.58514	1.58975	1.58538	1.58975	1.59439	1.58975	1.59439	1.58975	3	573568
573574	57.4	1.56619	1.56955	1.57259	1.57497	1.57953	1.58497	1.58951	1.58521	1.58951	1.59402	1.58951	1.59402	1.58951	3	573574
573574P	57.4P	1.56619	1.56955	1.57259	1.57497	1.57953	1.58497	1.58951	1.58521	1.58951	1.59402	1.58951	1.59402	1.58951	3	573574P
DENSE BARIUM CROWN																
588612	61.2	1.58184	1.58513	1.58811	1.59036	1.59474	1.59992	1.60424	1.60015	1.60424	1.60948	1.60424	1.60948	1.60424	1	588612
611572	57.2	1.60423	1.60785	1.61109	1.61364	1.61853	1.62438	1.62923	1.62462	1.62923	1.63417	1.62923	1.63417	1.62923	4	611572
611588	58.8	1.60439	1.60793	1.61109	1.61357	1.61832	1.62396	1.62867	1.62421	1.62867	1.63342	1.62867	1.63342	1.62867	4	611588
612595	59.5	1.60544	1.60886	1.61209	1.61455	1.61924	1.62484	1.62946	1.62508	1.62946	1.63404	1.62946	1.63404	1.62946	1	612595
617549	54.9	1.60995	1.61370	1.61710	1.61977	1.62493	1.63115	1.63541	1.63142	1.63541	1.64016	1.63541	1.64016	1.63541	4	617549
617551	55.1	1.60994	1.61371	1.61710	1.61976	1.62490	1.63104	1.63517	1.63130	1.63517	1.64031	1.63517	1.64031	1.63517	3	617551
620603	60.3	1.61342	1.61696	1.62011	1.62255	1.62724	1.63282	1.63748	1.63306	1.63748	1.64214	1.63748	1.64214	1.63748	3	620603
623569	58.9	1.61606	1.61978	1.62309	1.62571	1.63073	1.63675	1.64171	1.63701	1.64171	1.64671	1.64171	1.64671	1.64171	1	623569
638555	55.5	1.63074	1.63461	1.63810	1.64084	1.64611	1.65243	1.65772	1.65271	1.65772	1.66301	1.65772	1.66301	1.65772	3	638555
651558	55.8	1.64362	1.64757	1.65109	1.65389	1.65924	1.66563	1.67097	1.66591	1.67097	1.67627	1.67097	1.67627	1.67097	1	651558

Note 1: Available in Condenser quality only.
*Data not currently available.** Dissolves in HNO₃

Courtesy of BAUSCH & LOMB OPTICAL CO.

Table 2.4 - Excerpt from commercial glass catalog.

3 CONSIDERATIONS OF PHYSICAL OPTICS

3.1 INTRODUCTION

3.1.1 Diffraction nature of optical images.

3.1.1.1 The goal in designing a lens system on the basis of geometrical optics is to find a combination of lenses for which all rays in a specified cone of rays that diverges from an object point P are converged upon the corresponding image point P' such that the optical paths of all rays from P to P' are equal. Other requirements are added. For example, it may be required that points P and P' shall belong to a single object plane and a single image plane, respectively. Even when the design satisfies all these requirements to a high degree, the image P' of a self-luminous object point P is not a point but consists of a central bright spot surrounded by systematically distributed dark and bright fringes whose contour and width depend upon the contour and dimensions of the aperture of the lens. If, for example, the lens aperture is circular and if the self-luminous object point is located upon or near the optic axis, the image consists of a circular, central bright spot surrounded alternately by dark and bright rings. The central bright spot is called the Airy disk. Its diameter decreases as the diameter of the lens aperture is increased. The actual image of the object point is modified to such a degree by diffraction from the finite lens aperture that this image is appropriately called a diffraction image.

3.1.1.2 The diffractive nature of the image may not be so apparent with, for example, high-speed objectives in which compromises among the geometrical corrections and tolerable aberrations must be made. However, the image will generally exhibit effects due to diffraction, i. e., effects that cannot be explained from Snell's law of refraction or reflection alone. In any case, the image of a point will not be a point; an exact point by point similarity between object and image cannot be achieved. Resolution of details in the image of the object is restricted first by the degree of correction of the optical system and finally by the laws of diffraction, i. e., by the laws governing the bending of light rays from the paths consistent with Snell's law of refraction and reflection.

3.1.1.3 Whereas the action of most optical systems can be explained by the principles of geometrical optics, the action of other systems such as phase microscopy can be understood only as a proposition in diffraction. However, in any system, the ultimate resolving power and contrast in the fine-grained details of an image are determined by diffraction.

3.1.2 Diffraction and interference.

3.1.2.1 Broadly, diffraction is the phenomenon whereby waves are modified in direction, amplitude, and in phase by interaction with an object or obstacle. In its most general sense, diffraction includes the phenomena of refraction and reflection but these two phenomena are ordinarily considered apart from diffraction. However, when the dimensions of the object become comparable to the wavelength, the concepts of refraction and reflection become useless. With such small objects, even scattering becomes a direct aspect of diffraction.

3.1.2.2 Interference is the process by which two or more overlapping waves interact so as to re-enforce one another in some regions and to oppose one another in other regions. This process is essentially one of addition of the instantaneous amplitudes of the overlapping waves. It matters a great deal whether or not the overlapping waves are coherent. In case the added waves are incoherent, the time-averaged energy density is simply the sum of the time-average of the energy density associated with each wave, i. e., the resulting energy follows the law of superposition of energy. Conversely, it may be concluded that if the time-average of the energy densities follows the law of superposition of energy, the interfering waves are essentially incoherent. Interference includes the process by which a given wave is split or decomposed into two or more waves (often called component waves). These component waves are automatically coherent since they belong to the same wave-train. The action of interferometers can usually (but not always) be explained adequately by considering the sum of two or more waves.

3.1.2.3 Diffraction and interference are related processes, but diffraction is the more inclusive. In fact, diffraction effects can include interference effects as special cases. For example, in explaining the "interference fringes" produced with monochromatic light leaving two small pinholes that are illuminated coherently from a third pinhole, it is natural to regard the formation of the interference fringes as an interference effect, i. e., as a process of adding the two well defined spherical waves that emerge from the pair of pinholes. However, as the area of the pinholes is increased, the location of the origin of the spherical waves that leave different portions of the pinholes begins to matter. The process of summing the effects of the infinite many wavelets that leave the pinholes is now carried out most conveniently by means of integrals that characterize diffraction processes.

3.2 THE PHYSICAL NATURE OF LIGHT

3.2.1 The wave theory

3.2.1.1 Much evidence supports the view that light is propagated as electromagnetic waves whose wavelengths λ fall in the visible range from 0.38 to 0.76 microns. The transverse nature of electromagnetic waves is illustrated in Figure 3.1 in which E and H denote the electric and magnetic vectors, respectively. The electric

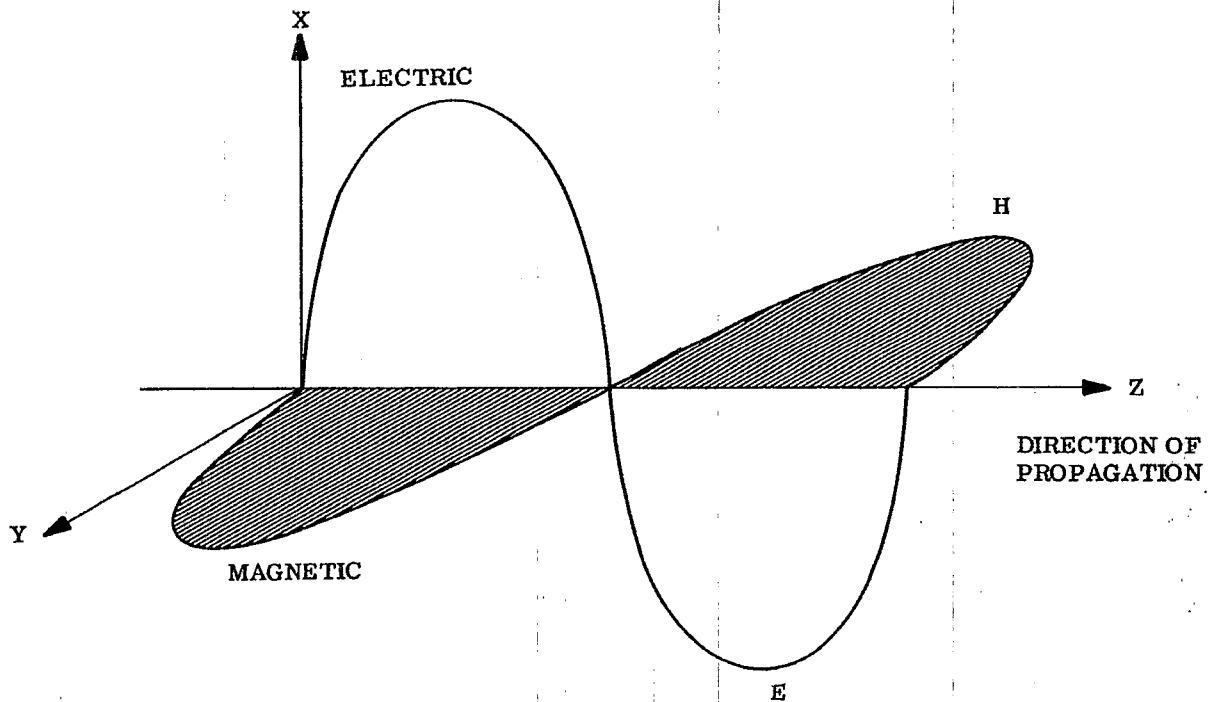


Figure 3. 1—The electromagnetic nature of a plane polarized light wave. The electric vector **E** and the magnetic vector **H** oscillate at right angles to the direction of propagation and at right angles to one another.

and magnetic vectors are ordinarily perpendicular to each other and to the direction of propagation. The electric vector describes an electric force field that will cause an electric charge to vibrate along the **E**-direction. Thus, the electric vector produces displacements of ions or electrons along the positive or negative **E**-direction, respectively. The vectors **E** and **H** are inseparable and are mutually dependent. For this reason it usually suffices to specify only the electric vector. The luminous flux can be computed whenever the radiant flux of the electromagnetic waves is known (as it is when the **E**-vector is specified).

3.2.1.2 The velocity of all electromagnetic waves in vacuum is a constant = $c = 299792.5$ kilometers per second. The velocity of monochromatic waves in non-vacuum media invariably depends upon the wavelength and is accordingly called the phase velocity to distinguish it from the group velocity of a group of monochromatic waves. The refractive index n of a medium is defined such that

$$n = \frac{\text{velocity in vacuum}}{\text{phase velocity in the medium}} \quad (1)$$

Let T denote the period of vibration of a monochromatic wave. Let $\nu = 1/T$ denote the frequency ν of vibration. Then if v denotes the phase velocity

$$v = \nu \lambda = \frac{c}{n} \quad (2)$$

As an electromagnetic wave moves from one medium into another, its frequency remains fixed. Hence its wavelength must change such that the wavelength λ in a medium of refractive index n varies according to the law

$$\lambda = \frac{c}{n\nu} = \frac{cT}{n} = \frac{\lambda_0}{n} \quad (3)$$

where $\lambda_0 = cT =$ wavelength in vacuum.

3.2.2 Plane-polarized light waves.

3.2.2.1 A plane-polarized light wave is one whose electric vector vibrates in a fixed plane (which we shall call the plane of polarization) in homogeneous media that do not rotate the plane of polarization. The wave illustrated in Figure 3. 1 is plane-polarized. If the direction of propagation is the Z-axis, the magnitude $E(z, t)$ of

the electric vector can be specified as the trigonometric function

$$E(z, t) = a \cos(knz + \phi - \omega t) \quad (4)$$

where

z = distance measured along Z
 t = time
 $k = 2\pi/\lambda$
 $\omega = 2\pi/T$
 λ = wavelength
 T = period for one complete vibration

ϕ = phase angle
 n = refractive index. It can be a function of z for variable media.
 a = amplitude of the wave. It is an exponential decreasing function of z for absorbing media.

The phase angle ϕ is needed for specifying the phase of one wave relative to another. If, for example,

$$E_1 = a_1 \cos(knz + \phi_1 - \omega t) \quad (5)$$

$$E_2 = a_2 \cos(knz + \phi_2 - \omega t) \quad (6)$$

the corresponding waves differ in phase by the amount $\phi_1 - \phi_2$ at like values of t and z .

3.2.2.2 The state of vibration or polarization is the same for all points that belong to a wavefront. On each wavefront

$$knz + \phi - \omega t = \text{constant} = w \quad (7)$$

where w is different for each wavefront. The wavefront moves so as to satisfy Equation (7). By differentiating the members of Equation (7) with respect to the time t , one finds that

$$\frac{dz}{dt} = v = \frac{\omega}{kn} = \frac{1}{n} \frac{\lambda}{T} = \frac{c}{n^2}; \quad \frac{\lambda}{T} = \frac{c}{n} \therefore \frac{1}{n} \cdot \frac{c}{n} = \frac{c}{n^2} \quad (8)$$

3.2.2.3 The wavefronts of the plane-polarized wave described by Equation (4) are perpendicular to the Z -axis, the direction of propagation. If the plane-polarized plane wave is propagated along an arbitrary direction OP , Figure 3.2, the magnitude E of the electric vector assumes the form

$$E = a \cos \left[kn(px + qy + rz) + \phi - \omega t \right] \quad (9)$$

where p , q and r are the direction cosines of OP with respect to X , Y , and Z , respectively. Thus,

$$p^2 + q^2 + r^2 = 1. \quad (10)$$

Equation (9) reduces to Equation (4) when the direction of propagation OP is the Z -direction only, for then $p = q = 0$ and $r = 1$. It is important to observe that the wave motion of Equations (4) and (9) is of the form

$$E = a \cos(\Phi - \omega t) \quad (11)$$

where

$$\Phi = kn(px + qy + rz) + \phi \quad (12)$$

with p , q , and r defined as the direction cosines of the direction of propagation of the plane-polarized, plane wave. The electric vector vibrates in the wavefront.

3.2.3 Energy in a single wave. The instantaneous energy, W_i (whether energy flux or energy density) in the wave is proportional to E^2 , where E denotes the instantaneous magnitude of the electric vector. We take the factor of proportionality as unity and write from Equation (11)

$$W_i = E^2 = a^2 \cos^2(\Phi - \omega t). \quad (13)$$

The oscillations of light waves are so rapid that the eye or other known detectors are unable to follow the in-

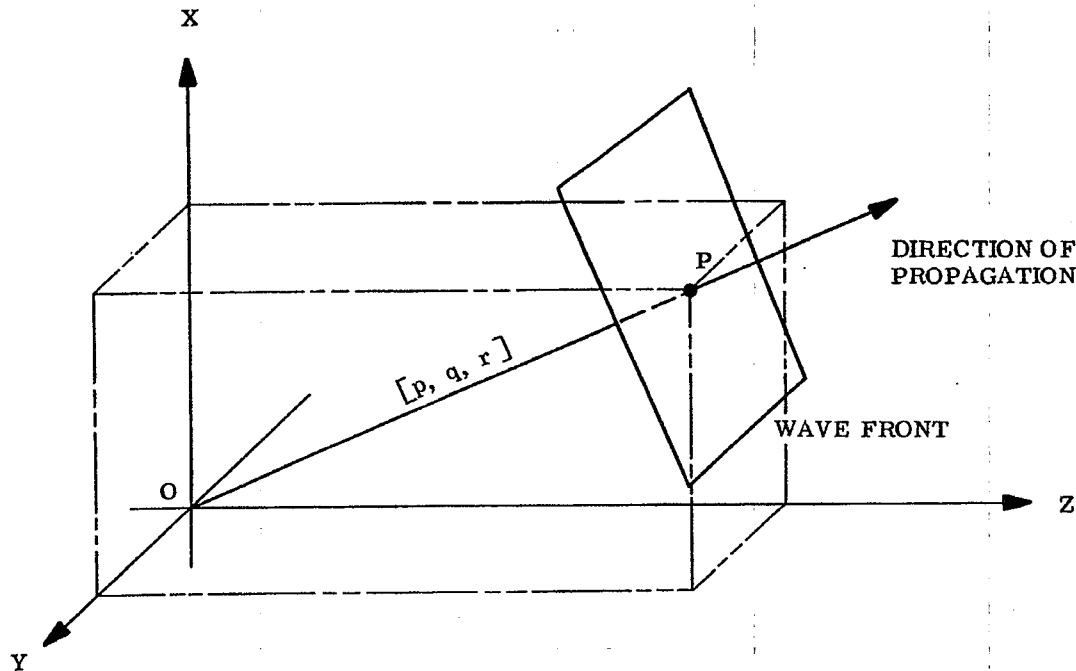


Figure 3.2—Notation with respect to the propagation of a plane wave.

stantaneous values. Rather, the time average W of W_i is detected and measured. It suffices to average over one period T of oscillation. Thus,

$$\begin{aligned} W &= \frac{1}{T} \int_0^T a^2 \cos^2 (\Phi - \omega t) dt \\ &= \frac{a^2}{2T} \int_0^T [1 + \cos 2 (\Phi - \omega t)] dt. \end{aligned} \quad (14)$$

Since $\omega = 2\pi/T$, it follows almost directly that

$$\int_0^T \cos 2 (\Phi - \omega t) dt = 0. \quad (15)$$

Hence,

$$W = a^2/2, \quad (16)$$

i. e. the time-averaged energy density or energy flux in a single wave is proportional to the square of its amplitude. W is independent of, for example, the phase angle Φ of the single plane wave.

3.3 INTERFERENCE BETWEEN WAVES

3.3.1 Collinear, coherent waves.

3.3.1.1 Two waves will be called collinear when they are propagated in the same direction. We consider the interference of two plane-polarized,* plane waves that are propagated in the same direction with a constant phase

* The electric vectors of these two plane polarized waves are assumed parallel, i. e., are assumed to vibrate in the same fixed plane.

difference δ . The magnitudes E_1 and E_2 of the electric vectors of two unlike plane waves assume from Equation (11) the form

$$E_1 = a_1 \cos (\Phi_1 - \omega t) ; \quad E_2 = a_2 (\Phi_2 - \omega t) . \quad (17)$$

From Equation (12)

$$\Phi_1 - \Phi_2 = \delta , \quad (18)$$

the phase difference between the two waves.

3.3.1.2 Let E denote the magnitude of the electric vector formed by the sum of E_1 and E_2 , i.e., formed by the interference of the two waves. Then,

$$E = a_1 \cos (\Phi_1 - \omega t) + a_2 \cos (\Phi_2 - \omega t) . \quad (19)$$

Let W be the time-averaged energy density formed by the two interfering waves. As in paragraph 3.2.3,

$$\begin{aligned} W &= \frac{1}{T} \int_0^T E^2 dt . \\ &= \frac{a_1^2}{T} \int_0^T \cos^2 (\Phi_1 - \omega t) dt + \frac{a_2^2}{T} \int_0^T \cos^2 (\Phi_2 - \omega t) dt \\ &\quad + \frac{2 a_1 a_2}{T} \int_0^T \cos (\Phi_1 - \omega t) \cos (\Phi_2 - \omega t) dt \\ &= \frac{a_1^2}{2} + \frac{a_2^2}{2} + 2 a_1 a_2 I \end{aligned} \quad (20)$$

where

$$I = \frac{1}{T} \int_0^T \cos (\Phi_1 - \omega t) \cos (\Phi_2 - \omega t) dt . \quad (21)$$

But

$$\cos (\Phi_1 - \omega t) \cos (\Phi_2 - \omega t) = \frac{1}{2} \left[\cos (\Phi_1 + \Phi_2 - 2\omega t) + \cos (\Phi_1 - \Phi_2) \right] . \quad (22)$$

As in Equation (15),

$$\int_0^T \cos (\Phi_1 + \Phi_2 - 2\omega t) dt = 0 .$$

Hence,

$$I = \frac{\cos (\Phi_1 - \Phi_2)}{2T} \int_0^T dt = \frac{\cos (\Phi_1 - \Phi_2)}{2} = \frac{\cos \delta}{2} \quad (23)$$

Finally, from Equations (23) and (20) we find that the time-averaged density, W , produced by the interference of two, plane-polarized, collinear, plane waves having amplitudes a_1 and a_2 and phase difference $(\Phi_1 - \Phi_2)$ is

$$W = \frac{1}{2} \left[a_1^2 + 2 a_1 a_2 \cos \delta + a_2^2 \right] . \quad (24)$$

3.3.1.3 For constructive interference, the phase difference $\Phi_1 - \Phi_2 = \delta$ between the two waves is 0, 2π , 4π , etc., so that

$$W = \frac{1}{2} (a_1 + a_2)^2 . \quad (25)$$

For destructive interference, $\delta = m\pi$ where m is an odd integer. Correspondingly,

$$W = \frac{1}{2} (a_1 - a_2)^2 . \quad (26)$$

It should be noted from Equation (26) that $W = 0$ when the two waves have equal amplitudes and are out of phase. Thus, two plane waves that are propagated in the same direction can cancel one another everywhere, or they can re-enforce one another everywhere provided that their phase difference δ is a suitably chosen constant. The

time-averaged energy density of the resultant wave is not merely the sum of the time-averaged energy densities of the two separate waves except in the special cases $\cos \delta = 0$. See Equations (25) and (16). The waves are coherent when δ is constant.

3.3.2 Collinear, incoherent waves.

3.3.2.1 One should expect that when light or any other radiation from two independent sources overlap, the resulting energy density is simply the sum of the overlapping energy densities, i. e., the law of superposition of energy should apply. The interfering waves ought to be incoherent. The following somewhat oversimplified argument brings to bear the essential physics underlying the interference of incoherent waves.

3.3.2.2 The time-averaged energy density, produced by two interfering waves that have amplitudes a_1 and a_2 and the phase difference δ , is given by Equation (24). We shall avoid considering the sum of a large number of waves having randomly distributed phase differences δ (as will occur with independent sources) by supposing that in a short interval of time the phase differences δ between the two interfering waves are distributed with equal probability in the interval $0 \leq \delta \leq 2\pi$. Then from Equation (24)

$$W = \frac{1}{2} \left[a_1^2 + 2 a_1 a_2 \overline{\cos \delta} + a_2^2 \right] \quad (27)$$

where $\overline{\cos \delta}$ is the average value of $\cos \delta$ when all values of δ are equally probable in the interval $0 \leq \delta \leq 2\pi$. One can show that

$$\overline{\cos \delta} = 0 \quad (28)$$

In this manner we conclude that

$$W = \frac{1}{2} (a_1^2 + a_2^2) \quad (29)$$

so that the interference between incoherent waves is of that degenerate variety to which the law of superposition of energy applies.

3.3.3. Non-collinear, coherent waves.

3.3.3.1 The theory of paragraph 3.3 is almost but not quite adequate for explaining and interpreting the interference fringes that appear in Twyman Green and other double-beam interferometers; for in these interferometers the mirrors are usually tilted so that the two interfering waves are not propagated in the same direction. It is well known that a series of straight and parallel interference fringes are seen when the interfering waves are not collinear and when the reflecting surfaces are optical flats.

3.3.3.2 We may suppose without essential loss of generality that one wave is propagated along the direction OP that makes any angle θ with Z but is oriented so that the direction cosine $q = 0$. The two interfering waves are described by Equation (17); but $\Phi_1 - \Phi_2$ will not be given by Equation (18). Instead,

$$\Phi_1 = knz + \phi_1 \quad (30)$$

$$\Phi_2 = kn (x \sin \theta + z \cos \theta) + \phi_2$$

so that

$$\Phi_1 - \Phi_2 = \phi_1 - \phi_2 - knx \sin \theta + knz (1 - \cos \theta) \quad (31)$$

From Equations (20) and (23) the time-averaged energy density formed by the two interfering, coherent waves is

$$W = \frac{1}{2} \left[a_1^2 + 2 a_1 a_2 \cos (\Phi_1 - \Phi_2) + a_2^2 \right] \quad (32)$$

Substituting $\Phi_1 - \Phi_2$ from Equation (31) and setting $\phi_1 - \phi_2 = \delta$, the fixed phase difference between the two waves, one obtains

$$2W = a_1^2 + a_2^2 + 2 a_1 a_2 \cos \left[\delta - knx \sin \theta + knz (1 - \cos \theta) \right] \quad (33)$$

in which θ is the angle indicated in Figure 3.3, $k = 2\pi/\lambda$ and n is the refractive index of the medium. δ is the phase difference between the two interfering waves having amplitude a_1 and a_2 at the point $x = 0$, $z = 0$.

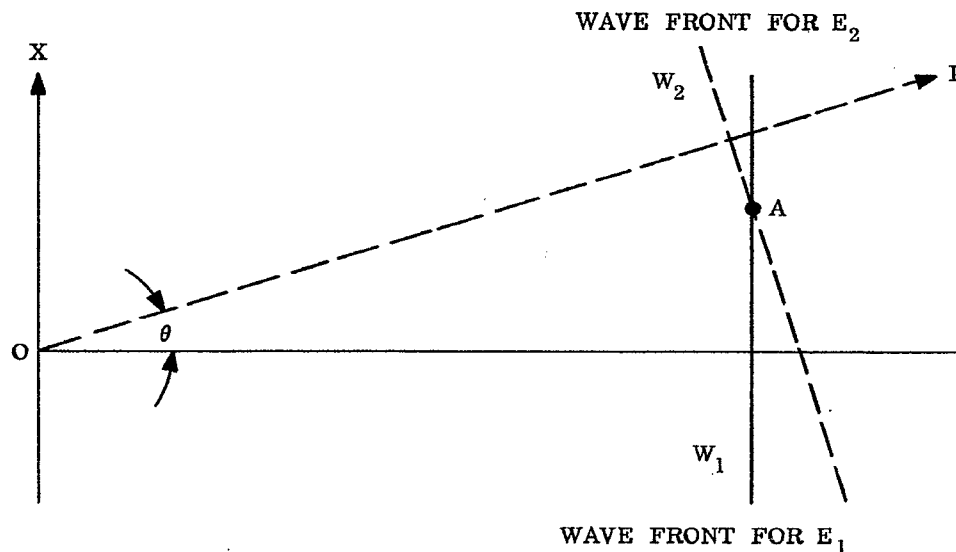


Figure 3.3— Interference between two plane wavefronts W_1 and W_2 that are propagated along different directions.

3.3.3.3 In double beam interferometry, the angle θ is usually so small that one can set $\sin \theta = \theta$ and $1 - \cos \theta = \theta^2 / 2$. If, then, one makes observations in planes z near $z = 0$, Figure 3.3, the term containing z in Equation (30) can be neglected. The approximation thus obtained is the usual interference formula

$$2W = a_1^2 + a_2^2 + 2 a_1 a_2 \cos (\delta - 2\pi n x \theta / \lambda) . \quad (34)$$

The fringes are repeated whenever x is increased by an amount Δx such that

$$kn\Delta x \sin \theta = 2\pi .$$

The fringe width h is therefore given by

$$h = \Delta x = \frac{\lambda}{n \sin \theta} . \quad (35)$$

The greater fringe widths belong to the longer wavelengths.

3.3.3.4 In case the fringes are photographed with a camera that images a plane into a plane, the interference fringes will be straight. Suppose, however, that the camera has curvature of field. In this case a plane $z = \text{constant}$ will not be focused upon the photographic plate. Consequently, one has to expect from Equation (33) that the photographed fringes will be curved and that the distortion of the fringes should increase as θ and z are increased.

4 VISUAL OPTICS

4.1 INTRODUCTION

4.1.1 Characteristics of the human eye. The design of an efficient optical instrument must include consideration for the use of the instrument. When the human eye is to be the translating instrument, the instrument must be designed for proper seeing. This section will call attention to some of the advantages and limitations of the human eye and seeing that are important for instrument design. The human eye is sensitive to radiant energy from 380 to about 740 $m\mu$ in wavelength. The limits of visibility for young eyes are about 313 - 900 $m\mu$, but for practical purposes the narrower range is adequate and representative for average eyes. Light is defined as radiant energy evaluated according to its capacity to produce visual sensation. A few quanta can stimulate the retina and be seen as light. To see an object, light of suitable quality (color) and intensity from the object must form an image on the retina of adequate size, contrast, and duration for the retina to transform the light energy into nerve energy, and the nerve impulses must be conducted to the brain and integrated into consciousness. Age, glare, state of adaptation and visual acuity will modify vision.

4.1.2 Seeing. Seeing is a learned ability and training can improve the individuals seeing to limits set by the eye and nervous system. Seeing is a perceptual process that is affected by and incorporates other sensations, emotions, association mechanisms simultaneously active with vision, education, and past experience. It varies with the condition of the individual and the entities must be statistical probabilities of seeing rather than absolute values.

4.1.3 Loss of vision. The eye and vision are disturbed by many conditions and diseases. Emmetropia refers to an average normal eye, ametropia indicates a defective eye and amblyopia an eye with little or no vision that appears normal. Additional defects of the eye are covered in paragraph 4.3.3.

4.2 ANATOMY OF THE EYE

4.2.1 Physcial structure. The human eye, as illustrated in Figure 4.1, is a nearly spherical organ held in shape by a tough, outer, whitish-sclerotic coat and the pressure of its viscous content. The cornea, the transparent front part of the sclera, protrudes slightly as it has a greater curvature. Inside the sclera is the choroid containing the blood vessels, the opaque pigment and the ciliary process. The ciliary process includes the iris and the muscles which focus the lens of the eye. The pupil is the opening in the center of the iris. The retina covers the inside of the choroid to the ora serrata near the ciliary process. The space between the cornea and the iris is called the anterior chamber and between the iris and the lens is a posterior chamber. Both are filled with the aqueous humor. The space back of the lens and ciliary process is filled with the vitreous humor. The lens is attached to the ciliary muscle by many fibers or suspensory ligaments. Except for the opening in the iris the pigmentation of the sclera and iris normally makes the eye light tight. A lack of eye pigmentation is called albinism and vision is impaired by glare from light leakage onto the retina.

4.2.2 Intraocular pressure. The internal pressure of the eye is maintained quite constant by a balance of the formation of the aqueous humor at the back part of the ciliary process, from which it passes out through the pupil into the anterior chamber, and drains through the canal of Schlemm.

4.2.3 Metabolism. The transparent media, cornea, lens and vitreous do not have blood vessels and receive their nourishment from the fluids surrounding them. The transparency of the cornea depends on its relative hydration. The front part of the retina contains blood vessels which furnish nourishment to it and to the adjacent vitreous.

4.2.3.1 The retina is one region of the body where it is possible to see (with the ophthalmoscope) the condition of the blood vascular system and recognize changes from many systemic diseases. The focussing ability of the eye is altered by a change in the blood sugar concentration from inadequately controlled diabetes. Glaucoma is a disease characterized by an increase in the pressure within the eye ball and unless arrested promptly will lead to mechanical damage and loss of sight.

4.2.4 Development. The eye is developed early and is fairly well formed by six weeks after conception. An outgrowth from the front of the brain becomes the optic nerve and the retina of the eye. When this cup-shaped formation nearly reaches the skin of the embryo, that part of the skin sinks below the surface and becomes modified to form the lens of the eye. The skin closes over to form the cornea and the sclera. The choroid and the ciliary process form between the sclera and the retina. Like the brain, the eye is relatively large at birth al-

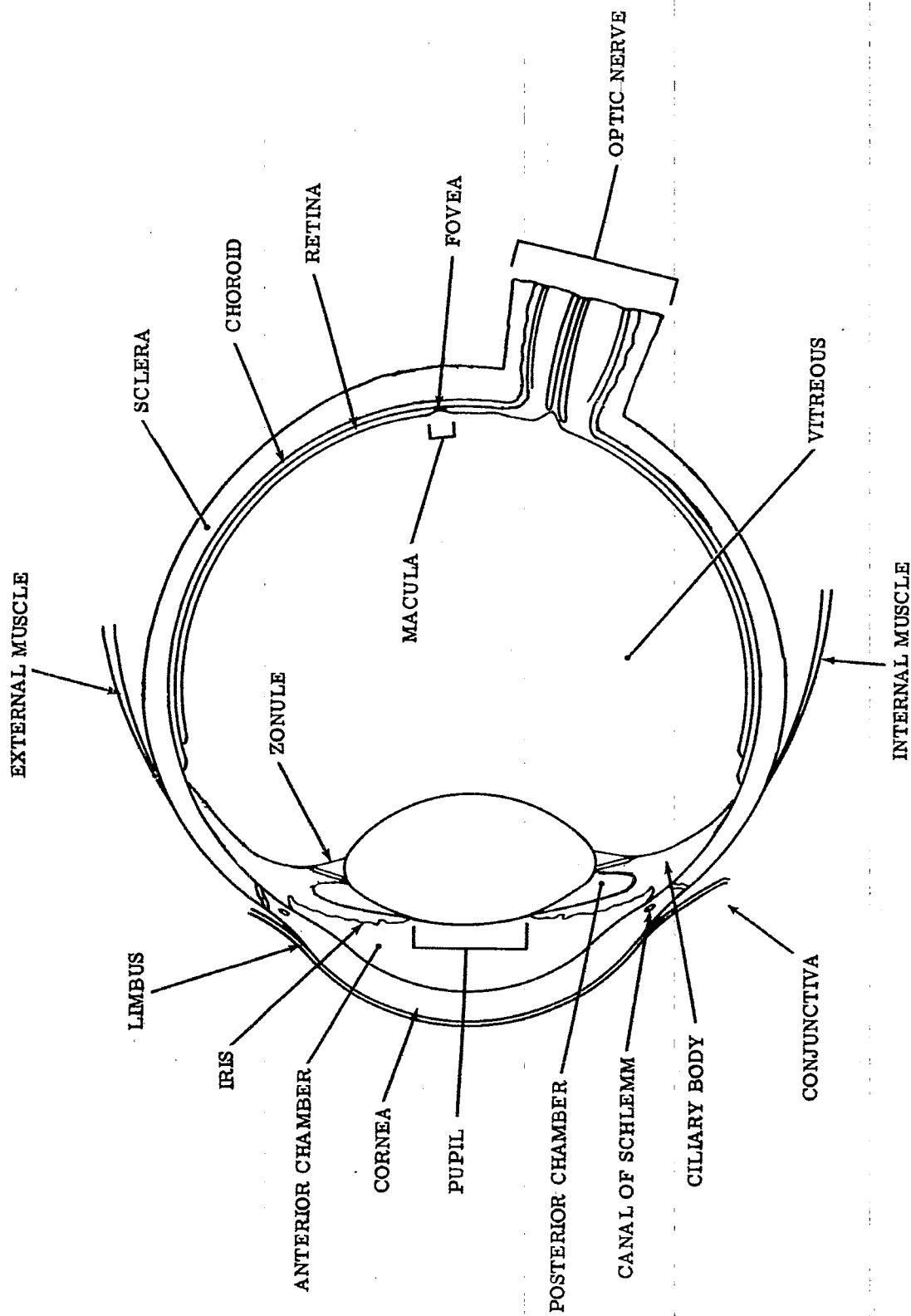


Figure 4.1. - Horizontal section of the right eye.

though vision then is imperfect and improves for several years. Color vision may not reach its greatest sensitivity until in the late teens. The various parts of the eye do not grow at the same rate and the eye and the body do not grow at the same rate. It is remarkable that the regulatory mechanisms tend to balance these different rates of growth to produce the emmetropic eye.

4.2.4.1 The muscles which control the eye will be described later. Briefly however, muscular action is usually a balance between opposing pairs of muscles which contain many contractile units and the resulting movement usually shows the action of the units in a stepwise progression, and fine oscillations when in equilibrium.

4.3 OPTICAL CONSTANTS OF THE EYE

4.3.1 Use of the "standard eye." As one would expect there are no universal dimensions for an eye, one finds instead, considerable variation in all dimensions. A good image formed on the retina may be the result of each part of the eye being perfect in form and refractive index, or the shapes and indices of the parts may have compensated for each others defects. Complete testing of each observer's eye would be time consuming and require special equipment. Instead a "standard" or typical eye is established and used as a standard observer for computational problems. Individual eyes can be examined to discover whether or not they correspond to the standard. There are several systems for "reduced" eyes, and a commonly used set of optical and mechanical characteristics for a typical eye is illustrated in Figure 4.2. Reduced is used here in the sense of an optically equivalent system.

4.3.2 Aberrations. Like other optical systems the eye is subject to the usual aberrations. The coordination of the focussing system and the retinal structure with sunlight over many years evolution has minimized some of the problems. Distortion and field curvature rarely bother in ordinary seeing, and chromatic aberration does not disturb vision. With the small pupils, of 3-4mm and average daylight, spherical aberration is minimal, although in dim light with large pupils it lessens vision.

4.3.3 Corrective lenses. The chief defects of the eye are myopia, hyperopia or hypermetropia, astigmatism, presbyopia and aniseikonia. The hyperopic eye focuses the image of a distant object behind the retina, and the myopic eye in front of the retina. In old age the focussing ability of the lens declines and this condition is termed presbyopia. Astigmatism results from asymmetry of the cornea. Aniseikonia will be discussed in paragraph 4.7 and aphakia will be discussed in paragraph 4.4.

4.3.3.1 Far sightedness, or hyperopia, can be due to the axial length of the eye being too short, or the focussing mechanism too weak, and is corrected by placing in front of the eye a plus lens of proper strength to replace the image on the retina. In near-sightedness, or myopia, the image is formed in the vitreous because the eye is too long, or the focussing mechanism is too strong, and the defect is corrected with a minus spectacle lens. Astigmatism due to irregular curvature of the cornea is corrected by a cylindrical spectacle lens.

4.3.3.2 Spectacles are usually fitted so that the back surface (vertex) of the lens is about 14 millimeters in front of the cornea although minus lenses for myopia may be set closer at 9 to 11 millimeters. Changing the position alters the effective power of the lens. Eyeglasses may be tilted slightly downward 4° to 12° for reading.

4.3.3.3 People with astigmatic corrections must wear their glasses for comfortable vision over long periods when using optical instruments. In recent years optical designers have made oculars with the eye point far enough from the lens so that the individual can see the whole field while wearing spectacles. The distance from the front of the spectacle lens to the cornea can vary from around 17 millimeters to 11 millimeters. If a substitute lens is mounted on the optical instrument to take the place of a spectacle lens, its power must be changed from that of the prescription when the substitute lens will be at a different position from the cornea than the spectacle lens. A substitute lens with cylindrical power must be mounted in proper orientation to the axis of the cylinder so it cannot rotate from the correct position.

4.3.3.4 People with only near or far sightedness (no astigmatism) usually remove their glasses when using optical instruments and refocus the instrument to correct for their defect. Therefore, focussing eyepieces should have sufficient range for the people intended to use them. A range of ± 1 diopter will include about 70 percent; ± 2 diopters will include about 85 percent; and ± 4 diopters about 98 percent of spectacle prescriptions.

4.3.3.5 Critical seeing can take place only when the image is located on the fovea at the center of the macula of the retina, as illustrated in Figures 4.1 and 4.2. This establishes a visual axis which is some $5-7^\circ$ from the optical axis of the eye. The retina is blind over the area of the optic disc where the nerve fibers enter the eye to distribute over the retina, and this blind spot subtends some 7° vertically and 5° horizontally.

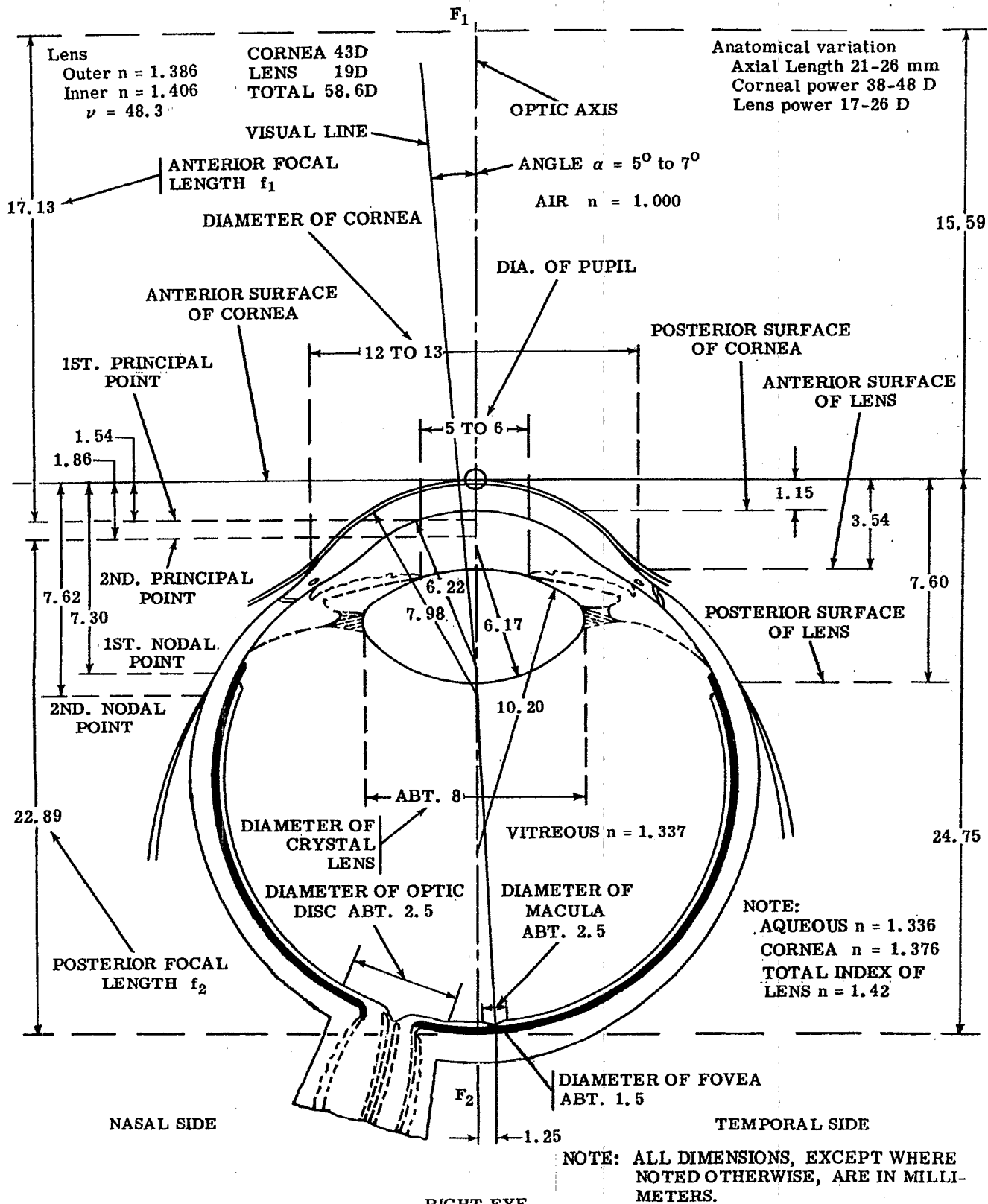


Figure 4.2 - Optical constants for a "standard eye."

4.4 IMAGE FORMATION AND THE RETINA

4.4.1 Cornea. The cornea is the first refracting surface for light entering the eye and is responsible for about 43 of a total of 58 diopters power of the eye. Normally the cornea is transparent and the refracting power is due to the curvature and refractive index difference between it and air on one side, and the aqueous humor on the other. The cornea in size averages 12 millimeters horizontally and 11 millimeters vertically.

4.4.1.1 A change in the hydration of the cornea can affect the light passing through it either by distortion or decreased transparency. The decrease caused by fluids and some early contact lenses, or from changes in old age, scatters light and haloes appear around light sources or small bright objects. Haloes from age changes are rarely reversible.

4.4.1.2 The two surfaces of the cornea usually are of similar curvature and have no lens effect on the entering light. Any deformity of the curvature of the cornea (astigmatism) distorts the image. Such changes are measured with a keratometer (ophthalmometer) and corrected by adding a corresponding cylinder of opposite sign into the spectacle lens for the eye. An extreme elongation of the center of the cornea (keratoconus) can be corrected by contact lenses. Astigmatism has some relation to the tension of the eye muscles and may change slowly from a vertical meridian to a horizontal meridian of greatest curvature during later life. There may be some residual astigmatism as well as that from the corneal surface.

4.4.1.3 Vision specialists sometimes refer to astigmatism with the rule (stronger power vertical) and against the rule (meridian of greatest curvature horizontal) based on the direction of movement of light reflected from the eye during skiascopic refraction.

4.4.1.4 Haidinger's Brushes are seen on looking at the blue sky (polarized), or at a uniform source of polarized blue light, as a diffuse cross. Some observers believe this phenomenon is due to the birefringence of the cornea. Other observers hold that it is due to neural structure or pigment arrangement in the retina. Attempts to use the Brushes for differential diagnosis of eye conditions has been unsuccessful so far.

4.4.2 Pupil. The pupil is the opening in the center of the iris as illustrated in Figures 4.1 and 4.2. In dim illumination the pupil opens to about 8 millimeters diameter in young eyes, and closes to about 2 millimeters diameter in intensely bright light. Under average conditions the pupil has a diameter of 3.5 to 4 millimeters. Resolution of the eye is decreased when the pupils are much larger or smaller than 3 to 4 millimeters. With ageing, the pupil remains smaller, and in extreme old age may not be more than 2 to 3 millimeters. The pupil is a stop, or diaphragm, in the dioptric system of the eye that affects image formation, illumination of the retina and the aberrations of the system. With small pupils (2 millimeters or less) diffraction becomes important.

4.4.2.1 The iris is composed of radial and circular muscle fibers and the size of the pupil is a resultant of these antagonistic muscles. Consequently the pupil shows continuous fine fluctuations in size, as well as opening and closing with changed luminance. The iris is not under voluntary control. Convergence of the eyes to a closer point in space also closes the pupil and this increases the depth of field.

4.4.2.2 Stimulation of the cornea, conjunctiva or eyelids, causes a slight dilation, followed by contraction of the pupil. Strong sensory stimulation, fear, and pain cause dilation via the psycho-sensory reflex. Many drugs effect the size of the pupil and some are used in the medical treatment of the eye to dilate (mydriasis) or contract (myosis) the pupil. Normally, both pupils respond together from the stimulation of either eye although the sizes may not be exactly the same. A marked difference in sizes indicates disease.

4.4.2.3 The pupil can decrease from 8 to 3 millimeters in 4 to 5 seconds. Dilation of the pupil from 3 to 6 millimeters takes 5 to 10 seconds and maximum dilation may take 5 to 10 minutes. Contraction at 5.5 to 7 millimeters per second and dilation at 3.0 to 4.5 millimeters per second is reported.

4.4.2.4 In designing optical instruments for visual use it should be kept in mind that the usable part of the exit pupil is no larger than the pupil of the eye. In order to decrease the precision with which the eye must be placed at the exit pupil in viewing, it is sometimes advisable to design the instrument so that the diameter of the exit pupil is considerably larger than any possible diameter of the pupil of the eye. In this case the portion of the exit pupil transmitting light to the observer's retina is limited to the size of the eye pupil, and the usable diameter of the entrance pupil (for axial bundles of rays) is equal to the diameter of the eye pupil multiplied by the magnification. However, if the exit pupil is smaller than the pupil of the eye the light entering the eye is limited by the exit pupil, and in instruments requiring maximum illumination on the retina every attempt should be made to provide an exit pupil diameter as large as the largest possible diameter of the pupil of the eye. Average pupil size for age and luminance are shown in Figure 4.3.

