

Elements of Modern Optical Design

DONALD C. O'SHEA

School of Physics
Georgia Institute of Technology
Atlanta, Georgia

A Wiley-Interscience Publication

John Wiley & Sons

New York / Chichester / Brisbane / Toronto / Singapore

A NOTE TO THE READER

This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

Copyright © 1985 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Section 107 or 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

O'Shea, Donald C.

Elements of modern optical design.

"A Wiley-Interscience publication."

Includes index.

1. Optical instruments—Design and construction.

I. Title.

QC372.2.D4084 1985 535 84-19708

ISBN 0-471-07796-8

Preface

The skill of optical design, like many other skills, is acquired with a great deal of practice. The purpose of this textbook is to illustrate the techniques needed in optical design and to provide practice in their use. As with any skill or craft, one cannot teach or explain how to design a specific new product. Instead, the basic elements of the skill are presented and some examples of good, or clever, work are shown with some analysis. It is up to the reader to work through the examples given in the text and other problems given by the instructor or presented on the job. Through practice, the designer will develop his or her unique approach to the process.

The great pleasure of design work is to propose a design, determine its properties, check these against the requirements for the real system, and modify the design to enhance its usefulness. It is the imaginative working out of the design within the constraints of the real world that results in the initial frustrations and final delight of a good design.

This textbook is intended to provide the reader with the elements of modern optical design. The basic process is that of tracing of light rays through the optical system and the evaluation of the results. Until two decades ago this was sufficient for most optical systems. Now that lasers, modulators, and scanners are common components in many systems, they must be considered as part of the design process. It doesn't get any easier, but today it's a lot more fun.

The organization of this text differs somewhat from that of most optics texts. In most cases one starts with an exact ray trace and, after having derived the correct equations, performs a calculation. Then approximations are made to simplify the equations, first to the paraxial case and then to the thin lens equation. While this approach leads to simplicity in the derivation of all subsequent equations once the original ones are obtained, it does not provide an approach that approximates the usual design process—simple first, complex later.

As will be noted in an examination of the table of contents, the text begins with a review of optics usually taught in a sophomore physics course. This material is, for most people, the first, last, and only optics they will ever have—unless they have to design an optical system. Then they will need to know considerably more optics. However, this does not mean that the thin lens

equation and other simple concepts have no place in optical design. On the contrary, it is these elementary ideas, already understood by most people, that provide the basis for quick back-of-the-envelope calculations. By building on this basic information we hope to give the reader some feel for rapid assessment of preliminary designs and a measure of confidence in his or her present abilities.

After developing a procedure for ray tracing through thin lenses, the ideas and terminology of stops are discussed. At this point the subject of radiometry is introduced with the belief that the sooner energy considerations are brought into the design of an optical system the better the overall design will be. Following this, the paraxial approximation is introduced to treat thick lenses and multiple-surface lens systems and color correction. Finally, the exact ray trace equations are derived and the effects of aberrations and their correction are explored.

Once these basic principles are in hand, we can progress to some optical design problems that are now important in today's technology. Gaussian beam propagation is such a problem, since many systems involve lasers as their primary light source. Another problem usually connected with lasers is deflecting and scanning systems. The final subjects in the text are the active components that are introduced into the optical system, such as modulators and detectors, and the design of optical systems.

There are a large number of subjects that are not covered in this text. They include optical fabrication, testing procedures, lasers, optical fibers, and off-axis systems, among other subjects. The field of optics is too large to encompass all possible design considerations in a single text. Additional references are provided at the end of most chapters. Even in those cases where we have provided information in a particular field, we have restricted it to the coverage needed to progress in mastering the design techniques. Since this information has been distributed throughout the text, the index should provide a means for a reader to review a particular area. Additional material on specific components and products can be obtained from manufacturers and suppliers. When this text is used in a course, it is a good idea for the instructor to distribute such material so that the student gets some feel for the type and thoroughness of the information (or lack of it) from suppliers.

I hope this text can serve both as a basis for a formal course in optical design and as a tutorial for individual readers. To the teacher, I offer one piece of advice: Give lots of homework and provide as much computer assistance as possible to prevent the work from becoming tedious. I ask for some assistance: Tell me of any errors in the text and let me know how the examples work. To the reader I extend my sincere wish that he or she will find the time spent with this text both profitable and enjoyable!

DONALD C. O'SHEA

Contents

Chapter 1. A Review of Elementary Optics	1
1.1 Light	1
1.2 Laws of Geometrical Optics	7
1.3 Mirrors and Lenses—Ray Sketching	10
1.4 Mirrors and Lenses—Calculations	19
1.5 Simple Optical Instruments	26
1.6 Polarization	32
1.7 Interference and Diffraction	41
Appendix	54
Chapter 2. Thin Lens Ray Tracing	56
2.1 Getting Started	56
2.2 Tracing Rays through a Series of Lenses	61
2.3 The y - u Trace	66
2.4 Stops, Pupils, and Windows	67
2.5 Evaluation of Systems Using the y - u Trace	73
2.6 Effective Focal Length	79
2.7 Cylindrical Lenses	81
Chapter 3. Radiometry	87
3.1 Basic Concepts and Terms	87
3.2 Radiometry in an Optical System	95
3.3 Radiometry of Extended Sources	100
3.4 Applications	107
3.5 Photometry and Spectrophotometry	114

Chapter 4. Mirrors and Prisms	119
4.1 Mirror Images	119
4.2 Plane Mirrors	123
4.3 Prisms	130
4.4 Fitting It In or Spreading It Out	134
Chapter 5. Paraxial Ray Tracing	144
5.1 The y - nu Trace	144
5.2 The Gaussian Constants	156
5.3 Chromatic Aberration	162
5.4 Paraxial Ray Tracing, Calculators, and Computers	169
Chapter 6. Exact Ray Traces	177
6.1 Ray Tracing Techniques	177
6.2 Aberrations	185
6.3 Correcting Aberrations	204
6.4 Optical Computation	219
6.5 Image Evaluation	223
Chapter 7. Gaussian Beams	230
7.1 Characteristics of a Gaussian Beam	230
7.2 Modifying a Gaussian Beam	234
7.3 Conjugate Distances	252
7.4 Laser Beam Illumination	256
7.5 Higher Order Modes	264
Chapter 8. Modulation and Scanning	270
8.1 Modulation of Light	270
8.2 Scanning	282
8.3 Mechanical Scanners	288
8.4 Acousto-optical Scanners	298
8.5 Holographic Scanners	302
8.6 Additional Topics	305

Chapter 9. Spectrometers	313
9.1 Filter Spectrometers	313
9.2 Prism Spectrometers	319
9.3 Grating Spectrometers	324
9.4 Spectrometer Systems	330
Chapter 10. Detectors	336
10.1 The Eye—The Ultimate Detector	337
10.2 Thermal Detectors	339
10.3 Photographic Materials	342
10.4 Photoemissive Detectors	347
10.5 Semiconductor Detectors	353
10.6 Photodiode Arrays	360
10.7 Detector Geometries	364
Chapter 11. The Design Process	368
11.1 Setting Down Objectives	369
11.2 Preliminary Design	372
11.3 The Breadboarding Process	374
11.4 Case Study—Laser Printer Design	375
11.5 Case Study—Slide Projector Design	384
11.6 Conclusions	392
Index	395

1

A Review of Elementary Optics

(WHAT YOU ALREADY KNEW,
BUT MAY HAVE FORGOTTEN)

Most people who have had a physics course in high school or college learn a little about optics that proves useful in their daily life. They understand how a lens works and how a mirror can focus the sun's rays. Some may remember the thin lens equation, Snell's Law, and simple ray tracing principles. They realize that the field of optics has expanded greatly since their last exposure. If they are uneasy at the prospect of trying to design an optical system, it is understandable, but unnecessary. All of the material on optics presented in basic physics courses is valid for optical design, and the simple approaches taught in these courses are particularly useful in the preliminary stages of a design. In this book, we begin with the ideas and facts of elementary optics and build on them. This chapter serves as a brief tutorial for those who have never had optics and as a review for those who have not thought about optics for some time. If you have a good idea of basic optics, skim through this chapter and do a few of the problems at the end of the chapter to check your understanding of the material. Notice that the material goes beyond geometrical optics and covers, briefly in Section 1.7, some topics in physical optics.

One of the methods introduced in this chapter is called *ray sketching*. In Chapter 2 we discuss the method of tracing light rays through a system of thin lenses, and in later chapters we cover more accurate methods of ray tracing. *Ray sketching*, as opposed to ray tracing, is a method for rapidly drawing rays to determine some of the distances and dimensions of a newly proposed system. It is a means of putting onto paper for the first time the ideas for an optical system for criticism, revision, and refinement. This means of *thinking graphically* rarely solves anything, but it does serve to clarify and restate many questions that must be answered by a successful design.

1.1. LIGHT

When we think of optics, we are usually thinking in terms of visible light. But light is just electromagnetic radiation that we can see. The laws of optics and the methods of optical design may be used with radiation at any wavelength,

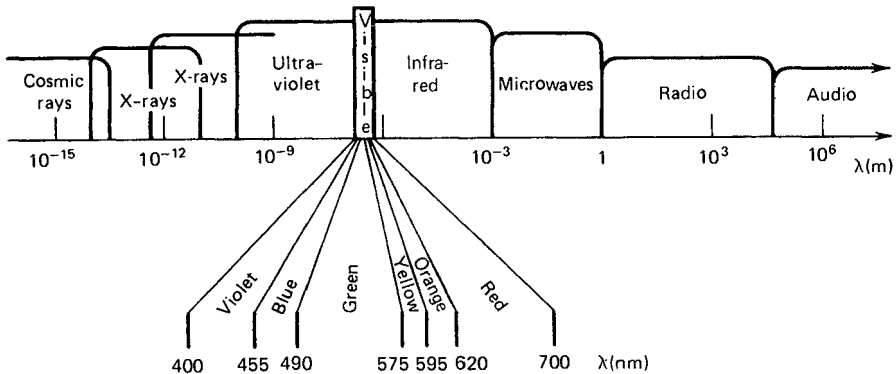


Figure 1.1. The electromagnetic spectrum.

even those radiations invisible to us. Figure 1.1 displays the extent of the electromagnetic spectrum and its regions with some arbitrary boundaries. In some of these regions the methods for focusing and reflecting may be somewhat exotic, but the laws still apply. For example, in the x-ray region there are no materials with substantial refracting power to be made into lenses, so all focusing and imaging is done with mirrors. Because of our basic prejudice in favor of visible light, most of the illustrations and calculations are chosen from this region of the spectrum, but the methods we discuss can be used across much of the spectrum.

The visible light spectrum extends in wavelength from about 400 nm (blue) to 700 nm (red). It is in the center of the spectrum (555 nm) that the sensitivity of the human eye is greatest. A continuous spectrum that spans the visible region and extends far into the infrared can be produced by heating a tungsten filament to 2900 K. A plot of intensity versus wavelength for a typical source is shown in Fig. 1.2.

Not all light sources emit in a broad continuum like the tungsten filament. An electrical discharge in a gas, like the glowing tube that makes up a neon sign, emits a series of discrete wavelengths. When a source such as a hydrogen discharge illuminates a slit and the light passing through the slit is broken up into its separate colors by passing it through a prism, the spectrum consists of a number of images of the slit, a series of lines of different colors. Because of this multiline appearance, the term “line” is used interchangeably with “wavelength” to designate a particular color of light.

In a sodium discharge, similar to those seen in the bright yellow lamps that illuminate many of our streets and highways, the emission is primarily in two lines at 589 and 590 nm. The hydrogen spectrum is made up of a series of strong lines that show up not only in a discharge tube, but also in the spectrum of the sun.* Some of these lines of hydrogen and sodium are

* They do not appear as bright lines in the spectrum of the sun, however, but as dark lines on the broad solar continuum. This is because these elements in the outer atmosphere of the sun are cooler and absorb light at these wavelengths.

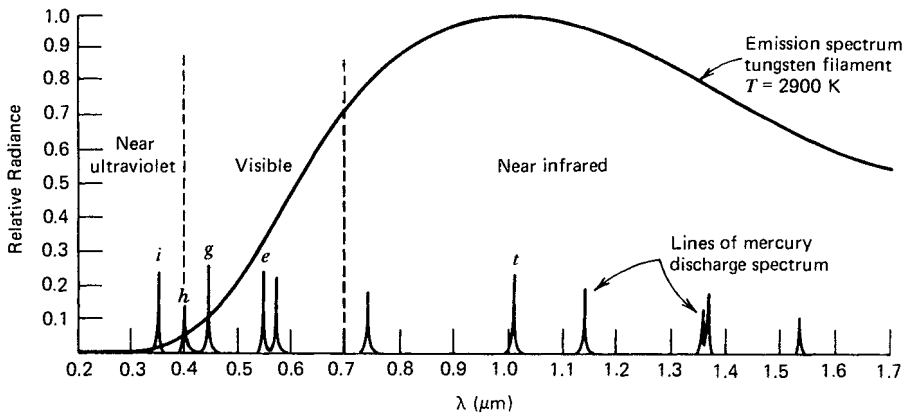


Figure 1.2. Spectra of blackbody radiation of 2870 K and mercury discharge emission.

Table 1.1. Spectral Lines and Refractive Indices of Two Glasses

Fraunhofer Lines			Crown (K-50) # 523601	Barium Flint (BaFN-10) # 670472
Designation	Element	$\lambda(\text{nm})$	n	n
<i>t</i>	Hg	1014.0	1.51266	1.65500
<i>b (or r)</i>	He	706.5	1.51839	1.66343
<i>C</i>	H	656.3	1.51992	1.66580
<i>C'</i>	Cd	643.8	1.52034	1.66647
<i>D</i>	Na	589.3	1.52249	1.66991
<i>d</i>	He	587.6	1.52257	1.67003
<i>e</i>	Hg	546.1	1.52464	1.67341
<i>F</i>	H	486.1	1.52861	1.68000
<i>F'</i>	Cd	480.0	1.52911	1.68083
<i>g</i>	Hg	435.8	1.53333	1.68801
<i>h</i>	Hg	404.7	1.53725	1.69483
<i>i</i>	Hg	365.0	1.54395	1.70686

prominent lines found in the solar spectrum. They were labeled by the solar scientist, Joseph von Fraunhofer, with letters of the alphabet. For example, the sodium lines listed above are often referred to as the *D* lines. In Table 1.1 we list a number of the important *Fraunhofer lines* along with their wavelengths, and we mark their position in the spectrum in Fig. 1.2. With the invention of the laser, sources are now available that can produce intense outputs at a single wavelength. In addition, the directional properties of lasers are quite different from other sources of light. We will use all of the sources just described to illuminate our designs and our problems.

The wavelengths for the Fraunhofer lines listed in Table 1.1 are those for light traveling in a vacuum at a speed of 3×10^8 m/sec. In any material medium, the speed and wavelength of light are less than the values in a vacuum. The quantity that remains constant in any medium is the frequency of the light. If we designate λ_0 as the vacuum wavelength and c as the speed of light in a vacuum, the relation between frequency and wavelength is

$$v\lambda_0 = c. \quad (1.1)$$

In optics, the properties of this medium at a particular wavelength can be described by a single constant,* the ratio of the vacuum light speed divided by the speed in the medium,

$$n = \frac{c}{v}. \quad (1.2)$$

Thus in a material medium, Eq. (1.1) becomes

$$v\lambda = v, \quad (1.3)$$

where λ , the wavelength in the medium, is equal to λ_0/n . [See Problem 1.1] This number n , the *index of refraction*, is all that is needed to calculate the behavior of light in the medium.

Because the velocity of light varies with the medium, it can be difficult to keep track of the progress of a wave as it travels through several media. Suppose x_0 is the distance that light travels in a vacuum in time t ($x_0 = ct$). In the same time interval t the wave travels a short distance x , in a medium with refractive index n_1 . That is,

$$x_1 = vt = \frac{c}{n_1}t$$

*This is true if there is no absorption of light at that wavelength. Otherwise the absorption coefficient of the medium must be included as a second constant.

For other media with refractive indices n_2, n_3, \dots , the distance traveled in a time t ,

$$x_2 = \frac{c}{n_2}t; \quad x_3 = \frac{c}{n_3}t; \dots$$

If all of these equations are multiplied by their respective refractive indices, then all of these quantities are equal to the vacuum distance,

$$x_0 = n_1 x_1 = n_2 x_2 = n_3 x_3 = \dots = ct. \quad (1.4)$$

Thus there are distances equivalent to the vacuum distance, called the *optical path length*. With this quantity, the distance of light traveled in any medium may be compared to that in any other medium.

The refractive index is not a constant but varies as a function of wavelength. The presence of this variation is called the *dispersion* of the material. This property is both useful and vexing. On one hand, it is dispersion that enables a prism to spread the light from a source into its spectrum, but, on the other hand, it causes color distortion of an image (chromatic aberration) formed by a single lens. In Table 1.1 the values of the refractive index at a number of wavelengths are listed for two common optical glasses. Not only are the values different, but the rate of change of the refractive index with wavelength is different for these two glasses. A number that indicates the amount of the dispersion for a glass is the *V-number*, or *Abbe number*, defined by

$$V = \frac{n_d - 1}{n_F - n_C}, \quad (1.5)$$

where n_d , n_F and n_C are the refractive indices at the Fraunhofer wavelengths indicated by the subscripts. (Since the *D* line of sodium actually is made up of two lines, the *d* line of helium is used to define V in many modern glass catalogs.) Because the dispersion is greatest when $n_F - n_C$ is largest, glasses with the strongest dispersion have the smallest V -numbers. The V -numbers for both of the glasses in Table 1.1 are given within the glass number (see below). Using Equation (1.5), check these values for yourself [Problem 1.2].

While many glasses have special names, all glasses can be labeled by their n_d and V . (See box, The Glass Map.) The *glass number* is a six-digit number whose first three digits are the three most significant digits of $n_d - 1$ and whose second three digits are 10 times the V -number. For example, the glass number for a crown glass ($n_d = 1.53200$ and $V = 58.7$) is 532587. Because each glass manufacturer has slightly different recipes for glasses with the same designation, glass numbers can vary from source to source.

Although the number of glasses is large, only a fraction of them are used regularly. Based on the cost, ease of optical fabrication of surfaces, thermal stability, and uniqueness (the fact there are no other glasses with a similar n_d and V in that region of the map), about 50 glasses can be chosen as "select" glass types [Ref. 5.2]. The various areas of the map are marked with the type of glass found in that area. Notice that all of the types end in either a "K" (designating a crown glass) or an "F" (designating a flint glass). The border between the two areas at $V = 50$ for $n_d > 1.60$ and at $V = 55$ for $n_d < 1.60$ has been drawn in. Another boundary imposed by the materials available is the lower boundary of values that extends from the low index and high V -numbers on the lower left to the high index and low V -numbers on the upper right. When all available glasses, not just the select ones plotted here, are marked on the map, about one third of the glasses cluster about this curve, known as the "glass line." A list of the glass numbers for these select types and the most common designation is given in Appendix A.

1.2. LAWS OF GEOMETRICAL OPTICS

There is no need to convince anyone that light travels in straight lines. When we see rays of sunlight pouring between the leaves of a tree in a light morning fog, we trust our sight. The idea of light rays traveling in straight lines through space is accurate as long as the wavelength of the radiation is much smaller than the windows, passages, and holes that can restrict the path of the light. When this is not true, the phenomenon of *diffraction* must be considered, and its effect upon the direction and pattern of the radiation must be calculated. However, to a first approximation, when diffraction can be ignored, we can consider that the progress of light through an optical system may be traced by following the straight line paths or *rays* of light through the system. This is the domain of geometrical optics.

Part of the beauty of optics, as it is for any good game, is that the rules are so simple, yet the consequences so varied and, at times, elaborate, that one never tires of playing. Geometrical optics can be expressed as a set of three laws:

1. *The Law of Transmission.* In a region of constant refractive index, light travels in a straight line.
2. *Law of Reflection.* Light incident on a plane surface at an angle θ_i with respect to the normal to the surface is reflected through an angle θ_r , equal to the incident angle (Fig. 1.3),

$$\theta_i = \theta_r. \quad (1.6)$$

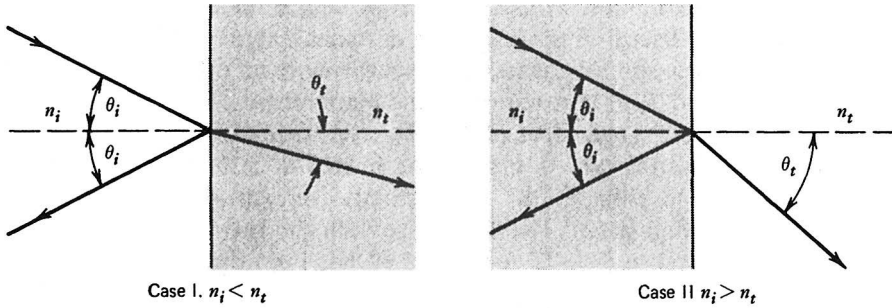


Figure 1.3. Reflection and refraction of light at an interface.

3. *Law of Refraction (Snell's Law).* Light in a medium at refractive index n_i incident on a plane surface at an angle θ_i with respect to the normal is refracted to an angle θ_t in a medium of refractive index n_t as (Fig. 1.3),

$$n_i \sin \theta_i = n_t \sin \theta_t. \quad (1.7)$$

A corollary to these three rules is that the incident, reflected, and transmitted rays, and the normal to the interface all lie in the same plane, called the *plane of incidence*, which is defined as the plane containing the surface normal and the direction of the incident ray.

Note that the second of these equations is not written as a ratio of sines, as you may have recalled it from your earlier studies, but rather as a product of $n \sin \theta$. This is because the equation is unambiguous as to which refractive index corresponds to which angle [Problems 1.3 and 1.4]. If you remember it in this form, you will never have any difficulty trying to determine which index goes where in solving for angles. In Example 1.A, we examine the dispersion of flint glass using Snell's Law.

A special case must be considered if the refractive index of the incident medium is greater than that of the transmitting medium, ($n_i > n_t$). Solving for θ_t , we get

$$\sin \theta_t = \frac{n_i}{n_t} \sin \theta_i.$$

In this case, $n_i/n_t > 1$, and $\sin \theta_i$ can range from 0 to 1. Thus, for large angles of θ_i , it would seem that we could have $\sin \theta_t > 1$. But $\sin \theta_t$ must also be less than one, so there is a *critical angle* $\theta_i = \theta_c$, where

$$\sin \theta_c = \frac{n_t}{n_i} \quad (1.8)$$

EXAMPLE 1.A. Consider a ray of white light in air ($n = 1$) incident on a piece of flint glass (670472) at an angle $\theta_i = 60^\circ$.

What are the directions of the rays for the *C*, *D*, and *g* lines?