4.4.3 Lens. The lens of the eye changes curvature to focus light onto the retina. The lens is a transparent elastic body with an outer capsule, a less dense cortex, and a denser inside core. The lens is held in position

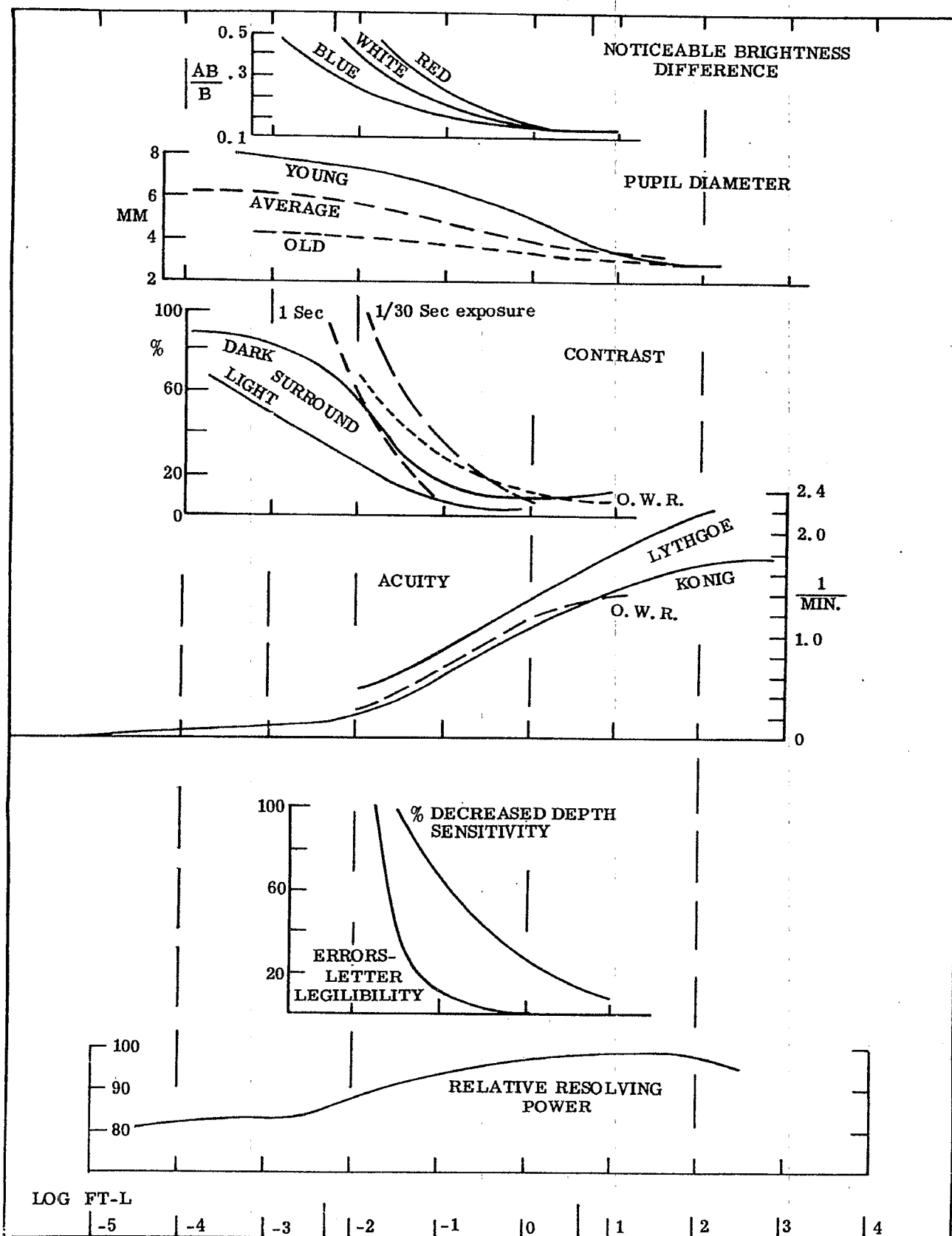


Figure 4.3. - Variation of some attributes of vision with Luminance.

by the suspensory ligaments, as shown in Figure 4.1. The ciliary process has circular, radial, and oblique muscle fibers which contract to pull on the fibers of the zonule and flatten the lens; or relax to lessen the tension and let the lens bulge to a more spherical form. Continuous fluctuations from muscle action take place producing amplitudes of 0.1 diopters focal change with a frequency of 4 to 8 cycles per second and smaller frequencies of 2 and 0.3 cycles per second. The lens has a total refracting power of some 19 diopters and the amplitude of accommodation of the lens varies from some 15 diopters in children to about 0.5 diopter in old age. The depth of field is about 0.5 diopter. However, to focus the eye from near to far requires 0.7 to 0.8 second, far to near 0.4 to 0.5 second, and near to far and back to near 1.15 to 1.25 seconds. When vision is less than 20/20, when exophoria exceeds 8 prism diopters at 33 centimeters, or when myopia, hyperphoria, or astigmatism are present, the time required to focus the eye will increase from that mentioned above.

4.4.4 Accommodation. The curvature of the front and back surfaces of the lens are different and the front surface is said to be hyperbolic in young people. The focussing of the lens is controlled by the sympathetic nervous system and cannot be altered voluntarily. There have been two theories advanced regarding accommodation. Helmholtz thought that relaxation of the tension on the suspensory ligaments permitted the elastic lens substance, which had been deformed in the unaccommodated state, to return to its more convex form. E. F. Fincham's experiments indicate that such relaxation allows a highly elastic capsule to deform the lens substance from its unaccommodated state to the greater convexity required. The variations in thickness in different parts of the capsule favors the latter theory.

4.4.4.1 When the eye sees only an empty field lacking detail the lens tends to focus, not at the 20 foot "infinity" of the vision specialists, but at about 1 meter. This near-sightedness is called empty field myopia for a bright empty field, and night myopia when the empty field is due to darkness. In the latter case the change in spherical aberration from the dilated pupil and the Purkinje shift also contribute to the total myopia. Changes in the curvature of the lens can be measured objectively from changes in the Purkinje-Samson images reflected from the surfaces of the lens, or with an optometer from changes in the retinal image, using either light or invisible infrared radiation.

4.4.4.2 At about forty years of age the focussing mechanism begins to gradually fail (presbyopia) and additional plus lens correction becomes necessary to see details at the usual reading distance. The lens also tends to become yellowish, blues are seen less well in old age, and less light gets to the retina. In some eyes the lens becomes opaque (cataract) and must be removed to restore vision. The eye lacking a lens is said to be aphakic and the spectacle lens correction must be increased to substitute for the lens. As the spectacle lens has a fixed focus the aphakic eye will be corrected only at one distance. When one eye is aphakic and the other is not, the difference in the size of the images on the retina precludes binocular vision.

4.4.4.3 Optical instruments with focusable eyepieces must be designed to have an adequate adjustment in power to permit older people to use them, and to provide at least -2 diopters when designed for night use.

4.4.4.4 The vitreous humor is a transparent gel of slightly greater refractive index than water, that fills the space between the lens and ciliary process and the retina. Sometimes particles of tissue (muscae volitantes) tend to hang or float in the vitreous and are seen when one is observing through optical instruments. These may be fragments left over as the vitreous formed, or that have broken away during life. Nothing can be done to remove these fragments and they should be ignored. In some diseases, parts of the vitreous become opaque and vision is lost to a corresponding extent.

4.4.4.5 The retina, covering most of the area behind the ciliary process, translates light energy into nervous energy and contains the first coordinating nerve cells in the visual system. The front part facing the lens is composed of blood vessels, nerve cells and fibers and connective tissues, and at the back of the retina are the light sensitive rod and cone cells and protective pigment layer. The entrance of the optic nerve forms a disc (a blind spot where there are no light sensitive cells) and the visual angles subtended by this disc are about 7° and 5° as illustrated in Figure 4.2. The disc is about 3.5 millimeters (15.5° to center) on the nasal side of the optical pole of the eye and 1.5 degrees below the horizontal meridian of the eye.

4.4.4.6 The retina thins at the visual axis, some 5° temporal to the optical pole, as there are no blood vessels or nerve fibers over the fovea. The macula subtends about 12° and is 2.5 to 3 millimeters in diameter. The fovea includes about 1.5 millimeters of the center of the macula, or about 5° of subtended arc and is the most sensitive part of the retina. Some anatomists recognize an area of about 0.35 millimeters in the center of the fovea called the foveola.

4.4.4.7 The center of the fovea only contains cones and those at the central region are longer, thinner and more densely packed than cones elsewhere in the retina. This rod-free area is about 0.5 millimeter in diameter and subtends about 50 minutes of arc. From here to the edge of the retina the number of cones per unit area decreases, and the number of rods increases. At 20° , as illustrated in Figure 4.4, the rod population is densest.

4.4.4.8 The sensitivity of the retina to light varies with the area stimulated as shown in Figure 4.5. The fovea is most sensitive and used for seeing fine detail and color. Color sensitivity varies with position on the retina.

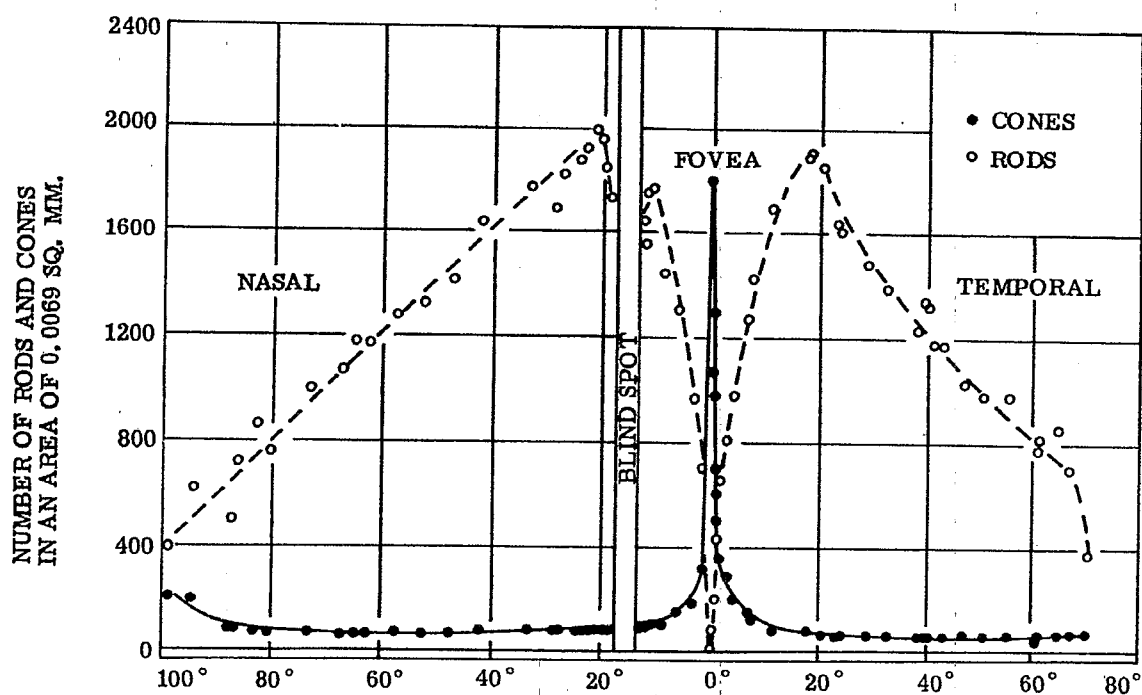


Figure 4.4 - The distribution of cones and rods across the retina (horizontal meridian).
(From National Research Council, A. Chapanis 1949)

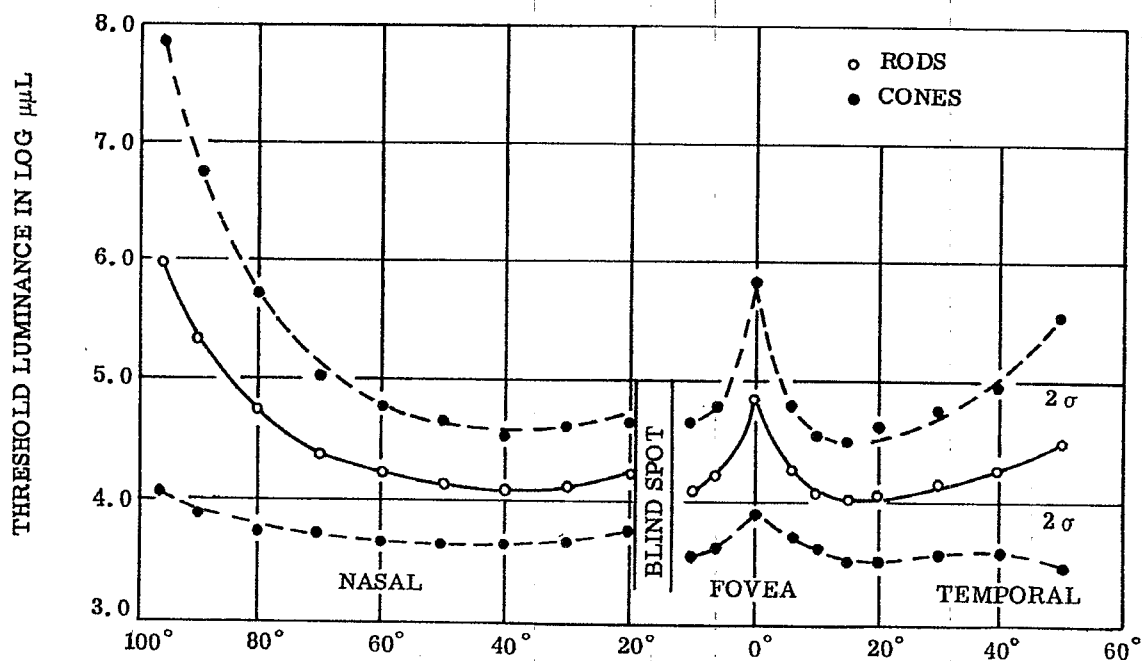


Figure 4.5 - Sensitivity to just perceptible luminance across the retina
(From National Research Council, A. Chapanis 1949)

4.4.4.9 The rods contain rhodopsin, which is bleached by light, and the products formed stimulate nerve conduction. Rods are sensitive to very small amounts of light and operate from a few quanta to a luminance of about that of moonlight (0.01 ft-L). The cones contain iodopsin and have a useful range from about 0.006 ft-L to 10,000 ft-L. Vision with the rod cells at low levels of light is called scotopic and cone cell vision at high levels is called photopic. The overlapping region (0.1-0.01 ft-L) is called mesopic vision. The structure of the rods and cones is complex and the exact mechanism of vision is not fully known. A nerve fibre conducts or it does not. Nerve fibers respond to stimulation after a latent period and are insensitive during the refractive period following conduction. Chemical action and electrical potentials accompany the impulse. These factors and the light intensity establish the timing of the impulses. The frequency rate of conduction, and the interconnections of the nerve cells, codes the light from the image on the retina into the brain and consciousness. The cones of the fovea are individually connected to a single nerve fiber and have a direct path into the optic nerve. Beyond the fovea, the rods and some cones are connected in groups by the retinal nerve cells, thereby facilitating pattern vision.

4.4.4.10 The nerve fibers from the right half of each eye cross at the point where the optic nerves join, and go to the right hemispheres of the brain. Those from the left halves of each retina go to the left hemisphere. What is seen in the right half of each visual field is connected to the left hemisphere of the cerebrum and vice versa. Cutting one optic nerve would blind that eye while damage to an optic tract would blind the same half of both eyes.

4.4.5 Resolution. The rods and cones give the retina a mosaic structure that determines resolution. Minimum resolution depends on three factors: retinal location of the image as illustrated in Figure 4.6; the nature of the image and the criterion used; and adequate time for stimulation. A very small light (bright on dark) will be seen when its image has enough quanta (2-8) to stimulate the retina, and the smallness of the bright spot depends solely on its brightness. Two small dark objects can be recognized as two when their images spread over or involve two cones providing the diffraction patterns are sufficiently separated. The arc subtense of a cone is about 1 minute (49 to 73 seconds from a gradient of 4 to 6 μ for the cones) and the average eye resolves details subtending 1 minute of arc at the eye (70 μ at 250 millimeters). An extended image (rather than point) can be seen when much smaller. For example, a telephone wire can be seen against the sky when it subtends only 0.5 second. Horizontally or vertically oriented wires are seen about equally well, but when at 60° or 120° to the horizontal they are only about one-third as visible. A break in a line, or the misalignment of two lines, one above the other, (e.g. scale and vernier) of 4 seconds is visible. Grating objects have different

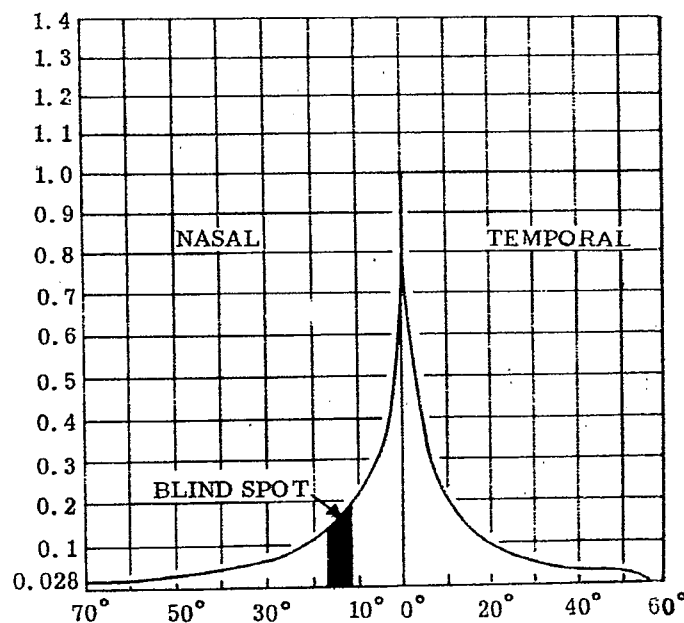


Figure 4.6 - Distribution of visual acuity across the retina expressed in degrees from the fovea

(From National Research Council, A. Chapanis 1949)

thresholds. The minimum separable for a grating in motion is reported to be about 2 minutes for a visual acuity of 1.0 for a 2° retinal area and an optical nystagmus criterion. Resolving power decreases with distance from the fovea, to 25% at 5° and only 7% of foveal resolution at 10° from the fovea. Thresholds, as illustrated in Figure 4.7, decrease linearly as the distance from the fovea increases.

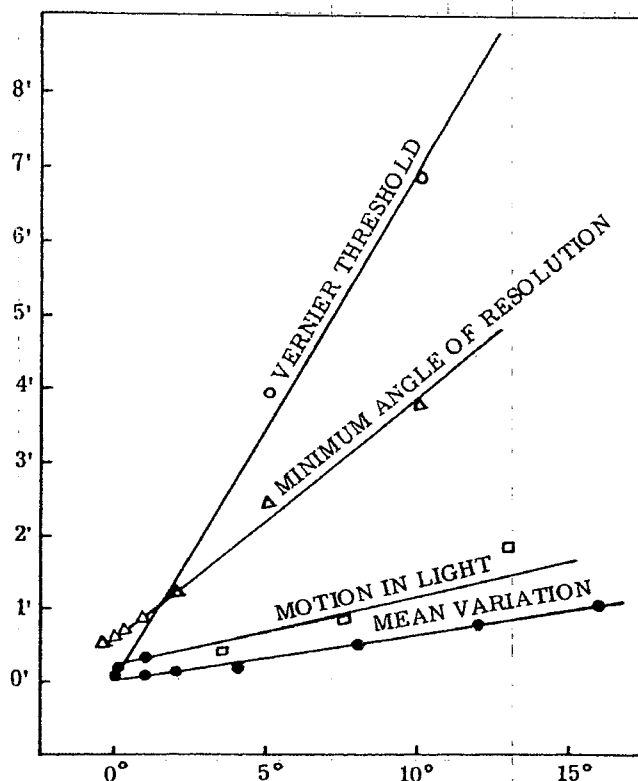


Figure 4.7 - Threshold decrease with distance (in degrees) from the fovea.
(From American Journal of Optometry No. 46, F.W. Weymouth, 1958)

4.4.5.1 Light entering the center of the pupil is more effective than light entering the edge of the pupil. This Stiles-Crawford effect is explained by the orientation of the cones within the retina since the effect occurs only in photopic vision (see paragraph 4.4.4.9). At about 1 millimeter from the center of the pupil there is a decrease to about 90%, at 2 millimeters 70%, 3 millimeters 40% and at 4 millimeters from the center of the pupil the effectiveness of the light is about 20% of that passing through the center of the pupil.

4.4.5.2 The light on the retina varies with the area of the pupil. The Troland (formerly called photon) is the unit of intensity of stimulus for 1mm^2 of pupil area and a luminance of 1c/m^2 . Luminance (mL) times $5d^2/2 = \text{Trolands}$, when d is the pupil diameter in millimeters. Correction may be required for the Stiles-Crawford effect and for the transparency of the eye should a value other than 0.5 be preferred.

4.4.5.3 Optical instruments for visual use should be designed to provide the best image on the retina, of a size and intensity resolvable by the retina. When measurements or judgments can be made by vernier acuity they will be most sensitive, e.g. when a scale value can be aligned to the specimen, the measurement will be more accurate than if the scale is superimposed on the specimen. Small linear detail is more readily seen when imaged horizontally or vertically on the retina, rather than at oblique angles.

4.5 SEEING

4.5.1 Sensitivity. Light of equal energy from different parts of the spectrum does not appear equally bright to the eye as illustrated in Figure 4.8. The yellow-green at $555\text{m}\mu$ is brightest and is ten times brighter than the blue of 470 or the red of $650\text{m}\mu$. The standard observer curve represents an internationally accepted sensitivity for use in calculations involving color and relative sensibility of the eye. Like the reduced eye discussed in paragraph 4.3.1, it is representative of average eyes and exact agreement is rarely found between it and an individual eye. Sensitivity curves for individual eyes reveal small departures from the standard observer curve that were averaged out of the standard.

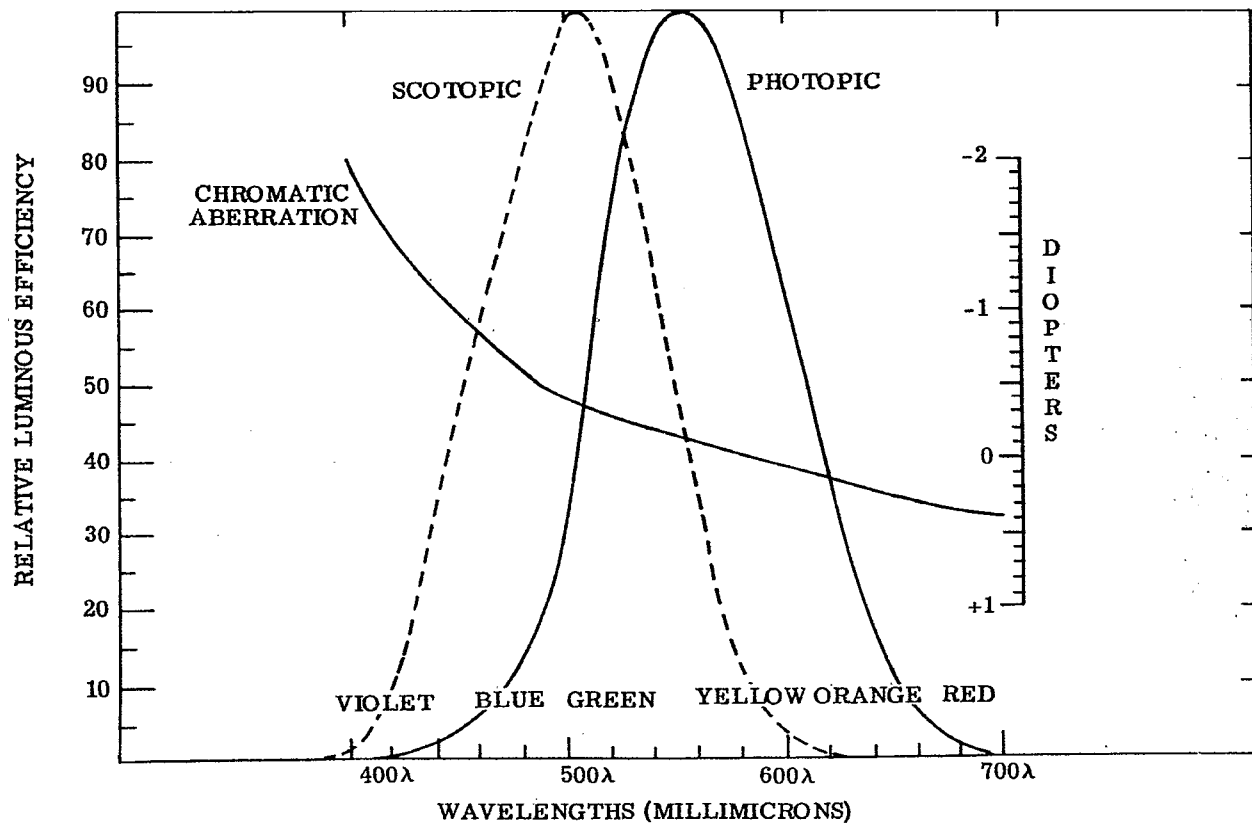


Figure 4.8. - Photopic and Scotopic standard observer curves and chromatic aberration of the eye.

4.5.2 Contrast and time. The eye can adapt itself to see over a wide range of light. The changes within the eye which make this possible involve the pigments of rods and cones and probably neural factors. The sensitivity of an eye in darkness increases rapidly for a few minutes, followed by a gradual increase for about ten minutes as illustrated in Figure 4.9. A further rapid increase of sensitivity (decrease of threshold) takes place until equilibrium is reached. While a further slow increase in sensitivity may take place for hours the amount is not large after one hour in the dark. The curve of Figure 4.9 is typical, and the change after ten minutes marks the end of the cone adaptation and the beginning of the dark adaptation of the rods. The shape of the curve depends on the adaptation state of the retina at the beginning of the dark period. The eye should be exposed for some minutes to a known light ($12 \log \mu\text{L}$) before measurement. This adaptation may be measured as the threshold at a given time, or as the time required to reach a known sensitivity. Wearing red glasses ($\lambda > 590\mu$) accomplishes some adaptation without being in total darkness.

4.5.2.1 After adaptation, the eye is more sensitive to blueish-green at 510μ , and the scotopic standard observer curve applies as illustrated in Figure 4.8. The change in the brightest region of the spectrum, from 555 to 510μ , is called the Purkinje shift. In the mesopic range, as the eye becomes dark adapted, blues appear brighter and reds darker until color vision fails at about 0.04ft-C of illumination.

4.5.2.2 Dark adaptation is effected by the amount of previous exposure and the physical condition of the individual. It is facilitated to a limited extent by an increase in the available oxygen and is decreased by malnutrition (especially vitamin A deficiency), some drugs, and various diseases. Night-blind individuals cannot adapt to lower light intensities and are disqualified from night operations. When the luminance is too low for the sensitivity of the cones, one has to look to one side of an object so that its image is not on the fovea. The retina is more sensitive for scotopic vision at about 20° from the fovea. This coincides with the greatest density of the rods.

4.5.3 Flicker. When the eye is illuminated by brief flashes of light, alternated with darkness, the eye sees a flickering until the rate reaches 10 to 30 cycles per second when the images fuse and appear continuous. This rate of fusion is called the critical flicker frequency (CFF) and slightly different values are obtained from increasing the rate than from decreasing the rate to fusion. The CFF increases with increased luminance. Talbot's law states that, "fluctuating and steady lights of the same energy content appear equally bright," although recent experimentation indicates that for brief exposures intermittent light is less efficient, while for long exposures fluctuations help. The difference is probably related to the small fluctuating movements of the eye. A great many factors affect the CFF and attempts to use it as a criterion of vision or health have not been very satisfactory.

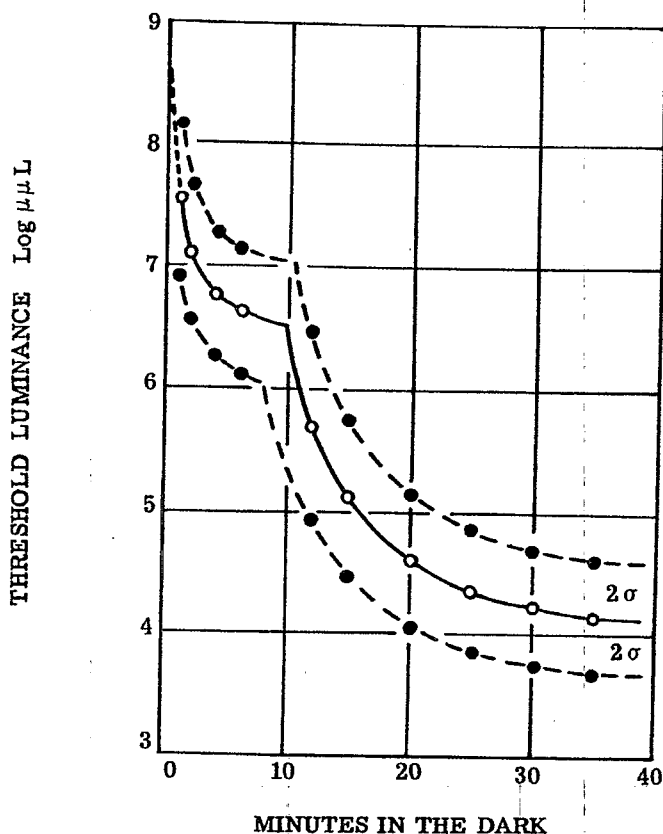


Figure 4.9. - A typical curve of dark adaptation.
(From National Research Council, A. Chapanis 1949)

4.5.3.1 When the image is stabilized on the same exact part of the retina, vision gradually fades and disappears. Continuous small fluctuating movements (30-80 cps) and slow drifting of the eye prevents loss of vision. After the image drifts too far from its original position, a quick motion returns the image to the more sensitive part of the retina. To avoid the effect of eye movements in vision research, it is necessary that the stimulus be exposed no longer than 1/100th of a second. During steady fixation for 3 to 4 seconds the image may move over 25 to 50 receptors.

4.5.4 Measuring vision. For the practical purposes of measuring vision for the prescription of spectacle lenses various types of test charts are used, usually consisting of letters of different sizes. The standard is a 5 minute square letter, the individual details of the letter subtending at the observer's eye 1 minute of arc. The reference line on the chart is made with details of a size for the viewing distance to be used. Ordinary Snellen letter charts are designed for use at 20 feet from the observer. Other lines on the chart have graded sizes of letters, e.g. the line marked 40 ft. on the chart would subtend details of 2 minutes at the eye. Visual acuity (VA) is expressed as a fraction, the numerator of which is the design distance for the chart (usually 20 ft.) and the denominator is the line which can be read at that distance. With such a chart 20/20 vision would be normal, 20/15 would be better than normal, and 20/80 would be about 1/4 normal vision (observers only able to read at 20 feet, the line normal observers would read at 80 feet). These charts have high contrast black on a white background. In Europe similar charts are based on 6 meters distance (very nearly 20 feet) and the corresponding acuities are written as 6/6, etc. The Landolt C, a circle of 5 minutes diameter with a break of 1 minute (equal to the width of the line of the character) is used also as a test character. The break can be turned up, down, etc., to test its recognition.

4.5.4.1 Different letters have different thresholds for recognition and the few letters of about equal difficulty restricts chart construction and explains why different charts give slightly different results. The differences are not great enough to be of concern in ordinary clinical practice, but can be important in research work.

4.5.4.2 Visual acuity for moving objects is different from that measured with static tests and is called dynamic visual acuity (DVA) to distinguish it from ordinary or static visual acuity (SVA). Acuity varies with the contrast of the test target and illumination, Figure 4.3. Contrast is expressed as the difference between the luminance of the object and the luminance of its surround divided by the luminance of surround. At any given intensity there is a minimum contrast which is visible. Some relations between contrast and illumination are shown on Figure 4.3.

4.5.5 Lighting, comfort, and glare. For a given intensity of illumination, contrast, and size of object, there is also a minimum time for vision. At any luminance level less time is required to see at higher luminance levels. The time relations are different for scotopic vision at low luminance levels than for photopic vision. Except on rapidly moving vehicles the time factor is usually too small during daylight to limit vision. However, in the present jet age the seeing reaction time of an individual is too great to avoid collision at the distances at which very rapidly moving aircraft can be seen.

4.5.5.1 Adequate lighting is necessary for comfortable seeing. Too little light is inadequate, leads to strain and fatigue, and with too much light (sunlight on snow or ice) temporary blindness occurs. Outdoors, the eye can see well in the shade with 100 to 400 ft-L brightness. Indoors, considerably less luminance is available (6-20 ft-L). Because of the adaptation of the eye, the indoor room appears bright at night. The amount of illumination required for seeing depends on the size and reflectivity (contrast) of the object. Sewing with black thread on black cloth requires many times the illumination needed for black thread on white cloth. Lighting recommendations of the Illuminating Engineering Society are available and a recent revision considers contrast and time for adequate vision.

4.5.5.2 Light reaching the retina other than in a useful image is called glare. Glare reduces vision most when the glare source is close to the object or is between the object and the viewer. Small amounts of glare make seeing difficult and are uncomfortable. Excessive glare disturbs the adaptive state of the eye, can prevent seeing and should be avoided. Methods for measurement and computation of glare effects are available.

4.5.6 Color vision. Color vision depends on the spectral distribution of the illumination and the wavelength range reflected or transmitted to the eye, the state of adaptation of the eye and the part of the retina involved. For example, a red object would reflect wavelengths greater than 640m μ , a blue object from 410 to 480m μ . A monochromatic yellow light (589m μ) from a sodium lamp falling on a blue object could not be reflected and the object would appear dark. A yellow can also include yellow, orange and red light. Subtractive color appears when parts of the spectrum are removed; additive color when more than one color is combined, as by projecting onto a screen. The brightness of colors depends on the energy in the light and the sensitivity of the eye, Figure 4.8. The spectral distribution of energy from different sources can be quite different, e.g. ordinary tungsten lamps are deficient in blue and produce an excess of red light as compared with sunlight. The term daylight is meaningless unless specified with respect to, time, place and direction. Average noon sunlight is nearly an equal energy spectrum, but light from a north sky has an excess of blue and a higher color temperature than direct sunlight. To avoid these ambiguities in color measurement, standard sources have been defined and internationally accepted, and any work on color vision or color comparisons should be made with standardized conditions.

4.5.6.1 The normal human eye can match any color with a mixture of three primary colors: red, green, and blue. Color blindness, that is having only gray visual sensations, is extremely rare in humans and only a few such people (achromats) have been measured and described. More common is the condition of deficient color vision, and one in ten men and one in one hundred women have more or less color vision deficiency. The most common deficiency is poor red-green discrimination, and relatively rare are defects in blue-yellow vision. A mild deficiency, or anomalous color vision, is indicated when the person requires more or less green than red to match a standard yellow, but still must have all three primaries for color matching. When the deficiency is in green, the individual is said to be deuteranomalous; when the deficiency is in the red, protanomalous. A more severe type of color deficiency is dichromatic vision. The dichromat can match any color with only two primaries. Green deficient dichromats are called deuteranopes, and the red deficient dichromats are protanopes.

4.5.6.2 The color deficient individual is unable to distinguish certain colors, and the type of color confusion points to the kind of anomaly. There are appropriate tests to determine color deficiency and such tests must be done under proper illumination. A protan who is red deficient would see red, brown, dull green, and blueish green as the same color when they have the same brightness. A green deficient deutan would confuse purplish red, brown, olive, and a green. A tritan, the rare yellow-blue deficiency, would be unable to distinguish a purple from a tan or a yellow.

4.5.6.3 Color vision may improve and reach maximum towards the end of adolescence. Thereafter, there is little change until old age. Color defectiveness is inherited and no cure or remedy is known. A mild deficiency is only a small handicap and may not even be known by the person. Medium deficiency would exclude a person from working where medium color discrimination is important, and seriously deficient individuals should be excluded from all occupations where color recognition is important. Color codes should use colors which have a minimum confusion. A good example is a green traffic light with enough added blue that it is ordinarily not confused with the red light by most color defective people. The seeing of colors is more difficult when they are small and thereby require excellent color vision ability.

4.5.6.4 The very center of the retina is color deficient for yellow. A yellow object, sufficiently far away that its image is small enough to fall in this region, appears light grey or white. Yellow has not been a very satisfactory color for air-sea rescue, because of its confusion with the white caps on the ocean. The most conspicuous color depends on the background against which it is seen and the color vision of the observer. A golden yellow, or orange is usually readily seen. Reds appear dark and may not be seen by protans.

4.5.6.5 Looking at a colored object through a complementary colored filter makes the object appear dark; conversely, through a filter of the same color it may not be seen at all. Colored glasses reduces the overall amount of light to the eye, and vision is reduced in proportion to the loss of light. With the rare exception when complementary color contrast can be used, and there is sufficient light, colored glasses will reduce seeing. This reduction is increased as dusk approaches, and no colored glass improves seeing at night. A neutral glass can reduce the intensity of light and, if not too dark, maintain color discrimination.

4.5.6.6 The appearance of many colors will change with changes in the viewing conditions. Increasing, or decreasing, the intensity of light will de-saturate some colors, and change others to a different hue. As dusk falls, a lemon yellow gradually changes to light grey or white and may not be distinguishable from a white object. For the normal eye, red is seen as red when seen as a color, but other dim colors may not be recognizable. Some colors also change in hue after being fixated for some time.

4.5.7 Perception. Perception has been defined as a complex appearing in the field of consciousness and made of sense impressions supplemented by memory. Outside of experimental projects most seeing is done at the perceptual level. The recognition of objects depends on their form and shape, and is supplemented by learning or training. It is also possible to make psychological scales, as it is possible to adjust two lights so that one appears to be twice, or half as bright as the other. The scale of equal steps in brightness can then be related to the energies measured as photometric luminances. A brightness scale increases at an exponential rate with respect to the stimulating energy.

4.5.7.1 The appearance of objects depends on their immediate surrounds, due to retinal irradiation. A series of discs cut out of the same grey paper, but placed on brighter or darker greys will not appear to be the same, but lighter or darker depending on the contrast with the surround. The appearance of color depends on the surround and on the immediately previous color adaptation. White paper looks white in daylight and will also look white at night under tungsten illumination, even though the tungsten light has more red and yellow, and the paper is reflecting more red and yellow to the eye, as the eye has adapted to and interprets the new illumination. After exposure to an intense stimulation there is seen a series of after-images. These will be in complementary color when the object is colored and they are seen against a neutral background. The after-images gradually fade and may or may not affect seeing, depending on their intensity.

4.5.7.2 Much work, during and following World War II, has discovered better form, size and arrangement for visual displays to aid the designer when scales or indicators are needed. Vision through instruments involves the same principles discussed in this section. Unless the instrument produces a sharp image of proper size, intensity, and contrast on the retina it cannot be resolved and seen. Glare should be avoided. Reticles and scales that appear in the field of view require careful planning as to size, contrast, and lighting if they are to be seen with comfort. When half shade plates, or comparison fields are used in an optical instrument, the dividing lines should become invisible and the areas compared should have the same size, otherwise a slightly larger lighter area may be equated with a slightly smaller darker area.

4.6 MOVEMENT OF THE EYES

4.6.1 General. Six muscles move the eye. The conjunctiva, Tenon's capsule, and the fat pads within the orbital cavity of the skull aid in positioning the moving eye. The center of rotation is about 13-15.5 millimeters behind the cornea. Since there are no inflexible mechanical axes, the center of rotation may vary a millimeter or so depending on the resultant of the muscular action. The muscles which turn the eye are coordinated with those of the other eye, by the muscular movements within the eye, by the movements of the eye lids, and also by the neck muscles which move the head via the nervous system.

4.6.2 Muscular action. The superior and inferior rectus muscles as illustrated in Figure 4.10, raise and lower the eye in a plane 23° from the plane of the medial orbital wall. This is the wall of the skull separating the nasal and orbital (eye) cavities. The medial (internal) and lateral rectus muscles rotate the eye toward or away from the nose in a horizontal plane, when the eye is in the primary position of looking straight ahead. The superior oblique muscle passes through tendon pulley and inserts into the upper, back side of the eye so that contraction of the muscle depresses the eye. The inferior oblique muscle is attached underneath the eye and on contraction raises the eye. The movement of the oblique muscles is in a plane through the center of rotation of the eye which slopes back about 129° from the medial orbital plane. The gaze must be directable to any place within its field of view, Figure 4.11, and maintain a horizontal reference on the retina corresponding with horizontal in the field of view. The superior oblique and the inferior rectus muscles working together

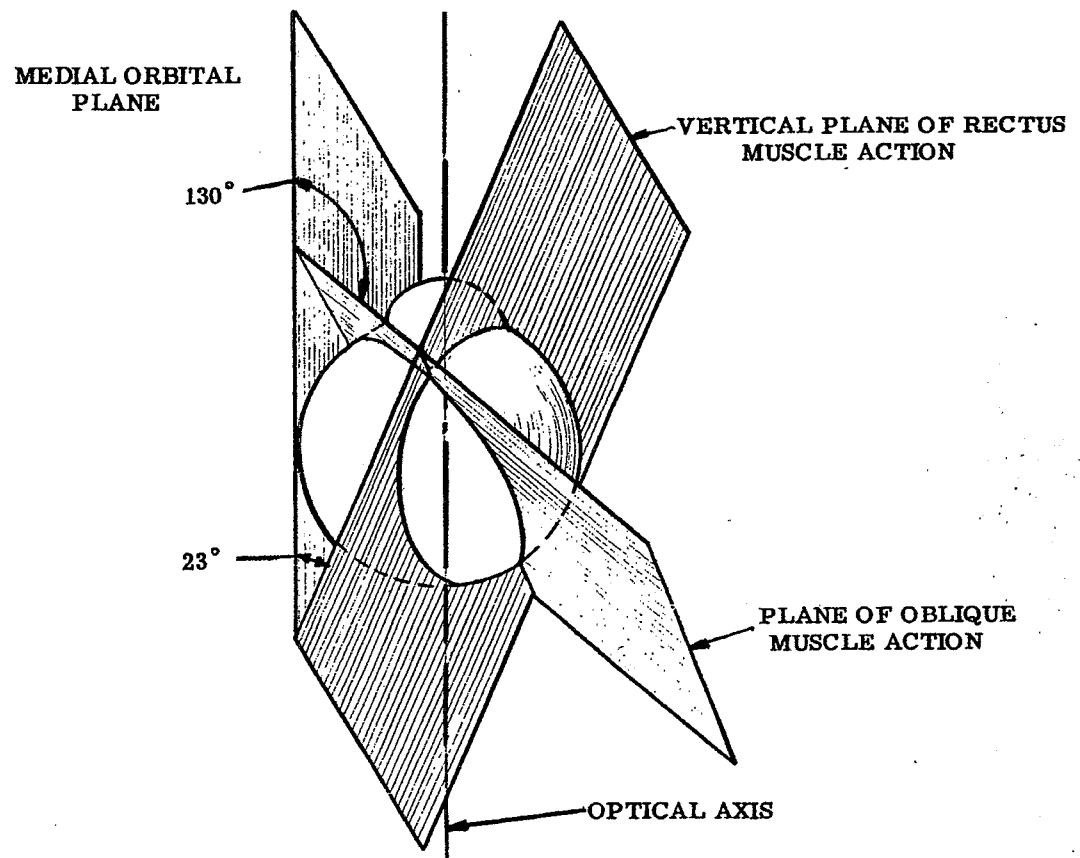


Figure 4. 10 - Planes of rotation of the external eye muscles.

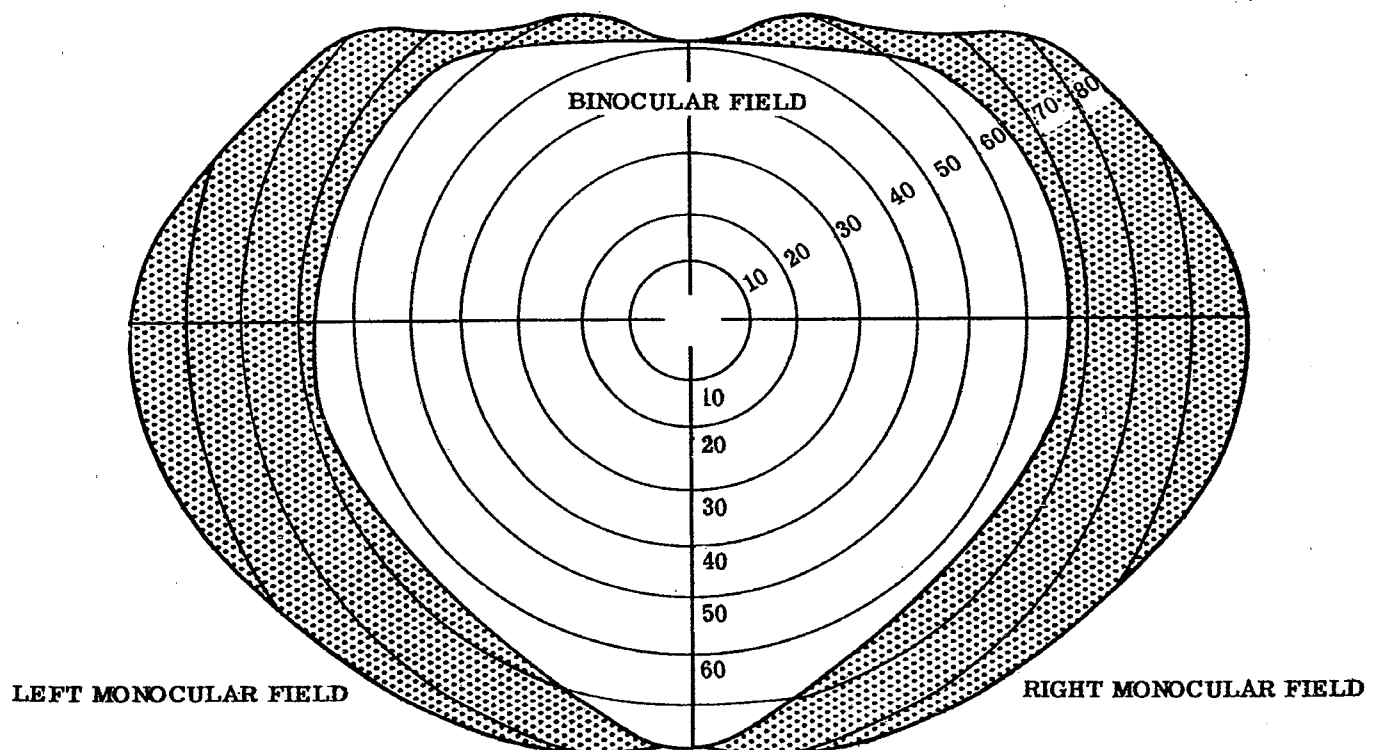


Figure 4. 11. - Monocular and binocular visual fields.

minimize a tendency for the eye to roll on an anterior posterior axis. Nevertheless, there is some torsion, or rolling, of the eye that can be mapped with the aid of after-images. Plotting the observations shows the visual field to have pincushion distortion.

4.6.3 Imbalance. The actual motions of the eye are complex. Adjustments of the eye to the left or right are made easily and up or down reasonably well, but the eye muscles are not arranged for movement of the eyes at oblique axes. Consequently, if the eyes are provided with more than slightly twisted images, they cannot adjust and fuse for single vision. The movement of the eye from one fixation to another is not smooth, and consists of movements of about 4 minutes subtended arc. The eye does not move directly to the point of fixation, but moves towards it and then approaches the fixation by a series of smaller movements. The following of a moving object by the eye also tends to go in small jumps rather than as a single smooth movement. The movements are the resultant of the contractions of one or more pairs of opposed muscles, and fluctuations are characteristic of neuromuscular mechanisms. Action potentials of the muscles can be recorded from electrodes placed around the eye, or within the muscles and their analysis is providing considerable new information on muscular movements. The eye follows a moving object as far as it can and then suddenly jumps back to a new fixation and this stepwise motion is called physiological nystagmus. Workers in mines under dim light develop a characteristic nystagmus.

4.6.4 Phorias and tropias. Two types of misalignment of the eyes have clinical importance. When one eye is covered and subsequently moves away from the fixation point, the condition is called a phoria. If the visual axes of the eyes are different when the eyes are open and uncovered, the condition is called strabismus or squint, and the direction is indicated by a tropia. Normal fixation is orthophoria or orthotropia and deviations would be heterophoria or heterotropia. The direction of the abnormal orientation is indicated by prefixes: eso- refers to movement toward the nose, exo- toward the temple, cyclo- a rotation, hyper- up, and hypo- down. Eso-tropia would indicate crossed eyes, while esophoria would indicate a moving toward the nose by a covered eye, or when the eye is dissociated from binocular vision.

4.7 BINOCULAR VISION

4.7.1 Advantages. The use of two eyes is a decided advantage in seeing. There is an apparent increase in brightness of about 20% when an object is seen with both eyes rather than with one eye alone. Normally the eye movements are equal and symmetrical and the sensory feed-back from the movements aids in balance and orientation of the organism.