Solution: We may solve the transmitted angle θ_t by:

$$n_{\text{air}} \sin \theta_i = n_{\text{flint}} \sin \theta_t,$$

$$\sin \theta_t = \frac{1 \cdot \sin 60^\circ}{n_{\text{flint}}} = \frac{\sqrt{3}}{2n_{\text{flint}}}.$$

From Table 1 we use the refractive indices given for glass number 670472 for the *D* line, $n_{\text{flint}} = 1.66991$; $\theta_t = \sin^{-1}(0.5186) = 31.24^\circ$, for the *C* line; $n_{\text{flint}} = 1.66580$; $\theta_t = \sin^{-1}(0.5199) = 31.32^\circ$, for the *g* line; $n_{\text{flint}} = 1.68801$; $\theta_t = \sin^{-1}(0.5130) = 30.87^\circ$. Thus the wavelength dependence of the refractive index resulted in a spreading (dispersing) of the beam into its various colors. The angle between the two extreme colors calculated here is 0.45° , or 27 minutes of arc. If we follow the rays for the different colors to a second interface of a prism [Problem 1.5], we see that the colors will be dispersed even more.

and $\sin \theta_t = 1$. This means the transmitted ray is traveling perpendicular to the normal (i.e., parallel to the interface), as shown in Fig. 1.4. For incident angles θ_i greater than $\theta_c = \sin^{-1}(n_t/n_i)$, no light is transmitted. Instead the light is totally reflected back into the incident medium. This effect is called *total internal reflection*. As illustrated in Fig. 1.5, prisms can provide highly reflecting nonabsorbing mirrors by exploiting total internal reflection [Problem 1.6].

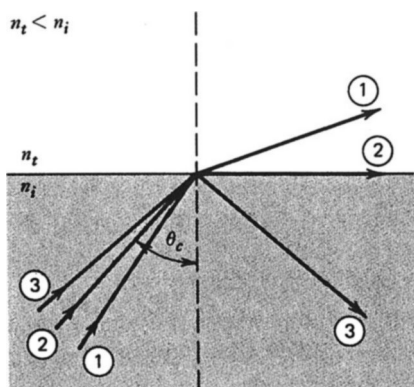


Figure 1.4. Three rays incident at angles near or at the critical angle. Ray ① is transmitted into medium 2; ray ② is incident at the critical angle, θ_c , the light traveling along the surface of the interface; ray ③ is incident at an angle greater than θ_c and is totally internally reflected.

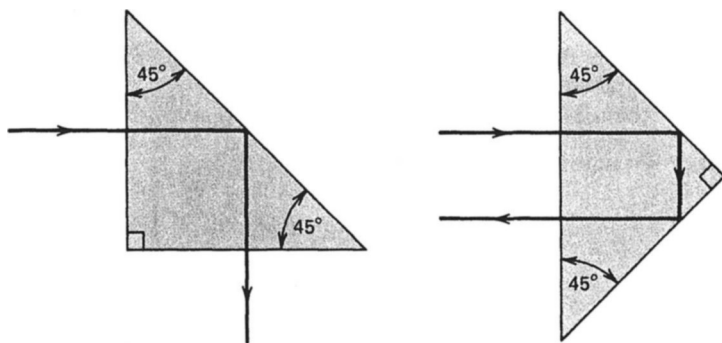


Figure 1.5. Total internal reflection from prisms.

Total internal reflection is the basis for the transmission of light through optical fibers. We do not cover the design of optical fiber systems in this text because the application has become highly specialized and more closely linked with modern communications theory than geometrical optics. This is not to say that the launching, transmission, and collection of light with optical fibers do not produce interesting problems, but only that they are not amenable to the design procedures discussed in this book.

1.3. MIRRORS AND LENSES—RAY SKETCHING

One of the most useful tools in optical design is a rapid study of the components of a system, their placement, and the light paths through them. It is analogous to the quick sketches that a painter makes before deciding on a final composition. In its most primitive form a preliminary design will begin with an objective, often unstated. Then a series of lenses, mirrors, prisms, modulators, detectors, and other optical components are laid out in a sequential fashion with a few important rays drawn through the components along paths that roughly approximate those obtained by a ray trace calculation.

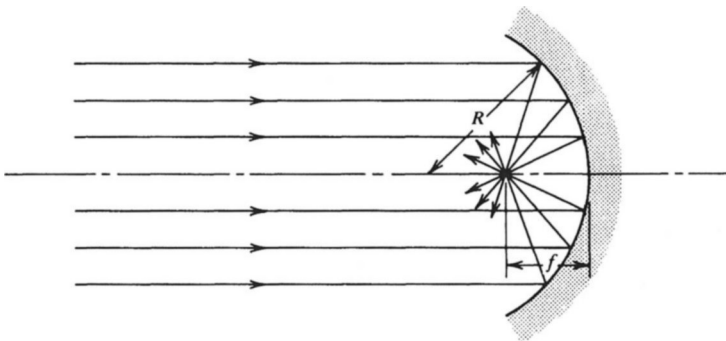
In most optical designs, the imaging components—the lenses and curved mirrors—are symmetric about a line, called the *optic axis*. This line is drawn first, in order to establish a reference line for the system. An optical surface, such as a concave mirror, is then drawn so that its center of curvature falls on the optic axis. (The point where a surface is cut by the optic axis is called the *vertex* of that surface.)

By drawing rays through a system, one can determine the advantages and disadvantages of the system. There are conventions for tracing rays; although not universally accepted, these conventions have sufficient usage that it is convenient to adopt them for sketches and calculations.

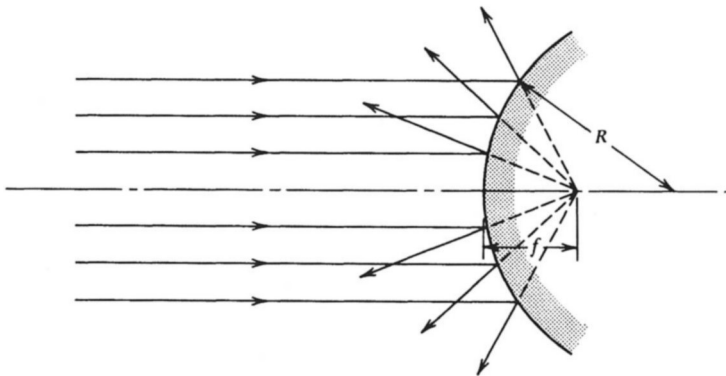
1. An object is placed to the left of the optical system. Light is traced through the system from *left to right* until a reflective component alters the general direction.

Although one could draw some recognizable object to be imaged by the system, the simplest object is a vertical arrow. (The arrow, imaged by the optical system, indicates if subsequent images are erect or inverted with respect to the original object and other images.) If we assume light from the object is sent in all directions, we can draw a sunburst of rays from any point on the arrow. An image is formed where all the rays from the point, that are redirected by the optical system, again converge to a point. In which case, any two rays can locate the image point. If all of the rays from the object point do not converge to a point, then the system has aberrations which may be minimized by additional design, as we shall see later.

A concave mirror is the simplest image-forming device. If the object is placed very far away (“at infinity” is the usual term), the rays from the object



(a) Concave mirror



(b) Convex mirror

Figure 1.6. Focusing of parallel light by spherical mirrors.

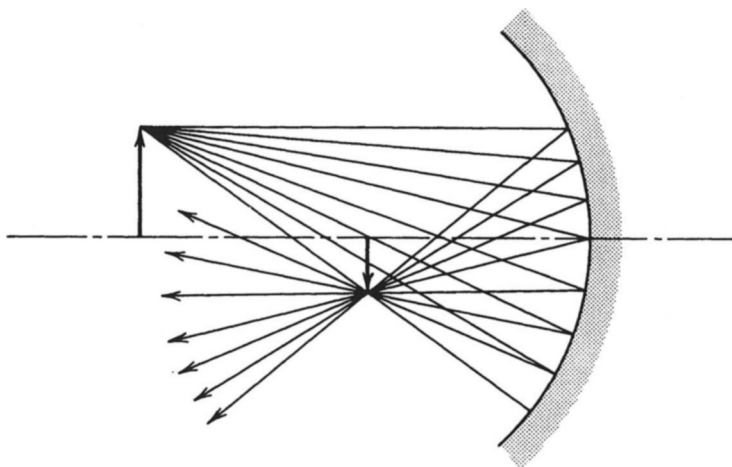


Figure 1.7. Imaging of an object point by a spherical mirror. A real inverted image with respect to the object is formed by the mirror.

are parallel to the optic axis and produce an image at the *focal point* of the mirror, a distance f from the mirror (the distance f is the *focal length* of the mirror), as shown in Fig. 1.6(a). The position of this focal point is one half of the radius of curvature of the mirror (Section 1.4). A convex mirror also has a focal point, as shown in Fig. 1.6(b). However, in this case, the parallel rays do not converge to a point, but instead appear to diverge from a point a distance f from the mirror.

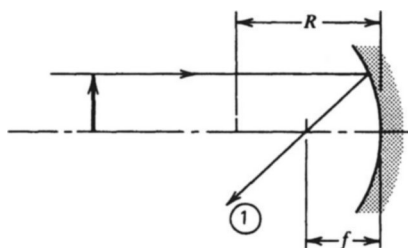
Although it is possible to draw many rays through the system, as in the elaborate drawing shown in Fig. 1.7, a few special rays, easily sketched, can give the same information. For a mirror, the rules are:

2. A light ray parallel to the optic axis of a mirror will pass, after reflection, through the focal point, a distance f from the vertex of the mirror.
3. Light rays through the focal point of a mirror will, after reflection, be reflected parallel to the optic axis.
4. A light ray reflected off the vertex point makes an equal angle with the optic axis, as does the incident ray.
5. A light ray directed through the center of curvature is reflected back on itself.

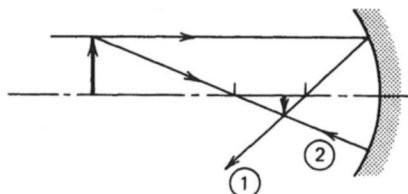
In Example 1.B the formation of an image by a concave mirror is shown. Notice that the rays cross at a point in space. If you were to put a screen at that point you would see the image in focus there. Because the image can be found at an accessible plane in space, it is called a *real image* [Problem 1.7]. For a convex mirror, the rays from an object do not cross after reflection from the surface, as shown in Fig. 1.8, but *appear* to come from some point behind

EXAMPLE 1.B. A concave mirror of focal length f ($= R/2$) images an object located at a distance of $4f$ from the mirror. Sketch the resulting image.

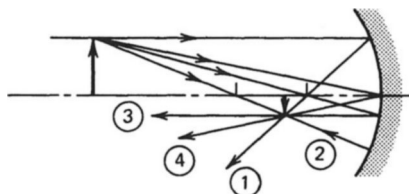
Solution: Draw the mirror and indicate its approximate center of curvature and *then* draw the optic axis. (See box, Hints On Ray Sketching). Draw a mirror normal (i.e., radius) and put tic marks on the optic axis for the focal point and center of curvature.



Draw in the parallel ray ① and reflect it through the focus. Send one ② through the center of curvature.



Although two rays are sufficient, you can also draw a ray ③ through the focus and reflect it parallel to the axis. You can also reflect a ray ④ off the vertex of the mirror. Any two of the four rays will locate the image.



Which rays you use will depend on the circumstances. It should be pointed out that the first two rays illustrated here intersect at an oblique angle and therefore locate the object with less chance of error when sketching than the last two, which meet at an acute angle.

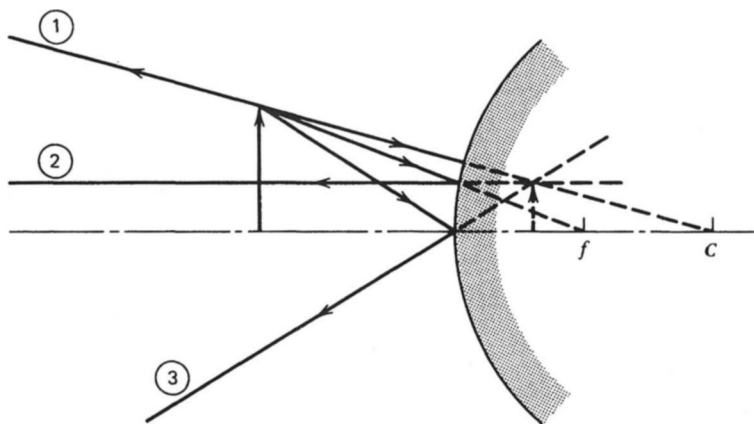


Figure 1.8. Ray trace of an object in front of a convex spherical mirror. Ray ① is directed toward the center of curvature of the mirror and therefore is normal to the surface and returns on itself. Ray ② is directed toward the focal point of the mirror (since the mirror is convex, it is negative and lies behind the mirror), is reflected parallel to the optic axis. Ray ③ is a single reflection off the vertex point.

the mirror. This image, which cannot be observed on a screen at some point in space, is called a *virtual image*. The image you see in the bathroom mirror in the morning is a virtual image [Problem 1.8]. One can also produce a virtual image with a concave mirror if the object is located between the vertex and focus of the mirror [Problem 1.9]. The labels, “real” and “virtual”, do not imply that one type of image is useful and the other is worthless. They simply indicate whether or not the rays redirected by the optical system actually cross.

Just as with mirrors, lenses can form both real and virtual images. In Fig. 1.9, a number of different lenses are labeled. Those lenses whose axial thickness is larger than their edge thickness will converge parallel light and have *positive* focal lengths. Those with a smaller axial thickness than edge thickness will diverge parallel light and have *negative* focal lengths.

Much of the nomenclature and methodology used for mirrors holds true for lenses. The rules for sketching rays are analogous to those for mirrors. We have used numbers with a prime to indicate the counterpart to the rule for a mirror.

- 2'. A light ray parallel to the optic axis of a lens will pass through the focal point, a distance f from the vertex of the lens.
- 3'. A light ray through the focal point of a lens will be refracted parallel to the optic axis.
- 4'. A light ray through the center of a lens is undeviated.

HINTS ON RAY SKETCHING

1. Use quadrille paper (a paper with a square grid printed on it) whenever possible. Not only does it reduce the need for straight edges, it also allows you to do some rough scaling of the system. If there is no quadrille paper available, use whatever is handy at the time.
2. Do not try to fold the design on the first sketch. Wait until you sketch in some marginal and chief rays (see next chapter) before trying a folded sketch.
3. Draw the lens surfaces first, then lay down the optic axis. It is easier to draw a straight line than a circle, so one can always “fit” an optic axis into the sketch after some arcs have been drawn and their center of curvatures estimated. (A convention followed in this text is that the optic axis is denoted by alternating long and short dashed lines, surfaces by solid lines, rays by solid lines with arrows, and ray extensions (used to find virtual images) by short dashed lines. Virtual objects are also drawn using short dashed lines.)
4. When an axis is redirected by a mirror or scanner, draw the new axis first, then bisect the angle between the axes. This provides a mirror normal and makes it easy to orient the mirror correctly.
5. When in doubt, use boxes (black ones, if you prefer). Do not nail down a ray sketch with too much detail before it is appropriate.
6. Except for specific errors in a sketch, do not erase. Generate additional sketches. Paper is cheaper than the time lost retracing a good idea overdrawn with a not-so-good idea.

There is no counterpart to Rule 5 for lenses. The application of these rules is shown in Example 1.C and explored in Problem 1.10.

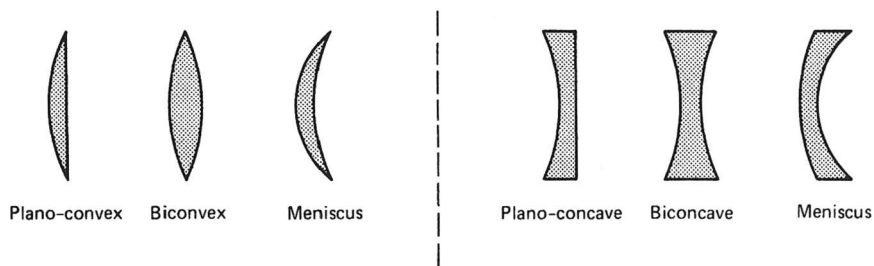
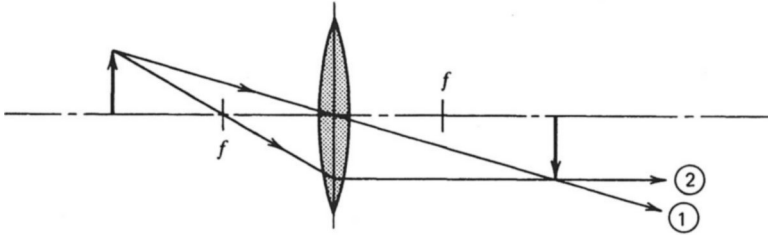


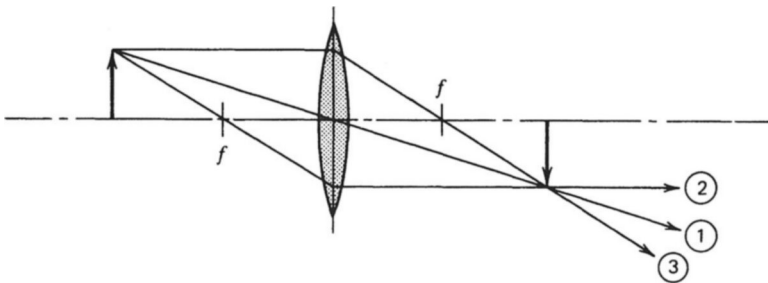
Figure 1.9. Examples of some common types of lenses.

EXAMPLE 1.C. Sketch the rays for two lenses of equal focal length but opposite sign. The object in both cases is located $2f$ from the lens.

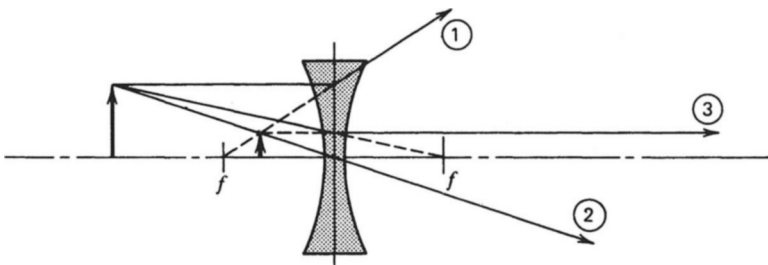
Solution: Draw the lens and then the optic axis with the object and focal length noted. Draw the undeviated ray ① through the center of the lens,



then draw the ray through the focus of the lens ② which becomes a parallel ray upon passing through the lens. The intersection of these two rays locates the object. The third ray ③, parallel to the axis will, after refraction by the lens, also go through the focus and then the image point.



The negative lens reverses the roles of the two focal points. Starting as before, we lay out the system, but the ray parallel to the optic axis ① diverges from the axis and *appears* to come from the focal point in the *object* space. Usually the extension of the apparent ray is drawn with a



dashed line, as shown. The undeviated ray ② crosses the extension of the deviated ray, locating the virtual image. The ray that is parallel to the optic axis after refraction ③ was directed at the other focus. The extension of this parallel ray back into object space also crosses the top of the virtual image, located by the first two rays.

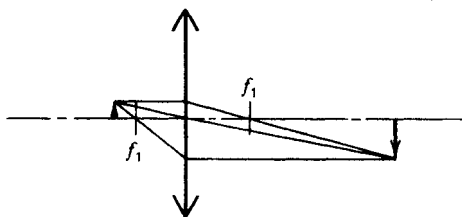
Most optical systems contain more than one lens or mirror. Combinations of elements are not difficult to handle [Problem 1.11] according to the following rule:

6. The image of the original object produced by the first element becomes the object for the second element. The object of each additional element is the image from the previous element.

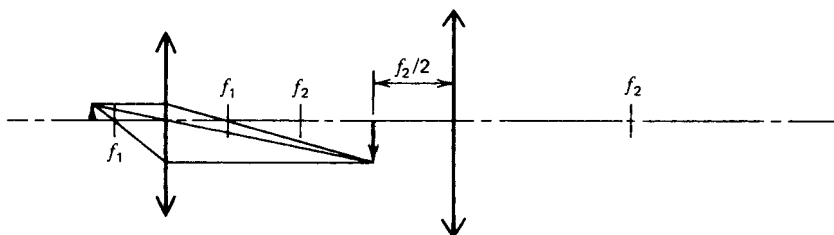
In Example 1.D is a sketch of the rays for a compound microscope. More elaborate systems can be handled in a similar manner. In many cases the elaborate systems can be broken down into simpler systems that can be handled separately, at first, then joined together later.

EXAMPLE 1.D. Given two positive lenses, focal lengths f_1 and f_2 . The first lens provides a three-power magnification and produces an intermediate image halfway between the front focal point of the second lens and its vertex. Sketch the rays for this system.

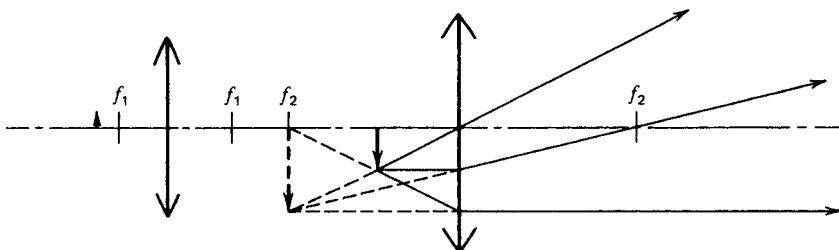
Solution: Draw the first lens (or if space is tight, just its centerline, as shown in the diagram) and its optic axis. Then draw the undeviated ray through the middle of the lens. Pick a point on the object side of the lens and then another point three times as far from the lens on the image side.



Draw parallel rays to the lens from the heads of the object and image. Then connect these to their respective image and object. The points where they cross the axis locate the focal points. Draw the second lens



and locate the front focal point and vertex of the second lens equal distances from the real image, then add the back focal point. Draw the undeviated ray through the second lens and draw a parallel ray to the second lens and through the back focus. Note that the rays diverge from



each other. Extend the rays forward, using dashed lines, to a point where they cross. This locates the virtual image, magnified and inverted with respect to the original object.

When a system requires large lengths to accomplish some design objective, it is sometimes necessary to “repackage” it into a more compact volume. It is then that folding mirrors and prisms can be used to turn an axis at right angles to itself or to fold it back on itself. There can be other constraints, such as changing a horizontal axis to a tilted axis for ease of observation, as in a microfilm projector shown in Fig. 1.10, or just “getting the optical system out of the way” to make room for mechanical components that cannot be so easily “rerouted.” Usually the design is done in an unfolded configuration. Then, noting those regions where a mirror would not interfere with the components of the system, the system is folded to fit the space. Sometimes this exercise forces one to reconsider the original design because the system will not fold conveniently. It is this interaction between elementary optics and practical considerations that presents some interesting problems for the designer in the early stages of the design.