4.7.2 Stereoscopy. A great advantage of two eye vision is the emergence of the experience of depth, or stereoscopic vision. Stereoscopic depth is a primary factor. Other factors which aid in the understanding of depth, such as superposition, are learned secondary factors. The basis of stereoscopic vision is horizontal dissimilarity of retinal images on corresponding points of the two retinas. In Figure 4.5, looking at the two points A and B which are at different distances from the eye, the images of the lines at A and B for the left eye are closer together than for the right eye. The fusion of these dissimilar images leads to the space perception that one is farther away from the other. Likewise, if one arranges drawings to give disparate images (within the physiological limits of the eye) when viewed through a stereoscope, the appearance of depth is produced. Stereopsis varies with the distance between the centers of the two eyes, the interpupillary distance (PD), and the spacing of the eyes alters the spatial visual geometry.

4.7.2.1 In designing binocular instruments, sufficient adjustment must be provided for the interpupillary distance of the intended observers. Formerly, 50 to 75 millimeters was considered adequate, but individuals are now growing larger and 76 millimeters maximum interpupillary distance have been used.

4.7.2.2 In stereoscopic depth the disparity between the retinal images for contours is probably more important than mere difference in size. There are limited areas on the retinas, within which objects can be placed on corresponding parts of the retinas, called Panum's areas. These areas are probably accounted for by the extent of the overlapping of the arborizations of the neurones from corresponding retinal points at the terminal areas of the cortex of the brain. The stereoscopic threshold is the smallest depth or disparity that can be experienced, and depends on the dimensions, contrast sensitivity of the retinal elements, and the sharpness of focus, i. e. the size of the blur circle on the retina. Stereoscopic acuity is less for individuals with less than 20/20 vision, but fails to increase with superior visual acuity. Stereoscopic vision is not limited to the macula and there is some evidence that it is maximal at an extra-foveal angle of 15-21 minutes. Useful stereoscopic depth is limited to about 1900 feet or a disparity angle of 24 seconds. For stereoscopic range finders the unit is about 12 seconds. The threshold for stereoscopic perception of depth increases with decreased illumination in dark adaptation, and shows a marked change which corresponds with the shift from photopic to scotopic vision.

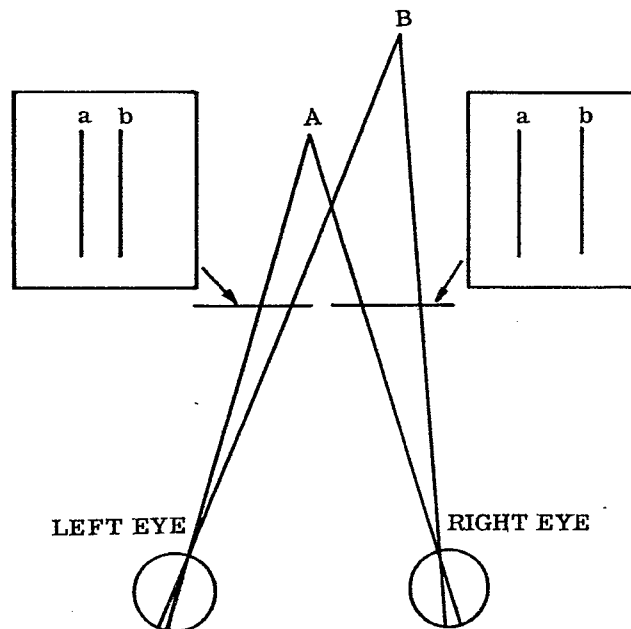


Figure 4. 12. - Stereoscopic vision with disparate images.

(From American Medical Association, Archives of Ophthalmology, No. 60, K. N. Ogle, 1958)

4.7.2.3 One of the main problems in vision is the interpretation of the geometry of what we see. This involves the two eyes, their separation, and the connections within the brain. If we use a neutral filter to absorb some of the light to one eye, little change is noted in the stereoscopic effect for static objects; but if we look at a pendulum we find that the apparent movement is no longer in a single plane, but the bob tends to swing around an ellipse. This Pulfrich illusion is explained as a result of the different reaction times for the eye with and without the filter.

4.7.3 Psychological and physical space variations. Psychological visual space is different from Euclidean physical space. If five lights are arranged in a dark room to be in a straight line they will be found to be in a curved line after the lights are turned on. When aligned at right angles to straight ahead gaze, one plane is found where the lights would be set in a straight line. Nearer than that, the lights would be in an arc concave toward the eye and farther away in an arc convex to the eye. Such experiments provide evidence that psychological visual space is hyperbolic or elliptical rather than Euclidean. The transformation equations between physical and psychological space have not been fully worked out.

4.7.4 Limitations. There are practical applications for instrument design. If the images of an object are different in each eye either a depth sensation or distorted space perception will occur. When the differences are due to unequal magnification in size the appearance is that of a distorted space, and space distortion from size differences in the images is aniseikonia. The tolerance of individuals to such differences varies, but differences of 1 to 2% or more usually result in visual strain and discomfort. Differences of 5% usually preclude binocular vision. The differences are not always those of the actual size of the images on the retina but rather are an overall size effect which involves the central nervous system. An Eikonometer is used for clinical measurement and the aniseikonia can be corrected by a special size lens for one eye. Differences in size are innate in some eyes. In others they are produced artificially by a considerable difference in the spectacle prescription for the two eyes. A common problem arises from unilateral aphakia, when a strong, plus-spectacle lens is needed to take the place of the lens of the eye. It may not be possible under these conditions to restore stereoscopic binocular vision.

4.7.5 Design considerations. The design of binocular instruments is challenging since comfortable viewing with two eyes presents difficulties that do not occur with monocular instruments. The coordinated motion of the two eyes must not be disturbed. A pupillary adjustment of 50 to at least 76 millimeters should be provided. Magnification differences to the two eyes should not exceed 2%. Some people cannot tolerate more than 0.5% while others may tolerate a little more than 2%. Oculars must be paired so that increased size differences will not occur. Beam splitters should be neutral, otherwise the light to the two eyes will cause discomfort from the chromatic

aberration of the eye. Should one eye receive a bluish light and the other eye a redish light the accommodation of each would have to be different, which would lead to strain and intolerable discomfort. The amount of light to the two eyes should be balanced, preferably within 10%. Vertical imbalance should not exceed 0.5 prism diopter. Horizontal imbalance need not be quite so small, but in excess of this value it would be fatiguing. Spectacle prescription practice holds to about 0.25 prism diopter. For low power instruments such as a bi-objective, binocular, microscope a 0.33 prism diopter difference may be tolerable. Any twist in the images should be kept to a minimum to avoid strain from complicated and difficult eye movements necessary to aline the images on the retinas. Since the light is divided to two eyes, more light will be required for binocular than for monocular instruments. In some types of binocular instruments, double mirrors or a large or diffusing mirror, may be necessary to direct the light to both eyes. When the objectives and the oculars of the instrument have different convergent angles, the appearance of depth can be made true (orthoscopic), or it can be increased or decreased (hyper- or hypostereoscopy), providing another variable for use by the instrument designer.

4.8 FATIGUE AND AGEING

4.8.1 Fatigue. Fatigue of the retinal processes is not likely at ordinary conditions. The usual "visual fatigue" (asthenopia) is muscular rather than retinal. Difficult seeing gradually involves 20 or more muscles, spreading to include those of the brow, cheek and lip. Greater mental effort is needed for getting and interpreting the visual information required. Uneven lighting results in one part of the retina needing more light and calling for pupil opening, while another is over stimulated and calling for a smaller pupil. The resultant conflict fatigues the ciliary process. Changes in illumination, too rapid for the accommodating ability of the eye, cause local and general fatigue. Continuous use of more than one-half of the available accommodative response, and close work necessitating strong convergence are fatiguing. Body tension increases during difficult seeing. An awareness of body sensation during difficult seeing, and the appearance of increasing hyper-reactivity, both increase general fatigue. A visual perceptual load, greater than can be assimilated, is also fatiguing. Visual fatigue is minimized with proper illumination, adequate contrast, form and time for seeing, proper arrangement for easy functioning of the eyes, and comfortable working conditions. An uncomfortable posture can cause eye strain and fatigue especially if seeing becomes difficult (dim light, fog, glare, etc.). An unpleasant task may make the eyes feel very tired, although instant recovery may occur on changing to an interesting visual task.

4.8.1.1 Any instrument that requires steady orientation of the eyes should be provided with a head rest, and heavy equipment should be properly supported in order to lessen fatigue. Instruments should be set up so that they are observed with a straight ahead position of the eyes, and when that is not feasible, the instrument should be adjusted to the head for comfortable vision, not the whole body of the observer cramped into a viewing position. Image brightness and convergence should be adjustable and no adjustments of the eyes beyond normal functional ability should be required by an optical instrument (unless designed to test a visual function).

4.8.2 Age. Seeing is probably at its best towards the end of adolescence. Some of the age changes are summarized in Figure 4.13. At about age 40 the accommodative mechanism begins to fail and the individual is no longer able to focus the eye on near objects. This is due to a decrease in the elasticity of the lens of the eye, although the focussing muscles may also be involved. The condition is called presbyopia and is corrected by adding positive spherical power to the spectacles, usually in the form of a bifocal, or trifocal addition. The trifocal addition has the further advantage of providing an intermediate distance of clear vision just beyond that of the near correction. The pupil of the eye does not open as far in the elderly, which fortunately increases the depth of field. Although less light gets to the retina and greater illumination is necessary for equal visual efficiency. One experimenter has found that the illumination should be doubled for each 13 years increase in age.

4.8.2.1 The eye media lose transparency, particularly the lens, which becomes yellowish as age increases. These changes effect color vision, and in addition, lessen the light available for image formation. Accommodation is slower in old age than in youth. The efficiency of the retina declines and resistance to glare becomes less. The fibers of the lens may become opaque and form a cataract. With developing cataracts, asymmetrical screening may improve the vision slightly by reducing glare. The balance between enough light for adequate seeing, and excess light or glare, is difficult and more critical in later life. When instruments are to be designed for use both by young and old people the limitations of the older eye should be kept in mind.

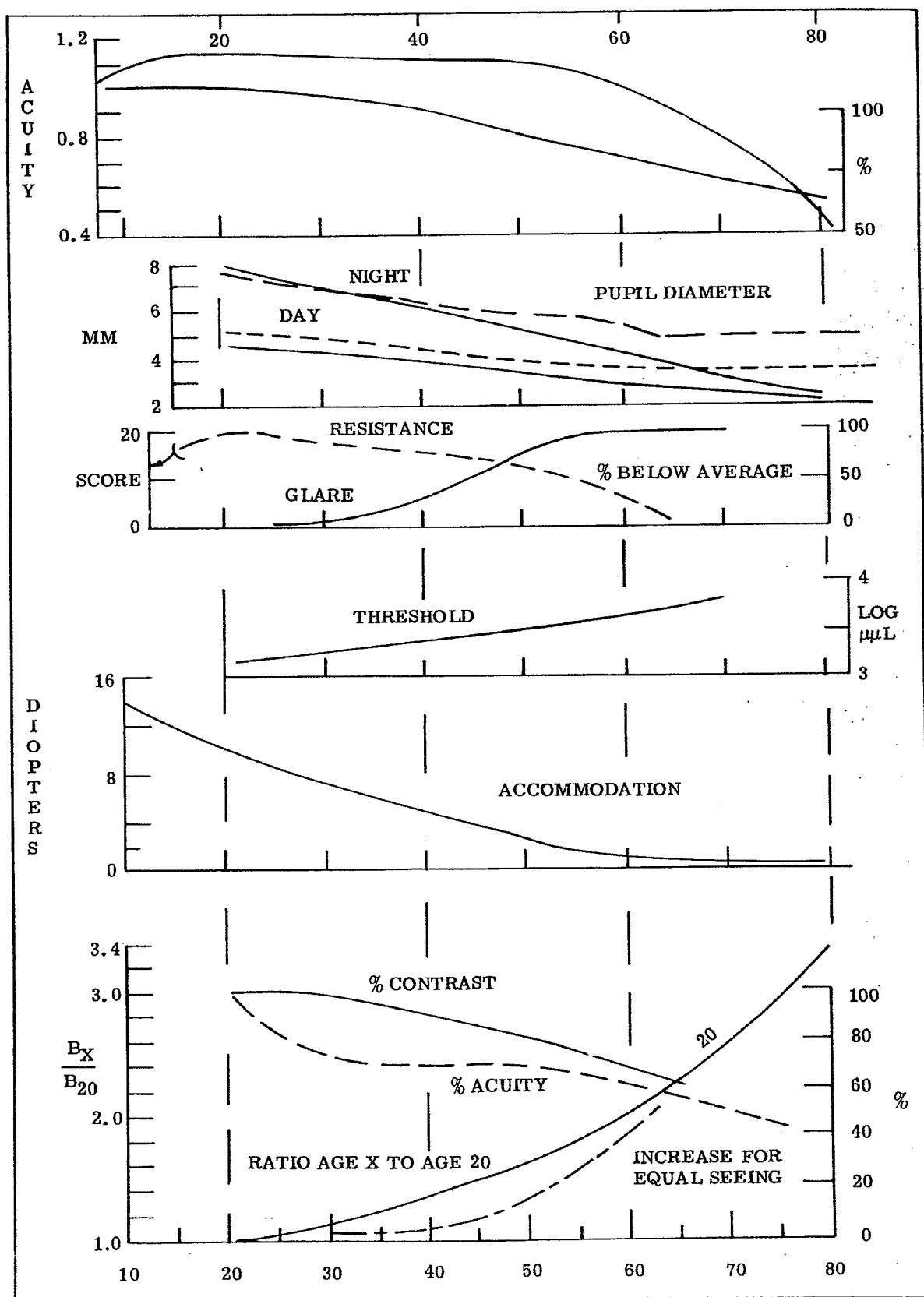
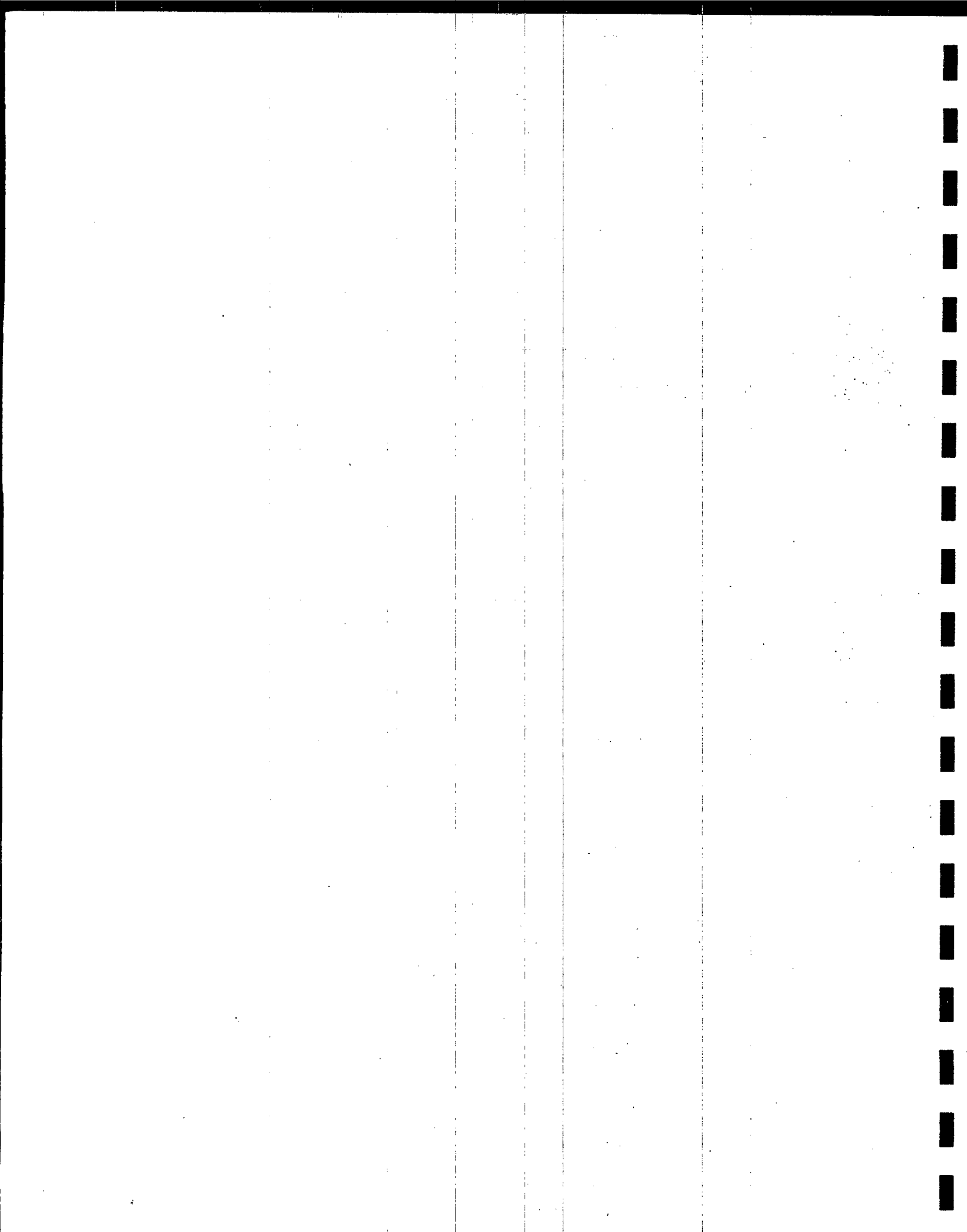


Figure 4.13 - Some age changes in vision.
(From American Journal of Optometry, O. W. Richards, 1958)



5 FUNDAMENTAL METHODS OF RAY TRACING

5.1 GENERAL

5.1.1 Basic optical system. Every optical system consists of one or more reflecting or refracting surfaces. The function of the system is to transform the diverging spherical wavefronts coming from object points in object space to converging spherical wavefronts going towards image points in image space. As mentioned in paragraph 2.1.2 the passage of the wavefronts through the optical system can be most easily discussed by utilizing the concept of rays. The passage of rays through an optical system may be determined by purely geometrical considerations, since it is correct to make the following assumptions:

- (1) A ray travels in a straight line in a homogeneous medium.
- (2) A ray reflected at an interface obeys the law of reflection.
- (3) A ray refracted at an interface obeys the law of refraction.

Computing the passage of rays through an optical system is a purely geometric problem best solved by the techniques of analytic geometry.

5.1.2 Centered optical systems.

5.1.2.1 Fortunately, nearly every theoretical optical system consists of centered refracting or reflecting surfaces. In a centered optical system all surfaces are rotationally symmetrical about a single axis. A cross-section view of a typical photographic lens is shown in Figure 5.1. In this case all the surfaces are spherical surfaces and the centers are assumed to lie on the optical axis. Herein lies one of the differences between theory and practice. In the design phase, the system is assumed to have an axis of symmetry. In practice the lenses may not be lined up perfectly so it will not be a centered optical system. If the lens is to perform according to the design, the lenses must be adjusted until they are centered. Procedures to assure centering of the elements are a prime consideration in the mechanical design of optical instruments.

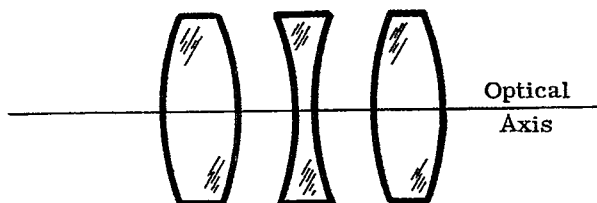


Figure 5.1 - A cross-section view of a photographic lens.

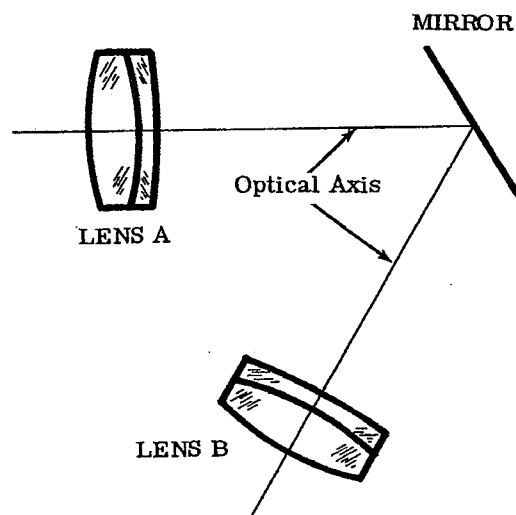


Figure 5.2 - An optical system containing a mirror.

5.1.2.2 The optical system shown in Figure 5.2 may not appear at first glance to be a centered optical system. The optical axes of the two lenses do not coincide. However, if properly constructed this may be a centered optical system. To understand this consider Figure 5.3, which shows how a system involving plane mirrors can be thought of as folded out. These ideas are treated in detail in Section 13.

5.1.2.3 Consideration of the law of reflection shows that the ray of light traveling along the optical axis from lens A is actually deflected, but can be thought to continue straight through the mirror. If the axis of lens B lies on the extended axis of lens A, then the system is a centered optical system. One can see that if lens B of Figure 5.3 is shifted to the left or right, there will be a corresponding shifting of lens B' up or down; the system will become decentered and lose its axial symmetry.

5.1.2.4 The sections on geometrical optics in this handbook consider centered systems. Decentered systems usually, when carefully analyzed, are seen to be part of some centered system. Hence if a final design calls for a decentered system, the preliminary design considers the centered system as a basic starting point.

5.1.3 Plane, spherical and aspheric surfaces.

5.1.3.1 Production techniques for generating plane and spherical surfaces on optical materials are well established and thus these are most commonly used. Aspheric surfaces, however, offer certain advantages, and recent advances in the generation of this type of surface, coupled with the need for the design refinements they offer, have resulted in more frequent design application of this type. Aspheric surfaces are also usually considered to have rotational symmetry about the optical axis.

5.1.3.2 In ray tracing, plane surfaces will be considered to be special cases of spherical surfaces, having radii equal to infinity; hence no special technique for plane surfaces will be developed in detail in this section. In Section 13, reflection from plane surfaces is considered more fully. The technique for treating aspheric surfaces is developed by extending the technique for spherical surfaces. In both cases, the surfaces are considered to be centered.

5.1.4 Ray tracing, the basic tool of optical design.

5.1.4.1 In order to understand clearly the kind of image formed by a system, and what must be done to im-

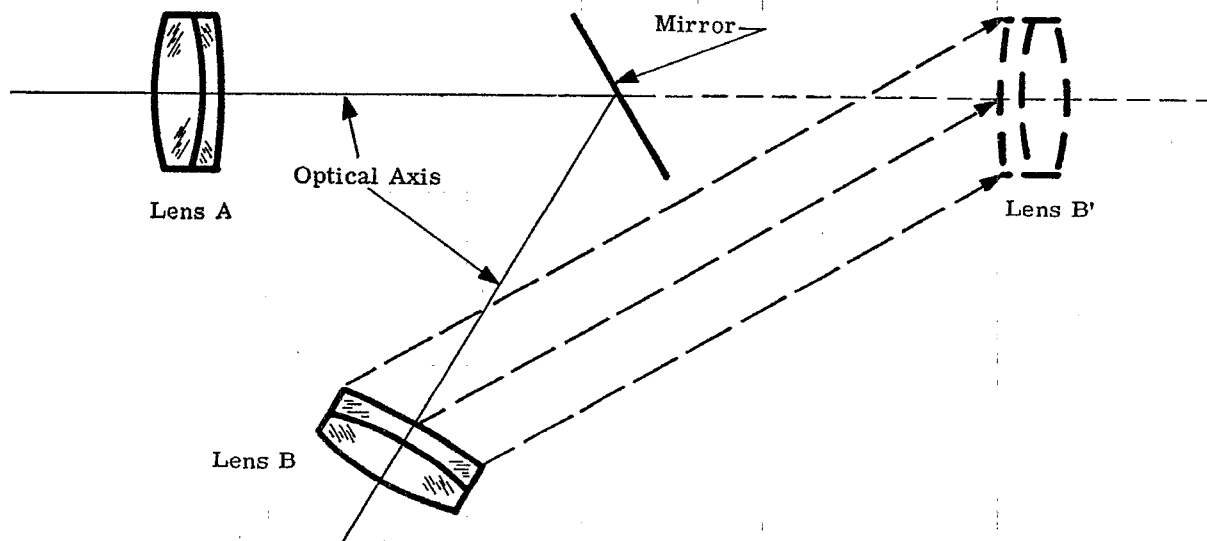


Figure 5.3 - Diagram showing "folding out" of an optical system containing a mirror.

prove this image, a certain number of rays must be determined in their passage through the system. This process of ray tracing involves the determination of the direction and location in space of each segment of a ray as it goes from object to image. Since the function of the system is to transfer light from an object surface to an image surface, the object surface and the image surface, although neither reflecting nor refracting, can be considered as surfaces of the optical system.

5.1.4.2 Figure 5.4 shows a cross-section view of a centered optical system. The ray, consisting of seven straight line segments, goes from the object point, O, on the object surface, to the image point, O', on the image surface, being refracted at six intermediate surfaces. The remainder of Section 5 will be concerned with numerical and graphical methods of determining the course of general and special rays through a general system.

5.2 DEFINITIONS AND CONVENTIONS

5.2.1 Need for specific conventions. The ray tracing formulae to be used for tracing a ray through a system involve parameters of more than a single surface or a single medium. Therefore, it is important to adopt a convention of notation which will clearly distinguish one surface from another and one medium from another. In addition, many optical systems employ mirrors, so that the rays sometimes proceed in a direction generally opposite to the incident rays. Our conventions should be such that a reflecting surface can be handled as any other general refracting surface. It is assumed that before applying these conventions the system has been folded out in the sense of Figure 5.3.

5.2.2 Statements of definitions and conventions. The following definitions and conventions, which are in agreement with those given in MIL-STD-34, will be used in Sections 2, and 5 through 15, inclusive. Reference to Figures 5.4 and 5.5 will indicate examples of some of these conventions.

- (1) It will be assumed that light initially travels from left to right.
- (2) An optical system will be regarded as a series of surfaces starting with an object surface and ending with an image surface. The surfaces will be numbered consecutively, in the order in which the light is incident on them, starting with zero for the object surface and ending with k for the image surface. A general surface will be called the jth surface.
- (3) All quantities between surfaces will be given the number of the immediately preceding surface.
- (4) A primed superscript will be used to denote quantities after refraction only when necessary.
- (5) r_j is the radius of the jth surface. It will be considered positive when the center of curvature lies to the right of the surface.
- (6) The curvature of the jth surface is $c_j = 1/r_j$. c_j has the same sign as r_j .
- (7) t_j is the axial thickness of the space between the jth and the $j + 1$ surface. It is positive if the $j + 1$ surface physically lies to the right of the jth surface. Otherwise it is negative.
- (8) n_j is the index of the material between the jth and the $j + 1$ surface. It is positive if the physical ray is traveling from left to right. Otherwise it is negative.
- (9) K_j , L_j , M_j are the products of n_j and the direction cosines (with respect to the X, Y, Z axes respectively) of a ray in the space between the jth and the $j + 1$ surface. They will be called the optical direction cosines.
- (10) The right-handed coordinate system shown in Figure 5.5 will be used. The optical axis will coincide with the Z axis. The light travels initially toward larger values of Z. Positive values of X are away from the reader in Figure 5.5.
- (11) X_j , Y_j , Z_j are the position coordinates of a ray where it intersects the jth surface.
- (12) In writing formulae where no confusion is likely to result, the j will be omitted from the subscript. Thus the curvature of the $j - 1$ surface will be written c_{-1} , the curvature of the jth surface will be written c and the curvature of the $j + 1$ surface will be written c_{+1} .

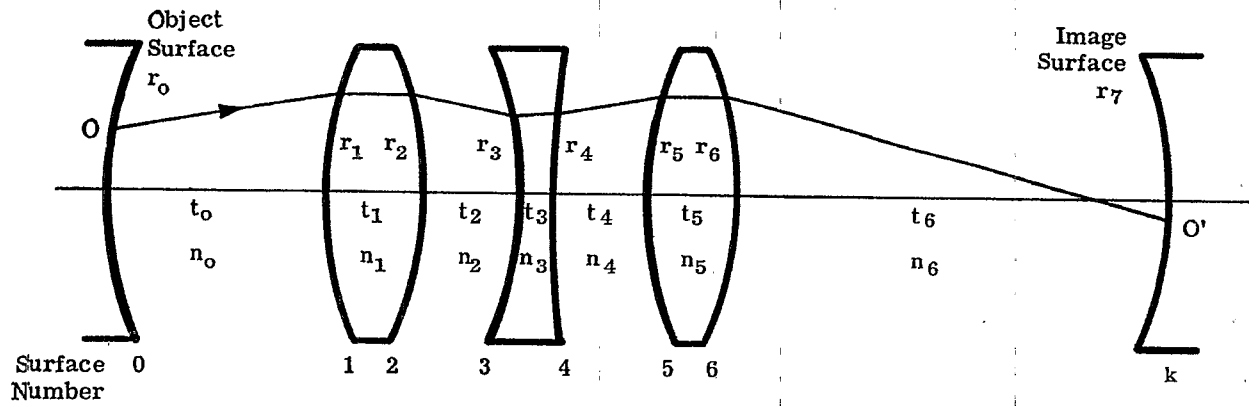


Figure 5.4—Cross-sectional view of a centered optical system.

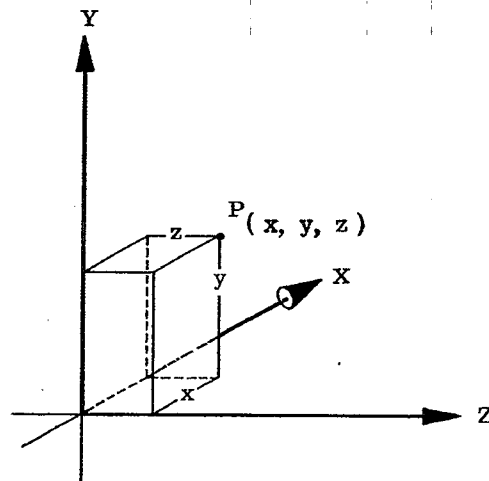


Figure 5.5- Right-handed coordinate axes.

5.3 BASIC RAY TRACE PROCEDURE

5.3.1 Transfer procedure. As can be seen from Figure 5.4 a ray travels in a straight line from a point on one surface to a point on the following surface. It is then refracted and proceeds to the next surface in a straight line. The ray tracing procedure then consists of two parts, the transfer procedure, and the refraction procedure. The transfer procedure involves computing the intersection point of the ray on the surface from the optical direction cosines and the intersection point data at the previous surface. That is, given K_{-1} , M_{-1} , L_{-1} and X_{-1} , Y_{-1} , Z_{-1} , compute X , Y , Z . The equations used are called the transfer equations.

5.3.2 Refraction procedure. The refraction procedure involves computing the optical direction cosines of a ray from the intersection point data and the optical direction cosines of the previous ray segment. That is, given X , Y , Z and K_{-1} , M_{-1} , L_{-1} , compute K , L , M . The equations used are called the refraction equations.

5.3.3 Repetition for successive surfaces. After having applied the two procedures, we have the initial data for the next application. The transfer equations will be used to compute X_{+1} , Y_{+1} , Z_{+1} and the refraction equations will be used to compute K_{+1} , L_{+1} , M_{+1} . It should be noted that it is often convenient to introduce fictitious or non-refracting surfaces to simplify the procedure. One example is the tangent plane, an XY plane tangent to a physical surface at the optical axis. Another example is a sphere, tangent to an aspheric surface at the optical axis. These fictitious surfaces are handled in exactly the same manner as a physical surface. Transfer equations are used to go to or from such a surface. The refraction equation reduces to $I = I'$, and the direction cosines of the refracted ray equal those of the incident ray, as would be expected at a non-refracting surface. Fictitious surfaces will be used in the next section.

5.4 SKEW RAY TRACE EQUATIONS FOR SPHERICAL SURFACES

5.4.1 Types of rays.

5.4.1.1 A general ray is any ray passing from any object point through the optical system to its image point on the image surface. A special ray that lies in a plane containing the optical axis and the object point is called a meridional ray. Any non-meridional ray is a skew ray. A ray close to the optical axis is a paraxial ray. Because of the approximation involved, a paraxial ray is a special type of meridional ray. A skew ray is considered to be non-paraxial since it is non-meridional. These distinctions will become apparent as the subject is developed.

5.4.1.2 Corresponding to the three types of rays, skew, meridional, and paraxial, we will develop three sets of ray trace equations and procedures. Because the three rays, in the order given here, become less general and more specialized, the equations relating to these types of rays become simpler as we proceed from skew through meridional to paraxial. One method of developing the subject would be to discuss the simplest case first (paraxial), then proceed to the more complicated (meridional) and finally to the most general (skew). This procedure would have the advantage of beginning with the simplest derivation. However, it would necessitate three separate derivations.

5.4.1.3 We will proceed in the other direction, beginning with the most general case, the skew ray trace. From this the meridional and paraxial equations follow by simplification; hence only one derivation is necessary, instead of three. The particular equations derived in Section 5.4 are set up in a form for an electronic computer. However they are completely satisfactory for use with a desk calculator, and represent a good starting point for the human computer who has not yet worked out his own equations.

5.4.2 Initial data for a skew ray. Figure 5.6 shows the skew ray as it traverses the space between two surfaces. At the right hand surface it is refracted, and a drawing corresponding to Figure 5.6 could show this ray as it traverses the space between the j th and j_{+1} spherical surfaces. Similarly, another drawing could show the ray before refraction at the j_{-1} surface. The initial data for the ray we are considering will consist of the emergence point with the left surface, and the direction of the ray in space. Hence we specify X_{-1} , Y_{-1} , and Z_{-1} , the coordinates on the j_{-1} surface, and K_{-1} , L_{-1} , and M_{-1} , the optical direction cosines of the ray. From these data we will determine the intersection of the ray with the next surface, and the optical direction cosines of the refracted ray. These values then become the initial data for the new ray, and the process is repeated until the image point is reached.

5.4.3 Transfer procedure, physical surface to next tangent plane.

5.4.3.1 The first part of the problem, namely the determination of the intersection of the ray with the j th spherical surface, will be divided into two parts: first, the intersection of the ray with a non-physical surface, the plane tangent to the spherical surface, and, second, the final intersection with the spherical sur-

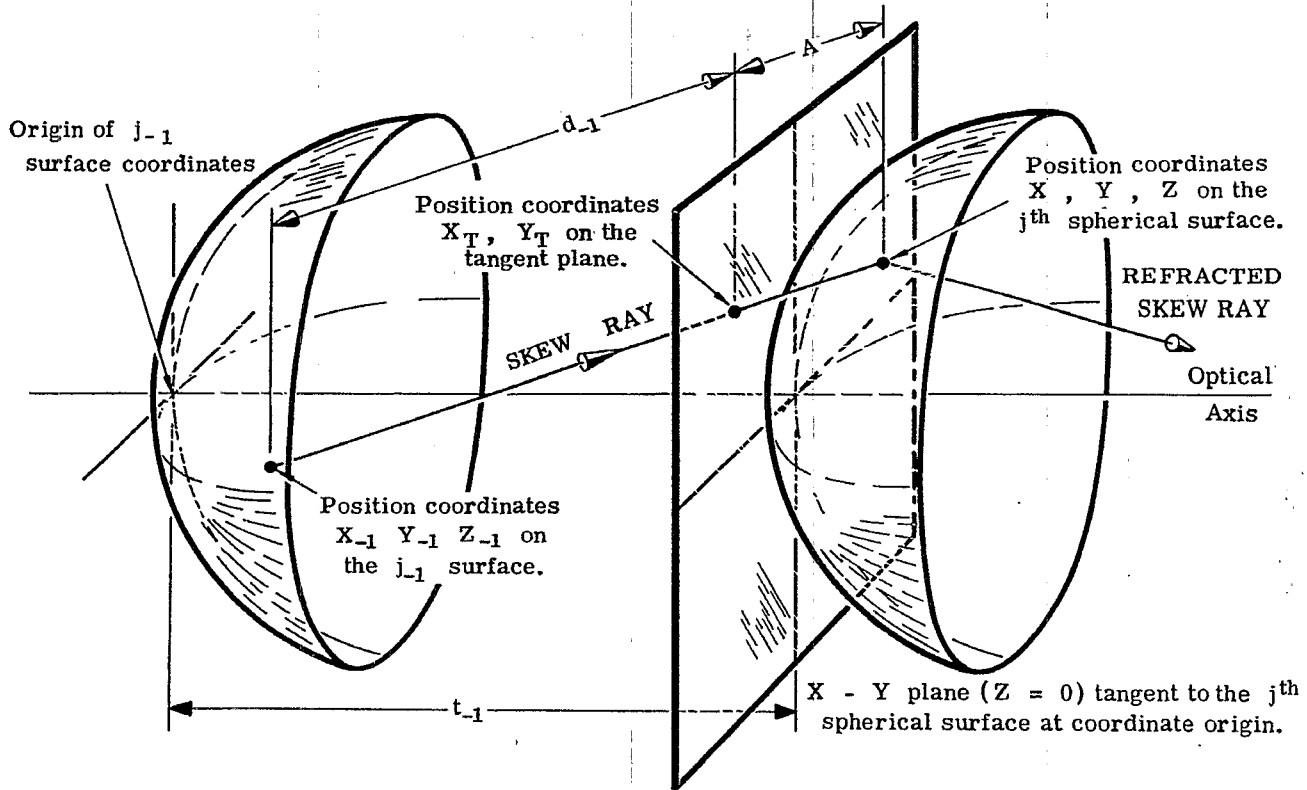


Figure 5.6 - Diagram of a skew ray in space between the $j-1$ surface and the j th surface.

face. In Section 5.4.3 we consider only the first part.

5.4.3.2 The origin of the position coordinates for points on the tangent plane is at the point of tangency, the optical axis. Hence $Z_T = 0$ for all points in the plane. The new value of X , X_T , is the old value, X_{-1} , plus the change in X , ΔX . The latter is the projection of the skew ray, of length d_{-1} , onto the X axis. Hence

$$X_T = X_{-1} + \Delta X = X_{-1} + d_{-1} \frac{K_{-1}}{n_{-1}},$$

since K_{-1}/n_{-1} is the direction cosine of the ray with respect to the X axis. There is a corresponding equation for Y_T .

5.4.3.3 The length of the ray, d_{-1} , between the left-hand surface and the tangent plane is not given; it must be calculated from the initial data. From Figure 5.6 the change in Z is given by

$$\Delta Z = t_{-1} - Z_{-1},$$

and this equals the projection of the ray along the Z axis. Therefore

$$\Delta Z = d_{-1} \frac{M_{-1}}{n_{-1}}.$$

5.4.3.4 It is now possible to summarize the three equations which are used to calculate the intersection of the ray with the tangent plane.

$$\frac{d_{-1}}{n_{-1}} = (t_{-1} - Z_{-1}) \frac{1}{M_{-1}}, \quad (1)$$

$$Y_T = Y_{-1} + \frac{d_{-1}}{n_{-1}} L_{-1}, \quad (2)$$

and

$$X_T = X_{-1} + \frac{d_{-1}}{n_{-1}} K_{-1}. \quad (3)$$

It should be pointed out that in addition to the initial data for the ray, we must be given the value t_{-1} , the

distance between the surfaces measured along the optical axis. It is not necessary, however, to know explicitly the value of n_{-1} at this time. The specific procedure followed is first, to use Equation (1) to calculate the numerical value of d_{-1}/n_{-1} ; second, to use the value thus obtained in Equations (2) and (3) to calculate Y_T and X_T respectively.

5.4.4 Transfer procedure, tangent plane to spherical surface.

5.4.4.1 The discussion in Section 5.4.3 treated the first part of the transfer problem. The following discussion treats the second part, transferring the ray coordinates on the tangent plane to those on the spherical surface.

5.4.4.2 Referring to Figure 5.6, since the tangent plane is not a refracting plane, the ray continues on to the sphere, for a distance A . The segment A has the same optical direction cosines as the segment d_{-1} . Therefore the new values of the coordinates, X , Y , and Z on the sphere, are determined from the values on the tangent plane, X_T , Y_T , and Z_T , by the process that was used to set up Equations (2) and (3). Remembering that Z_T is zero, we have

$$X = X_T + \frac{A}{n_{-1}} K_{-1}, \quad (4)$$

$$Y = Y_T + \frac{A}{n_{-1}} L_{-1}, \quad (5)$$

and

$$Z = \frac{A}{n_{-1}} M_{-1}. \quad (6)$$

5.4.4.3 In order to use Equations (4), (5), and (6), it is necessary to calculate the value of A . It is clear from Figure 5.6 that this value depends on the curvature of the j th spherical surface, the coordinates of the ray at the tangent plane, and the direction cosines of the ray. We will use a relation between X , Y , Z and c which depends on the properties of a sphere. This equation can be used with Equations (4), (5), and (6) to eliminate X , Y , and Z . The result will be an expression for A/n_{-1} in terms of known data.

5.4.4.4 Figure 5.7 shows a plane containing the optical axis and the intersection point (X, Y, Z) of the

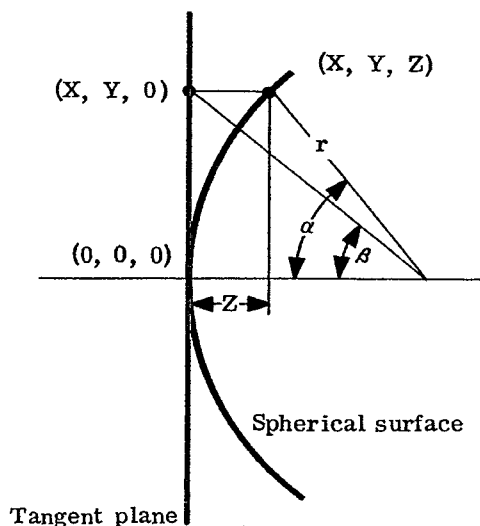


Figure 5.7 - Some properties of a spherical surface.

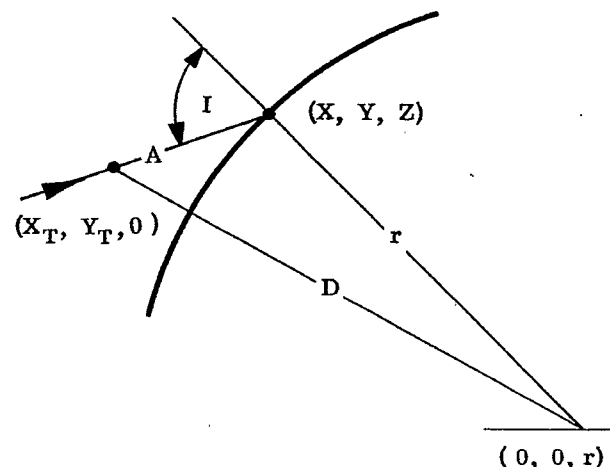


Figure 5.8 - Determination of $n_{-1} \cos I$.

ray on the spherical surface. From the figure, and recalling that $c = 1/r$, we have

$$Z = r - \left[r^2 - (X^2 + Y^2) \right]^{1/2} = \frac{1}{c} - \frac{1}{c} \left[1 - c^2 (X^2 + Y^2) \right]^{1/2},$$

which can be simplified, by transposing and squaring, to

$$c^2 (X^2 + Y^2 + Z^2) - 2cZ = 0.$$

Substituting into this equation the expressions for X , Y , and Z from Equations (4), (5), and (6). The result, on collecting terms, is

$$\left(\frac{A}{n_{-1}} \right)^2 c (K_{-1}^2 + L_{-1}^2 + M_{-1}^2) - 2 \left(\frac{A}{n_{-1}} \right) \left[M_{-1} - c (Y_T L_{-1} + X_T K_{-1}) \right] + c (X_T^2 + Y_T^2) = 0.$$

5.4.4.5 In this last simplification it was assumed that $c \neq 0$; the case of $c = 0$ will now be considered. Since the sum of the squares of the direction cosines is unity, the coefficient of $(A/n_{-1})^2$ is $c n_{-1}^2$. Calling the other coefficients $2B$ and H respectively, we have

$$c n_{-1}^2 \left(\frac{A}{n_{-1}} \right)^2 - 2B \left(\frac{A}{n_{-1}} \right) + H = 0,$$

which has the solutions,

$$\frac{A}{n_{-1}} = \frac{B \pm n_{-1} \left[\left(\frac{B}{n_{-1}} \right)^2 - cH \right]^{1/2}}{c n_{-1}^2}.$$

As $c \rightarrow 0$, that is as the spherical surface approaches a plane surface, $A \rightarrow 0$ as can be seen from Figure 5.6. To insure this we can use only the negative sign in the above solutions. A has the same sign as c ; this can be seen either by considering the expression for A/n_{-1} , or from Figure 5.6. When A is negative, the tangent plane lies to the right of the surface. The coefficients B and H were introduced for convenience in calculation. Their physical significance is not difficult to understand. From the definition of H , and from Figure 5.7, it is seen that

$$H = c (X_T^2 + Y_T^2) = r \left[\frac{X_T^2 + Y_T^2}{r^2} \right] = r \tan^2 \beta,$$

where β is the angle between the optical axis and a line drawn from the center of curvature to the intersection of the ray with the tangent plane. From this expression for H , and the result derived in paragraph 5.4.4.4, an expression for B in terms of n_{-1} , and angles I and β can be found.

5.4.4.6 Before simplifying the expression for $\frac{A}{n_{-1}}$ a discussion of the physical meaning of the square root, $\left[(B/n_{-1})^2 - cH \right]^{1/2}$, is in order. This term will be used by itself in the refraction procedure; it is convenient to put it in another form here. Consider Figure 5.8; all the lines are in the plane of incidence. Using the cosine law, it can be stated that

$$D^2 = X_T^2 + Y_T^2 + r^2 = A^2 + r^2 + 2Ar \cos I.$$

Solving for $\cos I$, and substituting H for $c(X_T^2 + Y_T^2)$ produces

$$n_{-1} \cos I = \frac{H - c n_{-1}^2 \left(\frac{A}{n_{-1}} \right)^2}{2 \frac{A}{n_{-1}}}.$$

Finally, substituting the expression for $\frac{A}{n_{-1}}$, with the negative sign, given in paragraph 5.4.4.5, gives

$$n_{-1} \cos I = n_{-1} \left[\left(\frac{B}{n_{-1}} \right)^2 - cH \right]^{1/2}. \quad (7)$$

5.4.4.7 Returning to the solution for $\frac{A}{n_{-1}}$ in paragraph 5.4.4.5, and using the expression for $n_{-1} \cos I$, we have

$$\frac{A}{n_{-1}} = \frac{B - n_{-1} \cos I}{c n_{-1}^2}.$$

But by using Equation (7)

$$c n_{-1}^2 = \frac{B^2 - n_{-1}^2 \cos^2 I}{H} = \frac{(B + n_{-1} \cos I)(B - n_{-1} \cos I)}{H},$$

and the final expression for $\frac{A}{n_{-1}}$ becomes,

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I}. \quad (8)$$

5.4.4.8 The four equations, then, which are used to calculate $\frac{A}{n_{-1}}$ are, in the order used,

$$H = c(X_T^2 + Y_T^2), \quad (9)$$

$$B = M_{-1} - c(Y_T L_{-1} + X_T K_{-1}), \quad (10)$$

$$n_{-1} \cos I = n_{-1} \left[\left(\frac{B}{n_{-1}} \right)^2 - c H \right]^{1/2}, \quad (7)$$

and

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I}. \quad (8)$$

Equations (4), (5), and (6) are then used to calculate X , Y , and Z .

5.4.5 Refraction procedure at the spherical surface.

5.4.5.1 Now that X , Y and Z have been calculated, these values together with initial data K_{-1} , L_{-1} , and M_{-1} , can be used to determine K , L , and M , which specify the direction of the ray after refraction. The basic equations which will be employed are Equations 2-(3) and 2-(4).

5.4.5.2 In Section 2 it was shown that Equation 2-(3) has the following meaning: if vectors are drawn (refer to Figure 2.3) from the intersection point, in the direction of the incident and refracted rays respectively, and these vectors have lengths equal to n_0 and n_1 , then the closing side of the triangle is parallel to the normal to the surface, and is of length Γ .