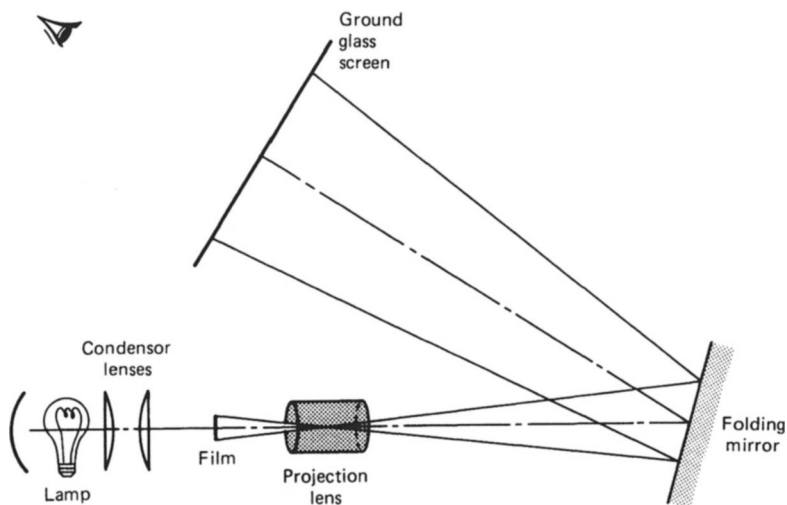


Figure 1.10. Folding a projection system to produce a more compact design and convenient screen orientation.

1.4. MIRRORS AND LENSES—CALCULATIONS

Thus far we have not put any numbers with the examples we have shown. While there are precise graphical methods for assessing an optical system, ray sketching is only used as a design shorthand. It is through calculation that we can determine if the system will do what we want it to. And it is only through these calculations that we can specify the necessary components, modify the initial values, and understand the limitations of the design. After settling on a sign convention for ray tracing, we shall derive some simple equations for mirrors and thin lenses.

Rays traced close to the optic axis of a system, those that have a small angle with respect to the axis, are most easily calculated because some simple approximations can be made in this region. This approximation is called the *paraxial approximation*, and the rays are called *paraxial rays*. This approximation can be used to derive the focal length of a spherical mirror.

A parallel ray close to the optic axis strikes the mirror at a point A , as shown in Fig. 1.11, at an angle θ to the normal to the mirror. The ray is reflected from the surface at angle θ and crosses the optic axis at the focal point C . The normal at the point A is a radius of the mirror centered at point O on the optic axis. Thus \overline{OA} is equal to the radius R of the mirror. Because the incident ray and the axis are parallel, the angle AOC is also equal to θ , thus triangle AOC is an isosceles triangle and $\overline{AC} = \overline{OC}$. As the angle θ gets smaller, $\overline{AC} = \overline{BC} = f$, the focal length and $\overline{BO} = \overline{AC} + \overline{OC} = 2\overline{AC} = 2f$. But $\overline{BO} = R$; therefore $f = R/2$, for rays close to the axis. For concave

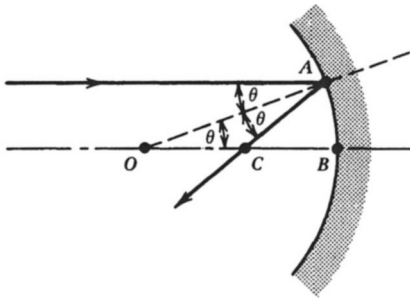


Figure 1.11. Geometry for the derivation of focal length for a concave spherical mirror.

mirrors the focal length is positive; for convex mirrors it is negative, as can be demonstrated by ray sketching with parallel incident rays.

As we examine parallel rays further from the optic axis, we find that after reflection they cross the axis at a point closer to the vertex of the mirror than the focus for small angles. This variation in the focal point with incoming ray height is a defect of a spherical mirror and is called *spherical aberration*. For the case of incident parallel rays it can be corrected by changing the shape of the mirror from a sphere to a parabola. The prescriptions for other object distances are different. Later on, we shall consider such aberrations; for the present, we assume that all our rays are paraxial rays, which exhibit no aberrations.

Before proceeding, a set of sign conventions should be set down for the thin lens calculations to be considered next and for the paraxial (Chapter 5) calculations to be done later. We use a standard right-handed coordinate system with light propagating generally along the z axis.

1. Light initially travels from left to right in a positive direction. (Refractive indices are positive in this case.)
2. Distances to the right of a component (or surface) are positive; to the left, negative.
3. Heights above the axis are positive; below the axis, negative.
4. The radius of curvature of a surface is positive if its center of curvature lies to the right of its vertex; to the left, negative.
5. Focal lengths of converging elements are positive; diverging elements have negative focal lengths.
6. Light traveling in a *negative* direction (i.e., in the z -direction), is traced in exactly the same manner except that (a) distances are reversed in sign (left, positive; right, negative) and (b) the signs of refractive indices are negative.

The change in the sign of distances holds both for inserting a value in an equation and for determining the image distance after calculation.

Having settled on the sign convention for tracing rays, it will be obvious after a re-examination of the derivation of the focal length of a spherical

mirror, that there is an error. It is that a mirror with a positive focal length has a negative radius of curvature (the center of curvature is to the left of the mirror). To be correct then, the conclusion to the derivation should be: $f = -R/2$. Note that this holds for a mirror with a negative focal length.

1.4.1. Mirror Equation

Next, we derive a simple equation relating the image distance to the object distance and the focal length of the mirror. In Fig. 1.12 is an object a distance t_0 in front of a mirror (t_0 is negative because of Rule 2 above). A ray at a small angle to the optic axis reflects off the mirror and crosses the axis at a distance t_1 from the mirror. (Note that t_1 is positive because of Rule 6.) Since a ray along the optic axis is reflected back along the axis, the two rays from the object form an image at t_1 . If u_0 is the angle of the incident ray to the axis and θ is its angle of incidence (and reflection) and y_1 is the distance of the reflection point above the axis, we may write for small angles [Problem 1.12]:

$$\tan u_0 \cong \sin u_0 \cong u_0 = \frac{y_1}{-t_0}. \quad (1.9)$$

Also

$$\tan u_1 \cong \sin u_1 \cong u_1 = \frac{y_1}{t_1}, \quad (1.10)$$

$$\tan \alpha \cong \sin \alpha \cong \alpha = \frac{y_1}{-R} \quad (1.11)$$

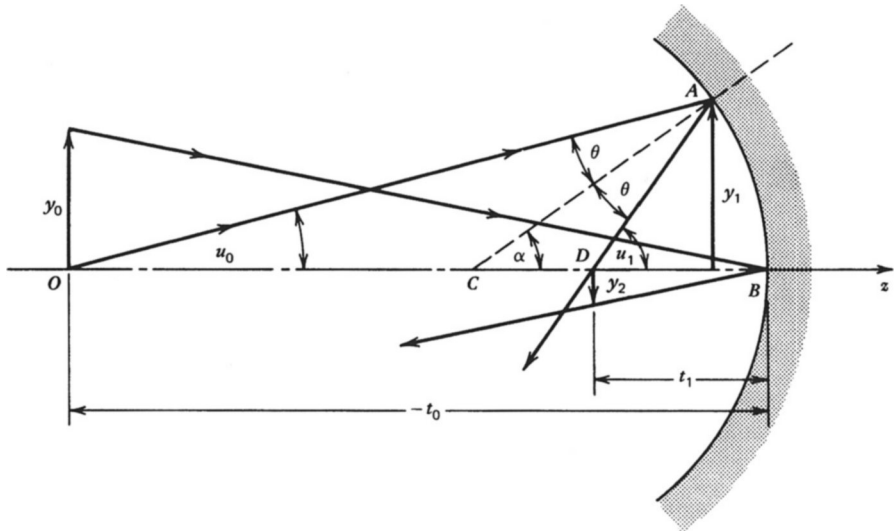


Figure 1.12. Geometry for a derivation of the mirror equation.

Since u_1 is the exterior angle to triangle ODA,

$$u_1 = 2\theta + u_0, \quad (1.12)$$

and u_1 is also the exterior angle for triangle CDA,

$$u_1 = \alpha + \theta.$$

Substituting for θ in Eq. (1.12),

$$u_1 = 2(u_1 - \alpha) + u_0,$$

or

$$u_1 + u_0 = 2\alpha,$$

and using Eqs. (1.9) to (1.11),

$$\frac{y_1}{t_1} - \frac{y_1}{t_0} = -2\frac{y_1}{R}.$$

The result is the *mirror equation*:

$$\frac{1}{t_1} - \frac{1}{t_0} = -\frac{2}{R} = \frac{1}{f}. \quad (1.13)$$

If the object has a height y_0 and a ray is traced to the vertex of the mirror ($y_1 = 0$), then the image at t_1 will have a height y_2 . From Fig. 1.12 it can be seen that the object and image heights are corresponding sides of similar triangles. Thus

$$\frac{y_0}{-t_0} = \frac{-y_2}{t_1}.$$

The *transverse* or *lateral magnification* of a mirror, M , the ratio of the image height to the object height, is simply the ratio of the image distance to the object distance:

$$M = \frac{y_2}{y_0} = \frac{t_1}{t_0} = \frac{\text{image distance}}{\text{object distance}}. \quad (1.14)$$

Not only does this ratio give the size of the final image, its sign also indicates the orientation of the image relative to the object. A negative sign indicates the image is inverted with respect to the object. The axial or *longitudinal magnifi-*

cation M_t can be shown to be the square of the lateral or transverse magnification M ,

$$M_t = M^2 = \frac{t_1^2}{t_0^2}. \quad (1.15)$$

In referring to transverse magnification, an unsubscripted M will be used. For any other magnification parameters a subscript will be provided.

1.4.2. Thin Lens Equation

We can derive the object-image relationship for a lens in a manner similar to that used for the spherical mirror. With reference to Fig. 1.13 let us use two rays from an off-axis object point, one parallel to the axis, and one through the front focal point. We have a set of similar triangles ABC and BCD . In ABC ,

$$\frac{y_0 - y_2}{-t_0} = \frac{-y_2}{f},$$

and in BCD

$$\frac{y_0 - y_2}{t_1} = \frac{y_0}{f}.$$

Adding these two equations and dividing through by $y_0 - y_2$ we obtain the *thin lens equation* [Problem 1.13].

$$\frac{1}{t_1} - \frac{1}{t_0} = \frac{1}{f}. \quad (1.16)$$

You may notice that this equation differs from the thin lens equation that you may have learned in a general physics course—the two terms on the left side

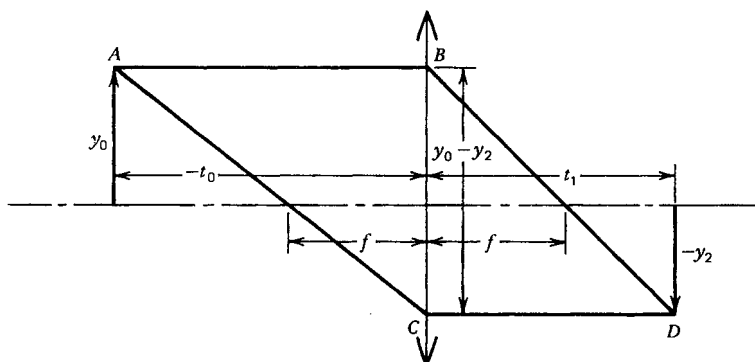


Figure 1.13. Geometry for a derivation of the thin lens equation.

are a difference rather than a sum, as in most elementary texts [Refs. 1.1 and 1.2]. This is only because the sign convention for object distances has been reversed. As you will see later, this convention gives a means of easily and consistently tracing rays through many surfaces.

The transverse magnification can be found from:

$$M = \frac{y_2}{y_0} = \frac{y_2}{f} \cdot \frac{f}{y_0} = \frac{y_0 - y_2}{t_0} \cdot \frac{t_1}{y_0 - y_2} = \frac{t_1}{t_0}. \quad (1.17)$$

It is the same as for mirrors. As in the case of the mirrors, the longitudinal magnification is the square of the transverse magnification, Eq. (1.15).

For a thin lens of refractive index n , the focal length of a lens can be shown to be

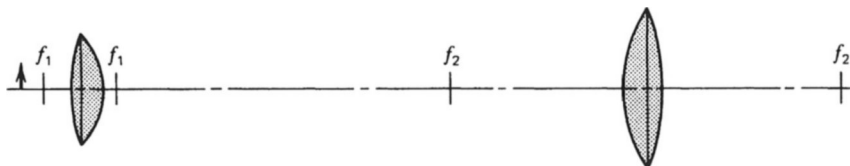
$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right), \quad (1.18)$$

where R_1 and R_2 are the radii of curvature of the first and second surfaces encountered by the light ray [Problem 1.14]. This relation is known as the *lensmaker's equation*. While it is possible to derive this equation using Snell's Law and some small angle approximations, we shall hold the derivation until the chapter on paraxial ray tracing, where it is easily derived and provides an effective example.

The relationship of an image to an object for a positive focal length lens is the same for all lenses. If we start with an object at infinity as it is moved toward the lens we find from Eq. 1.16 that for a positive lens a real image is located at the focal point of the lens ($1/t_0 = 0$, therefore $t_1 = f$), and as the object approaches the lens the image distance gets larger until it reaches a point $2f$ on the other side of the lens. At this point the object and images are the same size and the same distance from the lens. As the object is moved from $2f$ to f , the image moves from $2f$ to infinity. An object placed between the focal point and the lens gives a virtual, magnified image that decreases in magnification as the object approaches the lens. For a negative lens, the situation is simpler: starting with an object at infinity, a virtual image, demagnified, appears to be at the focal point on the same side of the lens as the object. As the object moves closer to the lens so does the image, until the image and object are equal in size at the lens [Problem 1.15].

The calculation for a combination of lenses is not much harder than that for a single lens. As indicated in the previous section on ray sketching, the image of the preceding lens becomes the object of the succeeding lens [Problem 1.6]. In Example 1.E we show a calculation for a two-lens system similar to the one sketched in Example 1.D, and in Example 1.F we derive the focal length for two thin lenses in contact using the same principle.

EXAMPLE 1.E. Given two lenses $f_1 = 10$ mm and $f_2 = 55$ mm, separated by 160 mm. What are the locations, magnifications, and orientation of the intermediate and final images if the object is located 11 mm in front of the first lens? This example is similar to the Example 1.D.



Solution: 1. Diagram the optical system

2. First lens calculation

$$t_0 = -11 \text{ mm.} \quad \frac{1}{t_1} = \frac{1}{f_1} + \frac{1}{t_0} = \frac{1}{10} - \frac{1}{11} = \frac{1}{110}.$$

$$t_1 = +110 \text{ mm.} \quad M_1 = \frac{t_1}{t_0} = \frac{110}{-11} = -10.$$

Since t_1 is positive, the image is to the right of the first lens and is therefore real. It is magnified 10 times, and because of the negative sign of the magnification, the image is inverted.

3. Second lens calculation. The object distance for the second lens $t'_1 = -160 + 110 = -50$. (Note t'_1 is negative since the intermediate image is to the left of the second lens)

$$\frac{1}{t_2} = \frac{1}{f_2} + \frac{1}{t'_1} = \frac{1}{55} - \frac{1}{50} = -\frac{1}{550},$$

$$t_2 = -550 \text{ mm,} \quad M_2 = \frac{t_2}{t'_1} = \frac{-550}{-50} = 11.$$

Since t_2 is negative, the final image is to the left of the second lens and therefore is virtual. It is magnified 11 times and not inverted with respect to the intermediate image, since the sign of the magnification was positive.

4. Combined calculation. Total magnification is

$$\begin{aligned} M &= M_1 \times M_2 = -10 \times 11 \\ &= -110 \text{ times.} \end{aligned}$$

The image is virtual, 110 times magnified, and inverted with respect to the original object.

EXAMPLE 1.F. Given two thin lenses in contact, f_1 and f_2 , show that their combined focal length is

$$\frac{1}{f_{\text{comb}}} = \frac{1}{f_1} + \frac{1}{f_2}.$$

Solution: Here the thin lens formula is used twice. Starting with an object distance for both the first lens and the combined system. The image distance is

$$\frac{1}{t_1} = \frac{1}{f_1} + \frac{1}{t_0}.$$

But this is also the object distance for the second lens, so that

$$t'_0 = +t_1 \quad \text{and} \quad \frac{1}{t'_1} = \frac{1}{f_2} + \frac{1}{t'_0} = \frac{1}{f_2} + \frac{1}{t_1}.$$

Substituting for $1/t_1$

$$\frac{1}{t'_1} = \frac{1}{f_2} + \frac{1}{f_1} + \frac{1}{t_0}$$

or

$$\frac{1}{t'_1} - \frac{1}{t_0} = \frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{f_{\text{comb}}}.$$

Since t_0 is the object distance for the combination and t'_1 is the image distance for the combination, therefore,

$$\frac{1}{f_1} + \frac{1}{f_2} = \frac{1}{f_{\text{comb}}}$$

acts as the focal length term for the combination.

1.5. SIMPLE OPTICAL INSTRUMENTS

Perhaps the simplest optical instrument that can be constructed is produced by placing a lens in front of a screen and moving it back and forth in front of the screen until an image is brought into focus, as shown in Fig. 1.14(a). When

the lens is placed in the wall of a black box with film on the opposite wall approximately one focal length away the result is a *camera*. By moving the lens in relation to the film plane we can focus on subjects at different distances from the box [Problem 1.17]. Through careful design of the lens, corrections can be made to improve the camera for the use at hand, whether it is for taking portraits or landscapes, for surveying at high altitudes, or for recording high-speed events.

1.5.1. The Eye

The eye is also a single-lens camera. As shown in Fig. 1.14(b), it consists of the front surface of the eye, the *cornea*, the *lens*, the *iris*, the *vitreous humor*, and the *retina*. In contrast to the conventional camera, where focusing is accomplished by moving the lens, the cornea and the lens of the eye remain fixed and the focal length is changed by a bending of the lens by the eye muscles [Problem 1.18]. The iris limits the amount of light entering the eye, closing down in

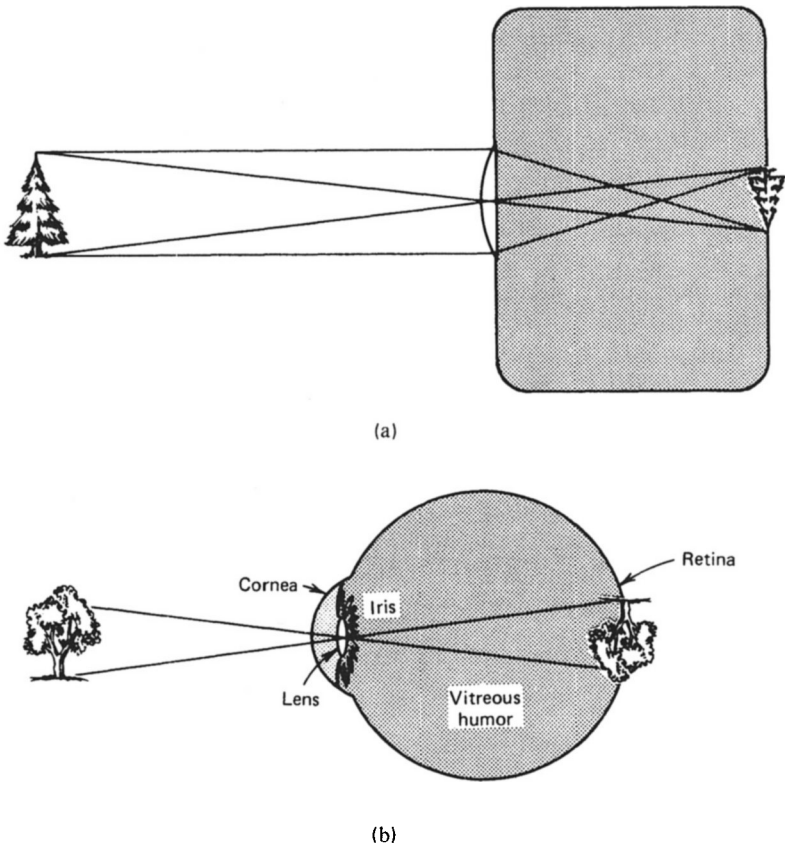


Figure 1.14. Two examples of optical instruments. (a) the camera and (b) the eye.

bright surroundings and opening in dim or dark environments. The maximum opening is about 7 mm in diameter. Under optimum seeing conditions the iris diameter is about 3.5 mm.

The eye muscles are most relaxed when the eye is looking at distant objects (i.e., “focused at infinity”), so that many optical calculations involving the eye are made under that assumption. However, for examining something critically at short distances, there is a closest point at which the eye can see an object, the *near point*. The distance from the eye to the near point is called the *distance of most distinct vision*. For young eyes this distance is about 25 cm, and it gets progressively longer as a person ages.

One of the terms used in the field of optics is *conjugate*. If you examine the thin lens equation (1.16), you will notice that an object located at t_0 from the lens and an image located at t_1 will satisfy the equation. The equation would be equally true if the object and image distances were interchanged. The two points are *conjugate* to one another. Unless they are equal, the image and object distances are sometimes referred to as the short conjugate or the long conjugate, depending on the relative sizes of the two distances. For example, in the case of the eye, the object distance is the long conjugate and the image distance is the short conjugate.

In the projector, the two conjugates are reversed. The object, usually a small transparency, is magnified on a screen, as shown earlier in Fig. 1.10. The image distance is now the long conjugate. The magnification of the projector is just a ratio of the conjugates of the system.

1.5.2. The Magnifier

If the slide transparency is placed inside the focal point of a lens, the lens acts as a simple magnifier providing the view with an erect, magnified virtual image. The magnification of an object by an optical instrument is calculated as an angular rather than linear magnification. The comparison is not between the sizes of the image and object; it is instead a comparison of the apparent sizes of the object as seen through the magnifier to the same object located at the near point by the unaided eye. This magnification is an increase in the angle subtended by the virtual image over the angle subtended by the object at the near point. If D is the distance of most distinct vision and h is the object height (Fig. 1.15(a)), then the angle subtended by the object before magnification is

$$\alpha = \frac{h}{D}.$$

When the object is viewed by a single thin lens the linear magnification is

$$M = \frac{t_1}{t_0} = \frac{h'}{h}.$$

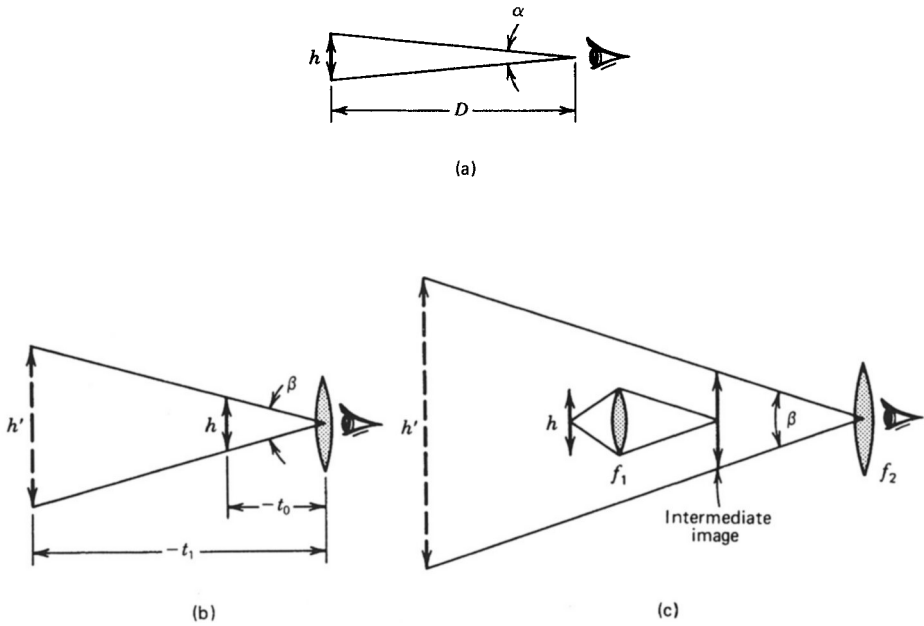


Figure 1.15. Magnifiers. (a) Object of height h subtends an angle α at the near point, a distance D from the eye. (b) A simple magnifier. (c) A compound microscope, f_1 is the objective lens and f_2 is the eyepiece.