5.4.5.3 We now redraw this figure considering the surface as the j th surface. This is shown in Figure 5.9, which is drawn in the plane of incidence. Thus, the radius of curvature of the surface is also in this plane. The line of length Γ is parallel to r . The unit vector \vec{M}_1 is the quotient of the vector parallel to the normal divided by r . Hence

$$\begin{aligned} \vec{M}_1 &= c \left[(0 - X) \vec{i} + (0 - Y) \vec{j} + (r - Z) \vec{k} \right] \\ &= c \left[-X \vec{i} - Y \vec{j} + (r - Z) \vec{k} \right], \end{aligned}$$

where \vec{i} , \vec{j} , \vec{k} are unit vectors along the coordinate axes. Using Equation 2-(3),

$$\vec{S}_1 - \vec{S}_0 = -c X \Gamma \vec{i} - c Y \Gamma \vec{j} + c (r - Z) \Gamma \vec{k}.$$

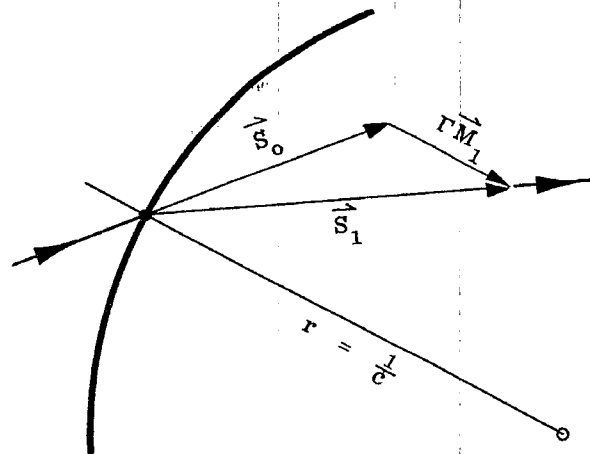


Figure 5.9 - Triangle for the law of refraction.

Now

$$\vec{S}_0 = n_{-1} \vec{Q}_0 = K_{-1} \vec{i} + L_{-1} \vec{j} + M_{-1} \vec{k},$$

and a similar equation holds for \vec{S}_1 . Hence

$$\vec{S}_1 - \vec{S}_0 = (K - K_{-1}) \vec{i} + (L - L_{-1}) \vec{j} + (M - M_{-1}) \vec{k}.$$

Equating like coefficients of \vec{i} , \vec{j} , and \vec{k} , we have relations between the old and new optical direction cosines.

5.4.5.4 There remains the calculation of Γ . This is done by using Equation 2-(4). We can now write down the five equations which are used in the order given to calculate K , L , and M from the initial data or from previously calculated results.

$$n \cos I' = n \left[\left(\frac{n_{-1}}{n} \cos I \right)^2 - \left(\frac{n_{-1}}{n} \right)^2 + 1 \right]^{1/2}, \quad (11)$$

$$\Gamma = n \cos I' - n_{-1} \cos I, \quad (12)$$

$$K = K_{-1} - X c \Gamma, \quad (13)$$

$$L = L_{-1} - Y c \Gamma, \quad (14)$$

and

$$M = M_{-1} - (Z c - 1) \Gamma. \quad (15)$$

5.4.6 Summary of ray trace equations.

5.4.6.1 In the previous sections there were derived the equations used to trace a skew ray from one surface through the following one. For convenience, the equations are now listed in the order of use. The initial ray data are X_{-1} , Y_{-1} , Z_{-1} , K_{-1} , L_{-1} and M_{-1} . The initial system data are t_{-1} , n_{-1} and c . Final values to be determined are X , Y , Z , K , L , and M .

$$\frac{d_{-1}}{n_{-1}} = (t_{-1} - Z_{-1}) \frac{1}{M_{-1}}, \quad (1)$$

$$Y_T = Y_{-1} + \frac{d_{-1}}{n_{-1}} L_{-1}, \quad (2)$$

$$X_T = X_{-1} + \frac{d_{-1}}{n_{-1}} K_{-1}, \quad (3)$$

$$H = c (X_T^2 + Y_T^2), \quad (9)$$

$$B = M_{-1} - c (Y_T L_{-1} + X_T K_{-1}), \quad (10)$$

$$n_{-1} \cos I = n_{-1} \left[\left(\frac{B}{n_{-1}} \right)^2 - c H \right]^{1/2}, \quad (7)$$

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I}, \quad (8)$$

$$X = X_T + \frac{A}{n_{-1}} K_{-1}, \quad (4)$$

$$Y = Y_T + \frac{A}{n_{-1}} L_{-1}, \quad (5)$$

$$Z = \frac{A}{n_{-1}} M_{-1}, \quad (6)$$

$$n \cos I' = n \left[\left(\frac{n_{-1}}{n} \cos I \right)^2 - \left(\frac{n_{-1}}{n} \right)^2 + 1 \right]^{1/2}, \quad (11)$$

$$\Gamma = n \cos I' - n_{-1} \cos I, \quad (12)$$

$$K = K_{-1} - X c \Gamma, \quad (13)$$

$$L = L_{-1} - Y c \Gamma, \quad (14)$$

and

$$M = M_{-1} - (Z c - 1) \Gamma. \quad (15)$$

5.4.6.2 The final calculated values, X , Y , Z , K , L , and M now become the initial ray data for the next calculation. The new system data, t , n , and c_{+1} must be given. These ray and system data are used with the above ray trace equations; in this way a given skew ray from any object surface can be traced through any number of spherical surfaces to the spherical image surface.

5.4.6.3 The equations listed in paragraph 5.4.6.1 are general, in that they also hold for plane surfaces. Referring to Figure 5.6, the physical result is that the j^{th} surface coincides with the tangent plane, hence the coordinates X_T , Y_T , Z_T equal X , Y , Z , and $A = 0$. These results follow mathematically by using $c = 0$ in the equations given in paragraph 5.4.6.1. Refraction at plane surfaces will be discussed in detail in Section 13.

5.4.7 Step by step ray tracing procedure.

5.4.7.1 The following table, Table 5.1, shows how these calculations can be made in a compact systematic manner. The surfaces are numbered 0, 1, 2, 3 beginning with 0 as the object surface. The initial system data are the values of the c , t , and n quantities indicated above the double line. In a numerical example (see Table 5.2) the values of these quantities are written in the places indicated. The letters in the left hand column have been defined in Section 5.2.2, or by the equations in Section 5.4.6.

5.4.7.2 The initial ray data are numerical values of X_0 , Y_0 , Z_0 , K_0 , L_0 , and M_0 , which would be written at the place indicated. Note that quantities pertaining to surfaces are written within the column for the corresponding surface; quantities pertaining to the space between surfaces are written in a break in the corresponding vertical line. The numbers running from 1 to 17 are the steps in the calculation in the order they are made. The steps, except (7) and (14), correspond to the 15 equations, listed in order of steps, in

Section 5.4.6.1. "Next step" indicates step No. 1 for the next ray segment. The table entries have been so chosen that a person using a desk calculator does not have to write down any number except those to be entered in the table.

SURFACE	0	1	2	3
c	c_0	c_1	c_2	
t	t_0	t_1	t_2	
n	n_0	n_1	n_2	
X	X_0	(9)		
Y	Y_0	(10)		
Z	Z_0	(11)		
K	K_0	(15)		
L	L_0	(16)		
M	M_0	(17)		
d_{-1}/n_{-1}	(1)	Next Step		
X_T		(2)		
Y_T		(3)		
H		(4)		
B		(5)		
$n_{-1} \cos I$		(6)		
$B + n_{-1} \cos I$		(7)		
A/n_{-1}		(8)		
$n \cos I'$		(12)		
Γ		(13)		
$c \Gamma$		(14)		

Table 5.1-Skew ray trace computing sheet.

SURFACE	0	1	2	3
c	0	0.25284872	-0.01473947	
t		-2.2	0.6	
n		1.0	1.62	1.0
X	1.48	1.48	1.43679417	
Y	0	-0.33445977	-0.29386784	
Z	0	0.30264162	-0.01585220	
K		0	-0.24330257	-0.25700617
L		0.17360000	0.22858306	0.23138586
M		0.98481625	1.58522985	0.93830084
d_{-1}/n_{-1}		-2.23391927	0.18758061	
X_T		1.48	1.43436116	
Y_T		-0.38780839	-0.29158202	
H		0.59186710	-0.03157802	
B		1.00183892	1.57910362	
$n_{-1} \cos I$		0.92413654	1.57871680	
$B + n_{-1} \cos I$		1.92597546	3.15782041	
A/n_{-1}		0.30730770	-0.00999994	
$n \cos I'$		1.57430250	0.93163659	
Γ		0.65016596	-0.64708020	
$c \Gamma$		0.16439363	0.00953762	

Table 5.2-Skew ray trace for three surfaces.

5.4.8 Numerical example.

5.4.8.1 Throughout the discussion of geometrical optics, lengthy explanations have been avoided by the inclusion of numerical examples showing the actual calculations. Table 5.2 is such an illustration. The calculations shown in this table can be made by an experienced person with a modern desk calculator without undue labor. In order for the calculations to be useful, at least six significant figures must be carried throughout. Since the introduction of the modern electronic computing machines, there is really very little justification for a human computer to carry out these calculations unless ray tracing is only done occasionally. The above equations can be programmed in a modern machine to make these calculations in less than one second per surface, with at least eight significant figures. The calculations shown in this and other numerical examples may not offer complete consistency in the number of significant figures for two reasons: (1) some were prepared

from automatic computer results where intermediate values were not available and had to be developed by hand computing; (2) others were prepared from a designer's work sheets where the aim was not eight-figure accuracy but only three or four-figure accuracy in which case the designer had merely entered results as they appeared on the hand calculator. No units appear in this and other numerical examples, because the equations are valid for any set of consistent units. As long as all lengths are in the same units, the numerical example will be correct for any units.

5.4.8.2 Some specific remarks should be made about Table 5.2. The initial data which are given to one, two, or three significant figures are assumed exact. From the initial system data it is apparent that we are considering a double convex lens, index 1.62, surrounded by air. The incident light first intersects the convex face of the lens. The lens thickness is about one quarter of the distance between lens and object surface, but no information is given (or needed for a ray trace) concerning the absolute magnitude of any distance. The object surface is to the right of the lens; therefore the object is virtual.

5.4.8.3 From the initial ray data we see that the (virtual) object point, that is the point towards which the ray is heading, is on the X axis, but not in the Y - Z plane. The initial ray is parallel to the Y - Z plane, hence $X_1 = X_0$. The ray is inclined upwards at an angle with the Z axis of 10° . The calculations indicate that the ray intersects both surfaces of the lens below the X - Z plane (because Y is negative), and intersects both surfaces at points "away from the reader" with respect to the Y - Z plane (because X is positive). The Z value at the first surface is positive because the curvature is positive; likewise the Z value at the second surface is negative.

5.5 SKEW RAY TRACE EQUATIONS FOR ASPHERIC SURFACES

5.5.1 General.

5.5.1.1 The discussion in Section 5.4 developed equations for, and demonstrated their use in, ray tracing procedures through spherical surfaces. Although spherical surfaces are still much easier to make, and hence are preferred by the lens maker, aspheric surfaces are readily handled by the lens designer who has access to an electronic computer. Aspheric surfaces afford the designer a great deal more latitude in the design, and in addition often permit better correction of aberrations. Aspheric surfaces are being used more and more, and their widespread use depends on inexpensive methods of production.

5.5.1.2 In the skew ray trace for spherical surfaces, it was convenient to effect the transfer from one physical surface to the next by introducing a non-physical tangent plane, and effecting the transfer in two steps. In the case of aspheric surfaces we introduce two non-physical surfaces, a plane and a sphere, both tangent to the physical aspheric surface at the optical axis. See Figure 5.10. The transfer between physical surfaces is now effected in three steps:

- (1) first surface to next tangent plane;
- (2) tangent plane to tangent sphere;
- (3) tangent sphere to physical (second) surface.

Steps (1) and (2) are carried out using the procedure already developed in Section 5.4.

5.5.2 Mathematical description of an aspheric surface.

5.5.2.1 We need to describe the aspheric surface in a way that indicates clearly its departure from the tangent sphere. This kind of description will not only be easily handled by the ray trace equations, but will also quickly and quantitatively show how close in form the aspheric is to the sphere.

5.5.2.2 In Paragraph 5.4.4.4 there is given an equation for Z; this quantity is called the sag of the sphere, an abbreviation of sagitta. Using $S^2 = X^2 + Y^2$, this equation is

$$Z = \frac{1}{c} \left[1 - (1 - c^2 S^2)^{1/2} \right].$$

By multiplying and dividing by $\left[1 + (1 - c^2 S^2)^{1/2} \right]$, we have

$$Z = \frac{c S^2}{1 + \sqrt{1 - c^2 S^2}}.$$

Because the shape of an aspheric surface (which is assumed to have rotational symmetry about the Z axis)

differs from that of the tangent sphere, the sag (Z) of the aspheric at any distance S from the axis may differ from the sag of the tangent sphere. This is indicated by expressing the difference in these two sags by a power series in S^2 . (The series is in powers of S^2 , and hence only even powers of S appear, because the aspheric has rotational symmetry about the Z axis.) The final expression for the sag is

$$Z = \frac{c S^2}{1 + \sqrt{1 - c^2 S^2}} + e S^4 + f S^6 + g S^8 + h S^{10} + O(S^{12})$$

5.5.2.3 Each of the numerical coefficients e , f , g , and h may be positive or negative. The term $O(S^{12})$ stands for the rest of the series, that is terms of order 12 and higher. In a numerical calculation, if the sag is given by this expression, $O(S^{12})$ would be assumed zero, and the calculations would involve only e , f , g , and h . The terms $e S^4$, $f S^6$, etc., are called deformation terms.

5.5.3 Initial data, and transfer from physical surface to next tangent sphere. Part of the transfer from one physical surface to the next has already been solved in Section 5.4. The initial ray data for the skew ray between aspheric surfaces is the same as given in Section 5.4.2, namely X_{-1} , Y_{-1} , Z_{-1} , K_{-1} , L_{-1} and M_{-1} . We determine the intersection of this ray with the non-physical sphere, tangent to the j th aspheric surface, by the procedure given in Sections 5.4.3 and 5.4.4. In other words we apply Equations (1), (2), (3), (9), (10), (7), (8), (4), (5) and (6) in that order. The only difference so far between this ray trace and the former is that in the previous case the sphere was a physical surface, while in the present case it is a purely fictitious surface. The equations do not know the difference between physical and non-physical surfaces; hence the same equations are used for both cases.

5.5.4 Transfer procedure, tangent sphere to aspheric surface.

5.5.4.1 In Paragraphs 5.4.4.4 and 5.4.4.5 an expression for $\frac{A}{n_{-1}}$ was derived using four equations, namely the equation for the sag, Z , of the sphere and Equations (4), (5), and (6). This value of $\frac{A}{n_{-1}}$ was then used in Equations (4), (5), and (6) to transfer from tangent plane to sphere. It would be perfectly possible to proceed similarly here. We would set up three equations, corresponding to (4), (5), and (6), but replacing A by $A + A'$. (See Figure 5.10). Using these three equations, and the equation for the

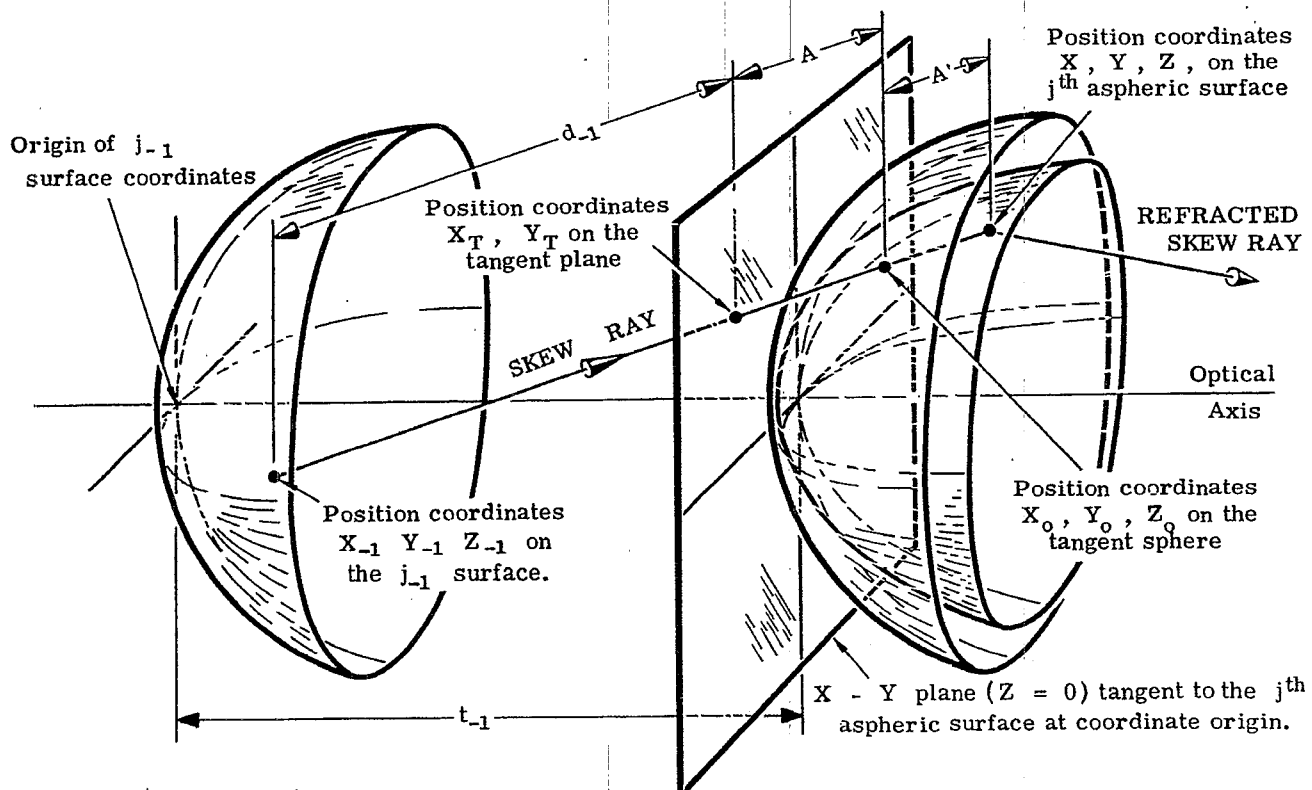


Figure 5.10 - Diagram of a skew ray in space between the j_{-1} surface and the j^{th} aspheric surface.

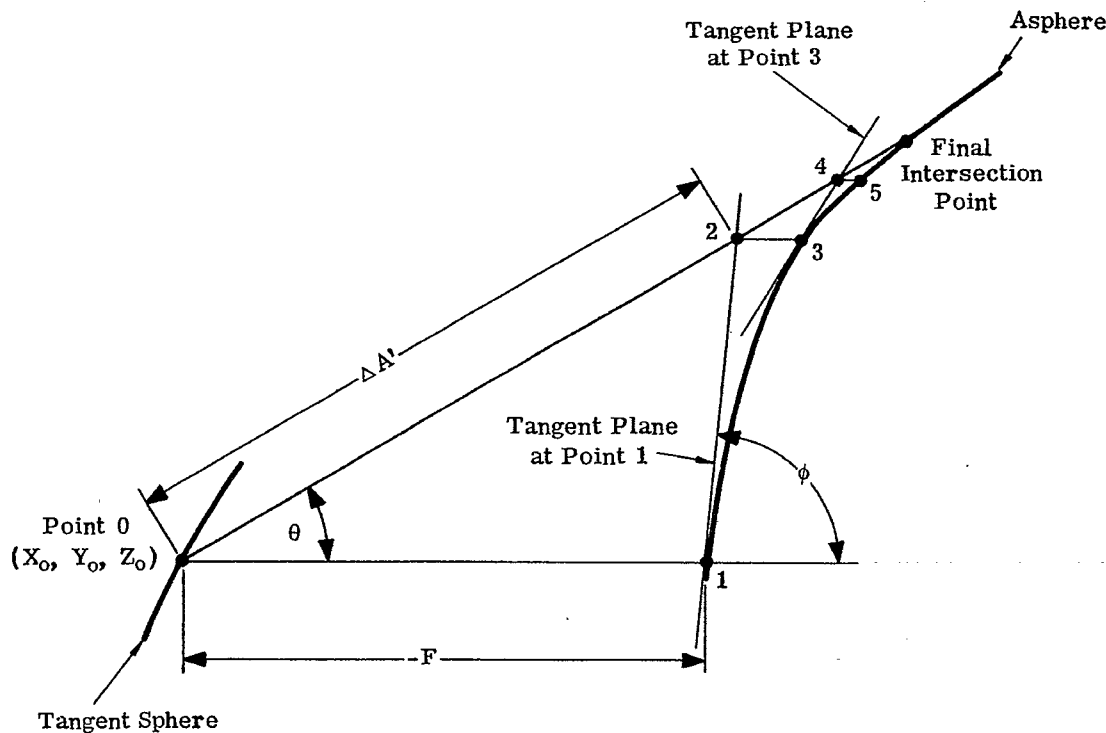


Figure 5.11- Step-wise approximations from tangent sphere intersection, point 0, to final intersection point.

sag of an aspheric surface, Paragraph 5.5.2.2, an expression for $\frac{A + A'}{n-1}$ could be found, and used to transfer directly from tangent sphere to aspheric. The resulting calculations are extremely involved and it is preferable to proceed otherwise.

5.5.4.2 The procedure to be employed makes use of the fact that transfer to the tangent sphere is fairly simple. The remaining transfer from tangent sphere to aspheric is effected in a step-wise procedure approaching the final intersection by successive approximations. The physical procedure is indicated in Figure 5.11; this figure represents the plane determined by the skew ray, and a line through this ray parallel to the Z axis.

5.5.4.3 Beginning at point 0, the intersection of the ray with the tangent sphere, the first approximation to the final point is point 1. Point 1 has the same X and Y values as point 0; its Z value differs from that of point 0 by the deformation terms evaluated at these particular values of X and Y. The second approximation is point 2, the intersection of the ray with a line tangent to the aspheric at 1. The tangent line is determined from the known coordinates of point 1 and the calculated curvature of the aspheric at point 1. Since the ray direction is known, its intersection with the tangent line, point 2, is determined. The procedure is now repeated. Point 3 has the same X and Y as point 2, and its Z value can be found from the deformation terms and the Z value at point 2. The fourth and fifth approximations are points 4 and 5, respectively. (The point 0, on the sphere, is correctly called the zeroth approximation to the final point.)

5.5.4.4 The various values of X, Y, and Z for points 0, 1, 2, ... will be referred to as X_n , Y_n , and Z_n where the n will stand for the order of approximation. Let us begin at any even-numbered point, that is a point on the ray; in practice, the calculations begin with point 0, but we wish to make the equations general so that n will stand for any even-numbered point. The next point, on the aspheric, will have coordinates X_n , Y_n , Z_m . Note that the X and Y values are the same as for the previous point. The S_n value now used to calculate the sag, Z_m , is

$$S_n^2 = X_n^2 + Y_n^2. \quad (16)$$

The change in Z, that is $Z_m - Z_n$, is the distance parallel to the Z axis between an even-numbered

point and an odd-numbered point. Calling this distance - F (see Figure 5.11), we have

$$F = Z_n - \left[\frac{c S^2}{1 + \sqrt{1 - c^2 S^2}} + e S^4 + f S^6 + g S^8 + h S^{10} \right]. \quad (17)$$

(The subscript n has been omitted. Henceforth all values of S are rigorously S_n .)

5.5.4.5 For notational purposes it is convenient to designate the square root in Equation (17) by W. Hence

$$W = \left[1 - c^2 S^2 \right]^{1/2}. \quad (18)$$

Referring to Figure 5.7, it can be seen that $W = \sqrt{1 - \sin^2 \alpha} = \cos \alpha$, where α is the angle between the normal to the surface and the optical axis. W therefore is the direction cosine, with respect to the Z axis, of a radius drawn from the sphere to the center.

5.5.4.6 From an odd-numbered point, whose coordinates we now know, we move along the tangent line to the ray. The coordinates of this new even-numbered point are X_{n+1} , Y_{n+1} , and Z_{n+1} . Calling the distance along the ray, between two even-numbered points, $\Delta A'$, we can write equations for the new coordinates similar to equations (4), (5), and (6). We have then

$$X_{n+1} = X_n + \frac{\Delta A'}{n_{-1}} K_{-1}, \quad (19)$$

$$Y_{n+1} = Y_n + \frac{\Delta A'}{n_{-1}} L_{-1}, \quad (20)$$

and

$$Z_{n+1} = Z_n + \frac{\Delta A'}{n_{-1}} M_{-1}. \quad (21)$$

We will consider the calculation of $\Delta A'$ presently. Once this is known, the new coordinates on the ray are known, and we repeat the calculations through two more steps until we get once again to the ray. This iteration procedure is continued until $\Delta A'/n_{-1}$ is less than any desired tolerance. In this manner we can approach the final point on the aspheric as closely as we choose.

5.5.4.7 The remaining problem in the transfer from tangent sphere to aspheric surface is the determination of $\Delta A'$. First we need the equation of the plane tangent to the aspheric surface at an odd-numbered point. From the equation for the sag of the surface, Paragraph 5.5.2.2, we can write

$$\psi(X, Y, Z) = Z - \left[\frac{c S^2}{1 + \sqrt{1 - c^2 S^2}} + e S^4 + f S^6 + g S^8 + h S^{10} \right] = 0,$$

where $\psi(X, Y, Z) = 0$ is the equation of the aspheric surface. Now a plane, tangent to the surface at the point X_n, Y_n, Z_m will coincide with the first approximation to the surface. Physically, if we restrict ourselves to points close to (X_n, Y_n, Z_m) the surface is a plane. To find the first approximation to the surface we expand $\psi(X, Y, Z)$ and keep only the zeroth and first order terms.

5.5.4.8 The equation of the tangent plane is then

$$\begin{aligned} \psi(X_n, Y_n, Z_m) + (X - X_n) \left[\frac{\partial \psi}{\partial X} \right]_{X_n, Y_n, Z_m} \\ + (Y - Y_n) \left[\frac{\partial \psi}{\partial Y} \right]_{X_n, Y_n, Z_m} + (Z - Z_m) \left[\frac{\partial \psi}{\partial Z} \right]_{X_n, Y_n, Z_m} = 0, \end{aligned}$$

where the first term is the zeroth order term, and the last three are the first order terms, in the expansion of $\psi(X, Y, Z)$. Using Equation (16) we have

$$\frac{\partial \psi}{\partial X} = \frac{\partial \psi}{\partial S} \frac{\partial S}{\partial X} = \frac{\partial \psi}{\partial S} \frac{X}{S},$$

$$\frac{\partial \psi}{\partial Y} = \frac{\partial \psi}{\partial S} \frac{Y}{S},$$

and

$$\frac{\partial \psi}{\partial Z} = 1.$$

Using the expression for $\psi(X, Y, Z)$ given in Paragraph 5.5.4.7, and Equation (18), we get

$$\frac{\partial \psi}{\partial S} = -\frac{S}{W} c - S \left[4eS^2 + 6fS^4 + 8gS^6 + 10hS^8 \right],$$

which can be simplified to $\frac{\partial \psi}{\partial S} = -\frac{S}{W} E$ by defining

$$E = c + W \left[4eS^2 + 6fS^4 + 8gS^6 + 10hS^8 \right]. \quad (22)$$

(If the deformation coefficients are small, E is approximately the curvature of the aspheric surface at the distance S_n from the optical axis.)

5.5.4.9 The equation for the sag of the aspheric surface given in Paragraph 5.5.2.2 is an equation for Z_m if $S^2 = X_n^2 + Y_n^2$. Because of this, $\psi(X_n, Y_n, Z_m)$ is zero, using the equation for ψ in Paragraph 5.5.4.7. The zeroth order term in the expansion is therefore zero. The equation of the plane becomes, using the above expressions for the partial derivatives,

$$-(X - X_n) \frac{X_n}{W} E - (Y - Y_n) \frac{Y_n}{W} E + (Z - Z_n) \frac{W}{W} + Z_n - Z_m = 0,$$

where we have separated the term $Z - Z_m$ into two terms. By Equation (17), $F = Z_n - Z_m$. We define here two quantities,

$$U = -XE', \quad (23)$$

$$V = -YE'. \quad (24)$$

5.5.4.10 With these substitutions the equation of the plane becomes

$$(X - X_n) U + (Y - Y_n) V + (Z - Z_n) W = -FW.$$

This equation holds for all values of X, Y , and Z , in particular X_{n+1}, Y_{n+1} , and Z_{n+1} . Instead of the difference $(X_{n+1} - X_n)$, we use $(\Delta A'/n_{-1})K_{-1}$ from Equation (19). Similarly, using Equations (20) and (21), and solving for $\Delta A'/n_{-1}$,

$$\frac{\Delta A'}{n_{-1}} = \frac{-FW}{K_{-1}U + L_{-1}V + M_{-1}W}. \quad (25)$$

From Figure 5.10, it can be seen that the distance, D_{-1} , along the ray is

$$D_{-1} = d_{-1} + A + A'. \quad (25a)$$

5.5.5 Refraction procedure at the aspheric surface.

5.5.5.1 Now that the intersection point, (X, Y, Z) , of the ray and the aspheric surface has been found, the refraction equation is used to determine the new direction of the ray. The procedure is basically the same as that used for refraction at spherical surfaces, discussed in Section 5.4.5. In that section Equation (11) was used to calculate $n \cos I'$, because $n_{-1} \cos I$ had already been calculated using Equation (7).

5.5.5.2 In the present case there is not yet a value for $n_{-1} \cos I$. To calculate this we use the fact that the cosine of an angle between two directed lines is equal to the sum of the products of their corresponding direction cosines. Since we are calculating $\cos I$, the two lines in question are the ray whose optical direction cosines are K_{-1}, L_{-1} , and M_{-1} , and the normal to the aspheric surface. Now the normal to the surface is just the normal to the tangent plane. The equation of this tangent plane is given in Paragraph 5.5.4.10, where X_n, Y_n , and Z_n are the coordinates of the final point on the ray, the intersection with the surface.

5.5.5.3 Given the equation of a plane, the direction cosines of the normal are proportional to the corresponding coefficients of X, Y , and Z . Hence the direction cosines of the normal are, in the usual order, $U/G, V/G$, and W/G , where G is a proportionality constant. Because the sum of the squares of the direction cosines is unity, we have

$$G^2 = U^2 + V^2 + W^2. \quad (26)$$

Using the direction cosines of the ray we get

$$\cos I = \frac{K_{-1}}{n_{-1}} \frac{U}{G} + \frac{L_{-1}}{n_{-1}} \frac{V}{G} + \frac{M_{-1}}{n_{-1}} \frac{W}{G} ,$$

which is rewritten in final form as

$$G n_{-1} \cos I = K_{-1} U + L_{-1} V + M_{-1} W . \quad (27)$$

5.5.5.4 Equation (11) is now used to determine $n \cos I'$. However, for calculation purposes, it is preferable to leave the G on both sides of the equation.

$$G n \cos I' = n \left[\left(G \frac{n_{-1}}{n} \cos I \right)^2 - G^2 \left(\frac{n_{-1}}{n} \right)^2 + G^2 \right]^{1/2} . \quad (28)$$

5.5.5.5 Returning to the equation in Paragraph 5.4.5.3, we write this vector equation as three scalar equations using the method of Paragraph 5.4.5.4. We get

$$K - K_{-1} = \Gamma \frac{U}{G} ,$$

$$L - L_{-1} = \Gamma \frac{V}{G} ,$$

and

$$M - M_{-1} = \Gamma \frac{W}{G} ,$$

because Γ is parallel to the normal to the surface and therefore has the same direction cosines. Introducing $P = \Gamma/G$, we have, using Equation (12),

$$P = (G n \cos I' - G n_{-1} \cos I) / G^2 . \quad (29)$$

Finally, K , L , and M are found from the equations,

$$K = K_{-1} + U P , \quad (30)$$

$$L = L_{-1} + V P , \quad (31)$$

and

$$M = M_{-1} + W P . \quad (32)$$

5.5.6 Summary of ray trace equations.

5.5.6.1 In the previous sections we have derived the equations used to trace a skew ray from a tangent sphere through the aspheric surface. For convenience we rewrite the equations in the order of use. The initial ray data are X_{-1} , Y_{-1} , Z_{-1} , K_{-1} , L_{-1} , and M_{-1} . The initial system data are t_{-1} , n_{-1} , c , and the deformation coefficients e , f , g , Final values to be determined are X , Y , Z , K , L , and M .

5.5.6.2 The position coordinates for the ray on the tangent sphere are calculated using the first ten equations listed in Section 5.4.6. Equations (16) through (32) are then used in the order listed below.

$$S_n^2 = X_n^2 + Y_n^2 , \quad (16)$$

$$W = \left[1 - c^2 S^2 \right]^{1/2} , \quad (18)$$

$$F = Z_n - \left[\frac{c S^2}{1 + \sqrt{1 - c^2 S^2}} + e S^4 + f S^6 + g S^8 + h S^{10} \right] , \quad (17)$$

$$E = c + W \left[4e S^2 + 6f S^4 + 8g S^6 + 10h S^8 \right] , \quad (22)$$

$$U = - X E , \quad (23)$$

$$V = - Y E , \quad (24)$$

$$\frac{\Delta A'}{n_{-1}} = \frac{-FW}{K_{-1}U + L_{-1}V + M_{-1}W} \quad (25)$$

$$X_{n+1} = X_n + \frac{\Delta A'}{n_{-1}} K_{-1} \quad (19)$$

$$Y_{n+1} = Y_n + \frac{\Delta A'}{n_{-1}} L_{-1} \quad (20)$$

$$Z_{n+1} = Z_n + \frac{\Delta A'}{n_{-1}} M_{-1} \quad (21)$$

$$G^2 = U^2 + V^2 + W^2 \quad (26)$$

$$G n_{-1} \cos I = K_{-1}U + L_{-1}V + M_{-1}W \quad (27)$$

$$G n \cos I' = n \left[\left(G \frac{n_{-1}}{n} \cos I \right)^2 - G^2 \left(\frac{n_{-1}}{n} \right)^2 + G^2 \right]^{1/2} \quad (28)$$

$$P = (G n \cos I' - G n_{-1} \cos I) / G^2 \quad (29)$$

$$K = K_{-1} + UP \quad (30)$$

$$L = L_{-1} + VP \quad (31)$$

and

$$M = M_{-1} + WP \quad (32)$$

5.5.6.3 The first ten of these equations are used in an iterative process until $\Delta A'/n_{-1}$ becomes as small as desired. The final values of U , V , and W are then used in the last seven equations (26) through (32). The final calculated values of X , Y , Z , K , L , and M become the initial ray data for the next calculation. These values, together with new system data, t , n , c_{+1} , and deformation terms, are used in a reapplication of the ray trace equations.

SURFACE	0	1	2	3
c	0	0.25284872	-0.01473947	
e		-0.005		
f		0.00001		
g		-0.0000005		
h		0		
t		-2.2	0.6	
n		1.0	1.62	1.0
X	1.48	1.48	1.44043943	
Y	0	-0.33905030	-0.29645624	
Z	0	0.27660001	-0.01594078	
K		0	-0.20481560	
L		0.17360000	0.22052072	
M		0.98481625	1.59179807	
d ₋₁ /n ₋₁		-2.23391927		
X _T		1.48		
Y _T		-0.38780839		
H		0.59186710		
B		1.00183892		
n ₋₁ cos I		0.92413654		
B + n ₋₁ cos I		1.92597546		
A/n ₋₁		0.30730770		
n cos I'				
Γ		Enter X _n Y _n Z _n		
c Γ		in Table 5.4		

Table 5.3 - Skew ray trace through an aspheric surface. Part of the calculations are shown in Table 5.4.

5.5.6.4 Because a spherical surface is a special case of an aspheric surface for which the deformation terms are zero, the ray trace equations for aspheric surfaces should easily reduce to those for spherical surfaces. We see, for the case of a sphere ($e = f = g = h = \dots = 0$),

$$E = c,$$

$$U = -Xc,$$

$$V = -Yc,$$

$$W = -(Zc - 1), \quad (\text{holds for aspheric also})$$

$$G = 1,$$

$$n_{-1} \cos I = -c \left[XK_{-1} + YL_{-1} + ZM_{-1} \right],$$

$$P = \Gamma,$$

and equations (30), (31), and (32) become identical with equations (13), (14), and (15).

ITERATION	1	2	3
X_n	1.48000000	1.48000000	1.48000000
Y_n	-0.33445977	-0.33905071	-0.33905030
Z_n	0.30264163	0.27659764	0.27659999
S_n^2	2.30226334	2.30535538	2.30535510
$1 - c^2 S^2$	0.85281060	0.85261290	0.85261290
W	0.92347745	0.92337040	0.92337040
$c / (1 + W)$	0.13145395	0.13146127	0.13146127
hS^2	0.00000000	0.00000000	0.00000000
$hS^4 + gS^2$	-0.00000115	-0.00000115	-0.00000115
$hS^6 + gS^4 + fS^2$	0.00002037	0.00002040	0.00002040
$hS^8 + gS^6 + fS^4 + eS^2$	-0.01146441	-0.01147976	-0.01147976
$hS^{10} + gS^8 + fS^6 + eS^4 + \frac{cS^2}{1+W}$	0.27624751	0.27660004	0.27660000
$-F$	-0.02639412	-0.00000238	-0.00000002
$-10 hS^2$	0.00000000	0.00000000	0.00000000
$-10 hS^4 - 8 gS^2$	0.00000921	0.00000922	0.00000922
$-10 hS^6 - 8 gS^4 - 6 fS^2$	-0.00011693	-0.00011706	-0.00011706
$10 hS^8 + 8 gS^6 + 6 fS^4 + 4 eS^2$	-0.04577605	-0.04583724	-0.04583723
$-E$	-0.21057557	-0.21052397	-0.21052398
U	-0.31165184	-0.31157548	-0.31157549
V	0.07042906	0.07137830	0.07137822
$K_{-1}U + L_{-1}V + M_{-1}W$	0.92168208	0.92174144	0.92174143
$-FW$	-0.02437438	0.000002198	0.000000018
$\Delta A'/n_{-1}$	-0.02644553	0.000002384	0.000000020
X	1.48000000	1.48000000	1.48000000
Y	-0.33905071	-0.33905030	-0.33905030
Z	0.27659764	0.27659999	0.27660001
G^2			0.95478704
$G_n \cos I'$			1.54937517
P			0.65735466
K			-0.20481560
L			0.22052072
M			1.59179807

Table 5.4 - Skew ray trace iteration and refraction calculations. The table shows three iterations.

5.5.7 Numerical example.

5.5.7.1 A numerical example is shown in Tables 5.3 and 5.4. The system data is the same as the example shown in Table 5.2, except for the addition of three deformation coefficients e , f , and g . The coefficient

h is specifically listed in both tables as zero. This avoids possible error in not being certain whether or not a coefficient was erroneously omitted. The initial ray data is identical with the previous example; hence the calculations and results for transfer to the tangent sphere are the same. Thus steps 1 through 11 (see Table 5.1) are identical, except for the location of the results of step 11. These are placed in Table 5.4 and are the initial data for the iteration process.

5.5.7.2 Table 5.4 shows the iteration process by which $(\Delta A'/n_{-1}) < 0.00001$; this represents the criterion, set up prior to the calculations, to determine when the iteration process is to be stopped. It is noticed that the first value of $\Delta A'/n_{-1}$ is negative, the second positive, the third almost zero. This oscillation about the target value (< 0.00001) is typical of the method of successive approximations. This method will be used in later sections where aberrations are discussed.

5.5.7.3 The final values of X, Y, and Z, shown just above the double line in Table 5.4 in the column 3, are entered in Table 5.3 in the place for steps 9, 10, and 11. (The entire iteration process gives the results for steps 9, 10, and 11 for an aspheric surface.) These values are now part of the initial ray data for the next surface. The refraction calculations at the aspheric surface are given in Table 5.4 below the double line, and use the final results found above. The values of K, L, and M are now entered in Table 5.3 as the results of steps 15, 16, and 17. They will be used as initial data for the next surface.

5.6 MERIDIONAL RAYS

5.6.1 Definition. A meridional ray is any ray lying in a plane containing the optical axis. A meridional ray will remain in the same plane throughout an entire centered system. For this reason, the tracing of meridional rays is a two dimensional problem, while the tracing of skew rays, which do not lie in a plane containing the optical axis, is a three dimensional problem.

5.6.2 Use of skew ray trace equations. The skew ray formulae given in Sections 5.4 and 5.5 are designed for use on modern automatic computing machines. However, they are in a form which can be used with relative ease - for skew rays - on a desk calculator. Extensive skew ray tracing, which is essential in order to make a complete analysis of a lens system, should be done on a computing machine. In the preliminary design of a lens system it is usually convenient to trace a few selected meridional rays. These are often traced by hand. If the object point has coordinates $(X_o = 0, Y_o, Z_o)$ and the ray pierces the first surface at coordinates $(X_1 = 0, Y_1, Z_1)$ the ray is meridional and will remain in the YZ-plane all the way to the image surface. Meridional rays can be traced using the skew ray formulae given in Sections 5.4 and 5.5 by setting $X = 0$ and $K = 0$.

5.6.3 Meridional ray trace, spherical surfaces.

5.6.3.1 Meridional ray tracing can be done for spherical surfaces by using Equations (1) through (10), followed by either Equations (11) through (15) or Equations (16) through (32). For meridional rays, Equations (1) through (10) reduce to the following eight equations, in the order used:

$$\frac{d_{-1}}{n_{-1}} = (t_{-1} - Z_{-1}) \frac{1}{M_{-1}}, \quad (1)$$

$$Y_T = Y_{-1} + \frac{d_{-1}}{n_{-1}} L_{-1}, \quad (2)$$

$$H = c Y_T^2, \quad (9a)$$

$$B = M_{-1} - c Y_T L_{-1}, \quad (10a)$$

$$n_{-1} \cos I = n_{-1} \left[\left(\frac{B}{n_{-1}} \right)^2 - cH \right]^{1/2}, \quad (7)$$

$$\frac{A}{n_{-1}} = \frac{H}{B + n_{-1} \cos I}, \quad (8)$$

$$Y = Y_T + \frac{A}{n_{-1}} L_{-1}, \quad (5)$$

and

$$Z = \frac{A}{n_{-1}} M_{-1}. \quad (6)$$

Only eight equations are needed, the other two being $X_T = X = 0$. These eight equations trace a meridional ray from any surface to the next spherical surface.

5.6.3.2 Refraction at the spherical surface may be calculated by applying Equations (11), (12), (14), and (15) as written. Equation (13) becomes $K = 0$. (This procedure is referred to as the short form). On the other hand Equations (16) through (32) may be used. These are reduced to the following seven equations, in the order used:

$$W = \left[1 - c^2 Y^2 \right]^{1/2}, \quad (18a)$$

$$V = -Yc, \quad (24a)$$

$$n_{-1} \cos I = L_{-1} V + M_{-1} W, \quad (27a)$$

$$n \cos I' = n \left[\left(\frac{n_{-1}}{n} \cos I \right)^2 - \left(\frac{n_{-1}}{n} \right)^2 + 1 \right]^{1/2}, \quad (11)$$

$$\Gamma = n \cos I' - n_{-1} \cos I, \quad (12)$$

$$L = L_{-1} - Yc\Gamma, \quad (14)$$

and

$$M = M_{-1} - W\Gamma. \quad (32a)$$

Only seven equations are needed, the other ten being $S_n = Y_n$, $E = c$, $G = 1$, $Y_{n+1} = Y_n$, $Z_{n+1} = Z_n$, and $F = U = \Delta A' = X_{n+1} = K = 0$.

5.6.4 Meridional ray trace, aspheric surfaces. For meridional rays and aspheric surfaces, after applying the eight equations given in Paragraph 5.6.3.1, the Equations (16) through (32) are used. These reduce to the following thirteen equations, in the order used:

$$W = \left[1 - c^2 Y^2 \right]^{1/2}, \quad (18a)$$

$$F = Z_n - \left[\frac{c Y^2}{1 + W} + e Y^4 + f Y^6 + g Y^8 + h Y^{10} \right], \quad (17a)$$

$$E = c + W \left[4e Y^2 + 6f Y^4 + 8g Y^6 + 10h Y^8 \right], \quad (22a)$$

$$V = -YE, \quad (24)$$

$$\frac{\Delta A'}{n_{-1}} = \frac{-FW}{L_{-1} V + M_{-1} W}, \quad (25b)$$

$$Y_{n+1} = Y_n + \frac{\Delta A'}{n_{-1}} L_{-1}, \quad (20)$$

$$Z_{n+1} = Z_n + \frac{\Delta A'}{n_{-1}} M_{-1}, \quad (21)$$

$$G^2 = V^2 + W^2, \quad (26a)$$

$$G n_{-1} \cos I = L_{-1} V + M_{-1} W, \quad (27a)$$

$$G n \cos I' = n \left[\left(G \frac{n_{-1}}{n} \cos I \right)^2 - G^2 \left(\frac{n_{-1}}{n} \right)^2 + G^2 \right]^{1/2}, \quad (28)$$

$$P = (G n \cos I' - G n_{-1} \cos I) / G^2, \quad (29)$$

$$L = L_{-1} + VP, \quad (31)$$

and

$$M = M_{-1} + WP. \quad (32)$$

Only 13 equations are needed, the other four being $S_n = Y_n$, and $U = X_{n+1} = K = 0$.

5.6.5 Simplified meridional ray trace, spherical surfaces.

5.6.5.1 There are many other methods, involving different parameters, which are commonly used to trace meridional rays. One such method specifies the angle the ray makes with the optical axis, and the perpendicular distance from the center of curvature of the surface to the ray. Figure 5.12 indicates the two

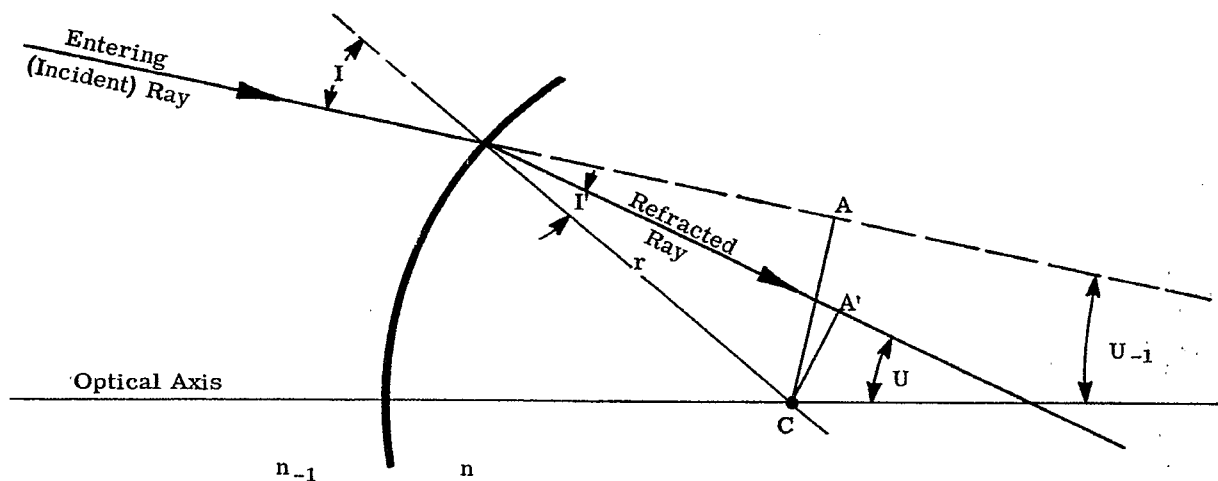


Figure 5.12—Ray tracing by the PR method.

quantities specified, U_{-1} and CA . The corresponding quantities, U and CA' , specify the refracted ray. This diagram involves the angles U and U_{-1} , called the slope angles. We will use a convention for the sign of a slope angle similar to that for incidence, reflection, and refraction angles. (See Section 2.2.2). If the ray must be rotated clockwise through the acute angle to bring it into coincidence with the optical axis the angle is called positive. Both U and U_{-1} are negative as drawn.