The angle subtended by the image is (Fig. 1.15(b))

$$\beta = -\frac{h'}{t_1} = -\frac{h}{t_0}$$

Using the thin lens equation

$$t_0 = \frac{ft_1}{f - t_1},$$

$$\beta = -\frac{h(f - t_1)}{ft_1}$$

The angular magnification is then

$$M_\theta = \frac{\beta}{\alpha} = -\frac{h(f - t_1)}{ft_1} \cdot \frac{D}{h} = D \left(\frac{1}{f} - \frac{1}{t_1} \right).$$

There are two cases. If the object is to be viewed without strain, the eye should be relaxed to view the image at infinity. In that case $t_1 = \infty$, and

$$M_\theta = \frac{D}{f}, \quad (\text{image at infinity}) \quad (1.19)$$

where D is the distance of most distinct vision. If the eye is focused to view the object at the near point ($t_1 = -D$), the angular magnification is [Problem 1.19]

$$M_\theta = \frac{D}{f} + 1. \quad (\text{image at near point}) \quad (1.20)$$

In all systems in which the image is magnified to provide better seeing, such calculations are required.

In some cases the magnification of a single lens is not sufficient. To overcome this limitation a second lens is inserted (Fig. 1.15(c)). The system resembles the one sketched in Example 1.D. The objective lens, a very short-focal-length lens, forms a real intermediate image magnified from 5 to 60 times. This real image is then examined using an eyepiece, which enables one to observe a virtual image as in the case of a simple magnifier under a combined magnification of 50 to 1000 times. It can be derived [Problem 1.20] that for the compound microscope the angular magnification is

$$M_\theta = \frac{t_1}{t_0} \cdot \frac{D}{f_2}. \quad (\text{image at infinity}) \quad (1.21)$$

for the image viewed at infinity or

$$M_\theta = \frac{t_1}{t_0} \cdot \left(\frac{D}{f_2} + 1 \right) \quad (\text{image at near point}) \quad (1.22)$$

for the image at the near point. A calculation for such a system is given in Example 1.E [Problem 1.21].

1.5.3. The Telescope

Just as a microscope makes something close look larger, the telescope attempts to do the same thing for objects far away. The simplest telescope consists of two optical components of different focal lengths, separated by the sum of their focal lengths. They may be either lenses or mirrors, as shown in Fig. 1.16. They are arranged so that the long-focal-length lens is the objective and the short-focal-length lens is the eyepiece. In comparison with most of the object distances, the focal lengths of the components used in most telescopes are small. Therefore the object may often be considered to be located at infinity

SIMPLE OPTICAL INSTRUMENTS

31

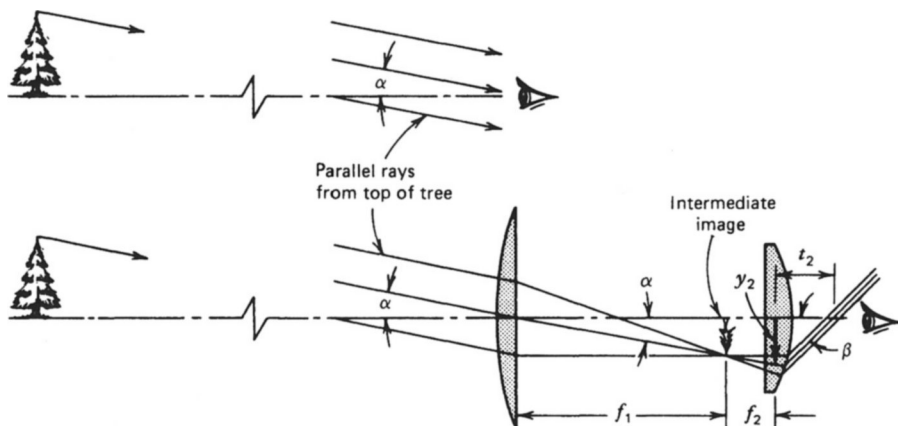


Figure 1.16. Geometry for a derivation of the angular magnification of a telescope. In the top picture the eye perceives the tree on the horizon as subtending an angle α . In the bottom, the tree subtends the same angle α with respect to the objective, but the focusing of the light for the tree to an intermediate image and the recollimation of the light by the second lens changes the angular subtense to β . To someone looking through the eyepiece (focal length, f_2) the tree would appear to be inverted and subtend an angle β .

[Problem 1.22]. Since the object distance is so large, the magnification of a telescope does not come from an increased image size, but from an increase in the subtended angle of the image as was discussed in the previous section. For example, a distant object such as a tree on the horizon will subtend a certain angle α when observed by the naked eye. When this object is observed through the telescope, the tree will appear to subtend an angle β that will be larger than the original angle. As illustrated in Fig. 1.16, the ray through the center of the objective subtends essentially the same angle α . The angle that the ray subtends after passing through the telescope is β . We can relate these two angles by using the thin lens equation for the eyepiece f_2 to find the point at which the ray through the center of the objective crosses the axis. Using the thin lens equation (1.16) on the second lens with an object distance equal to $f_1 + f_2$,

$$\frac{1}{t_2} = \frac{1}{f_2} - \frac{1}{f_1 + f_2}.$$

Now

$$\alpha = \frac{-y_2}{-(f_1 + f_2)} \quad \text{and} \quad \beta = \frac{-y_2}{t_2},$$

so that the *angular magnification* of the telescope is

$$\begin{aligned}
 M_{\text{telescope}} &= \frac{\beta}{\alpha} = \frac{-y_2}{t_2} \cdot \frac{f_1 + f_2}{y_2} = -\left(\frac{1}{f_2} - \frac{1}{f_1 + f_2}\right) \cdot (f_1 + f_2) \\
 &= -\frac{f_1 + f_2}{f_2} + 1 = \frac{-f_1}{f_2},
 \end{aligned} \tag{1.23}$$

the ratio of the focal lengths of the two components.

Since all the rays are by assumption coming from infinity, the image is formed at the focal plane of the first lens. And, since the focal point of the second lens is at the intermediate image, the final image is at infinity. Because both object and final image are located at infinity, the focal length of the telescope is indeterminate. This is called an *afocal system*. Another afocal system is the rangefinder. This is discussed in Chapter 4 as an example of an optical system consisting of prisms and mirrors.

1.6. POLARIZATION

Polarization has no part in geometrical optics. However, it is essential that we be able to understand and use the polarization properties of light in the design of optical systems. For example, we can control the amount and direction of light through the use of polarization. At this point we leave the realm of geometrical optics, and for the rest of this chapter, explore some of the concepts of physical optics that we will need for modern optical design.

Physical optics has as its basis the fact that electromagnetic radiation, including light, is a combination of electric and magnetic waves oscillating at a frequency ν and with a wavelength λ and propagating through a medium at a velocity v which satisfies Eq. 1.3. In its simplest form, a *monochromatic* (that is, single color) light wave consists of electric and magnetic fields that oscillate sinusoidally with a frequency that is related to its color. Since electric and magnetic fields are vector quantities, both their magnitude and direction must be specified. But, because the field directions are always perpendicular to one another in nonabsorbing media, only the direction of the electric field of a light wave is needed to specify the polarization of the light wave. Since this direction is perpendicular to the propagation direction of the light, light waves are transverse waves.

The form of polarization of light can be quite complex. However, for most design situations there are a limited number of types that are needed to describe the polarization of light in an optical system. Figure 1.17 shows the path traced by the electric field during one full cycle of oscillation of the wave ($T = 1/\nu$) for a number of different types of polarization. Figure 1.17(a) shows *linear polarization*, where orientation of the electric field vector of the wave does not change with time as the field amplitude oscillates from a maximum

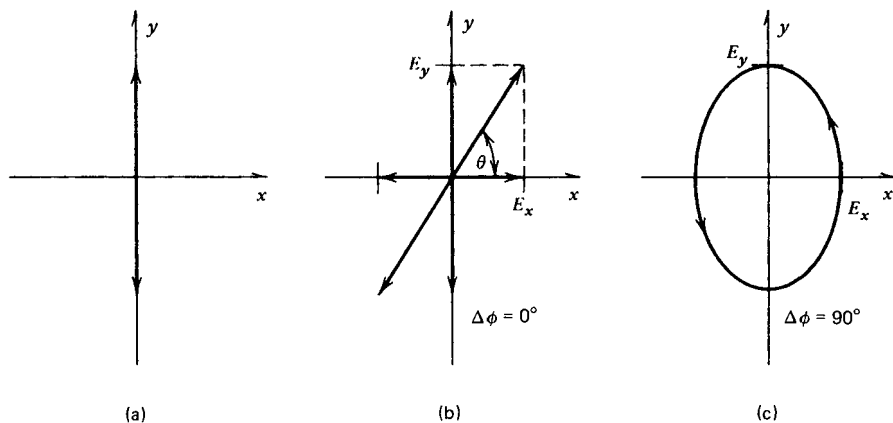


Figure 1.17. Three special polarization orientations: (a) linear, along a coordinate axis; (b) linear, components along coordinate axes are in phase ($\Delta\phi = 0$) and thus produce linear polarization; (c) same components, 90° out of phase, produce elliptical polarization.

value in one direction to a maximum value in the opposite direction. The orientation of the electric field is referenced to some axis perpendicular to the direction of propagation. In some cases, it may be a direction in the laboratory or optical system, and it is specified as horizontally or vertically polarized or polarized at some angle to a coordinate axis.

Because the electric field is a vector quantity, electric fields add as vectors. For example, two fields, E_x and E_y , linearly polarized at right angles to each other and oscillating in phase (maxima for both waves occur at the same time), will combine to give another linearly polarized wave, shown in Fig. 1.17(b), whose direction ($\tan \theta = E_y/E_x$) and amplitude ($\sqrt{E_x^2 + E_y^2}$) are found by addition of the two components. If these fields are 90° out of phase (the maximum in one field occurs when the other field is zero), the electric field of the combined fields traces out an ellipse during one cycle, as shown in Fig. 1.17(c). The result is called *elliptically polarized light*. The eccentricity of the ellipse is the ratio of the amplitudes of the two components. If the two components are equal, the trace is a circle. This polarization is called *circularly polarized*. Since the direction of rotation of the vector depends on the relative phases of the two components, this type of polarization has a handedness to be specified. If the electric field coming from a source toward the observer rotates counterclockwise, the polarization is said to be *left handed*. *Right-handed* polarization has the opposite sense, clockwise. This nomenclature applies to elliptical as well as circular polarization. Light whose direction of polarization does not follow a simple pattern such as the ones described here is sometimes referred to as unpolarized light. This can be somewhat misleading because the field has an instantaneous direction of polarization at all times, but it may not be easy to discover what the pattern is. A more descriptive term is *randomly polarized light*.

1.6.1. Methods of Polarization

Light from most natural sources tends to be randomly polarized. While there are a number of methods of converting it to linear polarization, only those that are commonly used in optical design will be covered. One method is reflection, since the amount of light reflected off a tilted surface is dependent on the orientation of the incident polarization and the normal to the surface. A geometry of particular interest is one in which the propagation direction of reflected and refracted rays at an interface are perpendicular to each other, as shown in Fig. 1.18. In this orientation the component of light polarized *parallel* to the *plane of incidence* (the plane containing the incident propagation vector and the surface normal, i.e., the plane of the page for Fig. 1.18) is 100% transmitted. There is *no reflection* for this polarization in this geometry. For the component of light perpendicular to the plane of incidence, there is some light reflected and the rest is transmitted. The angle of incidence at which this occurs is called *Brewster's angle*, θ_B , and is given by [Problem 1.23]:

$$\tan \theta_B = \frac{n_{\text{trans}}}{n_{\text{incident}}} . \quad (1.24)$$

As an example, for a crown glass, $n = 1.523$, and the Brewster angle is 56.7° .

Sometimes only a small amount of polarized light is needed, and the light reflected off of a single surface tilted at Brewster's angle may be enough to do the job. If nearly complete polarization of a beam is needed, one can construct a linear polarizer by stacking a number of glass slides (e.g., clean microscope slides) at Brewster's angle to the beam direction. As indicated in Fig. 1.19, each interface rejects a small amount of light polarized perpendicular to the plane of incidence.

The "pile of plates" polarizer just described is somewhat bulky and tends to get dirty, reducing its efficiency. Plastic polarizing films are easier to use and mount. These films selectively absorb more of one polarization component and transmit more of the other. The source of this polarization selection is

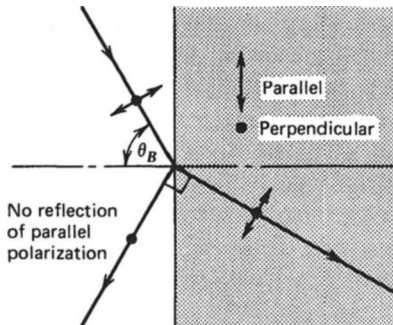


Figure 1.18. Geometry for the Brewster angle.

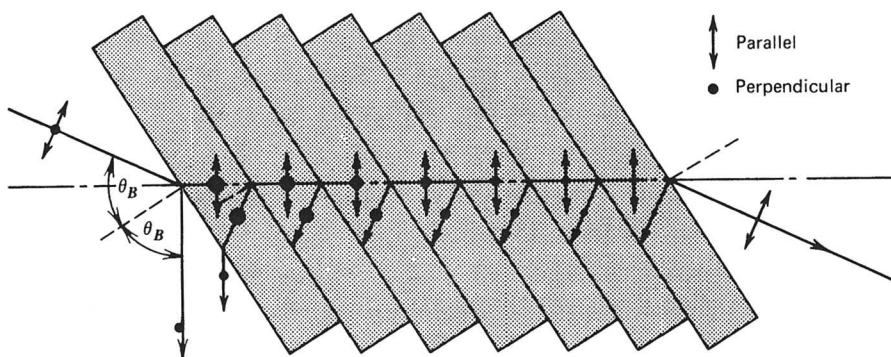


Figure 1.19. A “pile of plates” polarizer. This device working at the Brewster angle, reflects some portion of the perpendicular polarization (here depicted as a dot, indicating an electric field vector perpendicular to the page) and transmits all parallel polarization. After a number of transmissions most of the perpendicular polarization has been reflected away leaving a highly polarized parallel component.

many aligned linear chains of a polymer to which light-absorbing iodine molecules are attached. Light that is polarized parallel to the chains is easily absorbed, whereas light polarized perpendicular to the chains is mostly transmitted. The sheet polarizers made by Polaroid Corporation are labeled by their type and transmission. Three common linear polarizers are HN-22, HN-32, and HN-38, where the number following the HN indicates the percentage of incident unpolarized light that is transmitted through the polarizer as polarized light [Problem 1.24].

When you look through a crystal of calcite (calcium carbonate) at some writing on a page, you see a double image as illustrated in Fig. 1.20. If you rotate the calcite, keeping its surface on the page, one of the images rotates with the crystal while the other remains fixed. This phenomenon is known as *double refraction*. If we examine these images through a sheet polarizer, we find that each image has a definite polarization, and these polarizations are perpendicular to each other.

The calcite crystal is one of a whole class of *birefringent* crystals that exhibit double refraction. (Doubly refracting is the English equivalent for the Latin, *birefringent*.) The physical basis for this phenomenon is described in detail in most optics texts. For our purposes it is sufficient to know that the crystal has a refractive index that varies with the direction of propagation in the crystal *and* the direction of polarization. The *optical axis of the crystal* (no connection to the optical axis of a lens or a system) is a direction in the crystal to which all other directions are referenced. Light whose component of the polarization is *perpendicular* to the optic axis travels through the crystal as if it were an ordinary piece of glass with a single refractive index, n_o . Light of this polarization is called an *ordinary ray*. Light polarized *parallel* to a plane containing the optic axis has a refractive index that varies between n_o and a

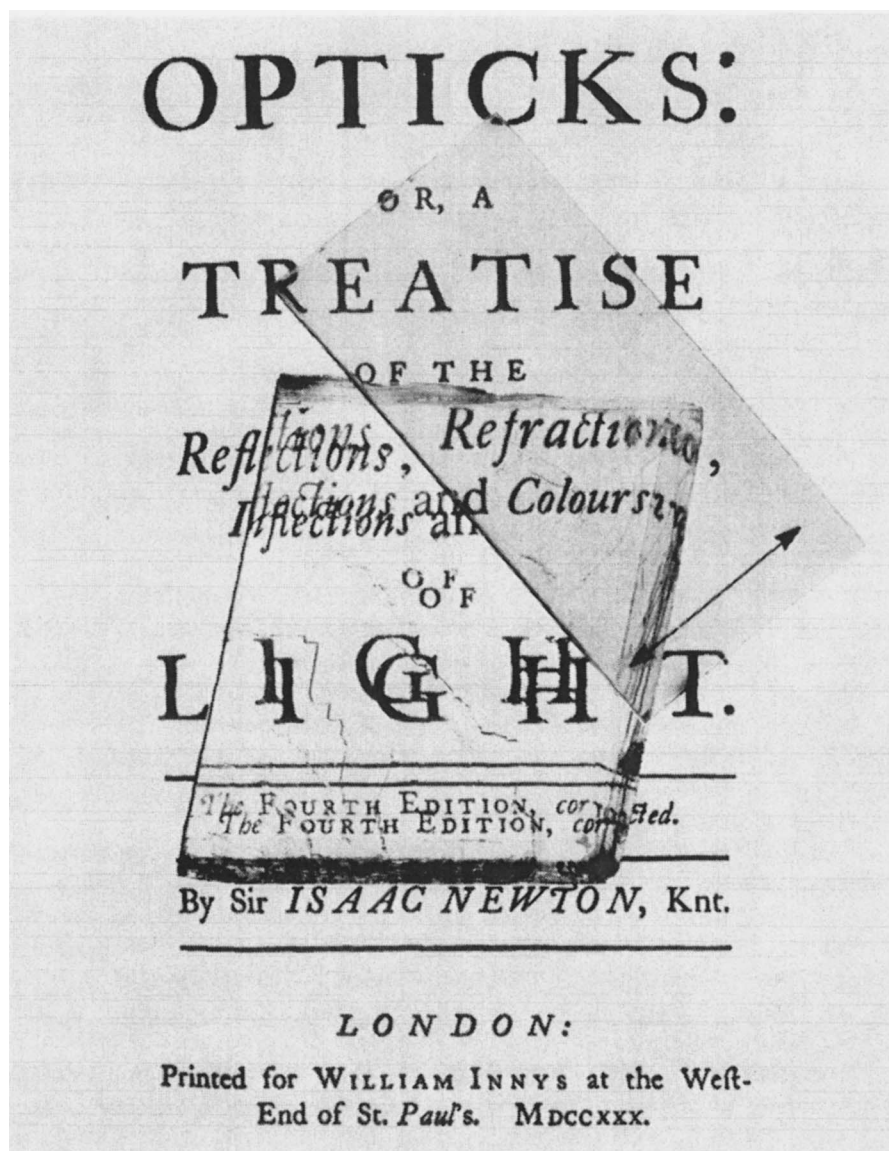


Figure 1.20. Polarization of light by a birefringent crystal. Note that through the polarizing sheet (dark rectangle marked with arrow to indicate polarization direction) only one image can be seen. This demonstrates that polarization of the two images are different. A rotation of the polarizer ray 90° would allow transmission of the other polarization only showing the polarization of the two images are orthogonal to each other. (Photo by Vincent Mallette.)

different value, n_e . The material exhibits a refractive index n_e where the field component is parallel to the optic axis and the direction of light propagation is perpendicular to the optic axis. Light of this polarization is called an *extraordinary ray*. The action of the crystal upon light of these two orthogonal polarization components provides the double images and the polarization of light by transmission through the crystals. If one of these components could be blocked or diverted while the other component is transmitted by the crystal, a high degree of polarization can be achieved. There are a number of designs for birefringent polarizers, many of which can be found in optics texts. To illustrate the method used, we show in Fig. 1.21 a *Nicol prism*, the first crystal polarizer to be made for that purpose. This polarizer consists of a piece of natural calcite cut in half at a particular angle and rejoined with a resinous cement from the Canada balsam tree. Light of all polarizations is incident on the crystal at an angle to the optic axis. At the first interface, double refraction splits the beam into components polarized parallel and perpendicular to the optic axis. Because of the different angles of refraction at the first crystal face, the two beams have different angles of incidence at the calcite–Canada balsam interface. Because the extraordinary refractive index of the crystal and refractive index of the Canada balsam are nearly the same, the extraordinary ray is transmitted into the second crystal with little reflection. The ordinary ray, however, is totally internally reflected since $\sin \theta_i > n_t/n_i$, as discussed in Section 1.2. In most cases, the reflected beam is absorbed by the material holding the crystal, while the transmitted beam exits through the other half of the crystal in a direction parallel to the original beam but slightly displaced from it. In some optical designs it is useful to keep the reflected beam, which is also highly polarized, and use it somewhere else in the system.

In many cases polarizers are used to provide information about a material that produces, in some manner, a change in the form of polarized light passing through it. The standard configuration, shown in Fig. 1.22, consists of a light source S, a polarizer P, the material M, another polarizer, called an analyzer

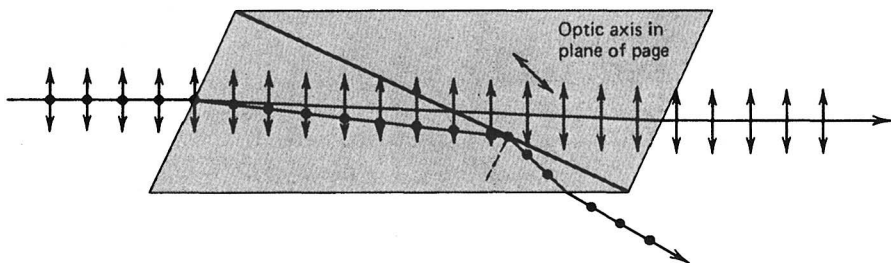


Figure 1.21. A Nicol prism. This set of two calcite crystals held together by Canada balsam cement serves as a polarizer of incident randomly polarized light, transmitting one polarization and total internally reflecting the other polarization.

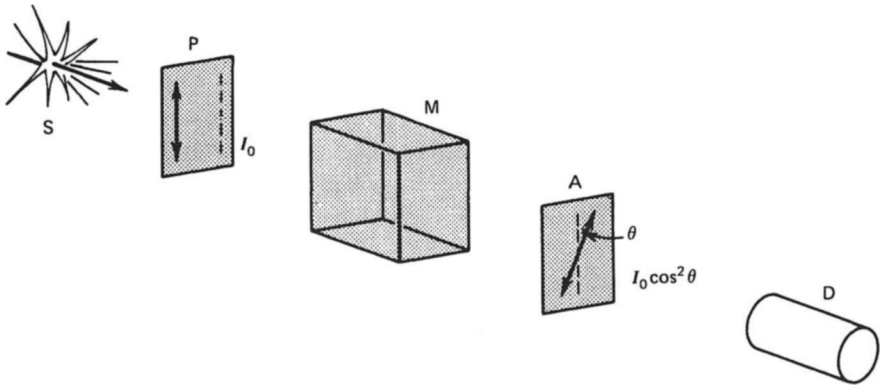


Figure 1.22. Analysis of polarized light. Randomly polarized light from source S is linearly polarized after passage through the polarizer P with irradiance I_0 . After passage through optically active material M, the polarization vector has been rotated through an angle θ . (The dashed lines on the polarizers P and A denote the transmission axes; the arrow indicates the polarization of the light). The light is analyzed by polarizer A, transmitting an amount $I_0 \cos^2 \theta$ that is detected by detector D.

A, and a detector D. Usually the polarizer is a linear polarizer, as is the analyzer. Sometimes, however, polarizers that produce other types of polarization are used.

The amount of light transmitted by a polarizer depends on the polarization of the incident beam and the quality of the polarizer. Let us take, for example, a perfect polarizer—one that transmits all of the light for one polarization and rejects (by absorption or reflection) all of the light of the other polarization. The direction of polarization of the transmitted light is the *polarization axis*, or simply the *axis* of the polarizer. Since randomly polarized light has no preferred polarization, there would be equal amounts of incident light for two orthogonal polarization directions. Thus, a perfect linear polarizer would have a Polaroid designation of HN-50, since it would pass half of the incident radiation and absorb the other half. The source in Fig. 1.22 is randomly polarized, and the polarizer pass is linearly polarized light of irradiance I_0 . If the material M changes the incident polarization by rotating it through an angle θ , what is the amount of light transmitted through an analyzer whose transmission axis is oriented parallel to the axis of the first polarizer? Since the electric field is a vector, we can decompose it into two components, one parallel to the axis of the analyzer, the other perpendicular to this axis. That is:

$$E = E_0 \cos \theta \hat{e}_{\parallel} + E_0 \sin \theta \hat{e}_{\perp}.$$

(Note that the parallel and perpendicular components here refer to their orientation with respect to the axis of the analyzer and not to the plane of incidence as in the case of the Brewster angle.) The transmitted field is the

parallel component, and the transmitted irradiance I_{trans} is the time average square of the electric field

$$I_{\text{trans}} = \langle E_0 \cos \theta \rangle^2 = \langle E_0 \rangle^2 \cos^2 \theta$$

or

$$I_{\text{trans}} = I_0 \cos^2 \theta. \quad (1.25)$$

This equation, which relates the irradiance of *polarized* light transmitted through a perfect polarizer to the irradiance of incident *polarized* light, is called the *Law of Malus*, after its discoverer, Etienne Malus, an engineer in the French army. For a nonperfect polarizer, I_0 must be replaced by αI_0 , where α is the fraction of the preferred polarization transmitted by the polarizer [Problem 1.24]. (For HN-22 film, $\alpha = 0.44$, as shown in Example 1.G.)

EXAMPLE 1.G. Given two HN-22 polarizers, what is the amount of naturally polarized light transmitted by these two polarizers whose axes are (a) parallel to one another and (b) at 45° to each other?

Solution: (a) Before we can do this problem we need to know how much light of the correct polarization will be *transmitted*. We know that 50% of the light is absorbed in the perpendicular direction. In the parallel direction 22% of the total (that is what HN-22 means) or 44% of the preferred polarization is transmitted. Therefore, when the second polarizer is put in front of the first polarizer 44% of the radiation that passed the first polarizer will be transmitted. Thus, two HN-22 polarizers whose axes are aligned will transmit 0.22×0.44 or about 10% of the initial unpolarized light.

(b) The amount transmitted by the two HN-22 polarizers at 45° to each other is then

$$\begin{aligned} I_{\text{trans}} &= 0.22 I_0 \times 0.44 \times \cos^2(45^\circ) \\ &= 0.22 \times 0.44 \times \frac{I_0}{2} = 0.0484 I_0 = 5\% I_0. \end{aligned}$$

1.6.2. Polarization Modifiers

Besides serving as linear polarizer, birefringent crystals can be used to change the type of polarization of a light beam. We shall describe the effect that these

polarization modifiers have on the beam and leave the explanation of their operation to a physical optics text.

In a birefringent crystal, light whose polarization is parallel to the optic axis travels at a speed of c/n_{\parallel} ; for a polarization perpendicular to that the speed is, c/n_{\perp} . In calcite $n_{\perp} > n_{\parallel}$, and therefore the speed of light polarized parallel to the optic axis, v_{\parallel} , is greater than v_{\perp} . Thus, for calcite, the optic axis is called the fast axis and a perpendicular axis is the slow axis. (In other crystals n_{\parallel} may be greater than n_{\perp} and the fast-slow designation would have to be reversed.)

The first device to be described is a *quarter-wave plate*. The plate consists of a birefringent crystal of a specific thickness d , cut so that the optic axis is parallel to the plane of the plate and perpendicular to the edge, whose length has been specified as d , as shown in Fig. 1.23. The plate is oriented so that its plane is perpendicular to the beam direction and its fast and slow axes are at 45° to the polarized direction of the incident linearly polarized light. Because of this 45° geometry, the incident light is split into slow and fast components of equal amplitude traveling through the crystal. The plate is cut so that the components, which were in phase at the entrance to the crystal, travel at different speeds through it and exit at the point when they are 90° , or a quarter wave, out of phase. This output of equal amplitude components, 90° out of phase, is then circularly polarized [Problem 1.25]. It can be shown that when circularly polarized light is passed through the same plate, linearly polarized light results. Also, it should be noted that if the 45° input geometry is not maintained, the output is elliptically polarized. The angle between the input polarization direction and the optic axis determines the eccentricity of the ellipse.

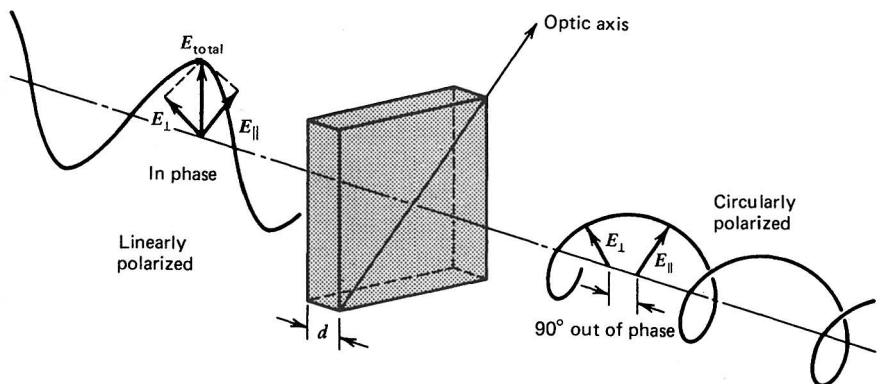


Figure 1.23. Quarter-wave plate. Incident linearly polarized light is oriented at 45° to the optic axis so that equal E_{\parallel} and E_{\perp} components are produced. The thickness of the plate is designed to produce a phase retardation of 90° of one component relative to the other. This produces circularly polarized light. At any other orientation elliptically polarized light is produced.

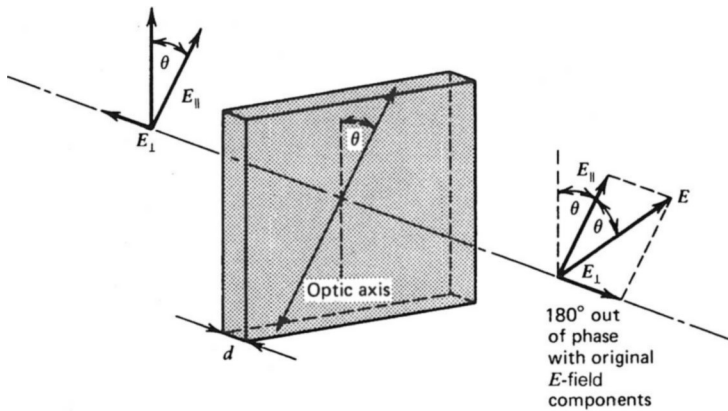


Figure 1.24. Half wave plate. The plate produces a 180° phase lag between the E_{\parallel} and E_{\perp} components of incident linearly polarized light. If the original polarization direction is at an angle θ to the optic axis, the transmitted polarization is rotated through 2θ from the original polarization direction.

If a crystal is cut that has twice the thickness of the quarter-wave plate, one has a *half-wave plate*. In this case, linearly polarized light at any angle θ with respect to the optic axis provides two perpendicular components which end up 180° out of phase upon passage through the crystal. This means that relative to one of the polarizations, the other polarization is 180° from its original direction. These components can be combined as shown in Fig. 1.24 to give a resultant whose direction has been rotated 2θ from the original polarization. Sometimes a half-wave plate is called a *polarization rotator*. It also changes the “handedness” of circular polarization from left to right or the reverse. This discussion of wave plates assumes that the crystal thickness d is correct only for the wavelength of the incident radiation. In practice, there is a range of wavelengths about the correct value for which these polarization modifiers work fairly well.

Although there may be no need to use polarized light as part of a particular design, the optical designer should always be aware that reflections from mirrors and windows may change the polarization content of the light as it passes through the system. Sometimes, inadvertently, a designer inserts a number of folding mirrors or tilted windows and beamsplitters in a system in such a manner that polarization selective surfaces reflect a good deal of the light out of the main path, starving the detectors in the system for light.

1.7. INTERFERENCE AND DIFFRACTION

Although the next few chapters treat light as rays propagating in straight lines, this picture does not fully describe the performance of an optical system. There are a number of additional concepts that are needed to explain certain

limitations of ray optics and to describe some of the techniques that allow us to analyze images and control the amplitude and direction of light. This section is a brief review of two important phenomena in physical optics, interference and diffraction. For a complete discussion of these and related subjects, the reader should consult one or more of the references (Refs. 1.3–1.4).

1.7.1. Wave Nature of Light

As was noted in the previous section on polarization, light is an electromagnetic wave. Since light from any source consists of waves of different wavelengths (even a laser!), it would seem that it would be difficult to analyze their resultant effect. But the effects of light made up of several colors can be analyzed by determining what happens for a monochromatic wave, then adding the irradiances of all the colors present. Thus, by analyzing these effects for monochromatic light, we are able to calculate what would happen in non-monochromatic cases.

A monochromatic wave propagating along the optic axis of an optical system, as shown graphically in Fig. 1.25, can be expressed mathematically as

$$E = A \cos\left(2\pi\nu t - \frac{2\pi z}{\lambda} + \phi\right), \quad (1.26)$$

where ν and λ are the frequency and wavelength of the light wave discussed at the beginning of this chapter and ϕ is an adjustable phase factor that gives the

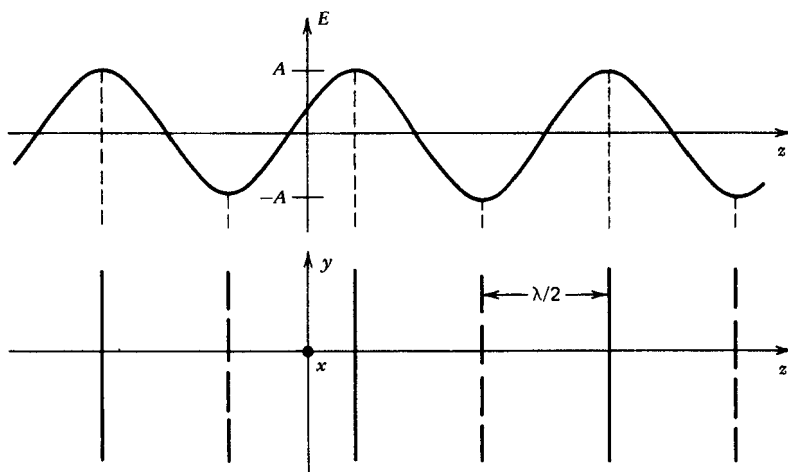


Figure 1.25. Monochromatic plane wave propagating along the z axis. For a plane wave, the electric field is constant in an x - y plane. The solid lines and dashed lines indicate maximum positive and negative field amplitudes.

correct amplitude at $z = 0$ and $t = 0$. Although it is necessary to compute the resultant wave fields to obtain qualitative expressions of light distributions, it is possible to gain an understanding of some effects through geometrical construction. For example, the wave expressed in Eq. 1.26 is a *plane wave* with sinusoidal variation in the z -direction. Since there is no variation in planes perpendicular to the propagation direction, the electric field is constant in any of these planes and equal to the value on the z axis. To represent this wave, we can draw the planes of maximum electric field strength, as shown in Fig. 1.25, where the solid lines represent planes in which the electric field vector is pointing in the positive y -direction and the dashed lines represent planes in which the field vector is pointing in the negative y -direction. The solid planes are separated by one wavelength, as are the dashed planes. The separation between the dashed and solid planes is $\lambda/2$.

Another useful form for the analysis of light waves is the *spherical wave*. A point source emits light as spherical wavefronts whose radii are equal to the distance from the source, as shown in Fig. 1.26(a). The expression for such a wave is

$$E = \frac{B}{r} \cos\left(2\pi\nu t - \frac{2\pi r}{\lambda} + \phi\right), \quad (1.27)$$

where r is the distance to the point source. Note that as the distance from the source increases, the radii become large and begin to approximate plane

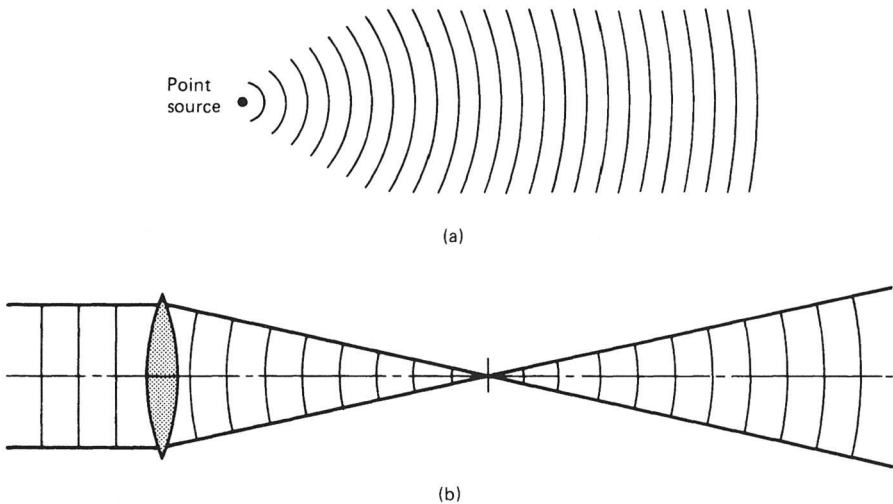


Figure 1.26. Spherical waves. (a) Generation of spherical waves from a point source. Note that a small section of the wave field far away from the point source approximates plane waves. (b) Generation of plane waves by focusing plane waves to a point. Diffraction prevents an actual point focus.

waves. Another way to create spherical waves is to focus a plane wave. The drawing in Fig. 1.26(b) shows the spherical waves collapsing to a point and then expanding. The waves never collapse to a true point because of diffraction. There are other possible forms of wave fields, but these two are all that is needed for our discussion of interference.

1.7.2. Interference

Some of the most useful effects in the field of optics from diffraction gratings to holography is due to *interference*. As we shall see, an interference pattern is often connected with some simple geometry. Once the geometry is discovered, the interference is easily understood and analyzed.

1.7.2.1. Young's Experiment. In Fig. 1.27 the geometry and wave pattern for one of the simplest interference experiments, *Young's experiment*, is shown. Two small illuminated pinholes, serving as point sources, produce overlapping spherical waves. The figure shows a cross-sectional view of the wavefronts in a

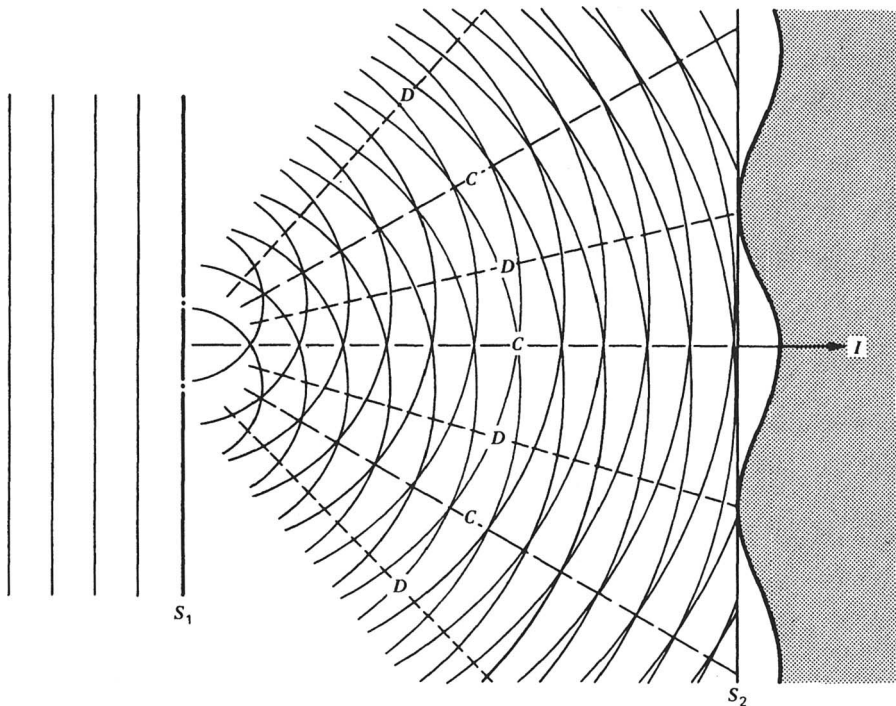


Figure 1.27. Young's experiment. Light diffracted through two pinholes in screen S_1 spreads out toward screen S_2 . Interference of the two spherical waves produces a variation in irradiance (interference fringes) on S_2 that is plotted to the right of the screen.

plane containing the pinholes. Notice that at points along a line equidistant from both pinholes, the waves from the two sources are always in phase (solid lines crossing solid lines and dashed lines crossing dashed lines). Thus, along the line marked *C* the electric fields always add to give a field that is twice that of a single field; the irradiance at a point along the line will be four times that due to a single pinhole. When electric fields add together to give a larger value it is referred to as *constructive interference*. There are other directions, such as those along the dotted lines marked *D*, in which the waves from the two sources are always 180° out of phase. That is, when one source has a maximum positive electric field, the other has the same negative value so the fields always cancel and no light is detected along these lines marked *D*, as long as both sources are present. This condition of canceling electric fields is called *destructive interference*. Between the two extremes of maximum constructive and destructive interference, the irradiance varies between four times the single pinhole irradiance and zero. It can be shown that the total energy falling on the surface of a screen placed in the interference pattern is neither more nor less than twice that of a single point source; it is just that interference causes the light distribution to be arranged differently!

1.7.2.2. The Michelson Interferometer. Another geometry that produces an interference pattern is shown in Fig. 1.28. This is a *Michelson interferometer*, which is constructed from a beamsplitter and two mirrors. The *beamsplitter* is a partially reflecting mirror that separates the light incident upon it into two beams of equal strength. After reflection off the mirrors, the two beams are recombined so that they both travel in the same direction when they reach the screen. If the two mirrors are the same distance ($L_1 = L_2$ in Fig. 1.28) from the

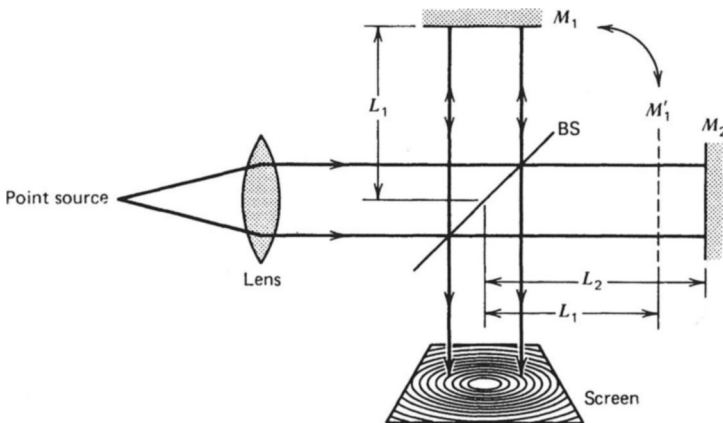


Figure 1.28. Michelson interferometer. By reflecting the mirror M_1 about the plane of the beamsplitter BS to location M_1' , one can see that a ray reflecting off mirror M_2 travels an additional distance $2(L_2 - L_1)$ over a ray reflecting off M_1 .

beamsplitter, then the two beams are always in phase once they are recombined, just as is the case along the line of constructive interference in Young's experiment. Now the condition of constructive and destructive interference depends on the difference between the paths traveled by the two beams. Since each beam must travel the distance from the beamsplitter to its respective mirror and back, the distance traveled by the beam is $2L$. If the path-length difference, $2L_1 - 2L_2$, is equal to an integral number of wavelengths, $m\lambda$, where m is an integer, then the two waves are in phase and the interference at the screen will be constructive.

$$L_1 - L_2 = \frac{m\lambda}{2} \quad (m = \dots, -1, 0, 1, 2, \dots). \quad (1.28)$$

If the path-length difference is an integral number of wavelengths plus a half wavelength, the interference on the screen will be destructive. This can be expressed as (Problem 1.22)

$$L_1 - L_2 = \frac{m\lambda}{4} \quad (m = \text{odd integers}). \quad (1.29)$$

In most cases the wavefronts of the two beams when they are recombined are not planar, but are spherical wavefronts with long radii of curvature. The interference pattern for two wavefronts of different curvature is a series of bright and dark rings. However, the above discussion still holds for any point on the screen. Usually, however, the center of the pattern is the point used for calculations.

This same path-length difference argument will be used in our discussions of highly reflective mirrors and anti-reflection coatings. In optical system design, interferometers, such as the Michelson interferometer, can be used to measure very small distances. For example, a movement of one of the mirrors by only one quarter wavelength (corresponding to a path-length change of one half wavelength) changes the detected irradiance at the screen from a maximum to a minimum. Thus, devices containing interferometers can be used to measure movements of a fraction of a wavelength. Applications of these principles are discussed in Sections 4.2.2 and 9.1.3.