5.6.5.2 The following equations are readily derived from the figure: *

$$\sin I = \frac{CA}{r},$$

and

$$\sin I' = \frac{CA'}{r}.$$

Therefore from Snell's law

$$n_{-1} \sin I = \frac{CA n_{-1}}{r} = \frac{CA' n}{r} = n \sin I'.$$

By definition:

$$P = CA n_{-1} = CA' n,$$

and

$$R = \frac{1}{n_{-1} r}, \quad R' = \frac{1}{n r}.$$

(Because of these two definitions, this method is referred to as the PR method.)

* The notation used in this simplified ray trace must not be confused with the skew ray formulae. There has been no attempt to avoid duplication of symbols.

The refraction equations then become

$$\sin I = PR, \quad (33)$$

$$\sin I' = PR', \quad (34)$$

and

$$U = U_{-1} - (I - I'). \quad (35)$$

The value of P is transferred from one surface to the next by the following equation:

$$P_{+1} = P - (r - r_{+1} - t) n \sin U. \quad (36)$$

5.6.5.3 Equation (36) is seen to follow from Figure 5.13. We have

$$P_{+1} = C_{+1} A_{+1} n = C A' n + C C_{+1} n \sin U,$$

because U is negative. The distance $CC_{+1} = t - r + r_{+1}$, and Equation (36) follows by rearrangement. The above ray tracing equations, (33) through (36), require a minimum of calculation and are ideal for hand computing. If several rays are to be calculated it is worth while to precalculate the lens constants R , R' , and $n(r - r_{+1} - t)$.

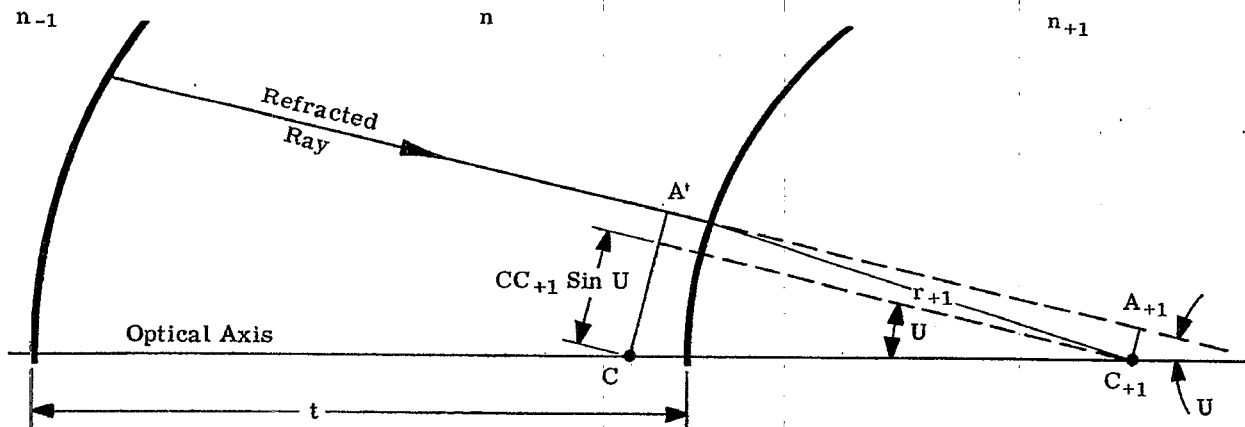


Figure 5.13 - Transfer procedure for the PR method.

5.6.5.4 A numerical example is shown in Table 5.5. The numbers above the double line are either given, such as r , t , and n , or are precalculated such as R , $-R'$, and $(r - r_{+1} - t)n$. The P below the line, surface 1, is calculated from initial ray data, CA . All other values in the table, below the double line, are calculated using Equations (33) through (36). The problem of finding angles I and I' from their sines, in order to use Equation (36), is discussed below.

SURFACE	1	2	3
r	19.23	-64.25	13.51
t	0.8	0.05	
n	1.51017	1	
R	0.0520021	-0.0103063	
-R'	-0.0344346	0.0155642	
n (r - r ₊₁ - t)	124.861	-77.81	
P	3.330000	10.708028	1.703612
sin I	0.173167	-0.110360	
-sin I'	-0.114667	0.166662	
U	0	-0.059124	-0.115983

Table 5.5 - Numerical example of ray tracing by the PR method.

5.6.5.5 One should note that the above formula, (36), cannot be used to transfer from a plane surface wherein $r \rightarrow \infty$, or to a plane surface wherein $r_{+1} \rightarrow \infty$. To deal with a plane, the procedure is to calculate the distance from the pole of the plane surface to the ray; see Figure 5.14.

Let

$OA n_{-1} = Q$ for the entering ray, and

$OA' n = Q'$ for the refracted ray.

Then, because $U_{-1} = I$, and $U = I'$, we have,

$$Q' = Q \frac{\tan U_{-1}}{\tan U}.$$

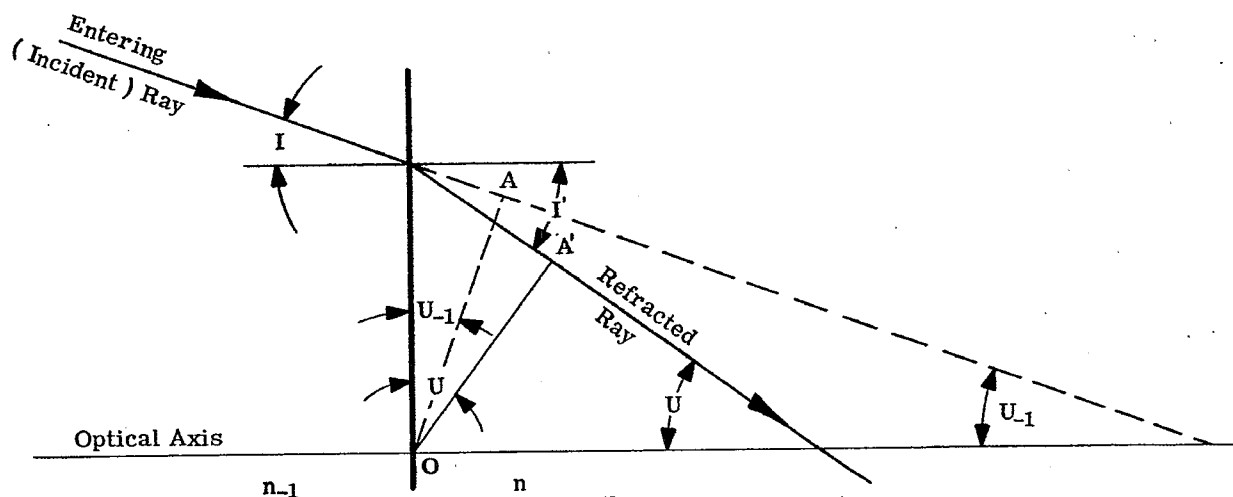


Figure 5.14 - Method of transfer for a plane surface.

To transfer from a spherical surface r , to a plane surface, $r_{+1} = \infty$, we use

$$Q_{+1} = P - (r - t) n \sin U.$$

To transfer from a plane surface, $r = \infty$, to a spherical surface, r_{+1} , the equation is

$$P_{+1} = Q' - (-r_{+1} - t) n \sin U.$$

5.6.5.6 To trace meridional rays through systems involving plane surfaces, Equations (33) through (36) are used until the plane surface is encountered. To transfer to a plane surface from a spherical one or vice versa, use one of the transfer equations in Paragraph 5.6.5.5. A transfer between two plane surfaces is calculated using

$$Q_{+1} = Q + t n \sin U.$$

Refraction at a plane surface is calculated using

$$\sin U = \frac{n_{-1}}{n} \sin U_{-1},$$

and

$$Q' = Q \frac{\tan U_{-1}}{\tan U},$$

where Q is either specified initially, calculated from initial data, or gotten by transfer from a previous surface. The calculations for plane surfaces are put into the same table (Table 5.5) as used for spherical surfaces. The values of $\sin U_{-1}$ and $\sin U$ are written opposite $\sin I$ and $\sin I'$ (which they equal respectively), and the values of Q and Q' are written opposite P . (The tangents need not be written down.)

5.6.5.7 One difficulty with the above formulation, Equation (36), is that if r_{+1} becomes large, but remains finite, P_{+1} becomes equal to the difference between a relatively small and a relatively large number. Hence unless a large number of significant figures are used for n , $\sin U$, and the coefficient of these terms, the value of P_{+1} will be independent of P . In doing hand computing one can readily notice this loss of precision. If this occurs, it is necessary to resort to other formulae, or reshape the lens so that the surface becomes plane. Another difficulty with Equation (36) arises if U becomes small, but remains finite. In this case the ray is almost parallel to the optical axis, and P_{+1} becomes equal to the difference of two nearly equal numbers. Hence unless both numbers are known to a large number of significant figures, the value of P_{+1} is quite inaccurate. In case the use of Equation (36) becomes difficult, the formulae given in Section 5.6.3 should be used.

5.6.5.8 In using the above equations it is necessary to convert sines to angles and to tangents, and to convert angles to sines. Tables are given in the Appendix. The tables convert from sine or tangent to the argument in radians and vice versa. They are designed for six place accuracy, and intervals are chosen for ease of interpolation. The first three digits of the function can always be found in the table and the last three digits are always multiplied by the interpolation constant and the product added to the tabular value. Interpolation therefore requires no mental arithmetic, and the process becomes completely automatic. By paying attention to such details a good human computer can trace rays through a lens at a speed of 40 to 60 seconds a surface. This method, in spite of its limitations, is an extremely useful method for hand computing meridional rays.

5.7 GRAPHICAL RAY TRACING PROCEDURE

5.7.1 Explanation of the method.

5.7.1.1 Rays may be traced graphically by means of a simple construction. The left side of Figure 5.15 shows a portion of two concentric circles whose radii are proportional to the indices n_{-1} and n . On the right side of the figure is shown the surface separating media of index n_{-1} and n . The angle of the refracted ray is determined from the diagram on the left. From this diagram $n_{-1} \sin I = n \sin I'$; thus, the construction solves Snell's law. Reference to Paragraph 5.4.5.3 will disclose that this is merely the graphical solution of the vector method.

5.7.1.2 The detailed procedure for tracing a ray is as follows. Draw a line through the center of the two circles parallel to the incident ray. Draw a line, parallel to the radius of curvature, through the intersection of the first line and the circle corresponding to the index of the object space. The line through the

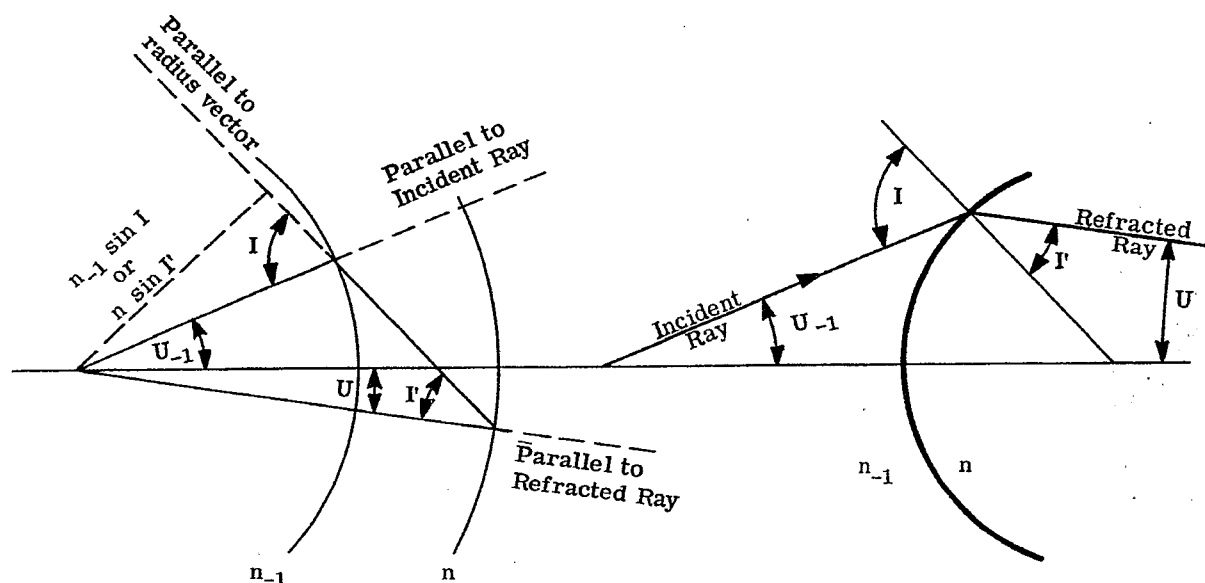


Figure 5.15- The method of tracing rays graphically.

center of the two concentric circles and the intersection of the second line with the other circle is the refracted ray. The incident and refracted rays can be drawn on the right hand diagram, but this is not necessary. The two diagrams may be superposed by placing the center of the concentric circles on the incident ray a distance n_{-1} (arbitrary units) to the left of the incidence point. This procedure makes unnecessary the drawing of the circle for n_{-1} , or the drawing of two lines each for the incident ray and the radius vector. The remainder of the construction is as given above.

5.7.2 Example using an air-spaced doublet. Figure 5.16 shows the graphical ray trace for a ray which is initially parallel to the axis (ray a). It is seen that the first surface of the second lens is a diverging surface; the other three surfaces are converging, because the ray is bent toward the optical axis. By measurement of the radii of the concentric circles, we see that $n_1 = 1.5$ and $n_2 = 1.7$. This combination of a converging crown lens, followed by a diverging flint lens is typical of a type of achromatic telescope objective. These lenses will be studied in detail in Section 11.

5.8 DIFFERENTIAL RAY TRACING PROCEDURE

5.8.1 Meaning of a differentially traced ray.

5.8.1.1 In the previous sections equations have been developed for tracing a general ray (skew or meridional) through a general surface having rotational symmetry. Once such a ray has been traced through the system, we have a baseline from which to find the path of neighboring rays. A differentially traced ray, sometimes referred to as a close ray, is a ray differing from the originally traced ray by small, first order quantities. This means that the change in direction cosines, dK_{-1} , dL_{-1} , dM_{-1} , and the change in the coordinates of the intersection point, dX , dY , dZ , are first order differentials.

5.8.1.2 The tracing of one ray gives us information about the one intersection point of that ray with the image surface. The tracing of several neighboring rays gives us their intersection points and hence information about the structure of the image formed by these rays. In addition to this useful information, differentially traced rays are generally easier to calculate than a single, general ray. Because of these advantages, the concepts and procedures of differential ray tracing are important.

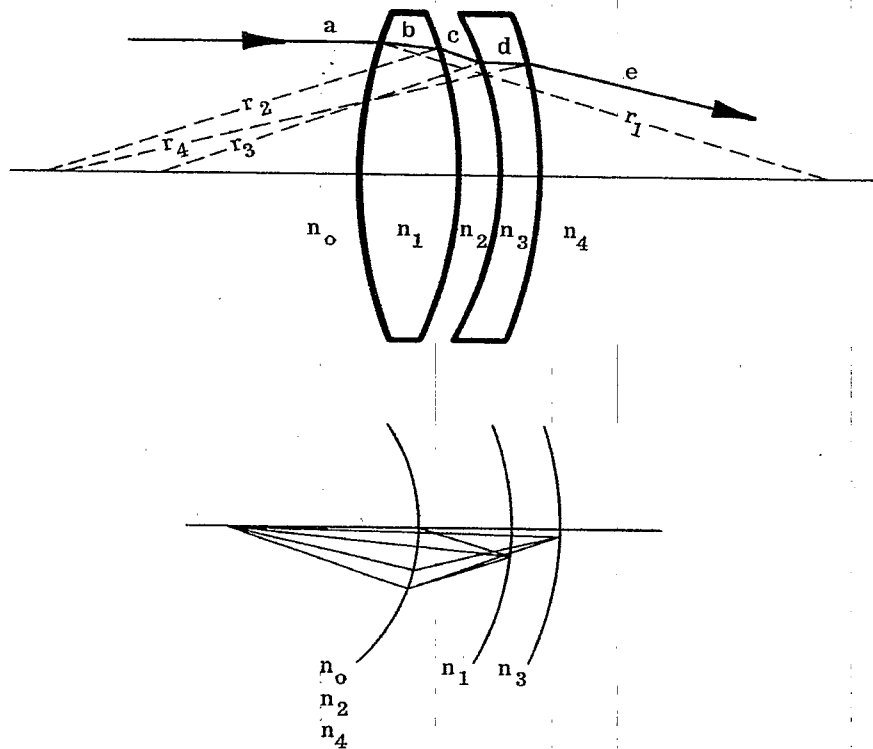


Figure 5.16 - Graphical ray trace of a doublet.

5.8.2 Differentially traced skew ray.

5.8.2.1 Once a skew ray has been traced through a lens system it is possible to trace the path of a ray differentially displaced from it. The skew ray trace provides the values of X , Y , Z on each surface, and K , L , M between surfaces. The values of X , Y , Z on adjacent surfaces are linked by the transfer equations

$$X = X_{-1} + \frac{D_{-1}}{n_{-1}} K_{-1}, \quad (37)$$

$$Y = Y_{-1} + \frac{D_{-1}}{n_{-1}} L_{-1}, \quad (38)$$

and

$$Z = Z_{-1} - t_{-1} + \frac{D_{-1}}{n_{-1}} M_{-1}, \quad (39)$$

where D_{-1} , given by Equation (25a), is the geometrical distance along the skew ray between the two surfaces. These equations follow from Paragraph 5.4.3.2 applied to any two surfaces.

5.8.2.2 A neighboring ray, in the sense of Paragraph 5.8.1.1, will have slightly different coordinates on the j th surface. The differences, dX , dY , and dZ are found by differentiating Equations (37), (38), and (39). We have

$$dX = dX_{-1} + \frac{D_{-1}}{n_{-1}} dK_{-1} + K_{-1} d\left(\frac{D_{-1}}{n_{-1}}\right), \quad (40)$$

$$dY = dY_{-1} + \frac{D_{-1}}{n_{-1}} dL_{-1} + L_{-1} d\left(\frac{D_{-1}}{n_{-1}}\right), \quad (41)$$

and

$$dZ = dZ_{-1} + \frac{D_{-1}}{n_{-1}} dM_{-1} + M_{-1} d\left(\frac{D_{-1}}{n_{-1}}\right). \quad (42)$$

These equations may be referred to as differential transfer equations, in that they are used to calculate the change in coordinates. The changes in coordinates at the previous surface have been determined by the previous application of these equations; the changes in optical direction cosines are calculated by differential refraction equations discussed below. The last term, involving the change in total ray length, must also be calculated. The remaining equations will first be derived; then their order of use will be summarized.

5.8.2.3 The procedure used to derive an equation for $d\left(\frac{D_{-1}}{n_{-1}}\right)$ will be quite similar to that used to derive an equation for $\frac{A}{n_{-1}}$. (See Paragraphs 5.4.4.3 - 5.4.4.5). In that case we used four equations, Equations (4), (5), (6) and the equation for the sag, Paragraph 5.4.4.4. These four equations were solved simultaneously for $\frac{A}{n_{-1}}$. The equation for the sag, Z , is the equation for the surface, in that case the sphere. Because the intersection point must lie on the surface, this equation is called an equation of constraint. In the present case, the four equations to be used are Equations (40), (41), (42) and the differential equation of constraint.

5.8.2.4 Although the physical j th surface is a general surface of revolution, this surface is replaced by the plane, tangent at the intersection point. The reason this must be done is that we have restricted the change in coordinates to first order differentials; as one moves away from a point on a surface by distances of the order of first differentials, the motion is constrained to the plane tangent to the surface. The equation of the tangent plane is given in Paragraph 5.5.4.10. Differentiating this to obtain the differential equation of constraint we have

$$UdX + VdY + WdZ = 0.$$

We now substitute into this equation the values of dX , dY , and dZ given by Equations (40), (41), and (42). Collecting terms, and using Equation (27), we have

$$d\left(\frac{D_{-1}}{n_{-1}}\right) = \frac{U(dX_{-1} + \frac{D_{-1}}{n_{-1}} dK_{-1}) + V(dY_{-1} + \frac{D_{-1}}{n_{-1}} dL_{-1}) + W(dZ_{-1} + \frac{D_{-1}}{n_{-1}} dM_{-1})}{-G n_{-1} \cos I} \quad (43)$$

5.8.2.5 Using Equation (43), and then Equations (40), (41), and (42), we will have completed the transfer of the differentially traced ray. The differential refraction equations, now to be derived, will be used to calculate dK , dL , and dM . Differentiating Equations (30) to (32) gives

$$dK = dK_{-1} + PdU + UdP, \quad (44)$$

$$dL = dL_{-1} + PdV + VdP, \quad (45)$$

and

$$dM = dM_{-1} + PdW + WdP. \quad (46)$$

5.8.2.6 In differentiating the equation for the tangent plane we kept U , V , and W constant and thereby obtained the differential equation of constraint. Physically this means that at any point on this tangent plane the ratio of the direction cosines of the normal, $U : V : W$, is the same as at any other point. (See Paragraph 5.5.5.3). Justifiably it may be asked why U , V , W were not held constant in differentiating Equations (30), (31), and (32). The answer is that though the tangent plane and surface differ by second order differentials, at the new intersection point, the normals to the two tangent planes, erected at the two intersection points, have direction cosines differing by first order differentials. Hence, since refraction involves the normal at the intersection point, dU , dV , and dW are not necessarily zero in Equations (44), (45), and (46).

5.8.2.7 Differentiating Equations (23), (24), and (18), we get

$$dU = -XdE - EdX, \quad (47)$$

$$dV = -YdE - EdY, \quad (48)$$

and

$$dW = -\frac{c^2}{W} (XdX + YdY). \quad (49)$$

dE may be found by differentiating Equation (22), remembering that

$$dE = \frac{\partial E}{\partial X} dX + \frac{\partial E}{\partial Y} dY + \frac{\partial E}{\partial Z} dZ,$$

thus

$$dE = - \left[\frac{c - E}{W} + \frac{2W^2}{c^2} (4e + 12fS^2 + 24gS^4 + 40hS^6) \right] dW. \quad (50)$$

5.8.2.8 The one remaining problem is the determination of dP to be used in Equations (44), (45), and (46). This is done by the same method used to derive Equation (43). The four equations used are Equations (44), (45), (46), and a differential equation of constraint. In each case, the fourth equation involves the differentials on the left-hand side of the first three equations. Because the sum of the squares of the direction cosines of a given line is unity, we have

$$KdK + LdL + MdM = 0$$

as the differential equation of constraint. Substituting Equations (44), (45), and (46) into this constraint, and remembering Equation (27), we have

$$dP = \frac{K(dK_{-1} + PdU) + L(dL_{-1} + PdV) + M(dM_{-1} + PdW)}{-Gn \cos I'} \quad (51)$$

5.8.2.9 We can now summarize the calculations, in the order made, used in tracing a differentially traced skew ray through aspheric surfaces. In addition to the results for the skew ray, available from tables such as 5.3 and 5.4, the initial ray data of the neighboring ray must be given. That is dX_{-1} , dY_{-1} , dZ_{-1} , dK_{-1} , dL_{-1} , and dM_{-1} must be specified. The following equations are used in the order given here: (43), (40), (41), (42), (49), (50), (47), (48), (51), (44), (45), and (46). With an automatic computer it does not seem to be worthwhile to trace close skew rays since the regular skew rays can be traced so rapidly. However for hand computing the close skew ray trace is a very valuable tool.

5.8.3 Differentially traced meridional ray. At first sight it might appear that it would take as much time to trace a differentially traced ray as a skew ray. Actually the equations are simple and no square roots are involved. An interesting application of the above equations occurs in connection with meridional rays. Assume a meridional ray has been traced from an object point ($X_o = 0$, Y_o , Z_o) through the lens system; let us now trace a ray from the same object point which will be differentially displaced by the amount dK_o . We also assume that $dL_o = 0$. Since

$$KdK + LdL + MdM = 0,$$

and the differential ray is to be traced around a meridional ray $K_o = 0$, $dM_o = 0$. (If originally we had assumed $dM_o = 0$, then it would follow that $dL_o = 0$). Equation (43) shows that $d(D_o/n_o) = 0$.

From Equations (40), (41) and (42),

$$dX_1 = \frac{D_o}{n_o} dK_o,$$

$$dY_1 = 0,$$

and

$$dZ_1 = 0.$$

Careful inspection of Equations (47) to (51) shows that,

$$dU = -EdX,$$

$$dV = 0,$$

$$dW = 0,$$

$$dE = 0,$$

and

$$dP = 0.$$

Substitution into Equation (44) gives the relation

$$dK_1 = dK_0 - \left[\frac{G n_1 \cos I' - G n_0 \cos I}{G^2} \right] E_1 dX_1.$$

It then follows, since $d\left(\frac{D_1}{n_1}\right) = 0$, that

$$dX_2 = dX_1 + \frac{D_1}{n_1} dK_1,$$

and

$$dK_2 = dK_1 - \left[\frac{G n_2 \cos I' - G n_1 \cos I}{G^2} \right] E_2 dX_2.$$

The close meridional ray may be traced through the system by successive application of the equations

$$dX = dX_{-1} + \frac{D_{-1}}{n_{-1}} dK_{-1} \quad (52)$$

and

$$dK = dK_{-1} - \left[\frac{G n \cos I' - G n_{-1} \cos I}{G^2} \right] E dX. \quad (53)$$

5.8.4 The Coddington equations.

5.8.4.1 The above two equations, (52) and (53), apply to a general surface having rotational symmetry. In case the surface is spherical, Equation (53) is simplified since $E = c$ and $G = 1$. If the close ray has $dL_0 = dM_0 = 0$, as in the above example, and if the traced meridional ray and the close ray intersect to form an image, these two rays obey one of the Coddington equations, namely,

$$\frac{n_0}{D_0} + \frac{n_1}{D_1} = c (n_1 \cos I' - n_0 \cos I).$$

Because the close ray was shifted in a way that resulted in $dY_1 = dZ_1 = 0$, the shift of the intersection point occurred parallel to the X axis, in other words in the sagittal plane. The resulting focus is referred to as the sagittal focus, or the skew focus, because the close ray is actually a skew ray.

5.8.4.2 The above Coddington equation can be derived from Equations (52) and (53) applied to a spherical surface, ($E = c$ and $G = 1$). Because we are dealing with a single object and single image point, $dX_0 = dX_2 = 0$. Applying these two equations to Equation (52), we have

$$dX_1 = \frac{D_0}{n_0} dK_0 = - \frac{D_1}{n_1} dK_1.$$

Using Equation (53), for a spherical surface,

$$dK_1 = dK_0 - (n_1 \cos I' - n_0 \cos I) c dX_1,$$

and, expressing dK_0 in terms of dK_1 , and dX_1 in terms of dK_1 , we get

$$dK_1 = - \frac{D_1}{n_1} \frac{n_0}{D_0} dK_1 + (n_1 \cos I' - n_0 \cos I) c \frac{D_1}{n_1} dK_1.$$

Simplification gives the above Coddington equation.

5.8.4.3 Instead of shifting the ray in a plane perpendicular to the meridional plane, the ray could have been shifted in the meridional plane. In this case, $dK = dX = 0$. In a manner similar to that used in Section 5.8.3, ray trace equations for dY and dL can be derived, corresponding to Equations (52) and (53). (We do not need specific equations for dZ and dM , because these are proportional to dY and dL respectively). For a single image to be formed by two close rays from a single object point, $dY_0 = dY_2 = 0$. The final result is the second Coddington equation involving the meridional or tangential focus,

$$\frac{n_0 \cos^2 I}{D_0} + \frac{n_1 \cos^2 I'}{D_1} = c (n_1 \cos I' - n_0 \cos I).$$

5.9 PARAXIAL RAYS

5.9.1 The paraxial ray concept. The previous section on differentially traced meridional rays provides a good way to introduce the concept of paraxial ray tracing and the meaning of paraxial rays. A ray passing directly along the optical axis of the system is a perfectly good ray to use as a base from which to trace a close, neighboring ray. Such a ray, differentially traced with respect to the optical axis, is a paraxial ray. Physically, paraxial rays are the rays that get through the system as the aperture of each lens, centered concentrically with respect to the optical axis, becomes very small. Because paraxial rays are fairly easy to visualize, and because the ray tracing equations become quite simple for these rays, the usefulness of paraxial rays in the preliminary design of optical elements cannot be overemphasized.

5.9.2 Ray trace equations.

5.9.2.1 For a ray coinciding with the optical axis, $\cos I$ and $\cos I'$ will be exactly equal to 1 on every surface and $D_{-1} = t_{-1}$, so Equations (52) and (53) become

$$dX = dX_{-1} + \frac{t_{-1}}{n_{-1}} dK_{-1} \quad (54)$$

and

$$dK = dK_{-1} - (n - n_{-1}) c dX. \quad (55)$$

Therefore a ray may be traced differentially close to the optical axis by applying the above equations. Since the original ray was the optical axis, there is no distinction between the X and Y axes, and these equations apply equally well for a close ray in the YZ plane. For such a ray, replace dX by dY and dK by dL, for each part of the system. It should be noted that these equations hold for aspheric as well as spherical surfaces. Mathematically this is so because for the optical axis, $X = Y = 0$; hence by Equations (18) and (26), $W = 1 = G$, and by Equation (22), $E = c$. Physically the aspheric and the sphere are tangent at the optical axis and have the same curvature; hence a ray close to the axis intersects a surface of curvature c.

5.9.2.2 Paraxial ray calculations will be used so extensively to build up an understanding of optical systems, that a special notation will be used to refer to paraxial data. It is customary to use lower case letters for paraxial rays. Equations (54) and (55) will be written for a ray in the YZ plane and become

$$y = y_{-1} + \frac{t_{-1}}{n_{-1}} (n_{-1} u_{-1}), \quad (56)$$

and

$$nu = n_{-1} u_{-1} + yc (n_{-1} - n). \quad (57)$$

The differentials have been replaced by small letters indicating paraxial ray data. One can see that dY has been replaced by y, indicating a small displacement perpendicular to the optical axis. dL, which replaces dK for a paraxial ray in the YZ plane, is the change in the optical direction cosine of the originally traced ray. Since the original ray is the axial ray, and the original $L = 0$, dL = new value of $L = n \cos \beta$, where β is the angle between the ray and the Y axis. Instead of $\cos \beta$, we can use $\sin U$, the angle between the ray and Z axis. Therefore $dL = n \sin U$. But U is a small angle, and we replace the $\sin U$ by U, the first order approximation. (See Section 5.11). Hence $dL = nU$, and using small letters, $dL = nu$. We see here why the term paraxial ray optics and first order optics are synonymous.

5.9.3 The use of finite angles and heights for paraxial rays.

5.9.3.1 Equations (56) and (57) were derived on the assumption that y and u are small, of the order of first order differentials. Physically, in order to form an image using paraxial rays, the actual rays must obey the condition that y and u are small. It is, however, both a remarkable and extremely useful fact that in ray tracing, we may use finite heights and angles, not necessarily small, for y and u. We will show this in the following paragraph.

5.9.3.2 Consider Figure 5.17 which indicates two rays from an axial object point O to the corresponding axial image point O'. Because u_{-1} and u in Equations (56) and (57) were assumed small, we can replace them by $\tan u_{-1}$ and $\tan u$ respectively. (The expansion of $\tan u$, in terms of u, shows that the first order approximation is $\tan u = u$, as in the case of $\sin u$. The third order approximation, how-

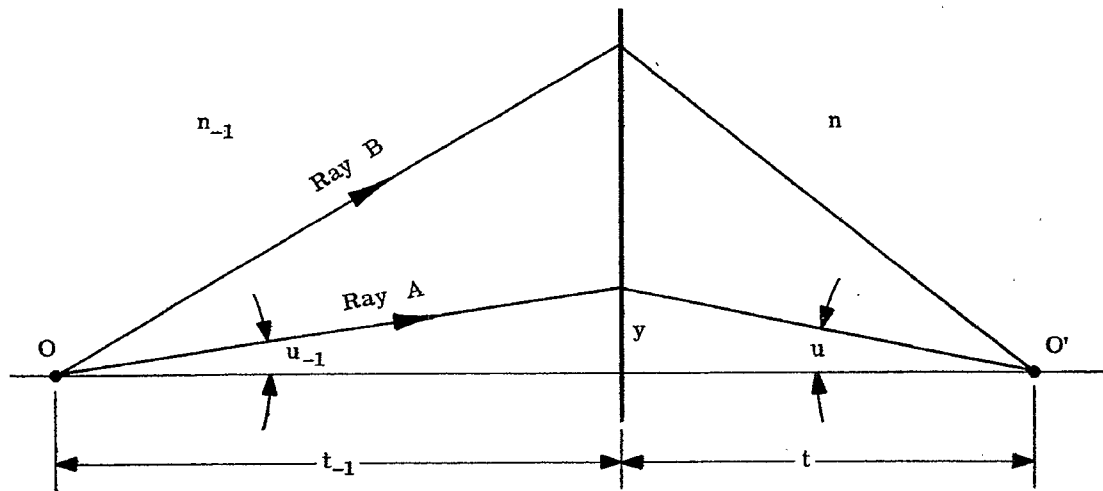


Figure 5.17. Paraxial rays through a single refracting surface.

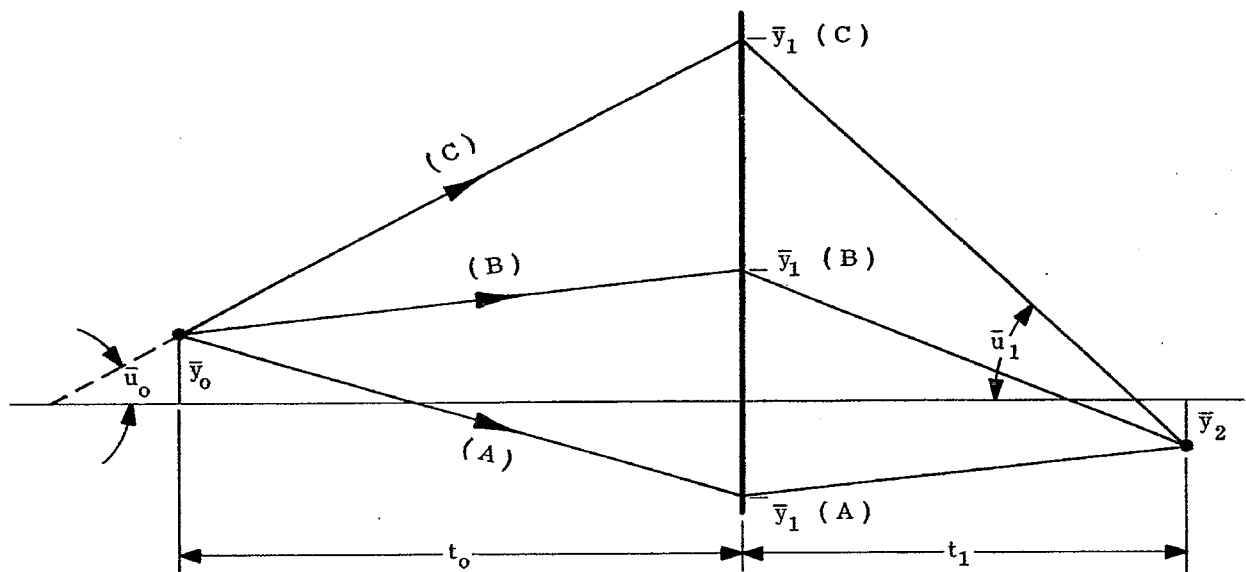


Figure 5.18 - Paraxial rays through a single refracting surface.

ever, differs from that of $\sin u$). Remembering that u is negative, we substitute into Equation (57) and find

$$n - \left(\frac{y}{t} \right) = n_{-1} \frac{y}{t_{-1}} + \frac{y}{r} (n_{-1} - n).$$

Upon rearrangement we get

$$\frac{n_{-1}}{t_{-1}} + \frac{n}{t} = \frac{n - n_{-1}}{r},$$

which is the familiar form of the paraxial equation for a single surface. The important thing to note is that u_{-1} , u , and y no longer appear in this equation. This fact is interpreted as meaning that mathematically we may consider the image O' formed by any ray leaving the object O . Thus both rays (A) and (B) intersect the axis at the same image point.

5.9.3.3 Figure 5.17 and the above paragraph apply to axial object and image points. The same conclusions concerning finite heights and angles hold for rays through off-axis object and image points. Hence, in Figure 5.18, all rays (A), (B), and (C) intersect at one image point. Neither the angles \bar{u}_0 or \bar{u}_1 , nor the heights \bar{y}_0 , \bar{y}_1 , or \bar{y}_2 , need be small.*

5.10 GRAPHICAL RAY TRACE FOR PARAXIAL RAYS

5.10.1 Specialization of the general graphical method.

5.10.1.1 Paraxial rays may be traced graphically through a lens system by a construction very similar to the construction shown in Section 5.7. This is done by replacing the refractive index circles by tangent planes through the vertices of the surfaces. The justification for these replacements will be given in Paragraph 5.10.1.3. For paraxial rays, the construction will appear as shown in Figure 5.19.

5.10.1.2 In the above paragraph we have indicated that Figure 5.19 is correct for paraxial rays.

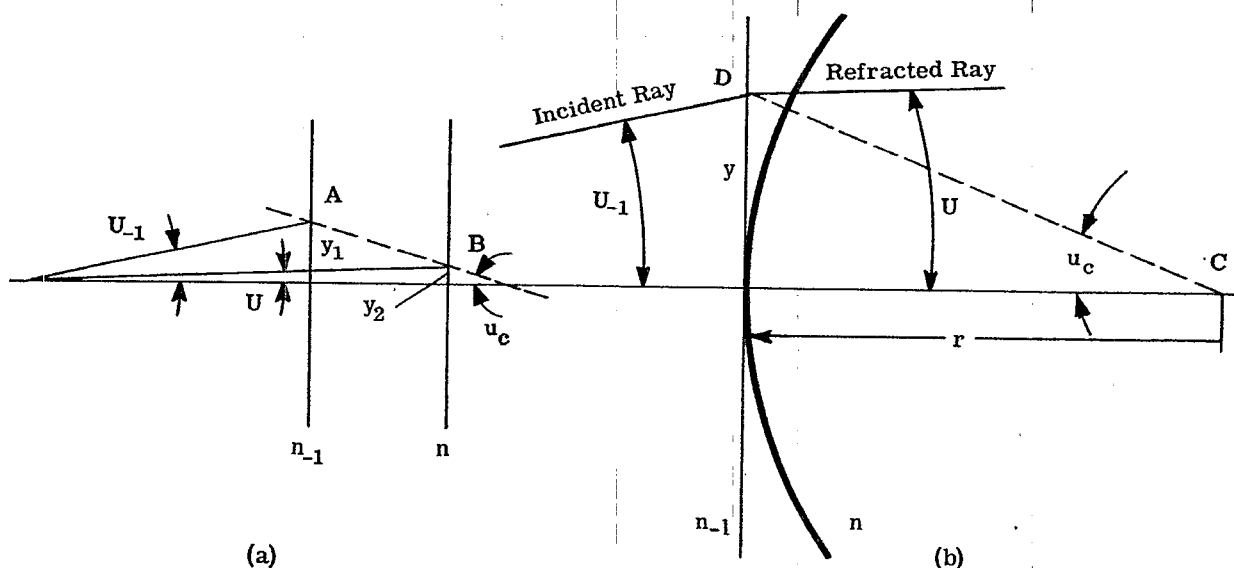


Figure 5.19 - The method for tracing paraxial rays graphically.

* The angles and heights corresponding to rays through off-axis object and image points are written with a line or bar over the symbol, as \bar{u} and \bar{y} .

Assuming this let us use the drawing to reexamine and extend the ideas discussed in Section 5.9.3. From Figure 5.19 it can be seen that

$$y_1 = n_{-1} \tan U_{-1}$$

and

$$y_2 = n \tan U.$$

Since the line connecting AB (a) is parallel to line DC (b) it is clear that, from similar triangles,

$$\frac{y_1 - y_2}{y} = \frac{n - n_{-1}}{r}.$$

By inserting the expressions for y_1 and y_2 into the last equation, we find on rearranging

$$n \tan U = n_{-1} \tan U_{-1} + y \left(\frac{n - n_{-1}}{r} \right).$$

5.10.1.3 This last equation, derived from Figure 5.19, is correct for small angles. This is easily seen, because when the $\tan U_{-1}$ and $\tan U$ are replaced by the angles u_{-1} and u respectively, Equation (57) results. However let us assume for the moment that both Figure 5.19 and the last equation are correct for any angles U and U_{-1} . In particular, these may be finite angles and do not have to be small. Now if we compare this equation to Equation (57), which is true for small angles u and u_{-1} , and therefore for paraxial rays, we see that $\tan U$ and $\tan U_{-1}$ correspond to u and u_{-1} , respectively. This indicates that the equation derived from Figure 5.19 can be used in connection with paraxial rays, provided the angles u and u_{-1} are replaced by $\tan U$ and $\tan U_{-1}$ respectively. Since U can have any value, $\tan U$ and therefore u can have any value. Equation (57) and Figure 5.19 can therefore be used for paraxial rays incident at any finite height and making any finite angle with the optical axis. Equation (56) and Figure 5.19 can also be used to accurately transfer the value of y from one surface to another for paraxial rays. This equation and figure can also be used with non-paraxial meridional rays to transfer between plane surfaces; in this case u_{-1} is replaced by $\tan U_{-1}$.

5.10.2 Two approaches to the treatment of paraxial rays.

5.10.2.1 We have shown that paraxial rays can be considered from either of two points of view:

- (1) We use small angles and finite curvatures for surfaces. This led to Equation (57).
- (2) We use finite angles and zero curvatures for surfaces. This led to Figure 5.19. It must be emphasized that we do not have to combine these and use small angles with plane surfaces.

5.10.2.2 It is convenient then to think of paraxial rays as passing through the optical system at finite heights, striking the surfaces on the tangent planes instead of the actual curved surfaces. Since the two Equations (56) and (57) are linear equations, and since the location of images are found for values of $y_k = 0$, it makes no difference what value of u is used. It is instructive to trace paraxial rays through a lens at heights equal to the actual ray heights, and note the difference in path for a paraxial ray and an actual ray. This is demonstrated for a single surface refraction in Figure 5.20. The ray traced through the curved surface crosses the axis at M , closer to the surface than the point P . The paraxial ray crosses at P , further away from the surface. This defect of focus is called spherical aberration.

5.11 THE DIFFERENT "ORDERS" OF OPTICS

5.11.1 Expansion of the sine function.

5.11.1.1 The fundamental equations which have been discussed and used in tracing rays are: (1) the transfer equations, and (2) the refraction equations. Both have been put into a form explicitly using the cosine function of various angles, such as the angles of incidence and refraction, and the angles which the ray makes with the coordinate axes. Both equations could have been written in terms of the sine function; so as to explain the meaning of the phrase orders of optics we will deal with the sine function.

5.11.1.2 The optical axis is a special ray for which both angles of incidence and refraction are zero. In addition the angles which this ray makes with the X , Y , and Z axes are 90° , 90° , and 0° respectively. For a meridional ray near the axis, the angles of incidence and refraction, and the angle with the Z axis, are small. The ray trace equations, therefore, involve the sines of small angles. As the meridional ray

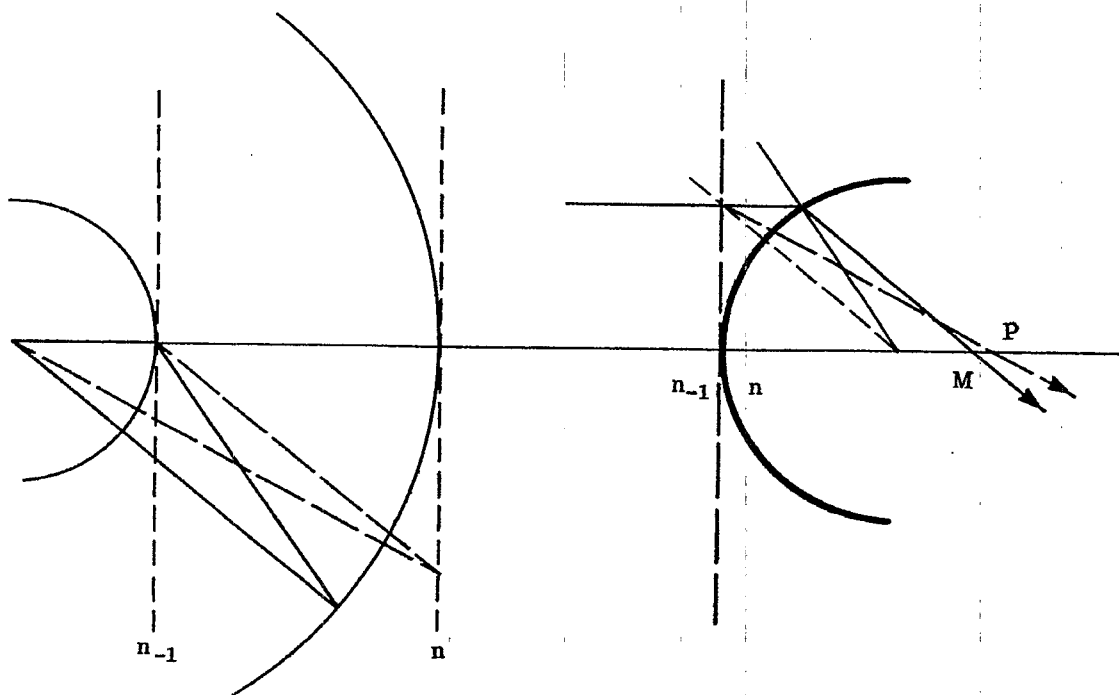


Figure 5.20 - Comparison between a paraxial and an actual ray showing spherical aberration.

makes larger and larger angles with the Z axis, we have to be concerned with the sines of larger and larger angles.

5.11.1.3 One reason the ray trace equations are complicated is that they involve the trigonometric functions of angles, instead of just the angles. (We have seen in Section 5.9 how the equations are greatly simplified when they can be expressed in terms of angles, instead of trigonometric functions). To relate the $\sin \alpha$ to the angle α , we expand the sine function in a series, thus

$$\sin \alpha = 0 + \alpha - 0 - \frac{\alpha^3}{3!} + 0 + \frac{\alpha^5}{5!} - 0 - \frac{\alpha^7}{7!} + \dots$$

5.11.1.4 The terms given explicitly as zero have been written down to clarify the situation. Whenever a function is expanded in a series the "first" term is called the zeroth approximation or the zeroth order, and successive terms are called the first, second, third, etc., orders. In the case of the expansion of the sine function, the zeroth, second, fourth, and all even order terms are identically zero; only the odd orders remain.

5.11.2 First order optics. If in ray trace equations the sine is replaced by the angle, we are using the zeroth and the first order terms in the above expansion. The resulting equations and design procedures are called first order optics, and the rays concerned are paraxial rays. One of the fascinating parts of geometrical optics is the extensive understanding of lens systems one can obtain by tracing two paraxial rays. With two paraxial rays one can predict the location and size of any image formed with paraxial rays, and by making further calculations based on these paraxial ray data it is possible to predict the approximate magnitude of image errors. The following sections, 6 and 7, will be devoted to the development and use of the equations of first order optics. This development will be based on the two simple equations, (56) and (57).

5.11.3 Third order optics.

5.11.3.1 If the first and third order terms in the expansion of the sine are retained, the resulting equations are part of third order optics. But this term has an added meaning, pertaining to aberrations, and it is usually in this latter sense that the term is used.