1.7.3. Diffraction

If interference provides some of the "magic" of optics, then it is diffraction that provides the baseline that tells us how far an optical technique can be extended. While certain diffraction effects prove useful, ultimately all optical performance is limited by diffraction, if there is sufficient signal, and by electrical or optical "noise," if there is not.

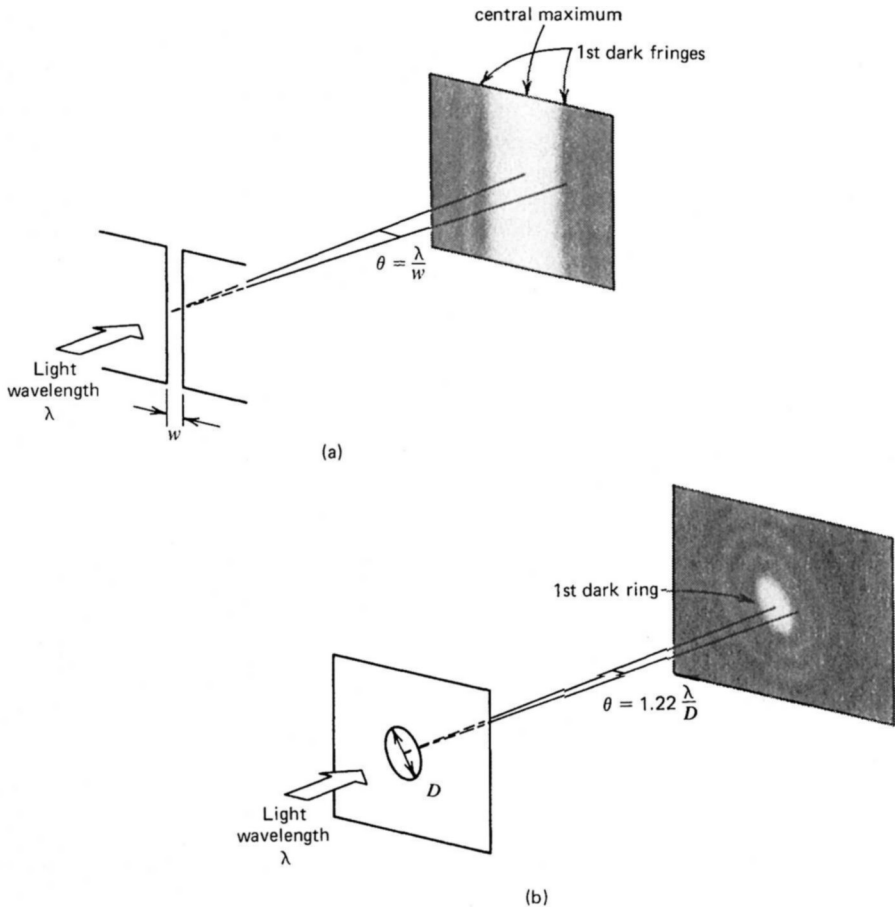


Figure 1.29. Diffraction of light by apertures. (a) Single slit. (b) Circular aperture.

Diffraction of light arises from the effects of apertures and interface boundaries on the propagation of light. In its simplest form, edges of lenses, apertures, and other optical components cause the light passing through the optical system to be directed out of the paths indicated by ray optics. When a plane wave illuminates a slit, the resulting wave pattern that passes the slit can be constructed by representing the wavefront in the slit as a collection of point sources all emitting in phase. *Diffraction* can be considered as interference between this collection of point sources. As in the Young's experiment, the resulting pattern on a screen contains maxima and minima, as shown in Fig. 1.29(a). A measure of the amount of diffraction is the spacing between the strong central maximum and the first dark fringe in the diffraction pattern. At distances far from the slit, the pattern does not change in shape, but only in

size. The fringe separation is expressed in terms of the sine of the angular separation between the central maximum and first dark fringe,

$$\sin \theta = \frac{\lambda}{w}, \quad (1.30)$$

where w is the slit width and λ is the wavelength of the light illuminating the slit. Note that as the width of the slit becomes smaller, the diffraction angle becomes larger. If the slit width is not too small, the sine can be replaced by its argument,

$$\theta = \frac{\lambda}{w}. \quad (1.31)$$

In the case of circular apertures, the diffraction pattern is also circular, as indicated in Fig. 1.29(b), and the angular separation between the central maximum and the first dark ring is given by

$$\sin \theta = 1.22 \frac{\lambda}{D} \quad \text{or} \quad \theta \doteq 1.22 \frac{\lambda}{D} \quad \text{for large } D, \quad (1.32)$$

where D is the diameter of the aperture. As in the case of the slit, for small values of λ/D , the sine can be replaced by its angle.

Although it is possible to examine these diffraction patterns on the screen far from the aperture, the same pattern can be observed at the rear focal plane of a lens. At that point the dimension of the diffraction pattern is given by $f\theta$, where we assume small angles, and θ is the diffraction angle given in Eqs. 1.31 and 1.32. In many cases the lens serves not only as the imaging element, but is also as a diffracting aperture. Thus, a point source imaged by a lens will not be imaged to a point source but to a diffraction pattern of that point source, with a first dark ring having a radius of $1.22\lambda f/D$, where the aperture is the lens itself and D is its diameter.

One good approximation to a point source is a bright star. A pair of stars close to one another can give a measure of the diffraction limits of a system. If the stars have the same brightness, the resolution of the system can be determined by the smallest angular separation between such sources that would still allow them to be resolved. This is provided that the aberrations of the optical system are sufficiently small and diffraction is the only limitation to resolving the images of these two point sources. Although it is somewhat artificial, a limit of resolution that has been used in many instances is that two point sources are just resolvable if the maximum of the diffraction pattern of one point source falls on the first dark ring of the pattern of the second point source, as illustrated in Fig. 1.30, then

$$\theta_R = 1.22 \frac{\lambda}{D}. \quad (1.33)$$

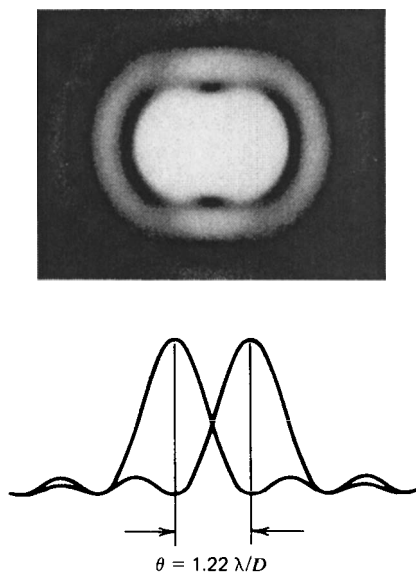


Figure 1.30. Rayleigh criterion. The plot of the intensity along a line between the centers of the two diffraction patterns is shown below a photo of two sources just resolved as specified by the Rayleigh criterion. (Photo by Vincent Mallette.)

This condition for resolution is called the *Rayleigh criterion*. It is used in other fields of optical design, such as specifying the resolution of a laser beam scanner (Section 8.2.1).

1.7.4. The Diffraction Grating

It is a somewhat confusing use of the term to call the item under discussion a *diffraction grating*. Although diffraction does indeed create the spreading of light from the slits, it is the combined interference of multiple beams that permits a diffraction grating to deflect and separate the light. In Fig. 1.31 a series of narrow slits, each separated from its neighboring slits by distance d , are illuminated by a plane wave. Each slit is then a point (actually a line) source in phase with all other slits. At some angle θ_d to the grating normal, the path-length difference between neighboring slits will be (see inset to Fig. 1.31)

$$\Delta x = d \sin(\theta_d).$$

Constructive interference will occur at that angle if the path-length difference Δx is equal to an integral number of wavelengths:

$$m\lambda = d \sin(\theta_d) \quad (m = \text{an integer}). \quad (1.34)$$

This equation, called the *grating equation*, holds for any wavelength. Since any grating has a constant slit separation d , light of different wavelengths will be diffracted at different angles. This is why a diffraction grating can be used in

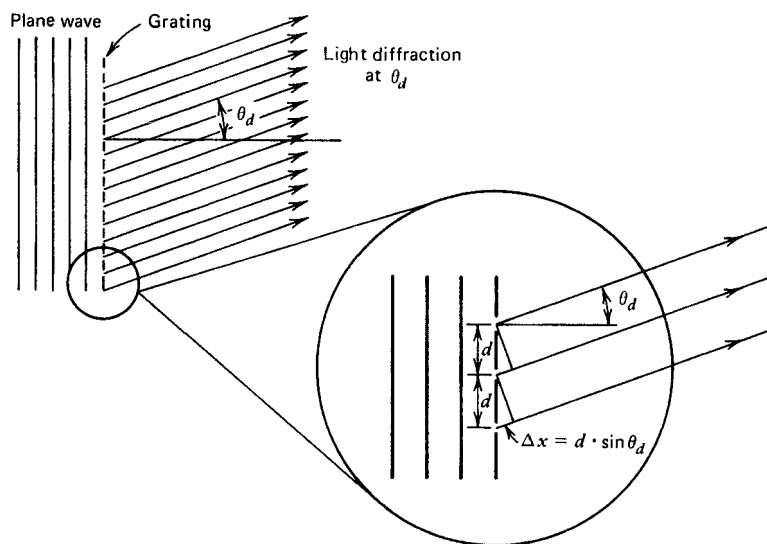


Figure 1.31. Diffraction of light by a diffraction grating.

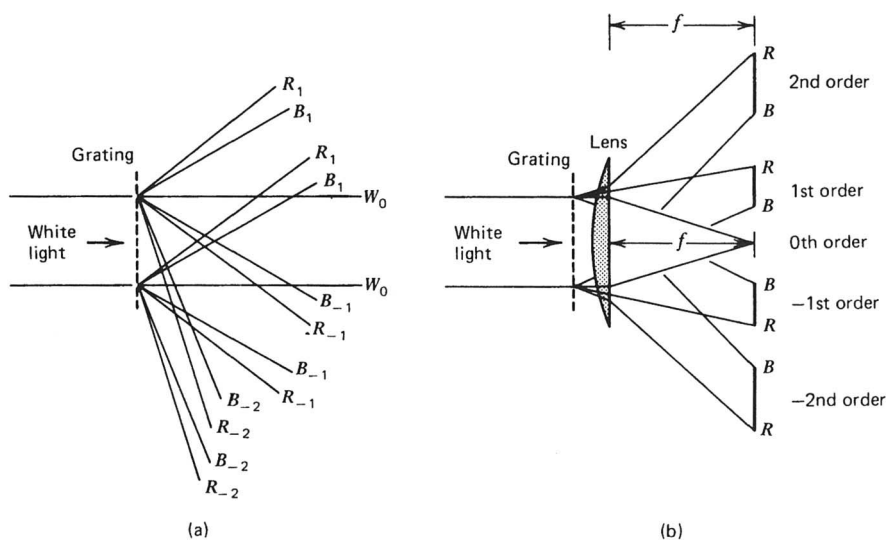


Figure 1.32. Orders of diffraction from a grating illuminated by white light. (a) Rays denoting the upper and lower bounds of diffracted beams for the red (R) and blue (B) ends of the spectrum; (b) spectra produced by focusing each collimated beam of wavelengths to a point in the focal plane.

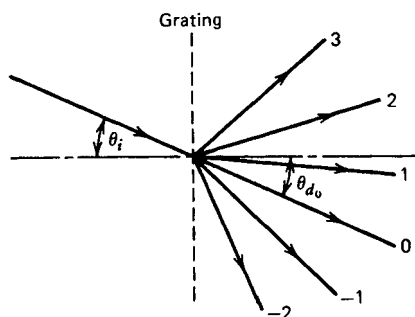


Figure 1.33. Diffraction of monochromatic light at non-normal incidence.

place of a prism to separate light into its colors. Because a number of integers can satisfy the grating equation, there are a number of angles into which monochromatic light will be diffracted. Therefore, when a grating is illuminated with white light, the light will be dispersed into a number of spectra corresponding to the integers $m = \dots, -2, -1, 0, 1, 2, \dots$, as illustrated in Fig. 1.32(a). By inserting a lens after the grating, the spectra can be displayed on a screen one focal length from the lens, Fig. 1.32(b). These are called the *orders* of the grating and are labeled by the value of m .

In those cases when the light is not incident normal to the grating surface, an additional term must be added to account for the phase difference between slits [Problem 1.26]. This gives rise to the general grating equation:

$$m\lambda = d(\sin \theta_d - \sin \theta_i) \quad (m = \text{an integer}), \quad (1.35)$$

where θ_i is the angle between the grating normal and the direction of incident light as shown in Fig. 1.33. The sign connection for the incident and diffracted angles is given in 6 (a) of Table 2.1.

The concept of the diffraction grating is very useful in optics. It will be used to explain acousto-optic modulation in Section 8.4, and grating spectroscopy in Section 9.3. As noted at the beginning of this section, this discussion is intended as a review and brief tutorial. Further understanding can be achieved by readings in the references.

PROBLEMS

- 1.1. If the frequency of light is 6.17×10^{14} hertz (cycles/second), what is its wavelength and velocity in crown glass ($n = 1.52929$) and in a light flint glass ($n = 1.58957$)?
- 1.2. Determine the V -number for both glasses listed in Table 1.1 from the indices given.
- 1.3. Light is incident upon a plane parallel slab of glass, $n = 1.55$ at an angle of 30° from the normal. Determine the direction of the beam (a) in

the glass and (b) on the other side of the plate. What would be the final direction of a beam relative to an original beam incident at 30° after it passed through three parallel 1-cm-thick slabs of glass $n_1 = 1.55$, $n_2 = 1.65$, and $n_3 = 1.75$?

- 1.4. (a) Derive an expression for the amount of sideways displacement d of a beam at an angle θ with respect to the normal by a plane parallel slab of glass in air whose thickness is t and refractive index is n . (b) If the glass is 15 mm thick and its refractive index is 1.517, how far is the beam displaced from its original path if the slab is tilted by 42° ?
- 1.5. In Example 1.A we calculated the amount of dispersion of a beam passing through the first surface of a prism. If the second surface makes an angle of 60° with the first surface (such that total internal reflection does not occur), what are the angles of transmission after the second interface relative to the incident beam for the D , C , and g lines?
- 1.6. If a prism is to be used as a reflector as pictured in Fig. 1.5, what is the minimum refractive index that will provide total internal reflection?
- 1.7. Sketch the rays and find the image of an object located at the center of curvature of a spherical mirror.
- 1.8. Sketch the rays and locate the image of an object a distance d in front of a plane mirror. (Because there is no curvature to the surface, the optic axis is any normal to the mirror.)
- 1.9. Sketch the rays and find the image of an object located halfway between the focus and vertex of a concave spherical mirror.
- 1.10. Sketch the rays and find the image of an object located halfway between the focal point and vertex of a positive lens. Estimate the magnification.
- 1.11. Sketch the rays and find the image location for two identical convex lenses separated by four times their focal lengths for an object two focal lengths from the first lens. What is the magnification? What would happen if the lenses were separated by twice their focal lengths?
- 1.12. Use a calculator or set of trigonometric tables with radian measure to determine at what angle (in radians) the approximations in (Eqs. 1.9–1.11) agree to within 1% of the actual trigonometric functions. Do this for the sine and tangent functions. Express the angles in degrees.
- 1.13. The form of the thin lens equation given as Eq. 1.16 is called the Gaussian form. If the object distance x_0 is measured from the front focal point and the image distance x_1 measured from the back focal point, show that the thin lens equation can be rewritten in Newtonian form; that is, as

$$x_1 x_0 = -f^2$$

- 1.14. What radii of curvature must be ground onto a piece of glass of refractive index 1.63 to produce an equiconvex lens of 105 mm focal length?
- 1.15. Plot the image distance versus object distance for (a) a lens with +100 mm focal length and (b) -100 mm focal length. Do some preliminary calculations so that you can choose the range of values of object distances to display the results compactly.
- 1.16. Calculate the location, magnifications, and orientations of the intermediate and final images for two lenses ($f_1 = 16$ mm and $f_2 = 32$ mm) separated by 115 mm if the object is 20 mm in front of the first lens.
- 1.17. A camera with a 50-mm lens can be used to focus on objects from 3 m to infinity onto a screen. How much must the lens be moved to focus on the extremes of this range?
- 1.18. The eye of a young person can focus from the near point (the distance of most distinct vision equals 25 cm from the eye) to infinity. Determine the range of focal lengths the eye can attain, if the lens is located 17 mm from the retina (the photo sensitive surface of the eye).
- 1.19. Show that the angular magnification of a simple magnifier is increased by 1 when the magnifier is refocused to give an image at the near point.
- 1.20. Show that the angular magnification of a compound microscope is as given by either Eq. 1.21 or 1.22.
- 1.21. If the focal lengths of the objective and eyepiece of a compound microscope are 8 mm and 10 mm, respectively, where must the object and the eyepiece be located, relative to the objective lens, if the overall power of angular magnification is -200 when the eye is focused for infinity?
- 1.22. If a telescope has an objective with a focal length of 10 cm, at what distance from the observer might an object be considered to be at "infinity"? (There is no single numerical answer for the question. You must make some assumptions and show what answer results.)
- 1.23. Derive Eq. 1.24 using Snell's Law and the orientation of the rays at the interface at Brewster's angle.
- 1.24. Natural light is sent through a perfect polarizer and then through a second perfect polarizer whose axis is at 30° to the first polarizer. What percent of the original light is transmitted by the pair of polarizers? What would be the value if these were HN-36 polarizers?
- 1.25. The refractive indices of calcite at the sodium D line are $n_o = 1.658$ and $n_e = 1.486$. What are the maximum and minimum velocities of light at

this wavelength through this crystal? How far will the light traveling perpendicular to the optic axis have to travel in the medium such that the maxima in the two polarizations are out of phase by 90° (i.e., separated by $\lambda/4$)?

- 1.26. Show that the equation describing the diffraction of light of wavelength λ by a grating whose slit separations are d is given by Eq. 1.35, assuming an angle of incidence, θ_i .

APPENDIX A. LIST OF SELECT GLASS TYPES [SEE REF. 5.2]

Glass Number	Designation ^a	Glass Number	Designation
487704	FK-5	618551	SSK-4
501564	K-10	622532	SSK-2
511605	K-7	623581	SK-15
517642	BK-7	624470	BaF-8
518651	PK-2	626390	BaSF-1
522595	K-5	636353	F-6
523515	KF-9	648338	SF-12
532488	LLF-6	652449	BaF-51
540597	BaK-2	652585	LaKN-7
547536	BaLF-5	664358	BaSF-2
548458	LLF-1	670471	BaFN-10
552635	PSK-3	673322	SF-5
557586	BaK-5	683445	BaF-50
564608	SK-11	689495	LaF-23
569561	BaK-4	697562	LaK-24
573576	BaK-1	717295	SF-1
580537	BaLF-4	720504	LaK-10
583465	BaF-3	723381	BaSF-51
589613	SK-5	740283	SF-3
596392	F-8	744448	LaFN-2
603380	F-5	762266	SF-14
606439	BaF-4	783371	LaF-22
609589	SK-3	785258	SF-11
613443	KZFS-1	808408	LaSFN-3
617366	F-4		

^a The designation is the catalog type given in the Schott glass catalog. This designation varies for each manufacturer, but the listed designations are the most common.

REFERENCES

- 1.1. D. Halliday and R. Resnick, *Physics*, 3rd ed. (Wiley, New York, 1978). Widely used physics textbook for undergraduate college physics. Eight chapters on optics.
- 1.2. F. W. Sears, M. W. Zemansky, and H. D. Young, *University Physics*, 6th ed. (Addison-Wesley, Reading, Mass., 1982). Another widely used textbook in college physics.
- 1.3. E. Hecht and A. Zajac, *Optics* (Addison-Wesley, Reading, Mass., 1974). This textbook covers much of the physical optics theory needed to explain many laser concepts. Includes short sections on lasers, holography, and nonlinear optics in the last chapter. Excellent illustrations.
- 1.4. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed. (McGraw-Hill, New York, 1976). The classic undergraduate optics text with a few sections updated and chapters on lasers and holography added.
- 1.5. J. R. Meyer-Arendt, *Introduction to Classical and Modern Optics* (Prentice-Hall, Englewood Cliffs, N.J., 1972). While not as detailed in its treatment of most topics, this text has considerable breadth.

2

Thin Lens Ray Tracing

(HOW TO START WHEN YOU ARE NOT TOO SURE WHERE TO START)

The thin lens equation and the other simple optical concepts that are taught in elementary physics courses and that were reviewed in Chapter 1 serve as a beginning for more elaborate calculations. In this chapter we discuss the starting point for solving an optical design problem and explore a method for producing rapid preliminary calculations on the system.

2.1. GETTING STARTED

Let us suppose that you were given the following problem: A large amount of computer information is to be put on microfiche (a transparent film containing demagnified reproductions of a large number of pages or lines of information). An example is shown in Fig. 2.1. A standard microfiche measures $101.6 \text{ mm} \times 152.4 \text{ mm}$ (4 in. \times 6 in.) and holds about 84 frames, or pages, of information. You are asked to design a camera that will record the output of a computer terminal cathode ray tube (CRT) directly onto film to create a microfiche master. To increase the rapidity of searching, and to provide for some automation of the microfiche viewers, a computer generated bar code is to be included at the edge of each frame of information. Due to the distortion in most terminal screens, the code, which is to be optically readable, cannot be put directly onto the screen. A system is needed to accomplish this task. Because the camera will fit over the entire screen, some provision must be made to allow inspection of the screen and the code.

Where do you start? If you are sitting at a desk, feet on the floor, and feeling industrious, you might write "Objective" at the top of a clean sheet of paper, followed by a statement to the effect that you want to reduce the image from a computer terminal to that of a microfiche frame and to provide an optical access for a frame code and for visual inspection of the frame contents. This step is to be encouraged. However, many times the initial design takes place at a lunch table, in a lab, or at a cocktail party. In such cases, the objective is unwritten or even unspoken. Still, a few seconds of thought about what it is that you want to do can save you a great deal of time later. It is always good to understand the

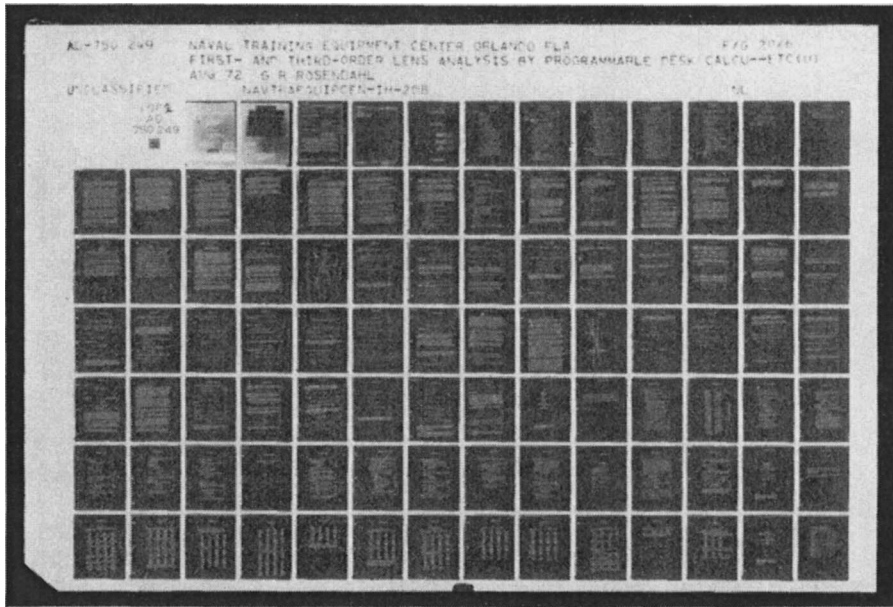
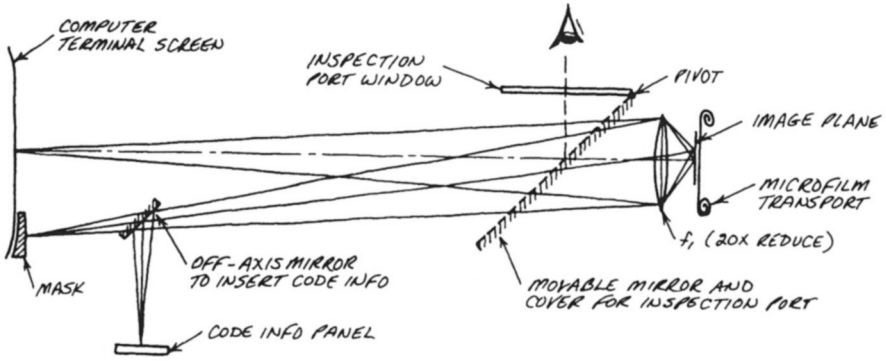


Figure 2.1. A microfiche (actual size).

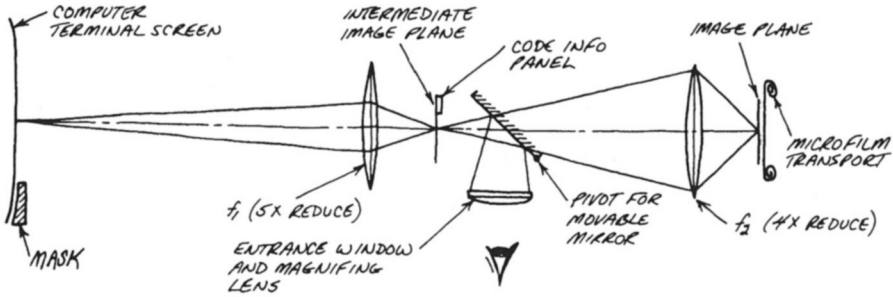
nature of the problem before you try to solve it. The next thing to do is to analyze the system in terms of its major operations. From this you see that there is really one system with two subsystems attached. The main system, a reduction camera, should be outlined first, with the subsystems—coding and inspection—providing constraints on the design.

Then what? Probably the next step is the most enjoyable. Using your sheet of paper (or your lunch bag, lab notebook, or cocktail napkin) with the objective sitting at the top you can begin to ray sketch any number of designs. Here is where experience has an advantage. Choices can be quickly made and kept or rejected. A general direction becomes evident as the possibilities from these sketches are exhausted. Figure 2.2 is a preliminary ray sketch for two systems. Since we must produce a reduced image on film, we want a demagnified real image. One method is to demagnify in one step, inserting the frame code on a separate panel next to it. The other method reduces the image in two stages. The intermediate stage could be used to insert the frame code and to inspect the frame and code. The final image, in either case, is recorded by the microfiche master which is exposed and then stepped to a new position for the next frame. We will consider the single-lens system first, then look at the two-stage system.

What is the focal length of the imaging lens in the first system? This is where the real world and the thin lens equation come together. You need to know two things: the size of the data area on the terminal screen and the size of the



(a) SINGLE LENS SYSTEM



(b) TWO-LENS SYSTEM

Figure 2.2. Ray sketches of two possible configurations for a computer terminal-microfiche recording system.

microfiche frame. From an examination of a terminal we find the data area to have dimensions of, 20 cm long and 14 cm high. The size of a standard microfiche frame, which usually reproduces an 8 × 10 in. area from a type-written original, is 10 mm long and 12.5 mm high. Because the letters on a computer display are larger than those on an ordinary typewriter, we can reduce the image somewhat more than the typescript and still have room at the bottom of the frame for a computer code. At this point one can do a little juggling of dimensions to provide the best possible display in the tightest microfiche format [Problem 2.1]. From the horizontal dimensions of the object and image, we choose an overall reduction of 20 times.

We can derive some simple relations between the focal length and the object and image distances once the magnification has been set. For a single lens

$$M = \frac{t_1}{t_0}. \quad (1.14)$$

Here $M = -1/20$ since this is a demagnification. Note the magnification, as well as t_0 and t_1 , is a signed quantity. Therefore,

$$t_0 = \frac{t_1}{M}. \quad (2.1)$$

Inserting this into the thin lens formula we obtain

$$\frac{1}{f} = \frac{1}{t_1} - \frac{M}{t_1} = -\frac{(M-1)}{t_1}$$

or

$$t_1 = -(M-1)f \quad (2.2)$$

and

$$t_0 = -\frac{(M-1)f}{M}. \quad (2.3)$$

The length of the system L is given by

$$L = t_1 - t_0 = \frac{-(M-1)^2}{M} f. \quad (2.4)$$

Thus, once the reduction is set, the length is directly proportional to the focal length of the lens chosen and if M is small, it is approximately equal to $-f/M$ (here, approximately $20f$).

One needs to know what range of focal lengths are available for use in such a system. Some can be found in lists of stock lenses available from manufacturers and suppliers. It should be noted here that such lists can serve only as a guide. Whether one of these lenses would actually perform well depends on the demands in image quality of your particular specifications. As we will discuss later, the choice and design of lenses are dictated by the optical errors (aberrations) that must be eliminated to provide an acceptable image. Camera lenses are usually designed for an object at infinity. Thus an actual camera lens might not do the job. Still, such a lens list (and price list) can give you some idea of what ranges of focal lengths are reasonable for a preliminary design. You might know that the focal length of an ordinary 35-mm camera lens is 50 mm. If we used a similar 50-mm lens in our system, the total length of the system, according to Eq. 2.4, would be 1000 mm—fairly long. Upon consulting a catalog you might find that wide-angle lenses are available with 28- and 35-mm focal lengths. These require system lengths of 617 mm and 772 mm. Let us choose a 35-mm lens and use it as a preliminary choice. The calculations are given in Example 2.A. Now we have an optical system with spacings and a lens

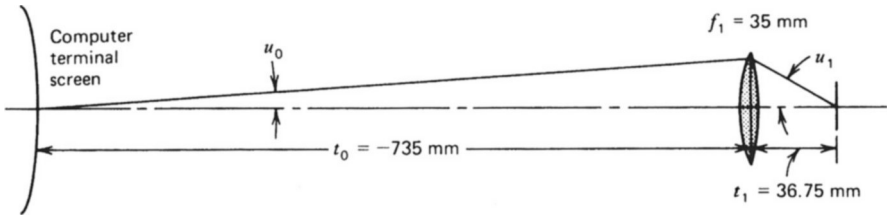


Figure 2.3. Single lens system with a 35-mm focal length lens.

as shown in Fig. 2.3. (The distances t_0 and t_1 in the figure should really be drawn with single-headed arrows to indicate that they are directed quantities. However, in order to show, with any degree of clarity, the various distances in the figures in this text, they are drawn with double-headed arrows.) Our thin lens equation has taken us about as far as it can. It will be necessary to revise it to be able to continue.

EXAMPLE 2.A. Based on a choice of a 35-mm focal length lens and a 20-fold reduction in the object size, find the image and object distances and the total length of the system.

Solution: Using Eq. 2.1

$$M = -\frac{1}{20} = \frac{t_1}{t_0}.$$

Using Eq. 2.3

$$\begin{aligned} t_0 &= -\frac{(-1/20 - 1)}{-1/20} 35 \text{ mm} \\ &= -21 \times 35 \text{ mm} = -735 \text{ mm} \end{aligned}$$

$$t_1 = Mt_0 = -\frac{1}{20} \times -735 \text{ mm} = 36.75 \text{ mm}.$$

Using Eq. 2.4

$$\begin{aligned} L &= t_1 - t_0 = 36.75 - (-735) \text{ mm} \\ &= 771.75 \text{ mm}. \end{aligned}$$

2.2. TRACING RAYS THROUGH A SERIES OF LENSES

The preceding section illustrated how an optical system can be laid out using just the thin lens equation and some reasonable estimates as to focal lengths, overall length of the optical system, and magnification. This approach enables the designer to begin, but it will not provide a means of evaluating the performance of the system. For this, the most useful tool an optical designer has is ray tracing. By calculating the paths of one or more rays through a proposed optical system, the performance and imperfections of an optical system can be assessed before the system is built. Thus, poor or imperfect systems can be discarded on paper rather than waiting until many costly steps later.

Before proceeding any further, however, a set of sign conventions should be set down for use in all ray-trace calculations. Table 2.1 is a list of the sign conventions used in this text. They are similar to those used in most modern texts, articles, and ray-trace programs. Unfortunately, there is *no* universal sign convention accepted by the optics community. Please note that the slope angle of a ray, the angle that the ray makes with the optic axis, is positive if a *clockwise*

Table 2.1. Sign Conventions for Equations and Calculations

1. Until a reflecting surface intervenes, light travels from left to right and all refractive indices are positive. (For the exception of reflecting surfaces, see #7 below.)
2. Focal lengths of converging elements are positive; focal lengths of diverging elements are negative.
3. Ray heights above the axis are positive; below the axis, negative.
4. The distance to the right of the vertex of a component (or surface) is positive; to the left, negative.
5. The radius of curvature of a surface is positive if its center lies to the right of its vertex; if to the left of its vertex, negative.
6. (a) The angle of a ray with respect to a normal to a surface is *positive* if a *clockwise* rotation of the ray brings it parallel to the normal; it is negative if it takes a counterclockwise rotation to do the same thing. (b) The same is true for the slope angle, the angle between a ray and the optic axis. A clockwise rotation of a ray through an angle to bring it parallel to the optic axis is positive; a counterclockwise rotation of a ray through the slope angle is negative.
7. Upon reflection, the sign of all refractive indices is changed until the next reflective surface. All distances measured from the vertex of a component follow the same sign convention as indicated in #4 above. This means that after one reflection the surface separations and distance along the ray are negative. All other conventions (angles, heights, and focal lengths) remain the same.

rotation of the ray aligns it parallel to the optic axis. One should be certain when using someone else's calculator and computer programs to determine the convention being used. This is easily done by tracing a sloping ray a short distance to a surface. From the new angles and height of the ray after a reflection or refraction, one can determine the convention.

First, we will look at what happens to a ray between an object and a lens or between two lenses. As was stated in Section 1.2 as the Law of Transmission, light in a medium of constant refractive index travels in straight lines. If we draw a right-handed coordinate system, as shown in Fig. 2.4, with the z axis coincident with optic axis of the system, then $\tan u_k = (y_{k+1} - y_k)/t_k$. In the limit of small angles we can replace $\tan u_k$ by u_k and solve for the ray height at the $k + 1$ surface as

$$y_{k+1} = y_k + u_k t_k, \quad (2.5)$$

where t_k is the distance between the k and the $k + 1$ surface and u_k is the ray angle defined earlier. This is called the *transfer equation* since it transfers the ray from one surface to the next.

When the ray meets a thin lens it will be redirected onto a path with a new angle. Let the ray with an angle u_k with respect to the axis meet a lens at a height y_{k+1} , as shown in Fig. 2.5. The ray will be refracted to the new angle u_{k+1} . If we extend the rays forward and backward to where they cross the axis at z_k and z_{k+1} , we can apply the thin lens equation (Eq. 1.16)

$$\frac{1}{z_{k+1}} = \frac{1}{f_{k+1}} + \frac{1}{z_k}. \quad (2.6)$$

The slope angles of the rays are

$$\tan u_k = \frac{-y_{k+1}}{z_k} \quad \tan u_{k+1} = \frac{-y_{k+1}}{z_{k+1}}.$$

The minus sign in the first equation is due to the fact that the $\tan u_k$ and y_{k+1} are positive, but z_k is negative. Thus a negative sign must be included for both sides

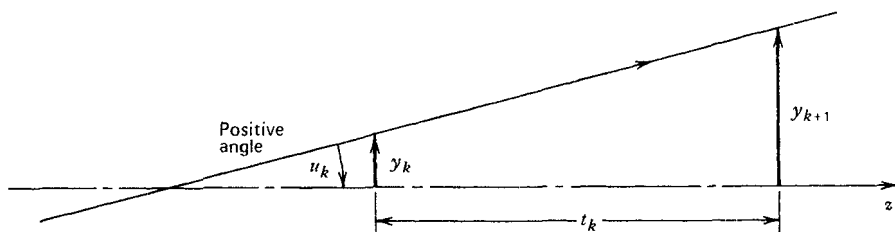


Figure 2.4. Ray trace of a ray between two surfaces separated by a distance t_k .

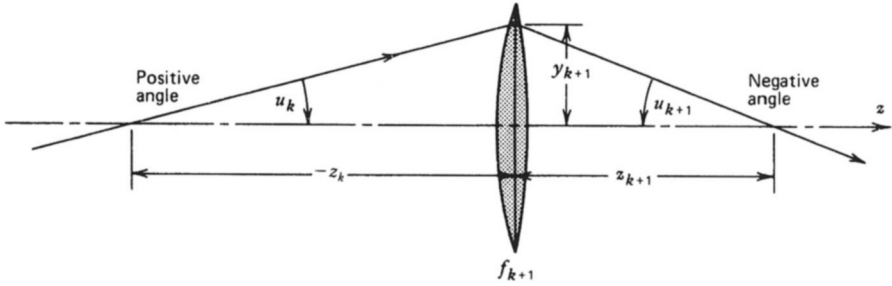


Figure 2.5. Ray trace of a ray through a thin lens of the focal length f_{k+1} .

of the equation to agree. In the second equation, the same argument is true, but in this case it is the angle u_{k+1} that is negative. In the limit of small angles (the same approximation used for the thin lens equation), $\tan u_k = u_k$ so that

$$z_k = \frac{-y_{k+1}}{u_k} \quad z_{k+1} = \frac{-y_{k+1}}{u_{k+1}}.$$

Substituting into Eq. 2.6,

$$-\frac{u_{k+1}}{y_{k+1}} = \frac{1}{f_{k+1}} - \frac{u_k}{y_{k+1}},$$

and multiplying by $-y_{k+1}$, we have an equation for the new angle,

$$u_{k+1} = u_k - \frac{y_{k+1}}{f_{k+1}}, \quad (2.7)$$

in terms of the ray height at the lens y_{k+1} and the incident ray angle u_k .

It is sometimes convenient to label a lens not by its focal length, but by its reciprocal, called the *power* of the lens. A lens with considerable refractive power is one that bends parallel light to a focus at a short focal distance, whereas a lens with a long focal distance (a low-power lens) hardly affects the light at all. If the focal length is expressed in meters, then the power is in units of reciprocal meters of *diopters*. For example, a 50-mm (or 0.05-m) camera lens has a power of 20 diopters; a flat piece of glass has an infinite focal length and a power of zero diopters. The typical powers of other lenses are listed in Table 2.2. If we represent the power of the k th element as $\phi_k = 1/f_k$, then Eq. 2.7 may be written as

$$u_{k+1} = u_k - y_{k+1}\phi_{k+1}. \quad (2.8)$$

Since the equation determines the new slope angle of a ray after it encounters an element of optical power ϕ , it may be called the *slope angle equation*.

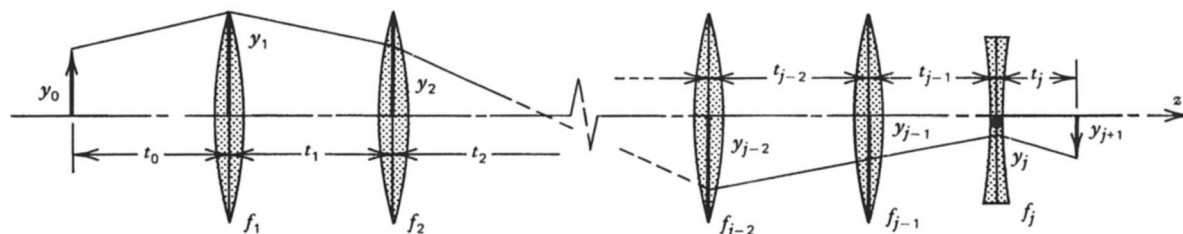


Figure 2.6. General notation for a ray trace through j elements.

Table 2.2. Powers of Lenses		
Type	Approximate f (mm)	Approximate ϕ (diopters or m^{-1})
Reading glass	150	7
Camera lens	50	20
Microscope } low power	25	40
Objectives } high power	2	500

Many times it is useful to locate a reference surface somewhere in the optical system. Since there is no beam deviating element there, we assign the surface zero power ($\phi = 0$). This is easier than assigning an infinite focal length at that surface. Please note the ϕ_k must have the inverse dimensions to those of y_k if their product is to yield a dimensionless angle; that is, if the ray height is given in millimeters, then the ϕ could not be entered in diopters, but would have to be converted to mm^{-1} .

We can set up a consistent labeling scheme that will serve us for all calculations to come. Starting with the object as the zeroth surface, all surfaces (lenses, images, or surfaces we need to examine, such as the stops, discussed later) are numbered sequentially through the optical system to the final image surface. All variables and constants in the space between surfaces are labeled with the number of the surface to the left. An illustration of the general numbering scheme is given in Fig. 2.6. A specific example, using the two lens designs we began in Section 2.1, is found in Fig. 2.7.

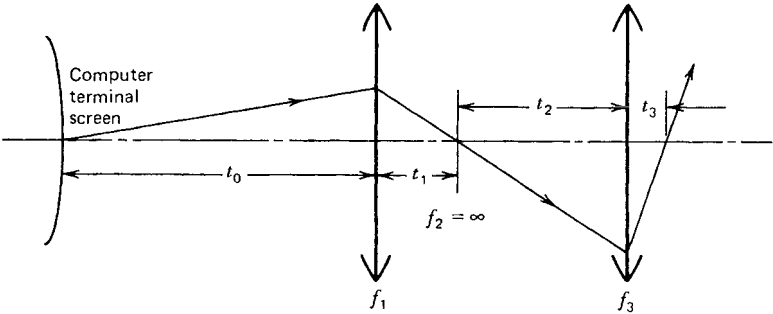


Figure 2.7. Notation for a two-lens system including the intermediate image.

2.3. THE y - u TRACE

A technique that is of great use for evaluating optical systems consisting of thin lenses is the y - u trace. It is a consistent, easy-to-use tabular arrangement of the initial specifications and the subsequent ray-trace calculations. Using the two equations

$$y_{k+1} = y_k + u_k t_k \quad (2.5)$$

$$u_{k+1} = u_k - y_{k+1} \phi_{k+1}, \quad (2.8)$$

we can trace any ray through the system given a starting ray coordinate (y_0, u_0) and the description of the system in distances t_k and powers ϕ_k or focal lengths f_k . By choosing only two rays we can learn a great deal more about an optical system.

Let us apply this y - u trace method to our computer terminal camera. We set up our system design as shown in Fig. 2.3 using the labeling method described in Section 2.2. From our choice of a 35-mm lens and Eqs. 2.1 and 2.2 we find that t_0 is 735 mm and t_1 is 36.75 mm, as shown in Example 2.A. As a first step we will arrange these values for the system in an array that can be conveniently used to calculate the path of different rays through this system. Starting with the image surface we list the ϕ_k (or f_k) and t_k in alternating columns until the image surface is reached, as shown in Table 2.3. The first ray to be traced is an *axial ray*. An axial ray starts at the point on the optic axis in the object plane and at an angle to the axis. How large an angle? As will be shown later, it does not really matter, but as a rule an angle of $u_0 = 0.1$ radian is reasonable and realistic for many ray traces. This set up of the initial ray angle and object height is referred to as

Table 2.3. Ray Trace of Single-Lens System^a

	k	0	1	2
System values	$\phi_k = 1/f_k$	0	1/35	0
	t_k	735	36.75	
Axial ray	y_k	0	73.5	0
	u_k	+0.1 ^b	-2.0	
Chief ray	\bar{y}_k	-73.5	0	3.675
	\bar{u}_k	0.1 ^b	0.1 ^b	

^a Calculated values are shaded.

^b Value set arbitrarily.

opening the ray trace. (For objects at infinity ($t_0 = \infty$), a ray parallel to the optic axis ($u_0 = 0$) is an axial ray. For such a *parallel ray* y_0 and y_1 are arbitrary; in this text 10 units is the standard value.) In a second array beneath the first, the axial ray is started and its progress recorded as it is traced through the system. All calculations made below are listed in Table 2.3 and shaded.

Given that $u_0 = +0.1$, $y_0 = 0$, and $t_0 = 735$, we use Eq. 2.5 to get

$$y_1 = 0 + (735)(0.1) = 73.5 \text{ mm.}$$

Putting this value into Eq. 2.8 with $\phi_1 = 1/35$ we obtain

$$u_1 = 0.1 - \left(\frac{73.5}{35} \right) = -2.0.$$

Notice that the sign of the slope angle is now negative, indicating that the lens has caused the light to bend enough so that it crosses the axis beyond the lens to give a real image. The point at which it crosses the axis locates the image distance. By setting $y_2 = 0$ (image height for an axial ray) we can find the image distance t_1 :

$$y_2 = 0 = 73.5 - 2t_1,$$

and solving for t_1 , we get $t_1 = 36.75$ mm. This termination of the ray trace is referred to as *closing*. For parallel rays, closing determines the focal point of the system. In this case we had already found the image distance using the thin lens equation in Example 2.A. Besides determining the image distance for a system, a ray trace provides the height of the rays at each element of an optical system, a quantity useful for identifying important components and surfaces. In systems with many elements one can rapidly assess a system using such an array [Problem 2.2]. The y - u trace method lends itself readily to programming on a small computer or hand-held calculator. While it might seem to be a good idea to entrust all of these calculations to a ray-trace program, until an understanding of ray-trace methods is gained, it is best to do these preliminary designs using pencil and paper.

2.4. STOPS, PUPILS, AND WINDOWS

Although for convenience of computation the initial slope angle $u_0 = +0.1$ was chosen, we could have chosen another value just as well. If the sizes of the optical components are given, we can adjust the slope angle to trace a ray at the outer edge of the optical system. As we increase u_0 , the ray will eventually hit the edge of one of the components of the system. For angles beyond this slope angle, the rays will not be imaged by the optical system. An axial ray traced near the outer edge of the system is termed a *marginal ray*. In our present case, we see that the amount of light to the image will be limited by the lens, since it is the only

optical component in the system between the object and image. The opening in an optical system that limits the amount of light that can be collected is the *aperture stop* of the system. After the image location, it is the most important quantity to be determined by a ray trace since all subsequent analysis depends on knowing the location of the aperture stop. In this case it is the lens itself, as shown in Fig. 2.8.