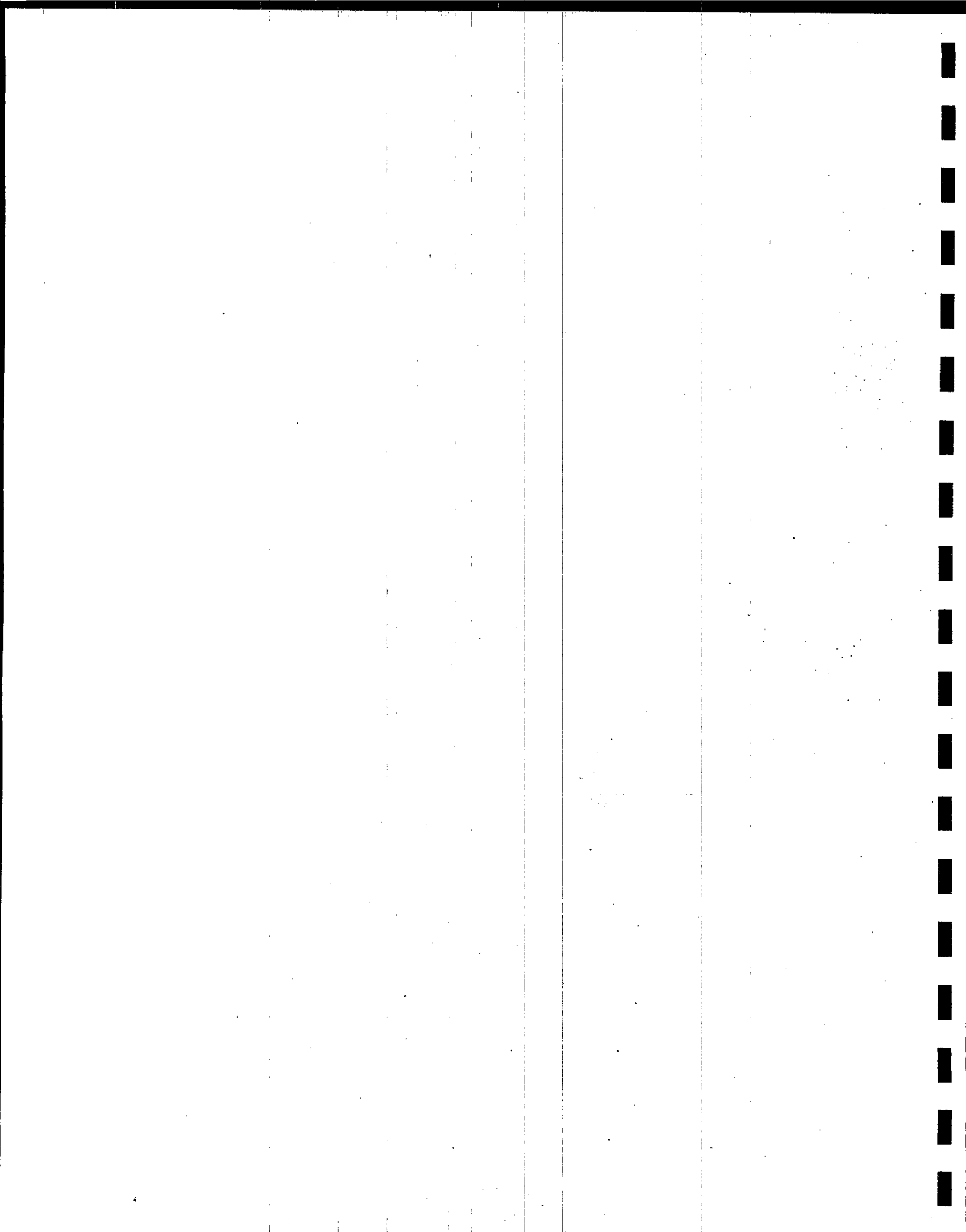
5.11.3.2 The intersection of a ray with the image surface locates the image. If the intersection has been computed using the skew ray trace equations, the intersection is the true image. If paraxial ray trace equations have been used, the resulting paraxial image will generally be displaced from the true image. The

difference between the true image and the first order approximation (paraxial approximation) is known by the general term aberration. (We are considering here monochromatic light only. Aberrations due to non-monochromatic light are considered in Paragraph 5.11.3.4.) In the same way that the sine was expanded in a series, the aberrations can be expanded. The first term in the expansion is known as the third order aberration. The reason for this is that it represents the first approximation to the total aberration, and hence can be considered as the difference between the paraxial image and the image using the third order approximation for the sine. Third order optics then has come to mean the equations and procedures dealing with the first approximation to the aberrations. It is fortunate, as will be evident in a later section (Section 8) that these third order aberrations can be calculated from first order (paraxial) ray trace data.

5.11.3.3 The next term in the expansion of the aberration, after the third order aberration, is called the fifth order aberration. Fifth order optics deals with the aberrations through the fifth order aberration term.* Hence fifth order optics deals with fifth order aberration, or the second approximation to the aberration.

5.11.3.4 Aberrations due to non-monochromatic light can also be expanded in a series. The first term gives the aberration appearing in paraxial images, hence is referred to as first order aberration. This is treated in Section 6, dealing with first order optics.

* In some countries other than the United States, for example England, the first, second, third, etc., terms in the aberration expansion are referred to as primary, secondary, tertiary, etc. aberration.



6 FIRST ORDER OPTICS

6.1 GENERAL

6.1.1 First order optics and paraxial rays. In Section 5.11.2 it was pointed out that when the sine of the angle is replaced by the angle, the resulting equations belong to the field of first order optics. In general, if any trigonometric function is replaced by its first approximation, we get first order equations, in the field of optics. In Sections 5.9.1 and 5.9.2 we defined a paraxial ray as one differentially displaced from the optical axis. Because of this definition we must use the first approximation to the trigonometric functions in the equations for a differentially traced ray. The resulting paraxial ray equations are hence identical to the first order equations.

6.1.2 Preliminary layout and graphical ray trace. The method of tracing paraxial rays graphically was explained in Section 5.10. Graphical ray tracing is extremely useful in the preliminary design stage, particularly for complicated systems, which cannot be visualized easily. The designer can thereby get a "feel" for the system, which a mere array of numbers often hides. Graphical ray tracing, however, is limited to an accuracy of about one percent. For additional accuracy, which is absolutely necessary in the calculation of aberrations, we must resort to numerical paraxial ray tracing. The methods and results of this type of ray tracing in the realm of first order optics will be discussed in Section 6.

6.2 NUMERICAL TRACING OF PARAXIAL RAYS

6.2.1 Importance of paraxial ray tracing. The accurate numerical tracing of paraxial rays is used extensively in the design of optical systems for three main reasons:

- (1) Tracing paraxial rays through the system is a simple mathematical procedure.
- (2) Images formed by paraxial rays provide very convenient reference planes.
- (3) Data obtained in paraxial ray calculations can be used to calculate the first approximation to image aberration.

For these reasons, a systematic method of numerical ray tracing of paraxial rays is a necessary tool for the designer, even today when large automatic computers are readily available. In this section such a method will be described; and it will be used extensively in the following sections to illustrate the vast amount of information made available by paraxial ray tracing.

6.2.2 Ray trace format.




6.2.2.1 The first step in tracing a paraxial ray is to lay out the system data in a form as shown in the top of Table 5.1. Then the two constants, $c(n_{-1} - n)$ and t/n , are computed for each surface and space respectively. (See Table 6.1). With these constants filled in, the paraxial ray may be traced by applying Equations (56) and (57) of Section 5.

6.2.2.2 As one uses this representation, its value becomes evident. These equations, and the way the data

SURFACE	0	1	2	3 etc.
c	c_0	c_1	c_2	c_3
t	t_0	t_1	t_2	
n	n_0	n_1	n_2	
$c(n_{-1}-n)$		$c_1(n_0-n_1)$	$c_2(n_1-n_2)$	
t/n	t_0/n_0	t_1/n_1	t_2/n_2	
y	y_0	y_1	y_2	
nu	$n_0 u_0$	$n_1 u_1$	$n_2 u_2$	

Table 6.1 - Recommended format for tracing paraxial rays through an optical system.

are laid out, make almost a perfect match with the requirements of a desk calculator. A few of these features are:

- (1) In calculating $c(n_{-1} - n)$ one obtains the data from a triangle of numbers, 
- (2) In tracing the ray, both equations are computed in the same way. First a number is multiplied by a number directly above it, then the product added to the number below the double line on the left, and the result written in the space on the right. This is indicated by the lines shown in the figure that appear as  and 
- (3) Many times, problems are worked backwards. For example, suppose $n_{-1} u_{-1}$, nu , and y are given, and the problem is to find c . The question is: how to remember what to do first, i.e., divide y by $(nu - n_{-1} u_{-1})$, or vice versa? It turns out that the correct method is always the easiest one to do on the calculating machine. Dividing $(nu - n_{-1} u_{-1})$ by y can be done without writing down $(nu - n_{-1} u_{-1})$. However, to calculate $y/(nu - n_{-1} u_{-1})$, the difference must be written down; therefore, we know that to calculate c , the result must be $(nu - n_{-1} u_{-1})/y$ divided by $(n_{-1} - n)$. As another example, suppose a value of y_1 , y_2 , and $n_1 u_1$ are given, what t_1/n_1 is needed? The formula can be remembered in the following way: first compute $y_2 - y_1$, and then divide by $n_1 u_1$. Therefore the formula is $t_1/n_1 = (y_2 - y_1)/n_1 u_1$.

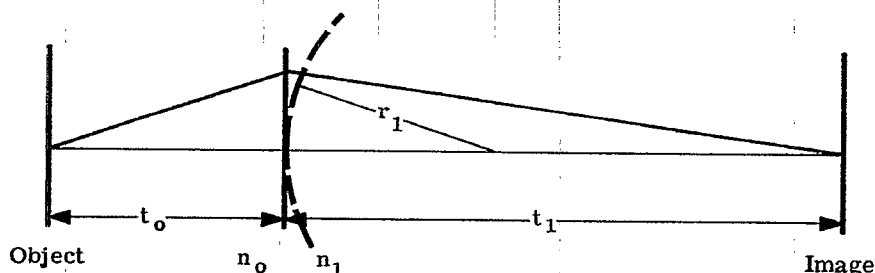


Figure 6.1 - Relation between image and object points.

6.2.3 Algebraic example. Table 6.1 may also be used to derive algebraic expressions useful in optics. One can very readily work out the equation for image and object distances for a single refracting surface. If a surface of radius r_1 separates two media of index n_o and n_1 , an object point will be imaged at a distance t_1 from the surface (see Figure 6.1). What is the relation between t_o and t_1 ? This is a three surface problem, the object surface, 0, the refracting surface, 1, and the image surface, 2. A paraxial ray at $y_o = 0$ will be imaged at $y_2 = 0$. It is therefore possible to fill out the calculations of Table 6.1 to the following extent. (See Table 6.2).

SURFACE	OBJECT	1	IMAGE
c	0	c_1	0
t	t_o	t_1	
n	n_o	n_1	
$c(n_{-1} - n)$	0	$c_1(n_o - n_1)$	0
t/n	t_o/n_o	t_1/n_1	
y	0		0
nu			

Table 6.2 - Single refracting surface, axial object and image points.

Now, as pointed out in Sections 5.9.3 and 5.10.1, the angle used to trace a paraxial ray does not affect the image position. Since the choice of y on the lens is arbitrary, let it be y_1 . The calculations to this point are shown in Table 6.3.

$c(n_{-1} - n)$ t/n	0	$c_1(n_o - n_1)$ t_o/n_o	t_1/n_1	0
y nu	0	$n_o u_o$	y_1 $n_1 u_1$	0

Table 6.3 - Continuation of Table 6.2.

With $y_o (=0)$ and y_1 (arbitrary) known, $n_o u_o = (y_1 - 0)/(t_o/n_o)$.

With $n_o u_o$ and y_1 known, $n_1 u_1 = n_o u_o + y_1 c_1 (n_o - n_1)$.

With $n_1 u_1$, y_1 and $y_2 (=0)$ known, $t_1/n_1 = (0 - y_1)/n_1 u_1$.

Therefore $t_1/n_1 = -y_1/n_1 u_1$, or

$$\frac{n_1}{t_1} = -\frac{n_1 u_1}{y_1} = \frac{-n_o u_o}{y_1} - c_1 (n_o - n_1) = -\frac{n_o}{t_o} + c_1 (n_1 - n_o).$$

This equation becomes the familiar refraction equation, derived in Paragraph 5.9.3.2,

$$\frac{n_1}{t_1} + \frac{n_o}{t_o} = c_1 (n_1 - n_o). \quad (1)$$

Notice how the y_1 has dropped out of the equation indicating that any value y_1 could have been used. The calculations will be finally filled out in Table 6.4 as follows:

y nu	0	y_1 $y_1 n_o/t_o$ $+ y_1 c_1(n_o - n_1)$	0
-------------	---	--	---

Table 6.4 - Conclusion of Table 6.3.

6.2.4 Numerical example. Equation (1) was given to show how the ray trace table can be used to derive a classical formula. Actually one will find very little occasion to use Equation (1) to calculate a numerical result, because problems can be solved much more readily using the format of Table 6.1. For example, suppose one is given the problem $c_1 = 0.10$, $n_o = 1$, $n_1 = 1.5$, $t_o = 10$. Rather than remember any special formula, go directly to the format as shown in Table 6.5.

SURFACE	OBJECT	1	IMAGE
c	0	0.10	0
t	10	t_1	
n	1	1.5	
$c(n_{-1} - n)$ t/n	0	-0.05 $t_1/1.5$	0
y nu	0	1 0.10	0 0.05

$$(0.05) \frac{t_1}{1.5} + 1 = 0$$

$$\frac{t_1}{1.5} = \frac{-1}{.05} = -20$$

$$t_1 = (1.5) (-20) = -30$$

Table 6.5 - Numerical example of a single refracting surface.

6.2.5 Ray trace for three element lens. Table 6.6 shows the data and ray trace results for a three element lens. All the material above the lowest double line has been discussed earlier in this chapter. The last two lines, involving \bar{y} and $n\bar{u}$, and the calculations of m (lateral magnification), f' (focal length), and Φ (optical invariant) will be discussed in the following sections.

SURFACE	OBJECT 0	1	2	3	4	5	6	IMAGE 7
c	0	0.25285	-0.01474	-0.19942	0.25973	0.05065	-0.24588	0
t		25.00000	0.60000	1.06541	0.15000	1.13691	0.60000	14.05015
n		1.00000	1.62000	1.00000	1.62100	1.00000	1.62000	1.00000
$c(n_{-1} - n)$		-0.15677	-0.00914	0.12384	0.16129	-0.03140	-0.15245	
t/n		25.00000	0.37037	1.06541	0.09254	1.13691	0.37037	14.05015
y	0	1.25000	1.19594	1.02879	1.02606	1.18070	1.21734	0
nu		0.05000	-0.14596	-0.15689	-0.02948	0.13601	0.09894	-0.08664
\bar{y}	-10.00000	-0.75000	-0.56942	-0.04440	0.00069	0.55481	0.72887	5.77084
$n\bar{u}$		0.37000	0.48758	0.49278	0.48728	0.48739	0.46997	0.35886

$$m = \frac{n_o u_o}{n_6 \bar{u}_6} = -0.57708 = \frac{\bar{y}_7}{\bar{y}_o} \quad \Phi = -0.50000$$

$$f' = -\frac{\Phi}{n_o (u_o \bar{u}_6 - \bar{u}_o u_6)} = \frac{0.5}{1 (0.05 \times 0.35886 + 0.37 \times 0.08664)} = 10.000$$

Table 6.6 Sample calculation of paraxial rays through a three element lens, using Equations 5-(56) and 5-(57).

6.2.6 Ray trace procedure for calculation of aberrations. Another way to trace paraxial rays is to use the following equations:

$$y = y_{-1} + t_{-1} u_{-1} \quad (2)$$

$$u = u_{-1} + i \left[(n_{-1}/n) - 1 \right] \quad (3)$$

This ray trace involves the new quantity, i , which is the limiting value of the angle of incidence, I , as the ray approaches the axis in the paraxial region. Equation (2) is merely Equation 5-(56) simplified. Equation (3) comes from Equation 5-(35), written for small angles, with the substitution $i' = i n_{-1}/n$; the latter is the law of refraction for small angles. Now from Figure 5.11, $I' - U =$ the acute angle between r and the optical axis. But for small angles this is $y/r = y c$. Hence, using Equation 5-(35) we have,

$$i = y c + u_{-1} \quad (4)$$

It will be shown later (Section 8) how the third order aberrations may be calculated from paraxial ray data. For these calculations it is easier to use Equations (2), (3) and (4), than Equations 5-(56) and 5-(57).

6.2.7 Numerical example.

6.2.7.1 To illustrate these equations, Table 6.7 includes paraxial rays traced through the same lens as used in Table 6.6. In this example, a different set of rays are traced through the lens. Below the lowest double line there are entries used in the calculation of chromatic aberration. These calculations will be explained in Section 6.10.

SURFACE	OBJECT 0	1	2	3	4	5	6	IMAGE 7
c	0	0.252850	-0.014740	-0.199420	0.259730	0.050650	-0.245880	
t	∞	0.600000	1.065410	0.150000	1.136910	0.600000	8.279369	
n	1.000000	1.620000	1.000000	1.621000	1.000000	1.620000	1.000000	
$(n_{-1} / n) - 1$		-0.382716	0.620000	-0.383097	0.621000	-0.382716	0.620000	
y	0	1.500000	1.412907	1.148619	1.138827	1.227353	1.241918	0
u	0	-0.145155	-0.248063	-0.065279	0.077866	0.024274	-0.150001	
i		0.379275	-0.165981	-0.477120	0.230508	0.140031	-0.281089	
dn/n	0	0.006370	0	0.010586	0	0.006370		$T_{Ach} = -0.0029$
$\Delta(dn/n)$		0.006370	-0.006370	0.010586	-0.010586	0.006370	-0.006370	$\Sigma a = -0.00044$
$a = -y n_{-1} i \Delta \frac{dn}{n}$		-0.00362	-0.00242	0.00580	0.00450	-0.00109	-0.00360	

Table 6.7 - A paraxial ray is traced through the same lens as used in Table 6.6. In this case Equations (2), (3), and (4) are used.

6.2.7.2 For use with a large computing machine there is no preference for either of these methods. For hand computing, unless aberrations are calculated, the method outlined in Table 6.1 is simpler. Therefore, all the paraxial ray theory given in Section 6.3-6.9 will be based on Equations 5-(56) and 5-(57).

6.3 THE OPTICAL INVARIANT

6.3.1 Axial and oblique rays. In Section 6.2 it was shown how images may be located along the axis of the optical system. The procedure is to trace a paraxial ray from where the object surface crosses the optical axis ($y_o = 0$). Such a ray is called an axial paraxial ray. An image surface is formed wherever this paraxial ray crosses the optical axis. By tracing a second ray from the object at a value of $y_o \neq 0$ it is possible also to determine the size of the image. Such a ray is called an oblique paraxial ray. The data for this second ray will be identified by writing y and \bar{u} . Table 6.6 shows a second ray traced through the lens. The second ray is commonly referred to as the oblique paraxial ray because it passes from an off-axis object point obliquely through the optical system to the image. If this ray passes through the center of the aperture stop it is called a chief ray. In tracing the oblique paraxial and the axial paraxial ray through the system, the following equations have been applied for each surface:

$$nu = n_{-1} u_{-1} + yc (n_{-1} - n) \quad \text{for the axial paraxial ray refraction.} \quad 5-(57a)$$

$$n\bar{u} = n_{-1} \bar{u}_{-1} + \bar{y}c (n_{-1} - n) \quad \text{for the oblique paraxial ray refraction.} \quad 5-(57b)$$

$$y = y_{-1} + \frac{t_{-1}}{n_{-1}} (n_{-1} u_{-1}) \quad \text{for the axial paraxial ray transfer.} \quad 5-(56a)$$

$$\bar{y} = \bar{y}_{-1} + \frac{t_{-1}}{n_{-1}} (n_{-1} \bar{u}_{-1}) \quad \text{for the oblique paraxial ray transfer.} \quad 5-(56b)$$

6.3.2 The optical invariant and its importance. We will use the last four equations, involving axial and oblique paraxial rays, to derive an expression called the optical invariant. This quantity, as its name implies, is a constant; as such it may be calculated in several ways and its value for a given system can be used in the calculation of various quantities. This invariant has a meaning for an optical system similar to momentum or energy for an isolated mechanical system.

6.3.3 The invariant for refraction. By transposition and division, using Equations 5-(57a) and 5-(57b), it is possible to equate the common term $c (n_{-1} - n)$ giving

$$\frac{nu - n_{-1} u_{-1}}{y} = \frac{n\bar{u} - n_{-1} \bar{u}_{-1}}{\bar{y}}$$

By rearranging, this may be written

$$\bar{y} (n_{-1} u_{-1}) - y (n_{-1} \bar{u}_{-1}) = \bar{y} (nu) - y (n\bar{u}) \quad (5)$$

The index and angle data on the left side of this equation refer to the space to the left of the surface, and the corresponding data on the right side refer to the space to the right of the surface. This equation shows that

$$\bar{y} (nu) - y (n\bar{u}) = \Phi \quad (6)$$

is an invariant for the refraction at any surface in the optical system. Φ is called the optical invariant.

6.3.4 The invariant for transfer. In a similar way Equations 5-(56a) and 5-(56b) may be combined to give the relation

$$\bar{y}_{-1} (n_{-1} u_{-1}) - y_{-1} (n_{-1} \bar{u}_{-1}) = \bar{y} (n_{-1} u_{-1}) - y (n_{-1} \bar{u}_{-1})$$

It is noted that the right hand side of this equation is equal to the left hand side of Equation (5), and hence is Φ , the optical invariant. Moreover both y values on the left apply to the surface to the left of the space, and both y values on the right refer to the surface to the right of the space. Therefore this equation shows that the optical invariant is also an invariant as the ray is transferred from one surface to the next.

6.3.5 The invariant for the entire system. We have shown above that there is a combination of y , n , u , \bar{y} , and \bar{u} , which has the same value on either side of a surface, that is, it is invariant across a surface between two spaces. We have also shown that this same combination of parameters is the same on either

side of a space, that is, it is invariant across a space between two surfaces. Hence the optical invariant is an invariant for an entire optical system. It is therefore possible to write down the optical invariant between any two surfaces (or any two spaces). For example, between the object surface and the image surface we can write

$$\Phi = \bar{y}_o (n_o u_o) - y_o (n_o \bar{u}_o) = \bar{y}_k (n_{k-1} u_{k-1}) - y_k (n_{k-1} \bar{u}_{k-1}).$$

The invariant may also be written in determinant form as

$$\Phi = \begin{vmatrix} \bar{y} & n\bar{u} \\ y & nu \end{vmatrix}.$$

6.3.6 Lateral magnification. If $y_o = 0$ on the object surface (the 0th surface), and $y_k = 0$ on the image surface (the kth surface), then the next to the last equation becomes

$$\Phi = \bar{y}_o (n_o u_o) = \bar{y}_k (n_{k-1} u_{k-1}).$$

This is illustrated in Figure 6.2.

Using the optical invariant then, it is possible to calculate the height of the image \bar{y}_k from the object height, \bar{y}_o . The lateral magnification, m , is defined as

$$m = \frac{\bar{y}_k}{\bar{y}_o} = \frac{(n_o u_o)}{(n_{k-1} u_{k-1})}. \quad (7)$$

This equation shows that the lateral magnification can be calculated by tracing a single paraxial ray from the base of an object to the base of the image, and by taking the ratio given in Equation (7). Physically, the lateral magnification is the ratio of the height of the image to the height of the object, both heights being measured perpendicularly to the optical axis. By defining lateral magnification by Equation (7), and remembering that y values of points below the optical axis have signs opposite to those above, we see that a posi-

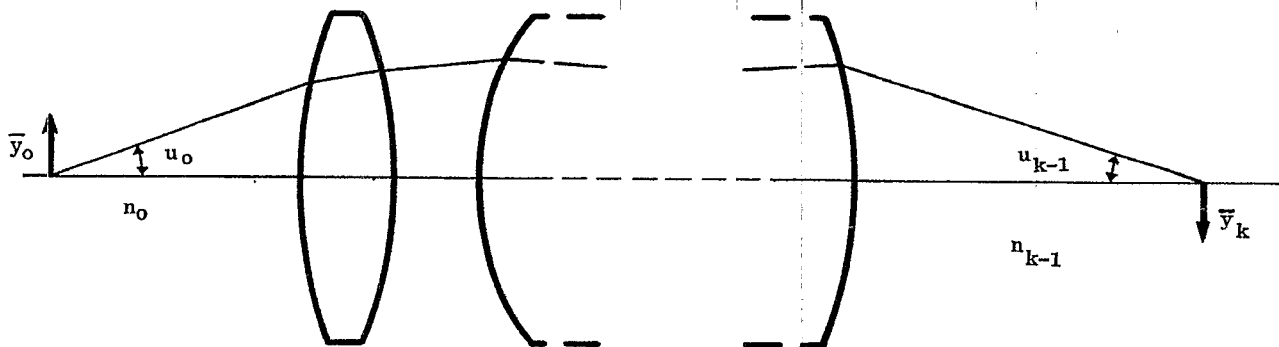


Figure 6.2 - Diagram illustrating the data used to compute the optical invariant.

tive value of m indicates an erect image. A negative value of m indicates an image inverted with respect to the object.

6.3.7 Angular magnification.

6.3.7.1 There are instruments which work with the object placed at a large distance t_o from the first surface of the lens or mirror. If this distance is great enough to assume it is infinite, then the ray coordinates on the first surface for the axial and oblique rays are: y_1 ; $u_o = 0$; \bar{y}_1 ; \bar{u}_o . The optical invariant, for the first surface (1) and the space to the left (0), becomes

$$\Phi = -y_1 (n_o \bar{u}_o).$$

In the image plane, $y_k = 0$, so

$$-y_1 (n_o \bar{u}_o) = \bar{y}_k (n_{k-1} u_{k-1}),$$

and

$$\bar{y}_k = \frac{-y_1}{(n_{k-1} u_{k-1})} (n_o \bar{u}_o). \quad (8)$$

In visual instruments, the image surface is usually at a great distance from the last optical surface ($k-1$). If the distance is assumed to be infinite, then $u_{k-1} = 0$, and

$$\Phi = -y_k (n_{k-1} \bar{u}_{k-1}).$$

When both the object and image surfaces are assumed to be at infinity we have a telescopic system and the optical invariant is

$$\Phi = -y_1 (n_o \bar{u}_o) = -y_k (n_{k-1} \bar{u}_{k-1}).$$

The most familiar example of a telescopic system is a telescope for which both object and image surfaces are at infinity; when so adjusted the telescope is said to be afocal. From the material to be presented in a

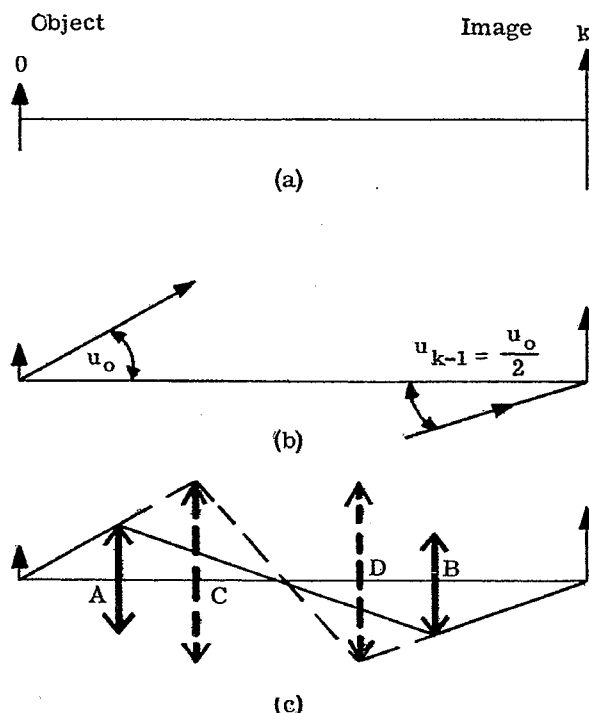


Figure 6.3 - Diagrams illustrating the use of the Smith Helmholtz equations. Thin positive lenses are represented by the symbol \updownarrow , thin negative lenses by χ .

later section we can say that such a telescope has its focal lengths equal to infinity and both focal points at infinity.

6.3.7.2 The angular magnification, α , is defined as the ratio \bar{u}_{k-1}/\bar{u}_0 . Therefore the angular magnification for a telescope in afocal adjustment is

$$\alpha = \frac{y_1 n_0}{y_k n_{k-1}} = MP. \quad (9)$$

For a telescope, the angular magnification is called the magnifying power (MP).

6.3.8 The Smith-Helmholtz and the Lagrange equations. Equations (7) and (8) can be rewritten as

$$\bar{y}_0 n_0 u_0 = \bar{y}_k n_{k-1} u_{k-1}$$

and

$$y_1 n_0 \bar{u}_0 = -\bar{y}_k n_{k-1} u_{k-1}.$$

These equations are referred to as the Smith-Helmholtz equations by some optical writers, and the LaGrange equations by others. Through the use of these equations, it is possible to decide rapidly what is needed to set up a given optical system. For example suppose we wish to form an erect image on surface k twice the size of the object on surface 0 . See Figure 6.3 (a). Equation (7) shows that if m is to be $+2$ then u_0 and u_{k-1} must have the same sign. This is illustrated in Figure 6.3 (b) for the case of $n_0 = n_{k-1}$. A ray emerging from the base of the object at an angle u_0 must pass through the optical system and emerge from below the optical axis at an angle $u_0/2$. As is shown in Figure 6.3 (c), this can be accomplished by any number of methods. A positive lens may be placed at A and be adjusted to refract the rays to cross the axis. At B a second positive lens refracts the rays to the final image. On the other hand two lenses could be used at C and D if desired, in which case the axial rays would refract as shown by the dotted lines.

6.4 LINEARITY OF THE PARAXIAL RAY TRACING EQUATIONS

6.4.1 General. In Sections 5.9.3 and 5.10 we have seen that finite heights and angles can be used with the paraxial ray trace equations. The basic reason for this is that these equations, 5-(56) and 5-(57), are linear. Another result of this linearity is that if two rays are traced through an optical system, it is possible to predict the path of any other paraxial ray. The proof of this fact will be developed below.

6.4.2 Proof of the theorem.

6.4.2.1 In order to prove the statements given above, let y and \bar{y} be the heights of any two paraxial rays on the j th surface. Corresponding to these two rays, u and \bar{u} are the angles between the rays and the optical axis. If \bar{y} and \bar{u} are the height and slope angle of any third ray, we wish to show that the equations

$$A \bar{y} + B y = \bar{y} \quad (10a)$$

and

$$A \bar{u} + B u = \bar{u} \quad (10b)$$

are valid for the entire optical system. We also must be able to calculate the values of A and B .

6.4.2.2 Equation (10a) applies to the j th surface. Using Equation 5-(56), we can show that an equation similar to (10a) applies to the $j+1$ surface. Substituting Equation 5-(56) into Equation (10a) gives

$$\bar{y}_{+1} - \frac{t}{n} (n\bar{u}) = A \bar{y}_{+1} - A \frac{t}{n} (n\bar{u}) + B y_{+1} - B \frac{t}{n} nu.$$

Collecting the terms involving n results in the expression $t(\bar{u} - A\bar{u} - Bu)$. But this equals zero by Equation (10b) so that

$$\bar{y}_{+1} = A \bar{y}_{+1} + B y_{+1}.$$

Hence Equation (10a) holds for the $j+1$ surface, and therefore, by induction, for any and all surfaces.

6.4.2.3 Similarly, we show that Equation (10b) holds for all spaces. Substituting Equation 5-(57) into Equation (10b), and collecting terms, we have

$$\frac{n+1}{n} \bar{\bar{u}}_{+1} = \frac{n+1}{n} A \bar{u}_{+1} + \frac{n+1}{n} B u_{+1} + \frac{c(n - n_{+1})}{n} (\bar{\bar{y}} - A \bar{y} - B y).$$

By Equation (10a) the last term equals zero. Hence

$$\bar{\bar{u}}_{+1} = A \bar{u}_{+1} + B u_{+1},$$

and Equation (10b) applies to any and all spaces.

6.4.2.4 We have shown that Equations (10a) and (10b) apply to all surfaces and all spaces respectively and hence to the entire optical system. Solving these equations for A and B gives

$$A = \frac{\bar{\bar{y}} u - \bar{u} y}{\bar{y} u - \bar{u} y} = n (\bar{\bar{y}} u - \bar{u} y) / \Phi$$

and

$$B = \frac{\bar{y} \bar{\bar{u}} - \bar{u} \bar{\bar{y}}}{\bar{y} u - \bar{u} y} = n (\bar{y} \bar{\bar{u}} - \bar{u} \bar{\bar{y}}) / \Phi.$$

These equations hold for any surface and the space to the right of that surface. In particular, we will use the expression for A for the object surface, and that for B for surface number 1.

6.4.3 Two particular rays.

6.4.3.1 Because the theorem proved in Section 6.4.2 holds for any three rays, we can choose these rays in

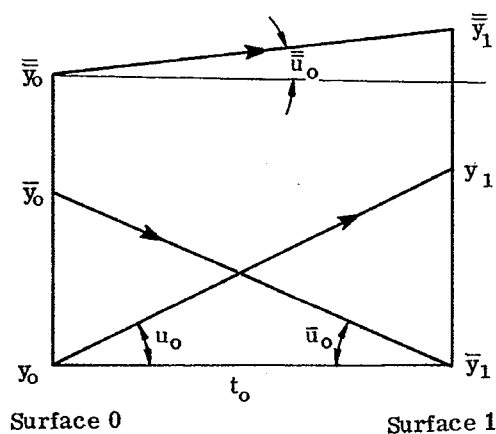


Figure 6.4 - Rays used to find simple expressions for A and B.

such a way as to simplify the calculation of A and B . The two particular rays we use are: (1) any ray from the center of the object surface ($x_o = y_o = 0$), and (2) any ray from the object ($\bar{y}_o \neq 0$) which intersects the axis at the center of the first surface ($\bar{x}_1 = \bar{y}_1 = 0$). These two particular rays, and any third ray, are shown in Figure 6.4. Using $y_o = \bar{y}_1 = 0$, the expressions for A and B reduce to

$$A = \bar{y}_o / \bar{y}_o ,$$

and

$$B = \bar{y}_1 / y_1 . \quad (11)$$

Using Figure 6.4, we have

$$y_1 = u_o t_o = -u_o \bar{y}_o / \bar{u}_o ,$$

and therefore

$$B = \frac{-\bar{y}_1 \bar{u}_o}{\bar{y}_o u_o} . \quad (12)$$

6.4.3.2 The two particular rays chosen are often specified more stringently. In order to get some idea as to the necessary diameters of the elements, the ray from the axial object ($y_o = 0$) is taken at a value of u_o so as to pass through the edge of the aperture stop. Such a ray is called a rim ray, or marginal ray; the value of u_o determines the energy passing through the system. The other ray is taken as coming from the top of the object. This gives some idea as to the diameters of the elements necessary to attain the desired field of view. We will specify later that this second ray ($\bar{y}_k \neq 0$) be the chief ray.

6.4.3.3 The above two paragraphs have specified the two particular rays ($y_o = 0$ and $\bar{y}_1 = 0$) be chosen so as to easily evaluate A and B from the known data and the initial third ray data. (It should be emphasized that this is not necessary; any two rays and the initial third ray data will suffice to determine A and B). Instead of choosing particular values of y_o and \bar{y}_1 , we could have chosen particular values of u_o and \bar{u}_1 , for example 0. This would result in $A = \bar{u}_o / u_o$ and $B = \bar{u}_1 / u_1$. Note the correspondence between these and the equations in Paragraph 6.4.3.1.

6.5 THE CARDINAL POINTS OF AN OPTICAL SYSTEM

6.5.1 General.

6.5.1.1 We have already seen, in Sections 5.9.3, 5.10, and 6.4, some important consequences of the linearity of the paraxial ray trace equations. Another consequence, to be discussed in Section 6.5, is the presence of certain special points which exist in any optical system. Six of these points, all lying on the optical axis and known as the cardinal points, are of great usefulness in analyzing an optical system. The reason why the linearity of the paraxial ray equations lead to the existence of the cardinal points will not be developed in detail. It may be mentioned here, however, that the equations which we will develop from the concept of the cardinal points can be derived directly from the ray trace equations. One such equation, for example, was derived in Paragraph 6.2.3. The fact that both the paraxial ray equations and the assumption of the existence of cardinal points lead to the same equations is indicative of the connection between Sections 6.4 and 6.5.

6.5.1.2 The cardinal points, and the letters used to designate them, are as follows:

- (a) The first and second focal points, F_1 and F_2 .
- (b) The first and second principal points, P_1 and P_2 .
- (c) The first and second nodal points, N_1 and N_2 .

Sometimes the words first and second are replaced by primary and secondary, or by object and image, respectively.

6.5.2 The second focal point and the second focal length. In the sample calculation shown in Table 6.6, if the axial ray is traced from an infinitely distant object, $t_o = \infty$ and $u_o = 0$. This ray will pass through the optical system and eventually cross the axis at what is called F_2 , the second focal point. (See

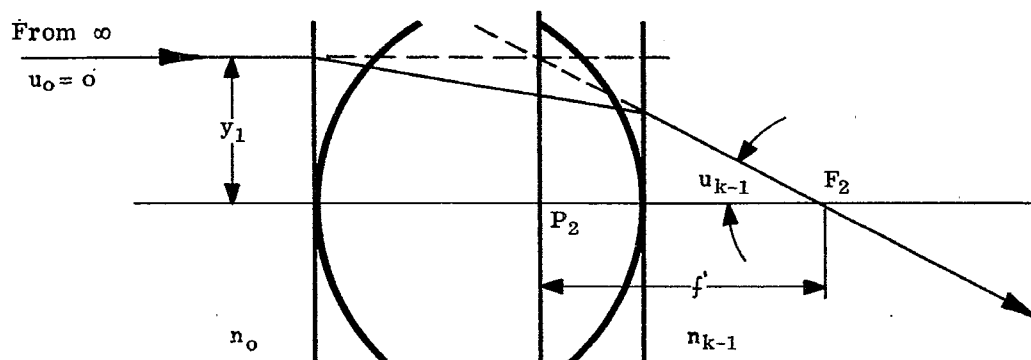


Figure 6.5 - Location of second focal point, second principal point, and second focal length.

Figure 6.5). The second focal point is, therefore, the intersection (in image space) of the optical axis and a ray which (in object space) was initially parallel to the optical axis. This cardinal point can also be considered as the axial image of an infinitely distant axial object. This is why it is sometimes referred to as the image focal point. Because the height of the axial ray, y_1 , is arbitrary, all rays parallel to the optical axis, coming from an object surface, intersect at the second focal point. We can think of an image surface, intersecting the axis at F_2 . This is the second focal surface, which for paraxial rays becomes the second focal plane. Then $y_k = 0$, and Equation (8) applies,

$$\bar{y}_k = -y_1 n_o \bar{u}_o / (n_{k-1} u_{k-1}) .$$

The second focal length is defined as,

$$f' = -y_1 / u_{k-1} . \quad (13)$$

Physically, the second focal length is the distance between the second focal point and the second principal point, defined below. The reason a telescope in afocal adjustment (see Paragraph 6.3.7.1) has an infinite (second) focal length is that $u_{k-1} = 0$. Hence the final axial ray is parallel to the axis, and F_2 is at infinity.

6.5.3 The second principal point. The second principal point is located by erecting a plane perpendicular to the optical axis at the point of intersection of the forward-extended entering ray and the backward-extended exit ray. The intersection of this plane (the second principal plane) with the optical axis is the second principal point, P_2 . From Figure 6.5 it can be seen that

$$f' = P_2 F_2 .$$

If the second principal point is to the left of the second focal point, f' is positive; otherwise it is negative.

6.5.4 The second nodal point. The second nodal point, N_2 , is also an axial point, as are F_2 and P_2 . It is a point such that the distance

$$N_2 F_2 = (P_2 F_2) \frac{n_o}{n_{k-1}} .$$

With this expression Equation (8) can then be written

$$\bar{y}_k = f' \frac{n_o \bar{u}_o}{n_{k-1}} = P_2 F_2 \frac{n_o \bar{u}_o}{n_{k-1}} = N_2 F_2 \bar{u}_o .$$

If $n_o = n_{k-1}$, then $P_2 F_2$ and $N_2 F_2$ are equal and the principal point P_2 and the nodal point N_2 coincide.

6.5.5 The first focal, principal and nodal points.

6.5.5.1 With similar arguments one can find a first focal point, F_1 , such that rays entering the system from F_1 will emerge from the last surface traveling parallel to the axis. For such an object point, $y_o = 0$, and $u_{k-1} = 0$. Therefore from the optical invariant equation,

$$\bar{y}_o = -y_k \frac{n_{k-1} \bar{u}_{k-1}}{n_o u_o} .$$

The first focal length f is now defined as,

$$f = \frac{y_k}{u_o} = F_1 P_1 . \quad (14)$$

Finally using $F_1 N_1 = (F_1 P_1) \frac{n_{k-1}}{n_o}$, we have

$$\bar{y}_o = -f \frac{n_{k-1} \bar{u}_{k-1}}{n_o} = -F_1 P_1 \frac{n_{k-1} \bar{u}_{k-1}}{n_o} = -F_1 N_1 \bar{u}_{k-1} .$$

6.5.5.2 The physical meanings of the first focal and principal points, and the first focal length, correspond to those discussed in Sections 6.5.2 and 6.5.3. The first focal point (see Figure 6.6) is the intersection of the optical axis and a ray which will be parallel to the axis when it leaves the system. It is also the axial object whose axial image is infinitely distant. All rays parallel to the optical axis after emerging from the

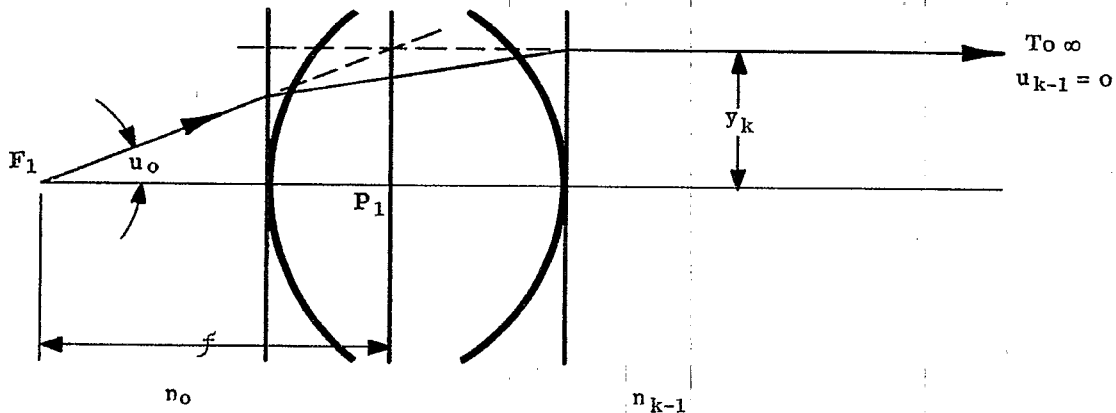


Figure 6.6 - Location of first focal point, first principal point, and first focal length.

system have passed through the first focal point. The plane perpendicular to the axis at F_1 is the first focal plane.

6.5.5.3 The first principal plane is a plane perpendicular to the optical axis passing through the intersection of the forward-extended ray through F_1 and the backward-extended ray emerging from the system parallel to the axis. The intersection of this plane with the axis is the first principal point. The first focal length is the distance between the first focal point and the first principal point, and is positive if F_1 lies to the left of P_1 .

6.5.6 Object and image positions with respect to focal and principal points.

6.5.6.1 The previous sections, in connection with Figures 6.5 and 6.6, have explained the meaning of the focal and principal points, and the principal planes, from a graphical point of view. First, these ideas will be used to derive some well known relations between object and image positions. These relations will then be used to indicate additional characteristics of principal planes and nodal points.

6.5.6.2 Consider Figure 6.7 which indicates an object of height \bar{y}_o at an arbitrary position. It should be emphasized here that Figure 6.7 indicates a general optical system, without reference to specific positions of refracting or reflecting surfaces. (Figures 6.5 and 6.6 show two refracting surfaces merely for concreteness; the ideas involved in those figures apply to the general system, as does the whole of Section 6.5). Of the infinite number of rays that come from the top of the object, we choose two whose course through the system we know from Figures 6.5 and 6.6. An entering ray, parallel to the optical axis, passes through F_2 , and can be considered to be deviated only once, at the second principal plane. Similarly, a ray through F_1 exits parallel to the optical axis, and can be considered as having been deviated only once, at the first principal plane.

6.5.6.3 Four new distances are shown, Z , Z' , S , and S' . Sign conventions are then established such that all these distances shown, as well as f and f' , are positive. If any pair of points at the ends of the double arrows are reversed, the distance is negative. For example if the object is to the right of F_1 , Z

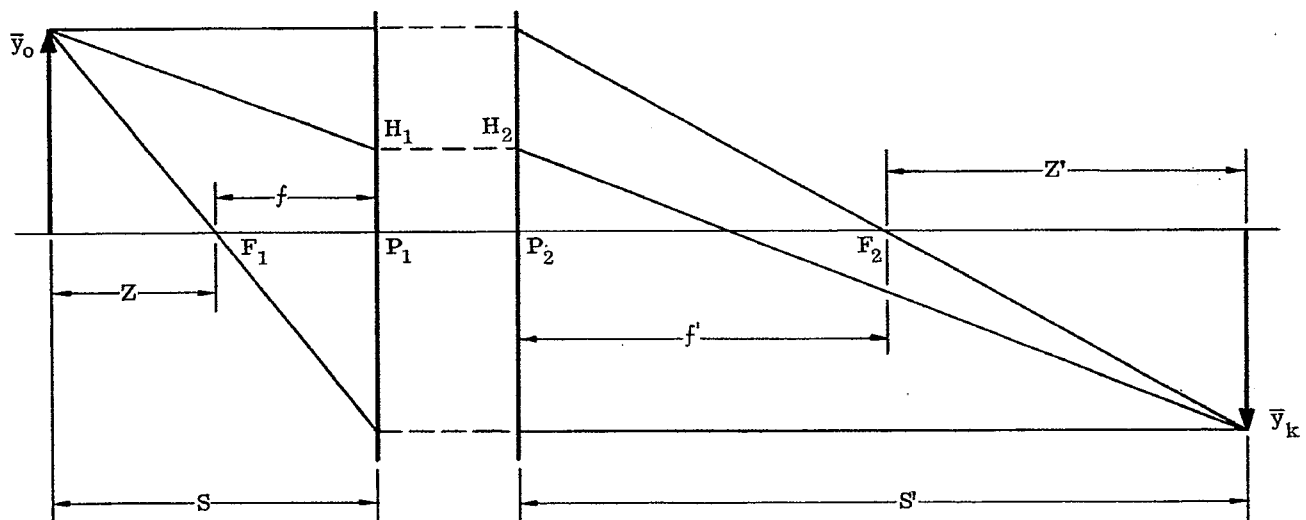


Figure 6.7 - Diagram showing object and image relations.

will be negative. From similar triangles, remembering that \bar{y}_k is negative,

$$\frac{\bar{y}_k}{\bar{y}_o} = - \frac{Z'}{f'} = - \frac{f}{Z} .$$

Using the definition of lateral magnification, $m = \bar{y}_k / \bar{y}_o$, we have

$$m = - \frac{Z'}{f'} = - \frac{f}{Z} . \quad (15)$$

Rearranging there follows

$$Z Z' = f f' . \quad (16)$$

Equations (15) and (16) are in the Newtonian form, in which object and image positions are measured from the focal points, F_1 and F_2 , respectively.

6.5.6.4 Another form of expressing these relations is the Gaussian form of these equations, in which object and image positions are measured from the principal points, P_1 and P_2 , respectively. From Figure 6.7, $Z = S - f$ and $Z' = S' - f'$. Substituting these expressions into (15) and (16) gives

$$m = - \frac{S' - f'}{f'} = - \frac{f}{S - f}$$

and

$$(S - f)(S' - f') = f f' .$$

Expanding the last equation and dividing by SS' , we have

$$\frac{f}{S} + \frac{f'}{S'} = 1 , \quad (17)$$

and using (17), the lateral magnification becomes

$$m = - \frac{f S'}{f' S} . \quad (18)$$

Equations (18) and (17) are in the Gaussian form and correspond to Equations (15) and (16). Whereas the latter pair does not involve S or S' , and the former pair does not involve Z or Z' , we may eliminate f and f' from Equation (16) by substituting $f = S - Z$ and $f' = S' - Z'$. The result is

$$\frac{Z}{S} + \frac{Z'}{S'} = 1 .$$

And using this with Equation (15), we have

$$m = - \frac{Z' S}{Z S'} .$$

6.5.6.5 It may be well to summarize here the specific meanings of the six distances used in the equations of Paragraphs 6.5.6.3 and 6.5.6.4. The sign conventions are included below if it is remembered that a distance measured to the right is positive.

- f is measured from F_1 to P_1 .
- f' is measured from P_2 to F_2 .
- Z is measured from the object plane to F_1 .
- Z' is measured from F_2 to the image plane.
- S is measured from the object plane to P_1 .
- S' is measured from P_2 to the image plane.

6.5.7 Additional characteristics of principal planes. Suppose the object is placed at the first principal plane. This means that $Z = -f$, and Equation (16) gives $Z' = -f'$. But this also means that the image is at the second principal plane; the two principal planes are therefore conjugate planes and P_1 and P_2 are conjugate points. (Equation (17) could have been used, with $S = 0$, giving $S' = 0$, which again locates the image at P_2). Using Equation (15) we find for this case $m = 1$. The two principal planes are therefore planes of unit positive magnification. This fact is very useful since it allows us to say that any point on the plane through P_1 is imaged at the same height on the plane through P_2 . Therefore any other ray (see Figure 6.7), entering the system so that it intersects the first principal plane at H_1 , exits from the system as if it came from H_2 , at the same distance from the axis.

6.5.8 Additional characteristics of nodal points.

6.5.8.1 There is an important relation between the focal lengths of any optical system, and the refractive indices of object and image space. Equation (7) can be rewritten, using Figure 6.7, to give

$$m = \frac{n_o u_o}{n_{k-1} u_{k-1}} = \frac{n_o}{n_{k-1}} \left(-\frac{S'}{S} \right).$$

Comparing this with Equation (18) we have

$$f/n_o = f'/n_{k-1}. \quad (19)$$

6.5.8.2 Equation (19) can be used to indicate a useful property of the nodal points. Using the expressions for $N_2 F_2$ and $F_1 N_1$ given in Sections 6.5.4 and 6.5.5, in connection with Figure 6.8, we have

$$P_1 N_1 = F_1 N_1 - F_1 P_1 = f \left(\frac{n_{k-1}}{n_o} - 1 \right) = \frac{f}{n_o} (n_{k-1} - n_o),$$

and

$$P_2 N_2 = P_2 F_2 - N_2 F_2 = f' \left(1 - \frac{n_o}{n_{k-1}} \right) = \frac{f'}{n_{k-1}} (n_{k-1} - n_o).$$

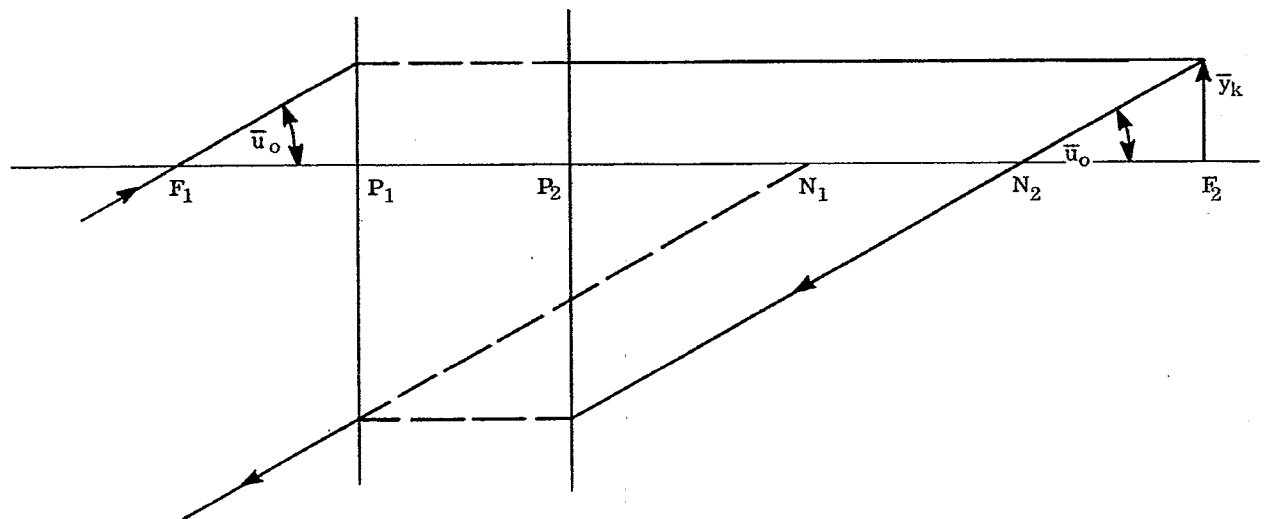


Figure 6.8 - Graphical construction to locate positions of nodal points.

Because of Equation (19), the following relations hold between the cardinal points.

$$P_1 N_1 = P_2 N_2 ,$$

$$P_1 P_2 = N_1 N_2 ,$$

$$F_1 N_1 = f' ,$$

$$N_2 F_2 = f .$$

And for object and image media the same, $n_o = n_{k-1}$, $f = f'$, and the principal and nodal points coincide, P_1 with N_1 , and P_2 with N_2 .

6.5.8.3 Because $P_1 P_2 = N_1 N_2$, two parallel lines, one through each nodal point, will intersect the principal planes in points equidistant from the axis. Hence these two rays are conjugate rays, and we have the important fact that any ray in object space which is heading toward N_1 will emerge from the system in the same direction from N_2 . This gives us a graphical method for locating the nodal points, shown in Figure 6.8. A ray is shown entering the system at an angle \bar{u}_o headed towards F_1 until it intersects the plane at P_1 . It then emerges from the plane at P_2 parallel to the axis at the image height \bar{y}_k . A ray then traced backwards at an angle \bar{u}_o with the axis must emerge anti-parallel to the entering ray as shown in the illustration, because all rays leaving a point on the focal plane are parallel to each other after emerging from the system. The two points N_1 and N_2 are the intersections with the axis of the two segments of this backwards traced ray.

6.5.9 Numerical example. A numerical example, represented in Figure 6.9, shows the location of the cardinal points of a lens with water on one side and air on the other. Given the three indices, two curvatures, and lens thickness, all other numerical values can be found using the equations already developed. An axial ray is traced through the system at $u_o = 0$ and y_1 arbitrary. t_2 can be found, using $y_3 = 0$. Therefore F_2 is located with respect to the second surface of the lens. A corresponding trace locates F_1 . Equations (13) and (19) give f' and f respectively. The principal points and nodal points can now be located.

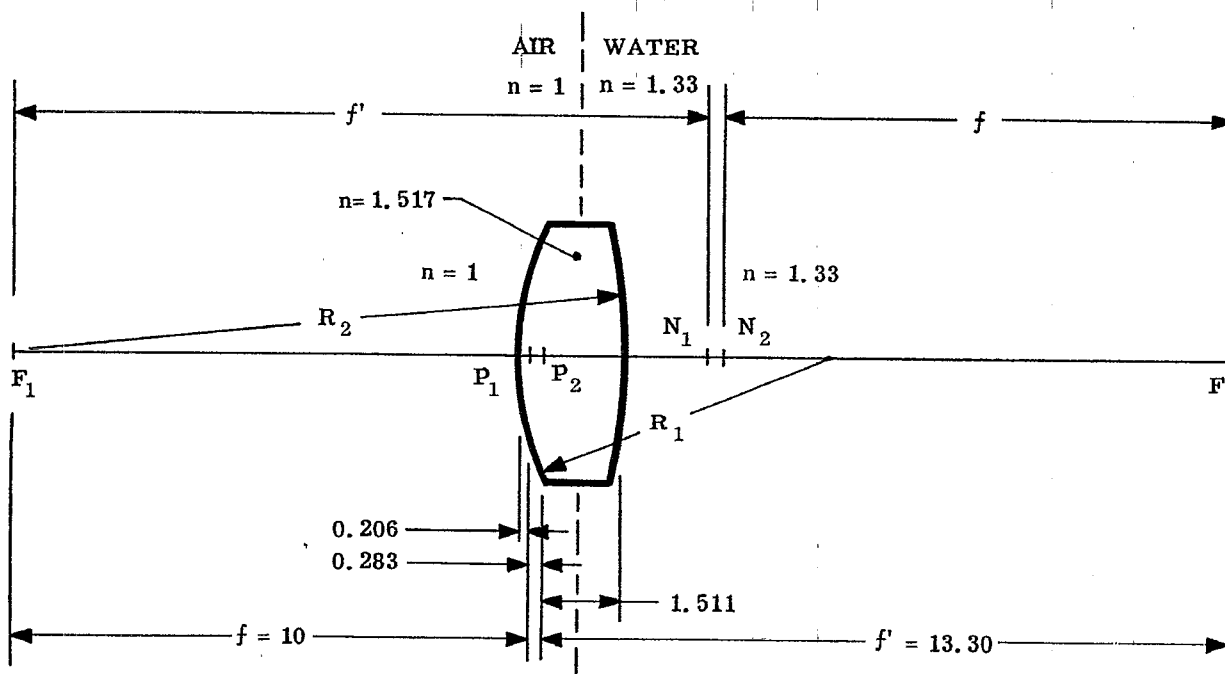


Figure 6.9- Numerical example showing location of the cardinal points for a lens with water on one side.

6.6 CALCULATION OF THE FOCAL LENGTH FROM FINITE CONJUGATE DATA

6.6.1 General. If an optical system is to be used at infinite conjugate, that is either the object or image or both are at infinity, then the entering axial ray is traced at $u_o = 0$, y_1 arbitrary. (For systems having the image at infinity for a finite object, the design is considered as if the rays went backwards through the system. Systems are therefore designed with the infinite conjugate as object, whether or not this agrees with the physical situation. The justification for this is that an optical system is reversible in the sense that rays traverse the same path in either direction). The ray trace automatically gives the focal length, f' , by using Equation (13).

6.6.2 Finite conjugates. However, if the system images a finite conjugate object, and an axial ray and an oblique paraxial have been traced, Equation (13) does not apply. It is possible, nevertheless, from the data obtained from these two rays, to calculate the focal length. If two rays have been traced as shown in Figure 6.4 and in the presentation of Table 6.6, then

$$A = \bar{y}_o / \bar{y}_1$$

and

$$B = -\bar{y}_1 \bar{u}_o / \bar{y}_o u_o, \text{ this latter being Equation (12).}$$

With these constants known, it is possible to predict the final \bar{u}_{k-1} for a ray entering the lens parallel to the axis. For then $\bar{u}_o = 0$ and $\bar{y}_o (= \bar{y}_1)$ are the initial conditions for the third ray.

Now writing Equation (10b) for the final angle,

$$\bar{u}_{k-1} = \frac{\bar{y}_o}{y_o} \bar{u}_{k-1} - \frac{\bar{y}_1 \bar{u}_o}{y_o u_o} u_{k-1}.$$

From this equation, and Equation (13) written for the third ray, we have

$$f' = - \frac{\Phi}{n_o (u_o \bar{u}_{k-1} - \bar{u}_o u_{k-1})} \quad (20)$$

where $\Phi = \bar{y}_o (n_o u_o)$ from Paragraph 6.3.6.

6.7 SYSTEMS OF THIN LENSES IN AIR

6.7.1 Concept of the thin lens.

6.7.1.1 None of the basic material presented so far presupposes any specific form of the optical system other than that it is a centered system. We now want to specialize the system somewhat and consider a single lens, an example of which is shown in Figure 6.9. In that example, $n_o \neq n_2$, so that $f \neq f'$. If $n_o = n_2 = 1$, the lens is in air, and $f = f'$. The nodal and principal points coincide as explained in Paragraph 6.5.8.2. Because of the equality of the two focal lengths, Equations (15) through (18) can be simplified.

6.7.1.2 An additional simplification can be attained by assuming that the axial lens thickness, t_1 in the above example, is small compared with t_o and t_2 . If t_1 can be neglected, the lens is called a thin lens. Since the two deviations of the ray are considered to occur at one point, for a thin lens, both principal planes coincide with the lens of zero thickness. For this case, S and S' are the distances measured to the intersection of the lens with the optical axis, and Equations (17) and (18) take the familiar form for a thin lens in air. The two nodal points also coincide with the lens; hence a ray directed towards the lens center will emerge from the same point in the same direction. In some special cases, such as high curvature meniscus lenses (highly warped lenses), the thickness may be small, but not completely negligible. In these cases the lens may be "thin" for certain applications (for example, calculation of focal length), but not "thin" for others (for example, calculation of principal points positions). In such intermediate cases, where the lens is neither completely thick or completely thin, the principal and nodal points do not necessarily coincide with the center of the lens.

6.7.2 Focal length and power of a thin lens in air. Many optical systems are made up of individual two-surface lenses separated by air. Paraxial rays can, of course, be traced through any system of this type by using Equations 5-(56) and 5-(57), but considerable simplification can be made if it can be assumed that the individual lenses are thin. In the layout shown in Table 6.8, an axial paraxial ray and an oblique paraxial

ray are traced through a thin, two-surface element in air. For the axial paraxial ray, we have

$$u_o = y/t_o ,$$

$$u_2 = y/t_o + y(1-n)c_1 + y(n-1)c_2 ,$$

and

$$u_2 = u_o - y(n-1)(c_1 - c_2) . \quad (21)$$

The focal length may be calculated from Equation (20), using $\Phi = \bar{y}_o (n_o u_o)$. If numerical calculations are made, the data are found in a table similar to Table 6.8. Therefore,

$$f' = - \frac{\Phi}{n_o (\bar{u}_2 u_o - \bar{u}_o u_2)} = \frac{t_o \bar{u}_o u_o}{(\bar{u}_o u_o - \bar{u}_o u_2)} ,$$

or

$$f' = \frac{t_o u_o}{u_o - u_2} = \frac{1}{(n-1)(c_1 - c_2)}$$

and

$$1/f' = (n-1)(c_1 - c_2) = \phi . \quad (22)$$

Equation (22) is the well known formula for the focal length of a thin lens in air. It is more convenient to use it in the latter form, where ϕ is called the power of the thin lens.

SURFACE	Object	1	2	3
c	c_o	c_1	c_2	c_3
t	t_o	0	t_2	
n	1	n	1	
$(n_1 - n)c$	0	$(1-n)c_1$	$(n-1)c_2$	
t/n	t_o	0	t_2	
y	0	y	y	0
nu	y/t_o	$y(1-n)c_1 + y/t_o$	$y(n-1)c_2 + y(1-n)c_1 + y/t_o$	
\bar{y}	$-t_o \bar{u}_o$	0	0	
$n\bar{u}$	\bar{u}_o	\bar{u}_1	\bar{u}_2	

$$\bar{u}_o = \bar{u}_1 = \bar{u}_2$$

Table 6.8- Paraxial rays traced through a thin lens.

6.7.3 Ray trace equations for thin lens systems in air.

6.7.3.1 Equation (21) can be written

$$u_2 = u_o - y \phi .$$

The similarity between this and Equation 5-(57) is now apparent. Equation 5-(56) can be used to transfer between lenses. We have then the transfer and refraction equations for thin lens systems. These equations, (23) and (24), are written for a general thin lens j .

$$y = y_{-1} + t_{-1} u_{-1} , \quad (23)$$

$$u = u_{-1} + y(-\phi) . \quad (24)$$

Table 6.9 illustrates a method using Equations (23) and (24) for calculating the familiar expression for the

focal length of a dialyte, i.e., two thin lenses separated by the distance d .

SURFACE	LENS (a)	LENS (b)	IMAGE
$-\phi$ d	$-\phi a$	$-\phi b$	
y u	y_1 0	$(1-d\phi a)y_1$ $-\phi a y_1$	$(-\phi a - \phi b + d\phi a \phi b)y_1$

$$\frac{1}{f'} = \phi = - \frac{u_{k-1}}{y_1} = \phi a + \phi b - d \phi a \phi b$$

Table 6.9 - Tracing a paraxial ray, $u_0 = 0$ and y_1 arbitrary through two thin lenses.

6.7.3.2 The tracing of paraxial rays through thin lens systems is probably the one remaining calculation that lens designers do on desk calculators. In optical design work, a great deal of time and thought must necessarily go into the preliminary layout work. The designer must decide where to place the lenses, and what focal lengths are to be used. He needs to know approximately the sizes of lenses needed, and the approximate path of rays as they pass through the system. All these calculations can be made assuming thin lenses, and it is a problem so varied that it does not lend itself well to a large computer. Experience shows that desk calculators or slide rules are preferred at this stage of the design.

6.8 OPTICAL SYSTEMS INVOLVING MIRRORS

6.8.1 Sign conventions. It was pointed out in Section 2.3.3 that the equation of refraction could be used for reflection by merely writing

$$n_{+1} = -n.$$

If this is done in all the refraction equations, they can be used for reflection. If a mirror is inserted in an optical system, it reflects the ray backwards so that if the light was originally traveling from left to right, it will travel from right to left after reflection. It is possible to treat reflecting surfaces in exactly the same way as refraction surfaces by adopting the following rules:

- (1) Write all the curvatures with the usual sign convention. If a single surface is encountered several times in a reflecting system, the radius is always considered to have the same sign.
- (2) Whenever the light travels from right to left, insert the index and thickness with a negative sign.

6.8.2 A mirror system and its ray tracing format. A typical mirror and lens system is shown in Figure 6.10. The proper way to lay out the data for ray tracing is shown in Table 6.10. Actual rays as well as paraxial rays can then be traced through this system exactly as though it were only a refracting lens. If the light travels from right to left in the j th space one must remember that the index of refraction (n_j) is negative.

6.8.3 First order imagery in a mirror.

6.8.3.1 By using the above procedure it is now possible to readily work out the first order optics of a single mirror. The problem is illustrated in Figure 6.11, and worked out in the presentation shown in Table 6.11. From the table, it is apparent, by applying Equation 5-(56), that

$$y_2 = 0 = 1 + (-t_1) \left(\frac{1}{t_0} + 2c \right).$$

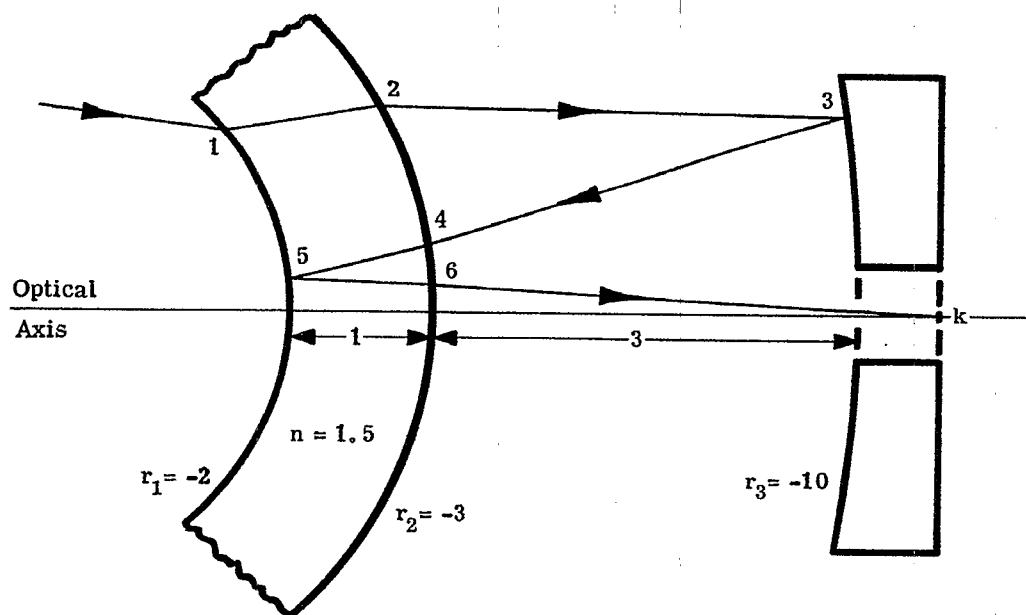


Figure 6.10 - The path of rays through a mirror system.

SURFACE	OBJECT	1	2	3	4	5	6	k
c	0	-0.500	-0.330	-0.100	-0.330	-0.500	-0.330	0
t	∞	1.000	3.000	-3.000	-1.000	1.000	1.000	
n	1.000	1.500	1.000	-1.000	-1.500	1.500	1.000	
$c(n_{-1} - n)$	0	0.250	-0.165	-0.200	-0.165	1.500	-0.165	0
t/n	∞	0.667	3.000	3.000	0.667	0.667		

Table 6.10 - Computing sheet format for mirror system illustrated above. Only the lens constants are included in the above table. The calculations, which are not given, are carried out as in Table 6.6.

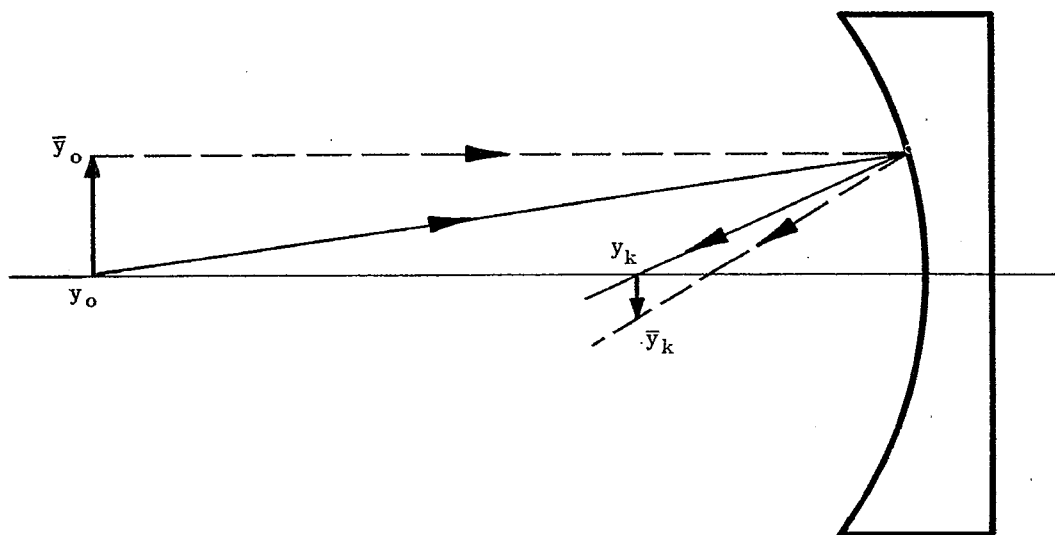


Figure 6.11 - Imaging an object in a concave mirror.

SURFACE	OBJECT	1	IMAGE
c	0	c	0
t	t_o	t_1	
n	1	-1	
$c(n_{-1} - n)$	0	$2c$	0
t/n	t_o	$-t_1$	
y	0	1	0
nu	$1/t_o$	$1/t_o + 2c$	
\bar{y}	1	1	$1 - t_1 2c$
$n\bar{u}$ *	0	$2c$	

* Ray traced parallel to axis to calculate focal length directly.

Table 6.11 - Ray tracing through a single mirror system.

Therefore

$$\frac{1}{t_1} = \frac{1}{t_o} + 2c = \frac{1}{t_o} + \frac{2}{r} . \quad (25)$$

For a numerical example assume $r = -10$ and $t_o = 20$. Then $t_1 = -20/3$. The minus sign indicates that the image surface lies to the left of the mirror surface, as shown in Figure 6.11. The same equation could have been derived using Equation (1),

$$\frac{n_1}{t_1} + \frac{n_o}{t_o} = c_1 (n_1 - n_o)$$

and setting

$$n_1 = -n_o .$$

The magnification for the mirror may be found from Equation (7),

$$m = \frac{n_o u_o}{n_1 u_1} = \frac{1/t_o}{(1/t_o) + 2c} = t_1 / t_o .$$

The same equation could have been derived from Equation (18), remembering that $f' = -f$ because $n_1 = -n_o$.

6.8.3.2 The focal length of the mirror may be found by tracing a paraxial ray through the mirror at $\bar{y}_1 = 1$ and $\bar{u}_o = 0$ as noted in the lower two lines in Table 6.11. Equation (13) can be written

$$f' = - \frac{y_1 n_{k-1}}{(n_{k-1} u_{k-1})} ,$$

and used with the ray at $\bar{u}_o = 0$.

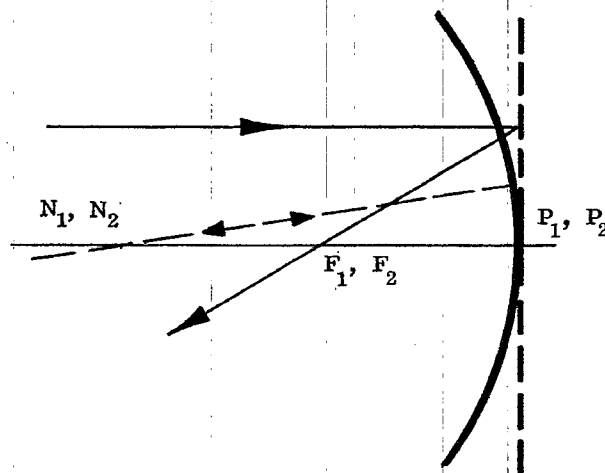


Figure 6.12 - The location of the principal points, focal points, and nodal points for a single mirror system.

Since $n_0 = -n_{k-1}$, $\bar{y}_1 = 1$, and $(n_{k-1} \bar{u}_{k-1}) = 2c$, we have,

$$f' = \frac{n_0}{2c} = n_0 \frac{r}{2}.$$

If r is negative as it is in Figure 6.11, f' is negative indicating that F_2 lies to the left of P_2 . The equation $P_2 F_2 n_0 = n_{k-1} N_2 F_2$ shows that for the mirror,

$$P_2 F_2 = -N_2 F_2.$$

Since $P_2 F_2$ is negative for the example shown in Figure 6.11, then $N_2 F_2$ is positive. The location of P_1 , N_1 , F_1 , P_2 , N_2 and F_2 are shown in Figure 6.12. The nodal points are at the center of curvature.

6.9 DIFFERENTIAL CHANGES IN FIRST ORDER OPTICS

6.9.1 General.

6.9.1.1 The various steps followed in the design of an optical system are discussed in Section 9. The first two steps of the procedure are (1) selection of type of element for each part of the system, and (2) calculation of a first order thin lens solution. Step (2) involves the calculation of the focal lengths and separations of the individual elements, as well as first order aberrations which will be discussed in Section 6.10. The basic procedure for tracing paraxial rays, and therefore for determining focal lengths and spacings, have already been outlined in Section 6.

6.9.1.2 After the completion of step (2) the designer may feel that some changes are necessary so that the system meets more closely the required specifications. For example, he may have to change the focal length of the system. At the present stage of the system design (thin lens, paraxial rays), the designer can vary only the curvatures, the separations, and the indices of refraction. It therefore becomes important to know how changes in these three parameters affect the first order solution. In the remainder of Section 6.9 formulae will be given for computing the effects on first order optics for differential changes in the lens parameters.

6.9.2 Determination of the differential coefficients.

6.9.2.1 A change of any parameter, such as thickness, index of refraction, or curvature of a surface, will result in the paraxial ray changing its path to the next surface. Specifically, changes in t will change y_{+1} , and changes in n or c will change both u and y_{+1} . These changes will, in turn, cause changes on each surface up to and including the final image. The final changes, dy_k and du_{k-1} , which result from a change of any parameter associated with the j th surface, is certainly a function of changes dy_{j+1} and du_j . If the changes can be assumed to be differentials, it is possible to write

$$dy_k = \left(\frac{\partial y_k}{\partial y_{+1}} \right) dy_{+1} + \left(\frac{\partial y_k}{\partial u} \right) du \quad (26)$$

and

$$du_{k-1} = \left(\frac{\partial u_{k-1}}{\partial y_{+1}} \right) dy_{+1} + \left(\frac{\partial u_{k-1}}{\partial u} \right) du. \quad (27)$$

6.9.2.2 The partial derivatives in the above equations are called differential coefficients. If we trace two differential rays through the system, we have two values each for dy_{+1} and du (initial ray data) and two values each for dy_k and du_{k-1} (result of ray trace). Therefore, by tracing two differential rays near a given ray, it should be possible to determine the respective differential coefficients. It was shown in Section 5.9 that a paraxial ray is a differential ray traced near the optical axis. Therefore, we will use the axial paraxial ray and the oblique paraxial ray as the two differentially traced rays near the optical axis, taken as the given ray. It is possible then to evaluate the differential coefficients for changes in y_k and u_{k-1} , by making the following substitutions in Equations (26) and (27):

$$\begin{array}{llll} dy_k = y_k & dy_{+1} = y_{+1} & du = u & du_{k-1} = u_{k-1} \\ d\bar{y}_k = \bar{y}_k & d\bar{y}_{+1} = \bar{y}_{+1} & d\bar{u} = \bar{u} & d\bar{u}_{k-1} = \bar{u}_{k-1} \end{array}$$

Two sets of simultaneous equations are thereby obtained. These equations, when solved for the derivatives, give:

$$\frac{\partial y_k}{\partial y_{+1}} = \frac{(\bar{y}_k u - y_k \bar{u})}{(\bar{y}_{+1} u - y_{+1} \bar{u})} = \frac{n(\bar{y}_k u - y_k \bar{u})}{\Phi}, \quad (28)$$

$$\frac{\partial y_k}{\partial u} = \frac{(y_k \bar{y}_{+1} - \bar{y}_k y_{+1})}{(\bar{y}_{+1} u - y_{+1} \bar{u})} = \frac{n(y_k \bar{y}_{+1} - \bar{y}_k y_{+1})}{\Phi}, \quad (29)$$

$$\frac{\partial u_{k-1}}{\partial y_{+1}} = \frac{(\bar{u}_{k-1} u - u_{k-1} \bar{u})}{(\bar{y}_{+1} u - y_{+1} \bar{u})} = \frac{n(\bar{u}_{k-1} u - u_{k-1} \bar{u})}{\Phi}, \quad (30)$$

and

$$\frac{\partial u_{k-1}}{\partial u} = \frac{(\bar{y}_{+1} u_{k-1} - y_{+1} \bar{u}_{k-1})}{(\bar{y}_{+1} u - y_{+1} \bar{u})} = \frac{n(\bar{y}_{+1} u_{k-1} - y_{+1} \bar{u}_{k-1})}{\Phi}. \quad (31)$$

6.9.3 Effect of curvature change on focal length.

6.9.3.1 The change in focal length, df' , due to changes in curvature, thickness, and index is given by

$$df' = \left(\frac{\partial f'}{\partial c} \right) dc + \left(\frac{\partial f'}{\partial t} \right) dt + \left(\frac{\partial f'}{\partial n} \right) dn.$$

If the differential coefficients are known, then df' can be found for any small change in the system parameters. It will now be assumed that t and n are held constant.

6.9.3.2 Combining the transfer equation

$$y_{+1} = y + tu,$$

with the above substitutions we have, for the case of $t = \text{constant}$,

$$dy_{+1} = t du.$$

Using this and Equations (28) to (31), Equations (26) and (27) become

$$dy_k = \frac{n}{\Phi} (y_k \bar{y} - \bar{y}_k y) du, \quad (32)$$

and

$$du_{k-1} = \frac{n}{\Phi} (\bar{y} u_{k-1} - y \bar{u}_{k-1}) du. \quad (33)$$

6.9.3.3 Equation (13), defining the second focal length, assumes that the axial paraxial ray was traced at $u_o = 0$. Differentiating this equation, remembering that y_1 is arbitrary and hence independent of c , we have

$$\frac{df'}{dc} = - \left(\frac{f'}{u_{k-1}} \right) \left(\frac{du_{k-1}}{dc} \right) = - \left(\frac{f'}{u_{k-1}} \right) \left(\frac{du_{k-1}}{du} \right) \frac{du}{dc}.$$

Differentiating 5-(57) it follows that

$$\frac{du}{dc} = \frac{y(n_{-1} - n)}{n}.$$

Therefore, using Equation (33),

$$\frac{df'}{dc} = \frac{[-f'(\bar{y} u_{k-1} - y \bar{u}_{k-1})][y(n_{-1} - n)]}{\Phi u_{k-1}}.$$

6.9.4 Effect of curvature change on final angle. In Table 6.12 a calculation is shown for a change in curvature made on the fourth surface of the example given in Table 6.6. Comparing the new u_6 with the original one in Table 6.6 we have $\Delta u_6 = 0.00469$. Now we will compare this value with a calculated value using the equations for the differential coefficients. Since we are making a change in the curvature only, keeping the thickness and index constant, we calculate

$$\frac{du_{k-1}}{dc} = \frac{du_{k-1}}{du} \frac{du}{dc}.$$

From Equation (32), and data from Table 6.6, the following calculation may be made,

$$\begin{aligned} \frac{du_6}{dc_4} &= \frac{y_4 (n_3 - n_4)}{\Phi} (\bar{y}_4 u_6 - y_4 \bar{u}_6) \\ &= - \frac{1.026 \times .621}{.5} (-0.00069 \times 0.08664 - 1.02606 \times 0.35886) \\ &= 0.469. \end{aligned}$$

We have then that

$$\Delta u_6 = \frac{du_6}{dc_4} \Delta c_4 = (0.469)(0.01) = 0.00469.$$

This is in exact agreement with the result from Table 6.12.

6.9.5 Effect of thickness change on final angle. It is also possible to compute the change in the final angle from a change in any thickness t . If t is changed, then,

$$dy_{+1} = u dt.$$

SURFACE	4	5	6	7
c	0.26973	0.05065	-0.24588	0
t	1.13691	0.6	14.9709	
n	1.621	1	1.620	1
$c(n_{-1} - n)$	0.16750	-0.03140	-0.15245	0
t/n	1.13691	0.37037	14.9709	
y	1.02606	1.18794	1.22686	0
nu	-0.02948	0.14238	0.10508	-0.08195

$$\Delta u_6 = -0.08195 - (-0.08664) = 0.00469$$

Table 6.12 - Calculations showing the effect on u_{k-1} of a change of $\Delta c_4 = 0.01$ in the data in Table 6.6

Therefore, using Equation (27) with $du = 0$, and Equation (30),

$$\frac{du_{k-1}}{dt} = \frac{\partial u_{k-1}}{\partial y_{+1}} \frac{dy_{+1}}{dt},$$

and

$$\frac{du_{k-1}}{dt} = \frac{nu}{\Phi} \left[\bar{u}_{k-1} u - u_{k-1} \bar{u} \right].$$

6.10 CHROMATIC ABERRATION

6.10.1 The meaning of chromatic aberration. The variation of refractive indices with wavelength was discussed under the topic of dispersion in Section 2.6. The method of differential coefficients described in Section 6.9 can be used to calculate the effect of such a change in the index of refraction of the lenses. This change in index affects the refraction of each ray so that rays of different wavelengths pass through the system in slightly different paths. Generally these rays of different wavelengths give rise to more than a single image, a phenomenon called chromatic aberration. If the images are at different positions along the optical axis, the system exhibits longitudinal or axial chromatic aberration. If the images are of different lateral magnification, the system exhibits transverse or lateral chromatic aberration. Axial and lateral chromatic aberrations are sometimes referred to as axial color and lateral color, respectively.

6.10.2 Surface contributions.

6.10.2.1 As mentioned above, each surface introduces a certain amount of chromatic aberration appearing in the final image. The amount due to a particular surface is called the surface contribution. The general approach used to calculate first and third order aberrations is (1) determine the surface contribution, and (2) sum the contributions for all surfaces to find the total aberration. The individual contributions may be positive, negative, or zero. Hence the sum may be either positive, negative, or zero. In the last case the system would be free of this particular aberration.

6.10.2.2 The first order chromatic aberration contribution of any surface may be found by differentiating Equation 5-(57), assuming that $du_{-1} = 0$. This assumption means that the ray between the $j-1$ and j th surfaces is unaberrated; hence we are considering only the contribution of the j th surface. The assumption $du_{-1} = 0$ also leads to $dy = 0$, because the ray to the left of the j th surface retains its original path. We then have,

$$ndu + udn = u_{-1} dn_{-1} + yc(dn_{-1} - dn).$$

This can be put into a form more suitable for calculation. From Equation 5-(35), written for small angles, and Equation 6-(4), we have

$$i' = yc + u. \quad (33a)$$

Using this equation, Equation 2-(1) for small angles, and Equation 6-(4), it is possible to derive the expression

$$du = i \frac{n_{-1}}{n} \left[\left(\frac{dn_{-1}}{n_{-1}} \right) - \left(\frac{dn}{n} \right) \right].$$

6.10.2.3 Here, dn and dn_{-1} represent infinitesimal changes in index due to an infinitesimal change in wavelength λ . The change in u , due to a change of dn_{-1} and dn , will thus cause the ray to take a deviated path to the image. The change dy_k , in the final image, may then be calculated from

Equation (32). Since $y_k = 0$ for the axial ray, the value of dy_k is:

$$dy_k = - \frac{y \bar{y}_k n_{-1} i}{\Phi} \left[\left(\frac{dn_{-1}}{n_{-1}} \right) - \left(\frac{dn}{n} \right) \right] ;$$

$$dy_k = - \frac{y n_{-1} i}{(n_{k-1} u_{k-1})} \left[\left(\frac{dn_{-1}}{n_{-1}} \right) - \left(\frac{dn}{n} \right) \right] ;$$

$$dy_k = y n_{-1} i \left[\Delta \frac{dn}{n} \right] / (n_{k-1} u_{k-1}) ; * \quad (34)$$

or

$$dy_k = - a / (n_{k-1} u_{k-1}) .$$

The above derivation could have been equally well carried out for the oblique paraxial ray giving

$$d\bar{y}_k = y n_{-1} \bar{i} \left[\Delta \frac{dn}{n} \right] / (n_{k-1} u_{k-1}) = - b / (n_{k-1} u_{k-1}) \quad (35)$$

$$= dy_k \bar{i} / i . \quad (36)$$

6.10.3 Total chromatic aberration. Equation (34) gives the amount by which the image of an axial object point is displaced from the optical axis due to the j th surface. Similarly Equation (35) applies to the image of an object point off the axis. Both these equations give the transverse displacement in the final paraxial image plane due to changes dn_{-1} and dn . Now, if these changes are due to a change of wavelength $d\lambda$, changes dn and dn_{-1} occur at every surface in the lens. Each surface then contributes a dy_k and a $d\bar{y}_k$, and since they are all differentials, they are directly additive. The totals are

$$\text{total } dy_k = \text{TAch} = \frac{-1}{(n_{k-1} u_{k-1})} \sum_{j=1}^{j=k-1} a , \quad (37)$$

and

$$\text{total } d\bar{y}_k = \text{Tch} = \frac{-1}{(n_{k-1} u_{k-1})} \sum_{j=1}^{j=k-1} b , \quad (38)$$

where a and b are the chromatic surface coefficients. Note that Equation (34) has i while Equation (35) has \bar{i} . In all other terms the equations are identical. The symbols TAch and Tch have replaced dy_k and $d\bar{y}_k$ as descriptive terms to indicate the total transverse chromatic effects. TAch is the abbreviation for transverse axial chromatic aberration. Tch is the abbreviation for transverse chromatic aberration. A sample calculation for TAch is included in Table 6.7.

6.10.4 Particular wavelengths used to calculate chromatic aberration.

6.10.4.1 The first order chromatic aberration, strictly speaking, is the infinitesimal change, dy_k , resulting from a change dn which is due to a change $d\lambda$. Therefore, in order to calculate the infinitesimals, TAch and Tch , it is necessary to know the index at all wavelengths. As was discussed in Section 2.6.3, indices are measured at only certain standard wavelengths. It is possible to interpolate between standard wavelengths, using an appropriate dispersion formula, in order to calculate the index, and hence the chromatic aberration, at any wavelength.

6.10.4.2 However, in order to obtain accurate indices for ray tracing, it is customary to use only measured indices. Therefore in order to calculate dn , which is now considered a finite change, two wavelengths are chosen n_v and n_r . Then $dn_{v-r} = n_v - n_r$. [v and r indicate wavelengths at the ends (violet and red) of the visible region]. Then a wavelength λ_g between v and r is used as the reference index of refraction. λ_g is any wavelength in the middle part of the spectrum. The paraxial

* $\Delta(dn/n)$ is defined as $\left(\frac{dn}{n} \right) - \left(\frac{dn_{-1}}{n_{-1}} \right)$. The use of Δ is often used in optics to denote the difference between a quantity on the two sides of a refracting surface. For example, $\Delta n = (n - n_{-1})$.

rays are traced at wavelength λ_g . Therefore,

$$T\text{Ach}_{v-r} = (y_k)_v - (y_k)_r,$$

and

$$T\text{ch}_{v-r} = (\bar{y}_k)_v - (\bar{y}_k)_r.$$

The differences are measured in the paraxial image plane where $(y_k)_g = 0$. It should be pointed out that $T\text{Ach}_{v-r}$ and $T\text{ch}_{v-r}$ tell only the difference in y_k and \bar{y}_k for light at wavelengths λ_v and λ_r . In order to calculate other chromatic aberrations, for example $(y_k)_v - (y_k)_g$, the calculations are made with

$$dn_{v-g} = (n_v - n_g).$$

The wavelengths chosen for calculation, depend on the wavelength region of interest. Visual optical systems are usually calculated with

$$n_v = n_F,$$

$$n_g = n_D,$$

and

$$n_r = n_C.$$

6.10.5 Graphical interpretation of axial and lateral color.

6.10.5.1 In Figure 6.13 a simple lens is shown with an exaggerated amount of chromatic aberration. A simple converging lens, which is necessarily uncorrected for aberrations, is said to be undercorrected. When a particular aberration is made zero, or smaller than some predetermined tolerance, the lens system is said to be corrected. If the aberration of the system has a sign opposite to that of a simple converging

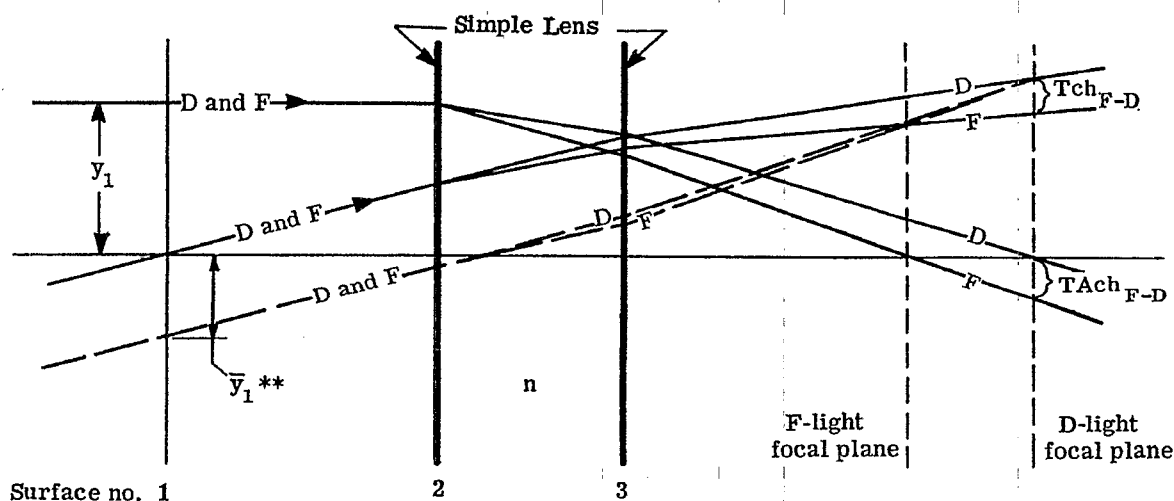


Figure 6.13 - Under-corrected chromatic aberration of axial and oblique rays in a simple lens.

lens, the system is over-corrected. The two surfaces of the lens in Figure 6.13 are labelled 2 and 3, and they appear as planes, as they should in the paraxial region. Axial and oblique rays in D and F light are shown as they pass through the lens. The oblique rays cross the axis at a reference surface #1. This reference plane will often coincide with the entrance pupil of the system. The pupils will be discussed in Section 6.11. With a positive lens, the F light image plane falls closer to the lens than the D light image plane. The chromatic blur, dy_{F-D} , is a linear function of y_1 , the height of the axial ray entering the system. This can be seen by considering Figure 6.13. All axial paraxial rays in D light pass through the same point on the optical axis, independent of y_1 . Hence all values of y_k for D light are zero, and therefore Figure 6.14 indicates a horizontal line for D light. Similarly all axial paraxial rays in F light pass through a common point on the optical axis, independent of y_1 . Hence the separation of the two focal planes for F light and D light is a constant, independent of y_1 . This separation is called the longitudinal axial chromatic aberration, and is denoted by LA_{F-D} . From Figure 6.13 it is seen that

$$Tach_{F-D} = (LA_{F-D}) u_{k-1} = - (LA_{F-D}) \frac{y_1}{f'}$$

Because the chromatic blur, $Tach_{F-D}$, is a linear function of y_1 , the line for F light in Figure 6.14 is straight and inclined to that for D light at the angle $(LA_{F-D})/f'$. Figure 6.14 shows a plot for $(y_k)_F$ and $(y_k)_D$ in the D light image plane, as a function of the height of the axial ray on the entrance pupil plane. This is a recommended way to indicate the transverse axial chromatic aberration of a system.

6.10.5.2 Figure 6.15 shows a plot of \bar{y}_k versus \bar{y}_1 for F and D light. The chromatic blur, $d\bar{y}_{F-D}$, is a linear function of \bar{y}_1 for a reason similar to that given in Paragraph 6.10.5.1. For all values of \bar{y}_1 , all D rays pass through a common point on the D light focal plane. Similarly, all F rays pass through a common point. Since the rays are paraxial, the oblique ray at $\bar{y}_1 = 0$ can be considered as an auxiliary axis; hence a ray parallel to it through a point $\bar{y}_1 \neq 0$ will make the same angle with the chief ray that an axial paraxial ray makes with the optical axis. The former angle is a linear function of \bar{y}_1 , as u_{k-1} is a linear function of y_1 . Hence the chromatic blur is a linear function of \bar{y}_1 , and the F light line is straight in Figure 6.15. The distance between the F and D chief rays in the D light image plane is as indicated in Figure 6.15. This is the value computed from Equation (38). The differ-

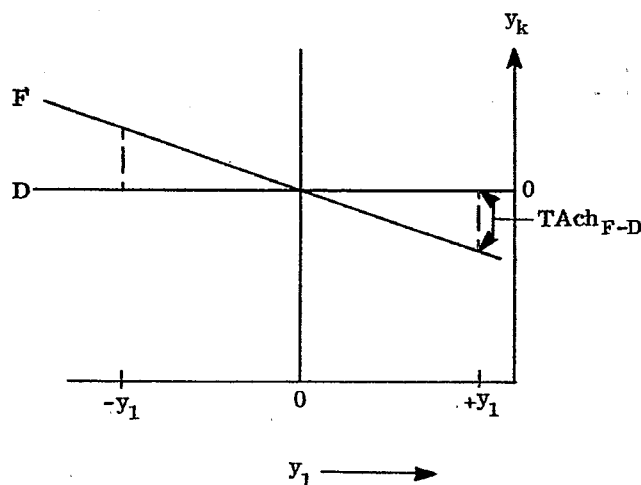


Figure 6.14 - A plot of y_k for F and D light versus the height y_1 of the axial paraxial rays on the entrance pupil plane.

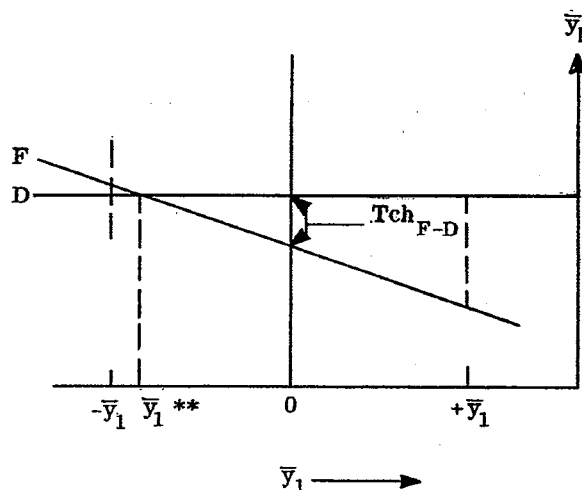


Figure 6.15 - A plot of \bar{y}_k for F and D light versus the height \bar{y}_1 of the oblique paraxial rays.

ence in slope between the F and D lines is the same as for the axial rays shown in Figure 6.14, because the proportionality constant between Tach and y_1 is identical to that between Tch and \bar{y}_1 . Figure 6.15 (and also Figure 6.13) shows that there is a value of \bar{y}_1^{**} such that $Tch_{F-D} = 0$. This means that if the oblique paraxial ray had been taken through the lens at a value of $\bar{y}_1 = \bar{y}_1^{**}$, instead of $\bar{y}_1 = 0$, then Tch_{F-D} would come out to be zero. In fact, in general, it can be said that

$$Tch^*_{F-D} = Tch_{F-D} + \frac{\bar{y}_1^*}{y_1} Tach.$$

6.10.5.3 This equation states that Tch_{F-D} can be calculated for any oblique ray striking the entrance pupil plane at \bar{y}_1^* with the above equation. The (*) is used to indicate the Tch for some oblique ray displaced from the ray passing through $\bar{y}_1 = 0$. Defining $\bar{y}_1^*/y_1 = Q$, the above equation may be written

$$Tch^*_{F-D} = Tch_{F-D} + Q Tach. \quad (39)$$

Again it can be seen that it is necessary to trace only two paraxial rays through a lens system. It is possible to compute Tach, and Tch for any other rays from the data on these two.

6.10.6 Basic concepts in correcting systems for chromatic aberrations.

6.10.6.1 If two wavelengths, F and D for example, come to focus in the same image plane, then $(y_k)_F = (y_k)_D = 0$. This equation gives the condition for correction of the axial color. However, this does not mean that $(u_{k-1})_F$ will necessarily be equal to $(u_{k-1})_D$. If these two angles are not equal, then the magnifications between the object and image will not be equal, and $(\bar{y}_k)_F \neq (\bar{y}_k)_D$. Therefore, the system will have residual lateral color. Hence if both axial and lateral color are to be corrected, the rays in F and D light should emerge from the system at the same value of y_{k-1} and u_{k-1} .

6.10.6.2 The usual achromatic doublet lens is corrected for axial and lateral color because the axial rays in the F and D light never become significantly separated. See Figure 6.16 (a). In the case of two separated lenses, Figure 6.16 (b), it is clear that both elements must be color corrected, to keep the rays together all the way to the final image. If any axial color is allowed in the front element the rear element would have to be thick enough and designed properly to get the two rays together again before emerging from the rear surface. It is possible, by using the proper lens power and glass dispersion, to correct for axial and lateral color in widely spaced lenses as shown in Figure 6.16 (c). This is the principle used in the design of the famous Taylor triplet photographic lens. As a general principle, however, it is always advisable to keep the color rays as close together as possible at all times. This means, if the system is to be made up of several components, each component should be made achromatic.

6.10.7 Chromatic aberration in a thin lens.

6.10.7.1 It is possible to apply Equations (37) and (38) to a thin lens immersed in a non-dispersive medium and simplify the equations because the values of y and of \bar{y} are the same on both surfaces. Suppose there is a thin lens in a system of thin lenses in air with values of y and \bar{y} for heights of the axial and oblique paraxial rays. (See Figure 6.17). This lens will contribute the following amounts of axial and lateral color to the final image.

$$Tach_{v-r} = \frac{1}{(n_{k-1} u_{k-1})} \left(y^2 \frac{\phi}{\nu_{v-r}} \right),$$

and

$$Tch_{v-r} = \frac{1}{(n_{k-1} u_{k-1})} \left(y \bar{y} \frac{\phi}{\nu_{v-r}} \right).$$

where ϕ is the power of the lens, and $\nu_{v-r} = (n_g - 1)/(n_v - n_r)$. These equations follow from Equations (37) and (38), with the use of Equations (4), (22), (33a) and 2-(1) for small angles.

6.10.7.2 Each of the thin lenses adds a contribution, so the final axial and lateral color for a system of η

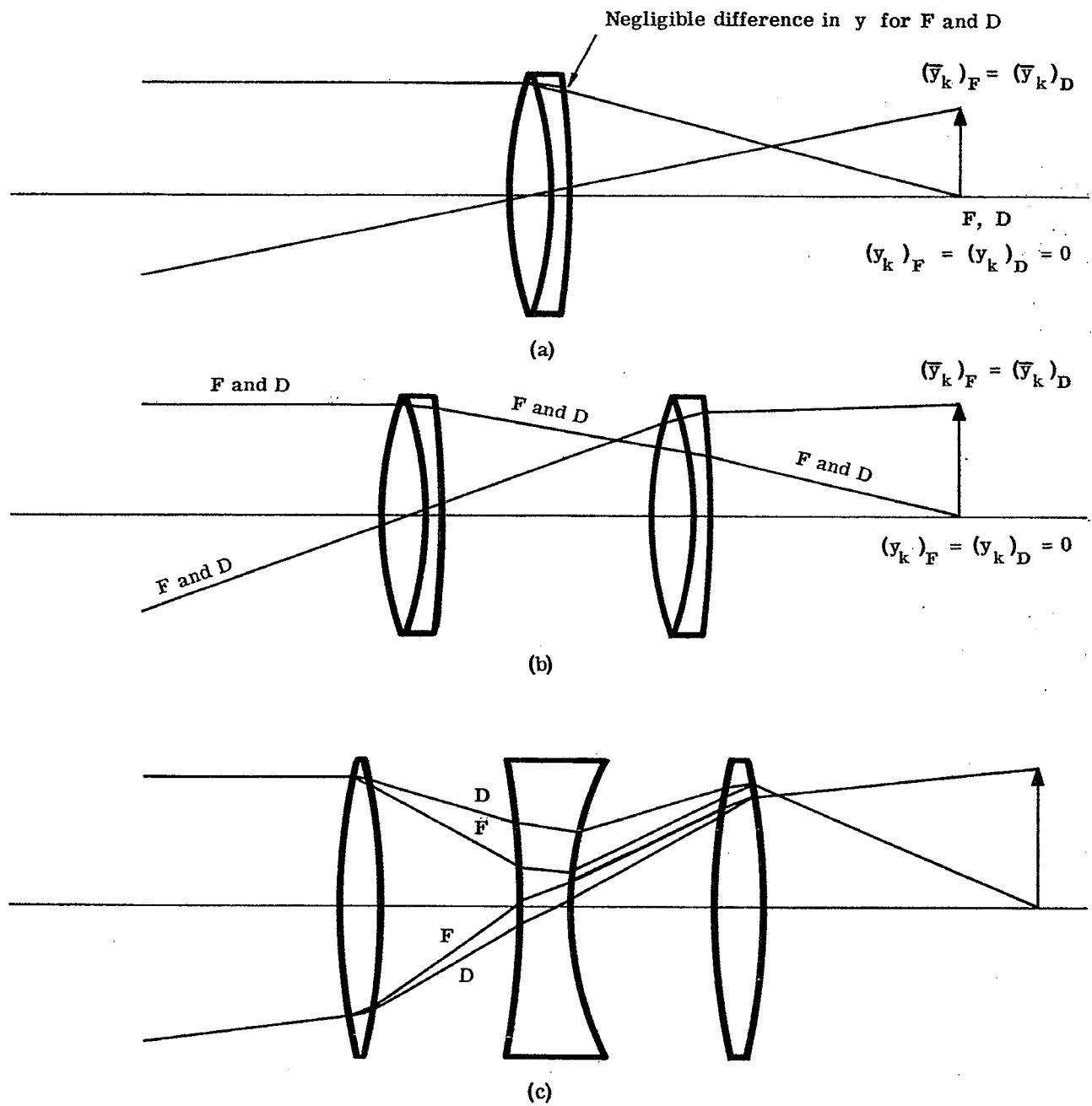


Figure 6.16 - Illustration of axial and lateral color correction for paraxial rays.

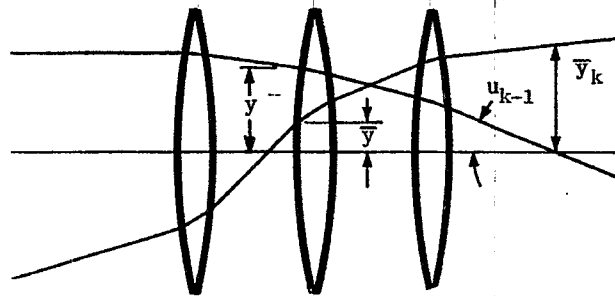


Figure 6.17 - A system of thin lenses.

SURFACE	(1, 2)	(3, 4)	(5, 6)	IMAGE
$-\phi$	-0.16537	0.28698	-0.18208	
t	1.4685	1.4868		
y	1.5	1.1357	1.2515	
u	0	-0.2481	0.0779	-0.1500
\bar{y}	-0.8	-0.07119	0.63633	
\bar{u}	0.364	0.49630	0.47587	0.36000
$\nu F-C$	60.3	36.2	60.3	
$a = -\frac{y^2 \phi}{\nu}$	-0.006171	0.010226	-0.004730	$\Sigma a = -0.000675$
$b = -\bar{y}y\phi/\nu$	0.003291	-0.000641	-0.002405	$\Sigma b = 0.000245$

$$T_{Ach} = \frac{-\Sigma a}{(n_{k-1}u_{k-1})} = -0.00450$$

$$T_{ch} = \frac{-\Sigma b}{(n_{k-1}u_{k-1})} = 0.00164$$

Table 6.13 - Thin lens computation of axial and lateral color for a triplet. ($f = 10$)

thin lenses is given by,

$$T\text{Ach}_{v-r} = \frac{1}{(n_{k-1} u_{k-1})} \sum_{j=1}^j \eta \left(y^2 \frac{\phi}{\nu_{v-r}} \right)_j = \frac{-\sum a}{(n_{k-1} u_{k-1})}, \quad (40)$$

and

$$T\text{ch}_{v-r} = \frac{1}{(n_{k-1} u_{k-1})} \sum_{j=1}^j \eta \left(y \bar{y} \frac{\phi}{\nu_{v-r}} \right)_j = \frac{-\sum b}{(n_{k-1} u_{k-1})}. \quad (41)$$

The use of these equations is illustrated in Table 6.13. The system used in the table is very close to the thin lens equivalent of the system shown in Table 6.7. Note how the angle u of the axial ray as it passes through the system is the same, to four decimal places, for both examples. The $T\text{Ach}$ for the equivalent lens is not exactly the same as for the thick lens due to the thicknesses of the elements.

6.10.8 Thin lens achromatic system.

6.10.8.1 If Equation (40) is written for two closely spaced lenses (a) and (b), and combined, there results

$$T\text{Ach}_{v-r} = \frac{1}{(n_{k-1} u_{k-1})} \left[y^2 \left(\frac{\phi}{\nu_{v-r}} \right)_a + y^2 \left(\frac{\phi}{\nu_{v-r}} \right)_b \right]. \quad (42)$$

This is an expression for the axial chromatic aberration of the doublet lens. In order to make $T\text{Ach}_{v-r} = 0$, it is necessary that,

$$\left(\frac{\phi}{\nu} \right)_a = - \left(\frac{\phi}{\nu} \right)_b. \quad (43)$$

In Table 6.9 it was shown that for two thin lenses in contact,

$$\phi = \phi_a + \phi_b.$$

Combining this equation with Equation (43) yields the relations,

$$\phi_a = \phi \frac{\nu_a}{\nu_a - \nu_b}, \quad (44)$$

and

$$\phi_b = -\phi \frac{\nu_b}{\nu_a - \nu_b}. \quad (45)$$

6.10.8.2 Equations (44) and (45) enable one to pick two glasses with different ν -values and calculate the powers of the two lenses to make an achromatic lens. It is important to realize that these equations reduce the transverse axial chromatic aberration to zero only for the two wavelengths λ_v and λ_r . These are the two wavelengths used to compute the value of ν for the glasses, where the ν -number of a glass is defined as,

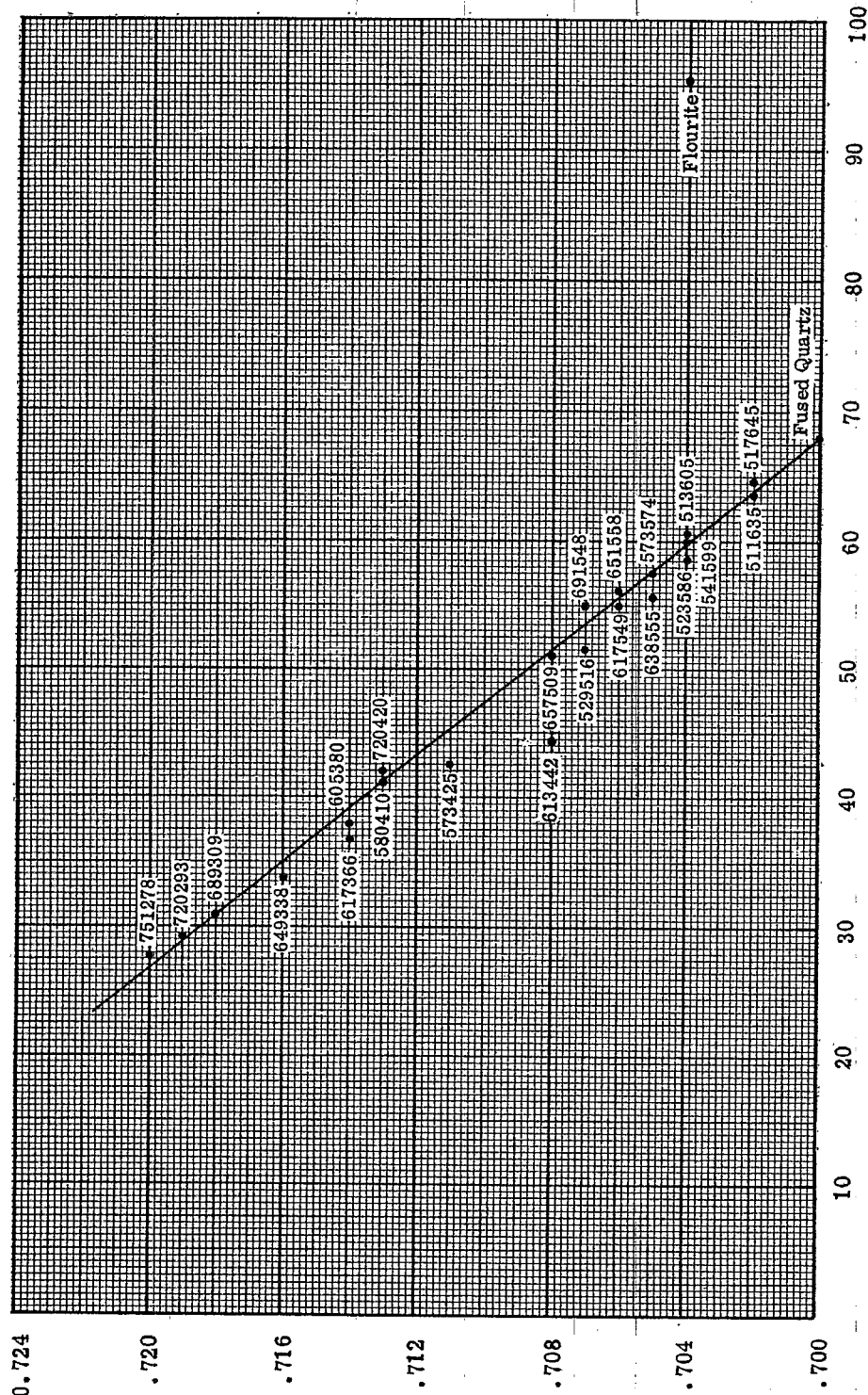
$$\nu_{(v-r)} = \frac{n_g - 1}{n_v - n_r}. \quad (46)$$

On the other hand, other wavelengths do not come to the same focus as λ_v and λ_r . The chromatic aberration $T\text{Ach}_{v-g}$ between an intermediate wavelength λ_g and λ_v may be calculated by substituting $\nu_{v-g} = \frac{n_g - 1}{n_v - n_g}$ for each element and inserting them in Equation (42). Then

$$T\text{Ach}_{v-g} = \frac{1}{(n_{k-1} u_{k-1})} \left[y^2 \left(\frac{\phi}{\nu_{v-g}} \right)_a + y^2 \left(\frac{\phi}{\nu_{v-g}} \right)_b \right]. \quad (47)$$

Since the lens was adjusted to be an achromat for λ_v and λ_r , then Equation (43) must also be satisfied. This equation can be readily inserted in Equation (47) by an obvious redefining of ν_{v-g} , as follows,

$$\nu_{v-g} = \left(\frac{n_g - 1}{n_v - n_g} \right) \left(\frac{n_v - n_r}{n_v - n_r} \right) = \nu_{v-r} \left(\frac{n_v - n_r}{n_v - n_g} \right).$$



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

Figure 6.18 - $\bar{\nu}$ versus $\nu_F - \nu_C$ for several glasses

$$\bar{\nu} = \frac{n_F - n_D}{n_F - n_C}$$

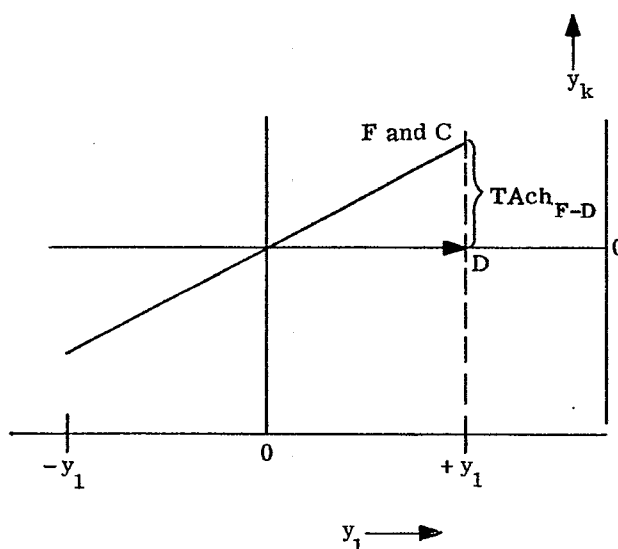


Figure 6.19 - Transverse axial chromatic aberration for an achromatic objective corrected for F and C light.

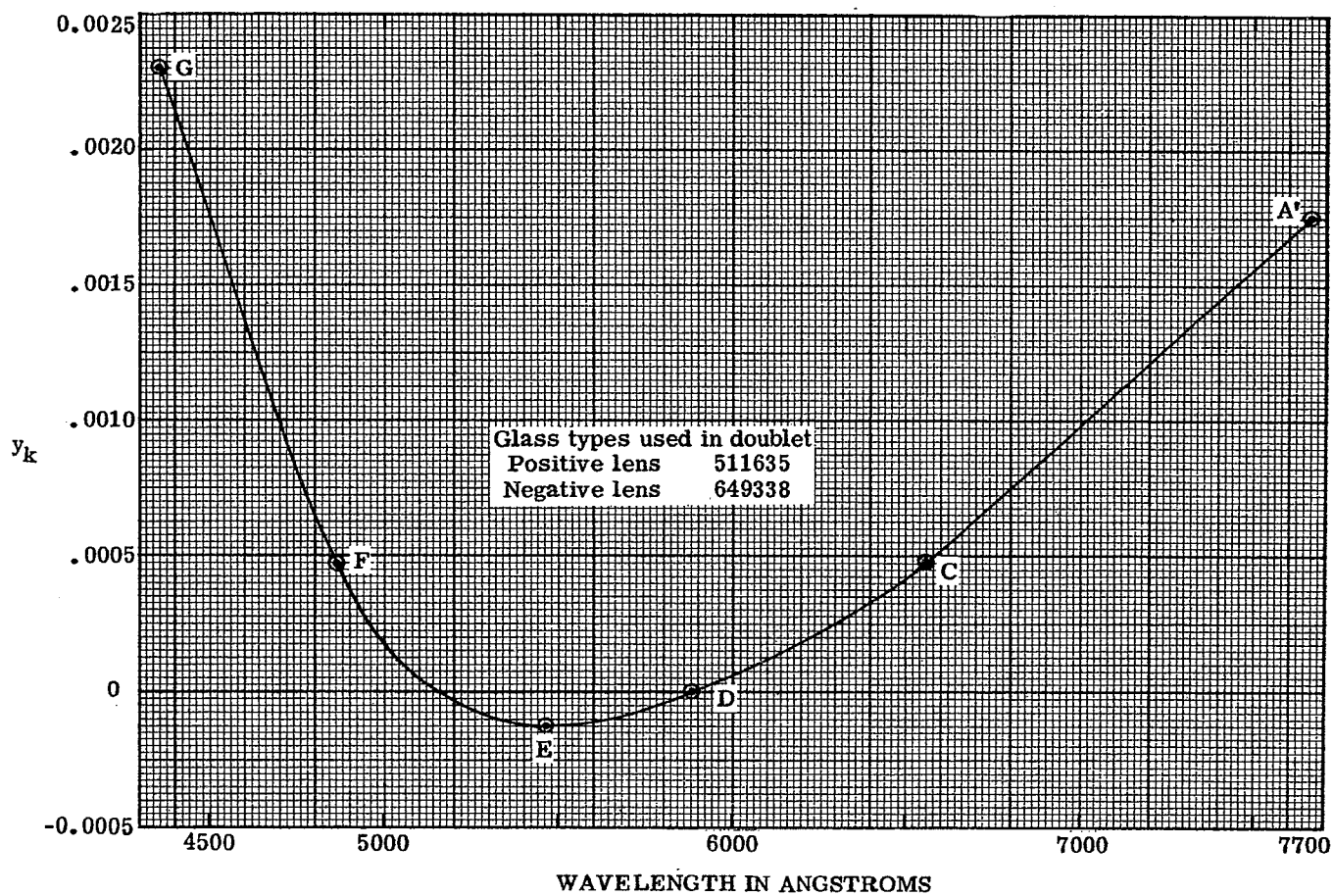


Figure 6.20 - Plot of y_k versus λ for an achromatic doublet.

Defining the partial dispersion ratio (see Paragraph 2.7.3),

$$\tilde{P} = \frac{(n_v - n_g)}{(n_v - n_r)},$$

we have

$$\nu_{v-g} = \nu_{v-r} / \tilde{P}. \quad (48)$$

Equation (47) then becomes, with the help of Equations (44) and (45),

$$T\text{Ach}_{v-g} = \frac{-y}{n_{k-1}} \left[\frac{\tilde{P}_a - \tilde{P}_b}{\nu_a - \nu_b} \right]. \quad (49)$$

6.10.8.3 Equation (49) gives the value of the transverse aberration between λ_v and λ_g , when λ_v and λ_r wavelengths are united. The equation indicates that if λ_v , λ_g , and λ_r are to be brought to focus simultaneously, then $\tilde{P}_a = \tilde{P}_b$. Most glass catalogs give values of \tilde{P} for many combinations of wavelength for each glass. In Figure 6.18 the value of \tilde{P} for $\frac{F-D}{F-C}$ is plotted against ν_{F-C} for several types of glass. As will be pointed out in later sections, a doublet should be designed with low powers of the individual elements. Equations (44) and (45) show that the powers of the (a) and (b) elements of a doublet may be kept small by selecting optical glasses with large differences in ν . Usually doublets should have ν differences larger than 20. As can be seen from the slope of Figure 6.18, for almost any combination of glasses one can select, the ratio of $(\tilde{P}_a - \tilde{P}_b)/(\nu_a - \nu_b)$ is a constant equal to $-1/2200$. When this number is substituted into Equation (49), $T\text{Ach}_{v-g}$ is positive for positive y . Reference to Figure 6.13 indicates that for positive $T\text{Ach}_{v-g}$ the axial ray in D light crosses the axis closer to the lens than the axial ray in F light. Using Equation (13) and noting that $(u_{k-1})_F = (u_{k-1})_D$ to this approximation, we see that if F and C wavelengths are united, then D light focuses closer to the lens by the amount $f'/2200$, if the lens is in air. In Figure 6.19 a plot similar to that of Figure 6.14 is shown for a typical achromatic doublet, corrected to unite F and C light. It is instructive to plot the transverse axial aberration as a function of wavelength. This has been done in Figure 6.20 for an achromatic lens. Note how the curve has a minimum near $\lambda = 5500\text{\AA}$. This is the wavelength at the peak of sensitivity for the eye, which is the reason F - C achromatism is considered to be proper for visual systems.

6.10.8.4 $T\text{Ach}_{F-D}$ is called the secondary spectrum or the secondary color. It is a very difficult aberration to eliminate with ordinary glass types, and often sets the limiting aperture for a lens. The following methods may be used to reduce the secondary spectrum in a lens system.

- (1) Use special materials with equal partial dispersions.
- (2) Use more than two types of glass.
- (3) Use proper combinations of lenses.

More information on the correction of the secondary spectrum will be given in Section 11 under the design of telescope objectives. One can use Equation (40) to compute the secondary color for more complex optical systems, such as air spaced doublets, triplets, or combinations of doublets; however, the algebra becomes so complicated that it is difficult to obtain useful equations like (49) for anything more complicated than a closely packed doublet. It can be shown however, that for a given pair of glasses, the secondary color increases as the air space increases. The relation between secondary spectrum and separation of the two elements is derived by a method similar to that used for Equation (49). First an equation analogous to Equation (42) is derived; this will involve the separation of the elements as well as the powers and ν - numbers. The condition for C - F achromatism, analogous to Equation (43) is then found. The total power for a dialyte from Table 6.9 is used with the achromatic condition to find the analogs of Equations (44) and (45). By the method given in 6.10.8.1, the equations analogous to (47) and (49) are then derived.

6.10.8.5 Although Section 6 deals with first order optics, and hence with the chromatic aberrations, we will mention here one of the third order aberrations, Petzval curvature, because of its close connection with the secondary spectrum. Petzval curvature, known also as curvature of field, has the following physical meaning. For monochromatic light, if spherical aberration, coma, and astigmatism are absent, the point images of point objects lie on a surface, generally curved. Near the optical axis this surface can be considered spherical with a curvature called the Petzval curvature. Flat-field systems have zero or very small Petzval curvature.

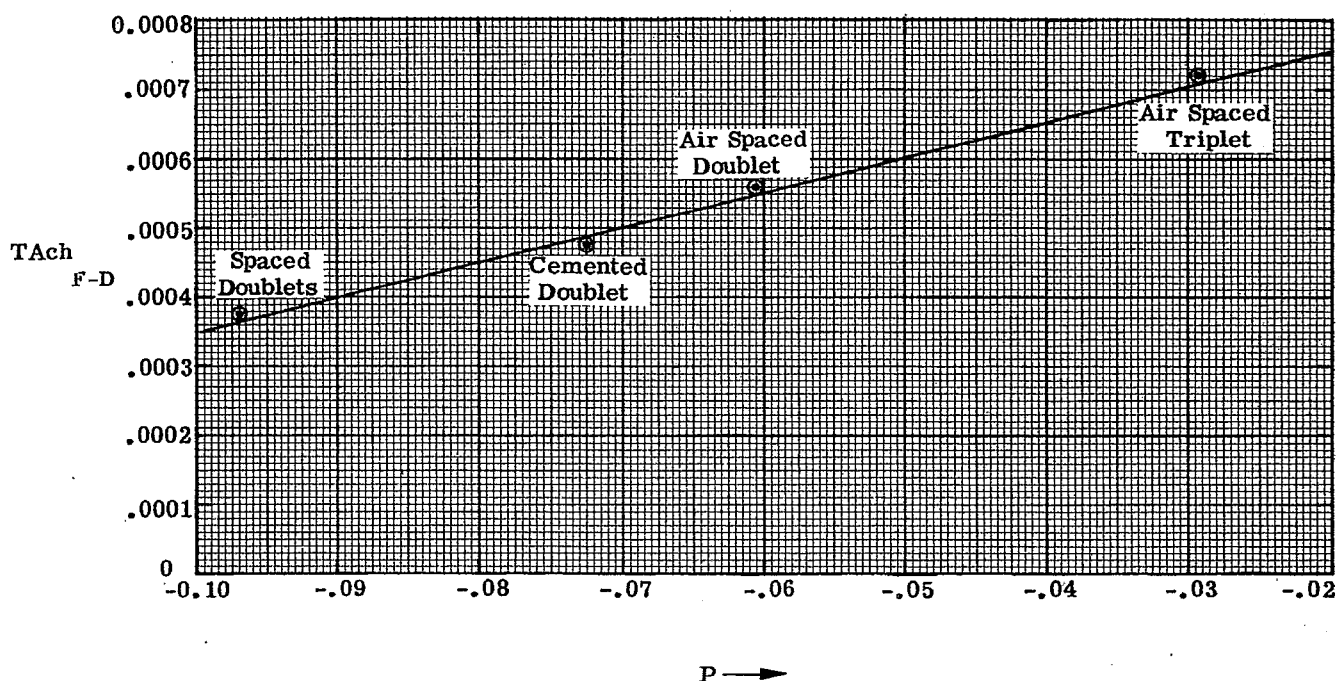


Figure 6.21 - Plot shows qualitative connection between Petzval curvature and secondary color. (Image is assumed in air).

6.10.8.6 Section 8 will discuss how the Petzval contribution for each surface is calculated. When the two surface contributions for a simple lens are added, as was done for the chromatic contributions in Paragraph 6.10.7.1, the Petzval contribution of a simple lens is $P = -\phi/n$. For a system of thin lenses in air, $-P$ is the sum of the power, ϕ , divided by the index for each lens. $P = -\sum_{j=1}^n \frac{\eta_j}{n_j} \left(\frac{\phi_j}{n_j} \right)$. If the

$T_{Ach\ F-D}$ is plotted versus P for lens types, the points lie along an approximate straight line. This is shown in Figure 6.21. To obtain the data for this curve, a zero spaced doublet, an air spaced doublet, a positive-negative-positive-triplet, and two widely spaced achromatic doublets (a Petzval lens) were set up for computation. Each system has an exact focal length of 10 and is corrected for zero $T_{Ach\ F-C}$. The axial paraxial ray was traced through at $y = 1.0$. All the positive lenses were of 511635 glass and all the negative lenses were of 649338 glass. This approximately linear relationship causes real difficulty in the design of flat-field lenses, since reduced Petzval curvature tends to accompany an increase in the amount of secondary color. This is a particularly serious problem in the design of periscope systems.

6.11 ENTRANCE AND EXIT PUPILS, THE CHIEF RAY AND VIGNETTING

6.11.1 General. As shown in Section 6.4, the complete analysis of the first order properties of an optical system can be found by tracing two rays through the optical system. Any two rays may be used, but it is convenient to pick the two rays with some care. In order to specify quantitatively which two rays are usually used, we must discuss the meanings of the pupils of an optical system.

6.11.2 The aperture stop. The bundle of rays, which proceed from an object point to the image point through an optical system, is limited in the sense that all the rays in the entire solid angle of 4π steradians do not get through the system. The aperture stop is the physical stop or diaphragm, as distinguished from an image of a stop, which limits the rays passing through the system. The aperture stop may be a lens or it may be an opening in an otherwise opaque surface. It is almost always circular; we will consider it as such since we are concerned with systems having rotational symmetry.

6.11.3 Entrance and exit pupils.

6.11.3.1 The pupils are images of the aperture stop. The entrance pupil is the image of the aperture stop in the part of the system preceding the aperture stop. Hence to locate the entrance pupil, given the position of the aperture stop, an axial paraxial ray is traced backwards through the system from the center of the

aperture stop. The point where it last crosses the axis is the entrance pupil point. The entrance pupil plane is a plane perpendicular to the axis at the entrance pupil point. If the diameter of the aperture stop is known, an oblique paraxial ray is traced backwards from the rim or margin of the aperture stop. The intersection of this ray with the entrance pupil plane gives the radius of the entrance pupil.

6.11.3.2 Similarly, the exit pupil is the image of the aperture stop in that part of the system following the aperture stop. By tracing an axial and oblique ray from the aperture stop, the exit pupil plane can be located, and the diameter of the exit pupil can be determined. It sometimes happens that the aperture stop precedes (or follows) the rest of the system. In this case the aperture stop coincides with the entrance pupil (or exit pupil).

6.11.4 The chief ray. The chief ray is an oblique ray from an off-axis object point, which intersects the axis at the entrance pupil point, the center of the aperture stop, and the exit pupil point. Because it passes through the centers of the pupils and the aperture stop, it is approximately the central ray of the conical bundle from the object point to the image point. Hence it is representative of the entire bundle.

6.11.5 Two convenient paraxial rays. The usual procedure is to trace one ray from the point on the object plane intersected by the optical axis ($y_o = 0$). The angle with the optical axis, u_o , should be chosen to equal one half the actual cone angle to be passed by the optical system. Hence this ray passes through the margin of the pupils and the aperture stop. u_o is the radius of the entrance pupil divided by the distance between object surface and entrance pupil plane. The second ray should be traced from a point \bar{y}_o in the object plane corresponding to an object near the maximum size to be accommodated by the lens system. This second ray is a chief ray from the object point chosen. Hence \bar{u}_o is \bar{y}_o divided by the distance between object surface and entrance pupil plane. (See Figure 6.22).

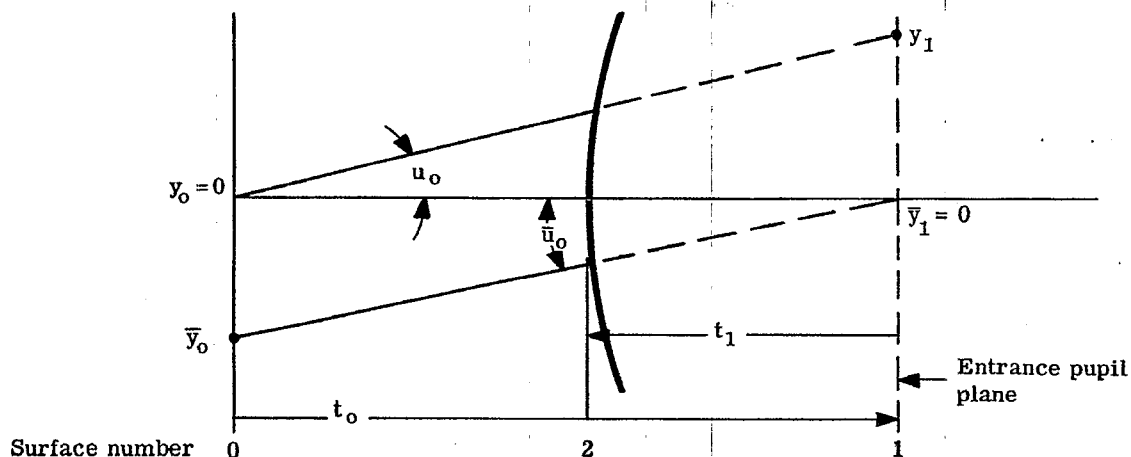


Figure 6.22 -Location of entrance pupil and numbering of surfaces.

6.11.6 Pupils as surfaces in the optical system. Many designers include the entrance pupil plane as a plane surface in the system. It is labelled number one. The actual first surface of the lens may be encountered before the chief ray reaches the entrance pupil. In this case the thickness t_1 is made negative, indicating the entrance pupil plane is actually virtual. As the chief ray passes through the lens it may cross the axis at several positions. Each position is called an aperture plane. After it finally emerges in the image space it can be extended until it crosses the axis. This position is the exit pupil plane of the system and is numbered the $(k - 1)$ surface. Although it is not necessary to include the entrance and exit pupil planes in the calculations of a lens, their inclusion is helpful because they are excellent planes of reference. It is convenient to describe aberration data by using the image coordinates plotted against their conjugate coordinates in the entrance pupil. (See Section 8).

6.11.7 Numerical example. As an example of the foregoing material, Figure 6.23 shows the pupils for a two-lens system. Table 6.14 shows the calculations for this system. In the example, the entrance pupil plane is found in the following way. As the chief ray is drawn, the lens (a) bends it up and the lens (b) bends it back down. It is nearly always true that $(\Delta u_a + \Delta u_b)$ should be close to zero. This tends to keep the distortion corrected. (Distortion is a monochromatic aberration which will be discussed in Section 8). Since $\Delta u = u - u_{-1}$, Equation (24) shows that to meet this condition, $y_a \phi_a + y_b \phi_b = 0$.

The chief ray, therefore, must cross the axis between the two lenses and divide the space in the ratio of ϕ_b / ϕ_a . A value of -1 for \bar{y}_a and $+1$ for \bar{y}_b may be selected for convenience in this problem, because ϕ_a and ϕ_b are equal. Since $t_2 = 5.0$, $\bar{u}_2 = 0.4$, and the chief ray can then be traced backwards to the object plane as shown in the example. The entrance and exit pupil planes are located by solving for t_1 and t_3 to make $\bar{y}_1 = \bar{y}_4 = 0$. Since $\bar{y}_a = -1$ was used for convenience, the object height may come out to be far different from the value to be used for the true object. If the designer wishes to have a ray traced from the true object height, it may be done by simply scaling all the ray data for the chief ray. In the sample \bar{y}_o came out -4 . A second ray was traced at $\bar{y}_o = -2$.

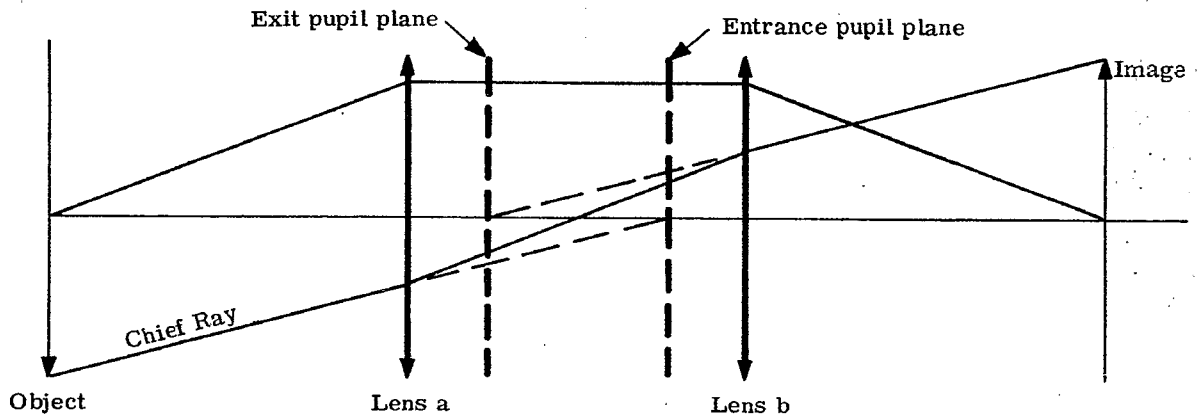


Figure 6.23 - Illustration of entrance and exit pupils.

Surface	Object Plane 0	Entrance Pupil Plane 1	Lens (a) 2	Lens (b) 3	Exit Pupil Plane 4	Image Plane 5
$-\phi$	0	0	-0.1	-0.1	0	0
t	13.33	-3.33	5	-3.33	13.33	
y	0	1.33	1	1	1.33	0
u	0.1	0.1	0	-0.1	-0.1	
\bar{y}	-4	0	-1	1	0	4
\bar{u}	0.3	0.3	0.4	0.3	0.3	
\bar{y}	-2	0	-0.5	0.5	0	2
\bar{u}	0.15	0.15	0.2	0.15	0.15	

Table 6.14 - Calculations showing location of entrance and exit pupil planes.

6.11.8 Vignetting.

6.11.8.1 In the above discussion on the aperture stop and pupils it was assumed that the aperture stop was circular. Hence the pupils are circular and a circular cone of rays passes through the system from an axial object point. For an off-axis object point, the cone of rays limited by the aperture stop will not be circular; and the entrance pupil will generally subtend at the object point a smaller solid angle than for an axial object point. This phenomenon is called vignetting; the oblique bundle of rays is said to be vignetted.

6.11.8.2 In the example shown in Table 6.14 the path of the chief ray has been calculated through a simple two-element lens. The next question is, what is the shape of the beam of light that passes through the optical system from the oblique object point? To answer this, it is necessary to project all the lens apertures in the system onto the entrance pupil plane. Since the path of any ray can be readily computed as a linear combination of two rays (see Section 6.4), it is possible to compute the coordinates in the entrance pupil plane for any ray from the object point of interest which passes through any part of any aperture of the system. For example, suppose we wish to find the coordinate on the entrance pupil plane of a ray from the object $y_o = -2$, which passes through the center of the (a) lens. Since two rays have been traced through the lens, a value of y and \bar{y} is known on each surface. Any other ray \bar{y} may be traced from the object point y_o with the use of Equation (10a)

$$A \bar{y}_j + B y_j = \bar{y}_j.$$

On the object plane $y_o = 0$, $\bar{y}_o = \bar{y}_o$. Therefore

$$A = \bar{y}_o / y_o = 1,$$

and for the i th surface,

$$B = \frac{\bar{y}_i - \bar{y}_i}{y_i}.$$

Finally then,

$$\left[\bar{y}_j - \bar{y}_j \right] = \left[\bar{y}_i - \bar{y}_i \right] \frac{y_j}{y_i}. \quad (50)$$

To calculate the coordinate of any ray on the entrance pupil plane, which has the coordinate \bar{y}_i on the i th surface, Equation (50) becomes

$$\bar{y}_1 = (\bar{y}_i - \bar{y}_i) \frac{y_1}{y_i},$$

since $\bar{y}_1 = 0$.

6.11.8.3 In the example shown in Figure 6.23 and Table 6.14, a ray from the object point $y_o = -2$ passing through the center of the (a) lens ($\bar{y}_2 = 0$) will project onto the entrance pupil plane at the value $\bar{y}_1 = (0 + 0.5)(1.33)/1 = 0.666$. The top edge of the (a) lens (assumed $\bar{y}_2 = 1$) will appear in the entrance pupil plane at $\bar{y}_1 = (1 + 0.5)(1.33) = 2$. The center of the (b) lens will project in the entrance pupil plane at $\bar{y}_1 = (0 - 0.5)(1.33) = -0.666$. The top edge of the (b) lens (assume $\bar{y}_3 = 1$) will appear in the entrance pupil plane at $\bar{y}_1 = (1 - 0.5)(1.33) = 0.666$.

6.11.8.4 Since the center and top edge of each lens, (a) and (b), are now projected on the entrance pupil plane, it is possible to construct circles indicating the complete aperture of the lenses as they appear in the entrance pupil plane. These apertures are shown in Figure 6.24. Only those rays passing through the area common to both circles will pass through the two lenses. In order to have the same aperture for the oblique beam as for the central beam, an aperture would have to be placed to appear as the inner circle shown in Figure 6.24. A circular aperture in the entrance pupil plane of radius 0.666 just fits in the common area of the two circles. Now in this case, the entrance pupil plane is virtual, so no physical stop can be placed in it. Since the chief ray does actually cross the axis at a point midway between the lenses, the physical aperture stop may be placed in this position and it will appear as a central stop in the entrance pupil plane. Using Equation (50), the size of the aperture stop can be calculated using the following data.

$$\bar{y}_1 = 0.666 = \text{height of edge of entrance pupil aperture.}$$

$$\bar{y}_i = \text{height of edge of aperture stop in the aperture plane.}$$

$$\bar{y}_i = 0 = \text{height of chief ray in the aperture stop plane.}$$

$$y_i = 1.0 = \text{height of axial ray in the aperture stop plane.}$$

$$y_1 = 1.33 = \text{height of axial ray in the entrance pupil plane.}$$

Therefore

$$\bar{y}_i = \frac{0.666}{1.33} = 0.5.$$

6.11.8.5 Usually some vignetting for the oblique beams is allowed, so the aperture stop is made larger than the largest circle included in the common area. Figure 6.25 shows the appearance of the aperture stop when it is made 0.75 in radius. The clear area is the common area for all the apertures, and its area is a measure of the total light passing through the system from the oblique object point. The common area is 67% of the area of the image of the (0.75) aperture stop in the entrance pupil plane. Therefore, the oblique beam is vignetted by 33%. All other factors remaining constant, the illumination at the image point, $\bar{y}_k = 2$, is 67% of the illumination at the point $y_k = 0$. In Figure 6.25, the aperture stop of radius 0.75 located midway between the (a) and (b) lens, is imaged in the entrance pupil plane with a radius of 1.0. The exit pupil also has a radius of 1.0.

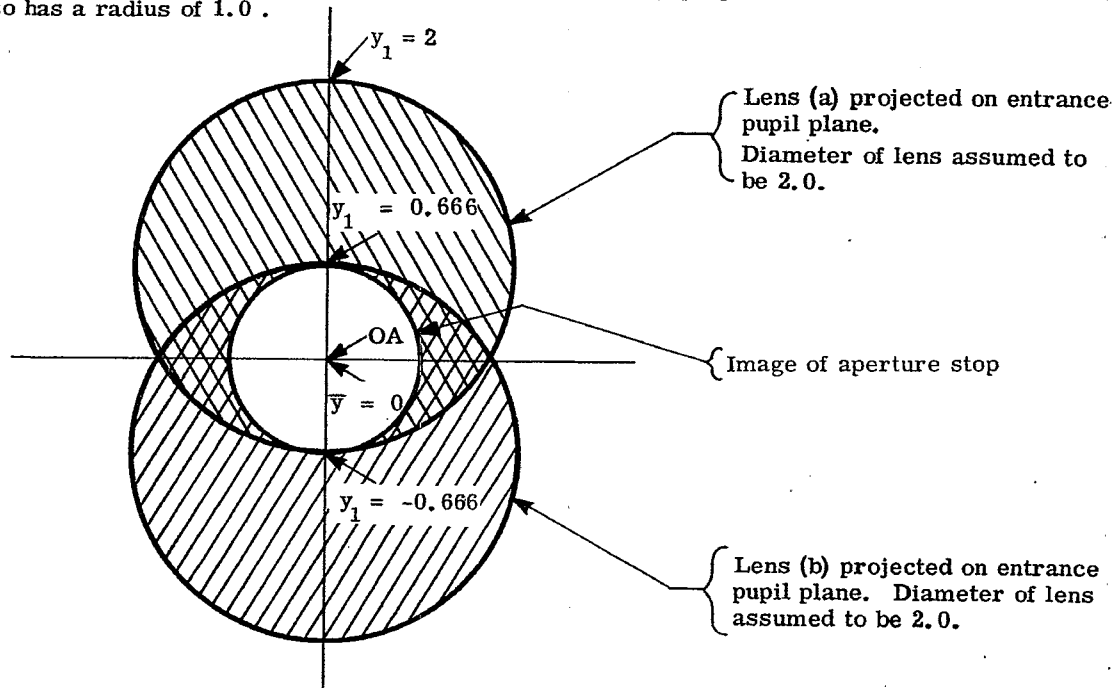


Figure 6.24 - Apertures of the (a) and (b) lenses and of the aperture stop projected onto the entrance pupil. The oblique beam is not vignetted.

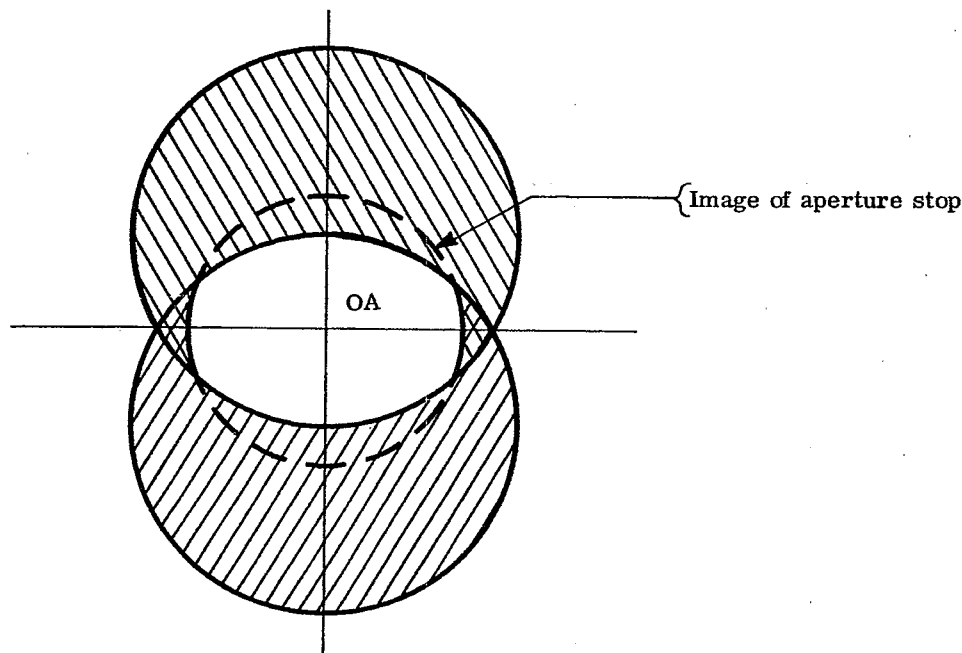


Figure 6.25 - Illustrating vignetting for the same system shown in Figure 6.24 but with a larger aperture stop.