One reason the aperture stop is a very important concept in optical system design is that it provides a measure of the light collecting ability of the system. If a system is built with an aperture stop that is too small, the detector, be it a film or photoelectric sensor, will not get enough light. Sometimes the aperture stop is in the middle of an optical system, inaccessible to the observer as in the case of Example 2.B. Yet it is necessary to be able to determine the size and location of its images in regions that are accessible. These regions are: *object space*, all the space to the left of the first element or surface of a system, and *image space*, the space to the right of the last optical element, assuming no reflecting surface. What must be done is to turn the aperture stop into an object plane and trace rays from this new object to determine the location and magnification of the images of the aperture stop as seen from object space and from image space. The image of the aperture stop *as seen from object space* is called the *entrance pupil* of the system. Likewise, the image of the aperture stop *as seen from image space* is known as the *exit pupil* of the system. In our microfiche example, the aperture stop is at the lens and accessible from both image and object space so it is also the pupils for this simple system. This is not the case in more complicated systems.

These three openings, the aperture stop, the entrance and exit pupils, their placement and size, have profound effects not only on the light collecting ability of the optical system, but also on certain aberrations. As we shall see, sometimes it is useful to shift the stop of a system along the optic axis to improve performance. This is another reason that the aperture stop is an important quantity in the analysis of an optical system. Many students have difficulty understanding the concepts of the aperture stop and its pupils. Perhaps this is because they expect the object to be in object space and the image to be in

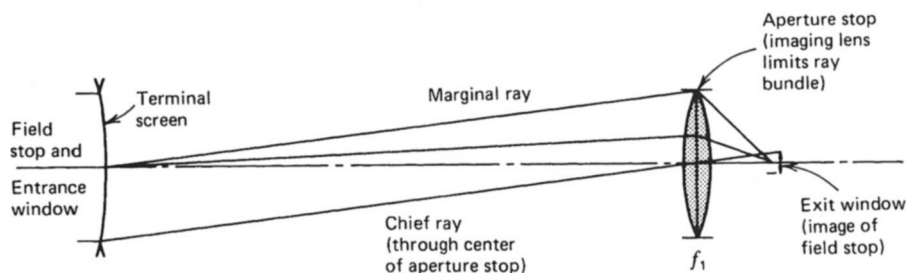
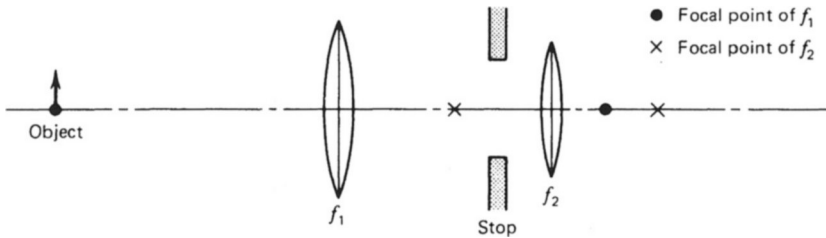
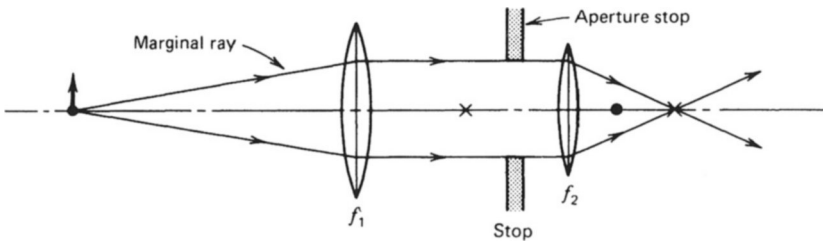


Figure 2.8. Ray trace of the single-lens example (see text). Stops and windows are identified. Since there is a single lens to limit the light, it is also the exit and entrance pupils.

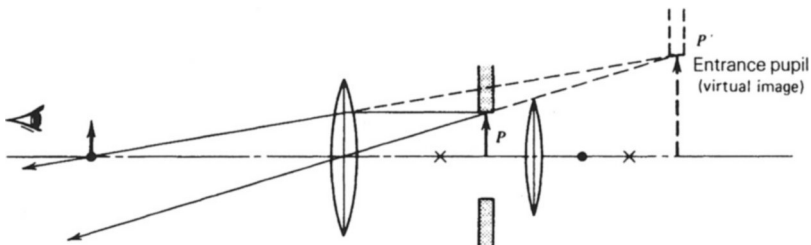
image space. But when they have to find the images of elements in an optical system by locating an object *in the middle of the system* and look for its images in object space and image space, some initial confusion is understandable. After you have examined a number of systems such as the one given in Example 2.B and worked most of the problems at the end of the chapter, the concepts of pupils (and windows, discussed below), will become a natural part of your analysis of any optical system. Examples of the usefulness of pupils in optical design are given in the next chapter and throughout the text.



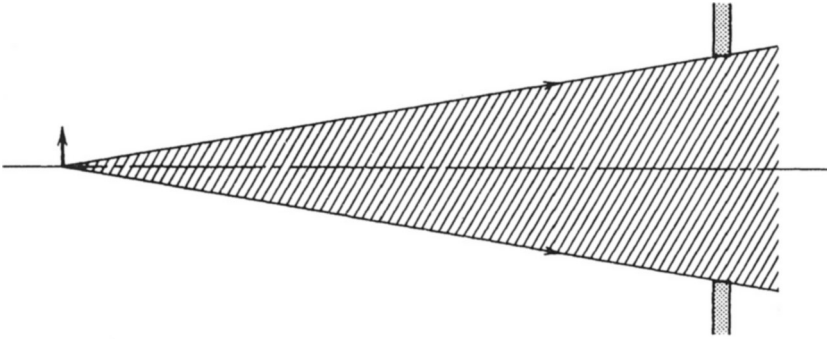
Solution: The object is at the focal point of the first lens. So it is easy to draw an axial ray. As the slope angle of the axial ray is increased, the ray is cut off first by the stop. This identifies the aperture stop.



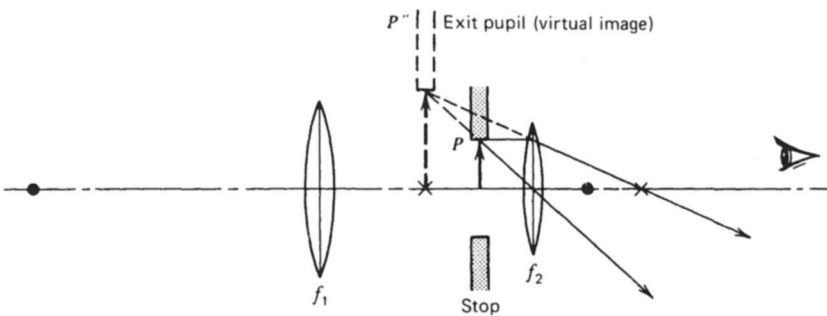
We can now locate the entrance pupil by using a point at the edge of the aperture stop (point P) as an object and tracing some rays to find its image as it is seen from object space.



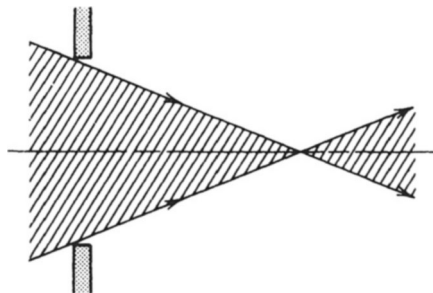
Thus, the aperture stop can be replaced by a hole located at the entrance pupil.



The exit pupil is located at P'' by finding the image of the edge of the aperture stop (point P) as it is seen from image space.

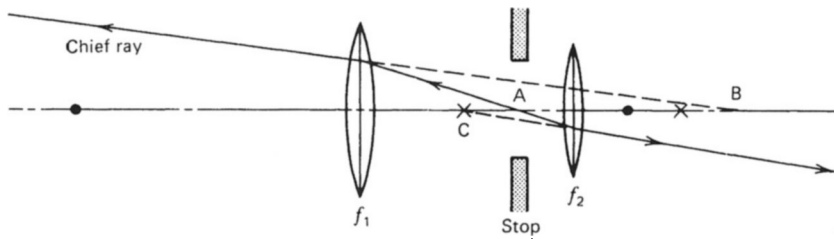


In image space the aperture stop and lens is replaced by the exit pupil.

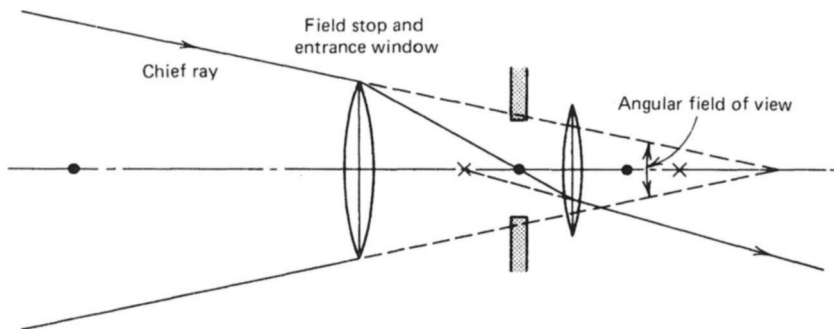


Thus, a cone of light filling the entrance pupil will be output as a cone of light emitted from the exit pupil. A ray centered at A in the aperture stop, the chief ray, can be traced in both directions. In object space the chief ray

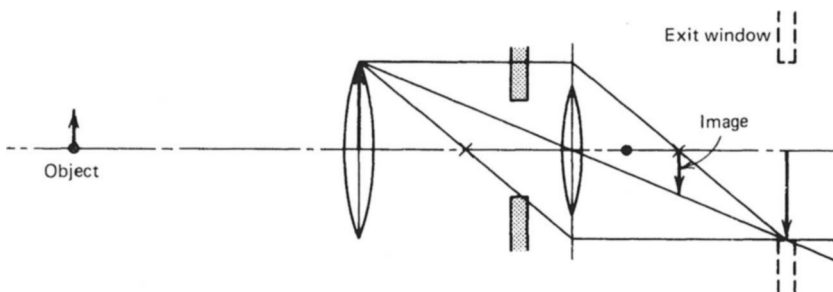
appears to come from the center of the entrance pupil B, and in image space it appears to come from the center of the exit pupil C.



The slope of the chief ray is increased until it is cut off by the edge of a component. This component is the field stop. In this case it is the first lens. Since the field stop is in object space, it is also the entrance window. The angle subtended by the marginal chief rays at the entrance pupil is the angular field of view of the system.



To find the exit window, the field stop (first lens) must be imaged through the second lens. Note that the exit window is not in the same plane as the image of the object. This means that the edges of the image field will be blurred and out of focus. Only if the field stop is at the object (as it is in the text example), at an intermediate image, or at the final image, will the edge of the image field be well defined.



The other ray that provides information on our system is a ray through the center of the aperture stop, called the *chief ray*. As shown in Table 2.3, we set $\bar{y}_1 = 0$ and $\bar{u}_0 = \bar{u}_1 = +0.1$ (since a ray through the center of a thin lens is undeviated). This yields values of \bar{y}_0 and \bar{y}_2 that indicate the 20:1 reduction, as shown in Table 2.3. (Note that all chief ray quantities have a bar over the symbol.) The question to be answered by these calculations is: What limits the extent of the image? (Obviously the diameter of the lens has no effect; it limits the axial cone of light.) The chief ray can be set at any arbitrary angle to the optic axis at the center of the lens. We can increase the angle of the chief ray until it is limited by some edge in the system. The limitation, in this case, must occur either at the object (size of data area on the terminal screen) or at the image (frame defining the image border at the microfiche film plane). The opening that limits the chief ray is the *field stop* of the system. To review: The terminal screen is 200 mm long \times 140 mm high. If we add a 50-mm band of computer code beneath the screen, the object dimensions are 200 mm \times 190 mm. When this area is reduced 20-fold by the lens it will be 10 mm \times 9.5 mm. Unless the border of the frame holding the microfiche film is smaller than this, the terminal screen opening is the limiting aperture and, therefore, the field stop of the system. Just as in the case of the aperture stop, the images of the field stop as seen from object and image space can be determined. They are called the *entrance* and *exit windows*, respectively. In our simple case, the field stop is already in object space, so it is also the entrance window. The image of the field stop behind the lens, the 10-mm \times 9.5-mm area just calculated, is the exit window, and it is located in the image plane, as shown in Fig. 2.8. In later calculations, the locations and size of the stops, pupils, and windows require a little more effort than was needed here (Example 2.B). In all calculations the location of the entrance quantities should be given relative to the first surface of the system and all the exit quantities to the last surface of the system.

One additional concept that is useful in describing an optical system is the *angular field of view*. This provides the designer with the size of the angular coverage of the system. In object space, the angular field of view is the angle at the entrance pupil defined by chief rays that touch the edges of the field stop. In our microfiche example, the terminal screen with an average dimension of 200 mm is located 735 mm from the entrance pupil. Thus, the angular field of view is $2 \tan^{-1} (100/735)$ or about 15.5° . In general, if \bar{u}_0 is a chief ray angle in object space, \bar{y}_i is the ray height at the field stop, the i th surface, and r_i is the radius of the field stop, the angular field of view is given by:

$$\theta_{fov} = 2 \tan^{-1} \left(-\bar{u}_0 \cdot \frac{r_i}{\bar{y}_i} \right).$$

It is possible to trace other rays for this system, but as long as we are using the small angle approximation, we will learn nothing new. It should be pointed out that all of these paraxial rays will scale with a constant multiplier. For example, if we start an axial ray with a new u_0 value equal to β times the old u_0 value, then

all subsequent ray values y_k and u_k would be multiplied by β . Similarly, if the height of the chief ray were multiplied by β , all subsequent values for this ray would be β times the original ray values. Any new ray having both nonzero y_0 and u_0 is just a sum of an axial ray and a chief ray with appropriate multipliers. Because Eqs. 2.5 and 2.8 are linear in y_k and u_k , the subsequent ray parameters are just the sum of the ray values for the separate rays (with correct multipliers) [Problem 2.3].

This might be an appropriate place to evaluate our system; it would be better, however, to perform the same analysis on the two-lens system. We can then compare advantages and drawbacks between the two systems.

2.5. EVALUATION OF SYSTEMS USING THE y - u TRACE

In the second system, the optical power is distributed over two lenses. Although we could assign each stage of the system equal magnification ($\sqrt{20}$), we shall instead make the first stage a 4-power demagnification ($M = -\frac{1}{4}$) and the second stage, 5-power ($M = -\frac{1}{5}$). From Eq. 2.4, we find that for two 50-mm lenses, the total system length is

$$L_1 + L_2 = \frac{(4 + 1)^2}{4} 50 + \frac{(5 + 1)^2}{5} 50 = 672.5 \text{ mm},$$

considerably shorter than the single-lens system. Using 28-mm and 35-mm lenses instead of a 50-mm lens yields system lengths of 377 and 471 mm, respectively. Let us again choose the 35-mm lenses and run a y - u trace on the system. From Eqs. 2.3 and 2.2 we determine the spacings for the system and describe the system in Fig. 2.9:

$$t_1 = -(M - 1)f_1 = \frac{5}{4} \times 35 \text{ mm} = 43.75 \text{ mm},$$

and

$$t_0 = \frac{t_1}{M} = -4 \times 43.75 \text{ mm} = -175 \text{ mm},$$

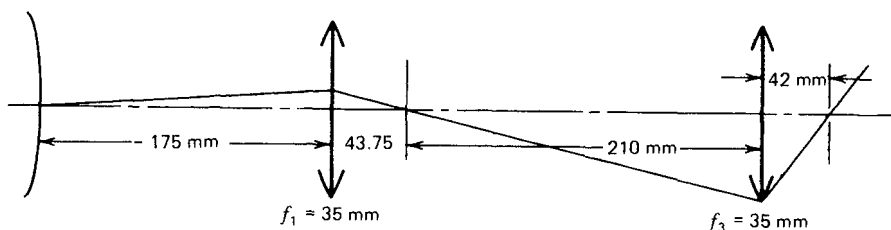


Figure 2.9. Two-lens system with 35-mm focal length lenses.

also

$$t_3 = -(M - 1)f_2 = \frac{6}{5} \times 35 \text{ mm} = 42 \text{ mm},$$

and

$$t_2 = -5 \times 42 \text{ mm} = -210 \text{ mm}.$$

In Table 2.4 the y - u trace of an axial ray for the two-lens system is given. An additional surface ($\neq 2$) has been introduced at the intermediate image point. Although it is not necessary in this example, it is not unusual to maintain surfaces at intermediate image locations. Based on this ray trace we can determine the aperture stop of the system. We note that an axial ray could be multiplied by a constant value and still preserve the same image and object distances; all that changes are the heights of the rays throughout the system. If we divide the clear aperture radius r_k of every component by the axial ray height y_k at the component, we can determine the value of r_k/y_k that would be needed to produce a ray that would touch the edge of that component. Thus, the larger the absolute value of r_k/y_k , the larger the initial and final slope angles required by a ray to touch the edge. Therefore, the component with the *smallest* $|r_k/y_k|$ will define the marginal ray, the ray that will pass through all the components. Since any axial ray larger than this will not pass through that limiting component, this serves to identify the aperture stop of the system. Thus, given any axial ray with ray height values y_k at each component surface, the component with the *smallest* $|r_k/y_k|$ is the aperture stop of the system. If we assume the two lenses are identical, then their clear aperture radius is the same ($r_2 = r_1$). Thus, the multipliers for these components would be $r_1/17.5$ for the first and $r_1/-84$ for the second. Since $|r_1/-84| < |r_1/17.5|$, the second lens serves as the aperture stop of the system. The location of the entrance pupil is found by calculating the image distance from the first lens with the second lens (aperture) as the object. This can be done either by treating this as a single lens problem and reversing object and image spaces or by tracing a ray *backward* through the system. We can trace a ray backward through a system by reversing the sign of the distances measured from a component: positive to the left, negative to the right (Rule 6 in Section 1.4). That is,

$$\frac{1}{t_{\text{ent. pup.}}} = \frac{1}{f_1} + \frac{1}{t_{\text{ap. stop}}},$$

where $t_{\text{ap. stop}} = t_2 + t_1 = -(210 + 43.75) \text{ mm}$ (negative, since the distances are to the right of the first lens) and $f_1 = 35 \text{ mm}$

$$\frac{1}{t_{\text{ent. pup.}}} = \frac{1}{35} - \frac{1}{253.75},$$

$$t_{\text{ent. pup.}} = 40.6 \text{ mm to the left of the first lens.}$$

Starting a chief ray through the aperture stop, we trace the ray in both directions. In the forward direction this is simple: $y_4 = 0 - 0.1 \times 42 = -4.2$.

Table 2.4. Ray Trace of the Two-Lens System^a

k	Object 0	1st Lens 1	Int. Image 2	2nd Lens 3	Final Image 4
t_k	175	43.75	210	42	
ϕ_k	0	1/35	0	1/35	0
r_k	200	r_1	∞	$r_2 = r_1$	10
y_k	0	17.5	0	-84	0
u_k	+0.1	-0.4	-0.4	+2	
$ r_k/y_k $	∞	$r_1/17.5$	∞	$r_1/84^b$	∞
\bar{y}_k	-84	25.375	21	0	-4.2
\bar{u}_k	+0.625	-0.1	-0.1	-0.1	
$ r_k/\bar{y}_k $	2.38	$r_1/25.4^c$	∞	∞	2.38

^a Calculated values are shaded.

^b Aperture stop.

^c Field stop (based on an unreasonably large clear aperture; see text). Tracing the chief ray backward through the system set $y_3 = 0$ and $u_2 = -0.1$ at the second lens (aperture stop)

$$\bar{y}_2 = \bar{y}_3 - \bar{u}_2 t_2 = 0 - (-0.1)(210) = 21,$$

$$\bar{u}_1 = \bar{u}_2 + \bar{y}_2 \phi_2 = -0.1 + (21) \cdot 0 = -0.1,$$

$$\bar{y}_1 = \bar{y}_2 - \bar{u}_1 t_1 = 21 - (-0.1)(43.75) = 25.375,$$

$$\bar{u}_0 = \bar{u}_1 + \bar{y}_1 \phi_1 = -0.1 + (25.375)/35 = 0.625,$$

$$\bar{y}_0 = \bar{y}_1 - \bar{u}_0 t_0 = 25.375 - (0.625)(175) = -84.$$

(Note that the ratio $\bar{y}_4/\bar{y}_0 = -4.2/-84 = \frac{1}{20}$, as required. Note also that $u_0/u_3 = 0.1/2 = \frac{1}{20}$.)

But how do you trace a ray in the backward direction? Equations 2.5 and 2.8 are still useful, we just solve for the y_k or u_k

$$u_k = u_{k+1} + y_{k+1} \phi_{k+1} \quad (\text{REVERSE}), \quad (2.10a)$$

and

$$y_k = y_{k+1} - u_k t_k \quad (\text{REVERSE}). \quad (2.10b)$$

The chief ray in the reverse direction is traced in Table 2.4. When tracing a ray backward, the same system parameters are used but the distance t_k to the k th surface from the $k + 1$ st surface is positive if it is to the left of the $k + 1$ st surface

and negative if it is to the right. A procedure for analyzing an optical system is given in Table 2.5.

An examination of the heights of the chief ray, tells us that either the object or the first-lens diameter will limit the size of the image and serve as the field stop of the system. But which? We have not specified the clear aperture of the lenses yet. From our chief-ray heights we know that for the object $|r_o/\bar{y}_o| = 2.38$ and that for the first lens $|r_1/\bar{y}_1| = r_1/25.4$. So depending upon the radius of the first lens, either the terminal screen edges or the clear aperture of the first lens would serve as the field stop. Let us assume that the terminal screen is the field stop. Then $|r_o/\bar{y}_o|$ must be less than $|r_1/\bar{y}_1|$. That is $2.38 < r_1/25.4$, or the clear aperture radius of the first lens, $r_1 > 60.3$ mm. This, as we shall see, is an unrealistic clear aperture radius for a lens with a focal length of 35 mm.

Here the real world intrudes. Given a choice between a large-diameter lens and a small one, one would usually choose a larger-diameter lens so that the lens could collect either more light (larger aperture stop) or provide a more extensive image (larger field stop). But, in most cases, the larger lens is more expensive. One measure of the size and light collecting ability is the optical *speed* of the lens given by its *f-number*. The *f-number* of a lens is its effective focal length (to be defined in Section 2.6) divided by its entrance pupil diameter. Thus a 50-mm

Table 2.5. Analyzing an Optical System

1. Set up the system array [distances, component locations, focal lengths (powers), and clear apertures].
2. Run an axial ray using a $y-u$ trace.
3. Find the aperture stop (calculate the ratio of clear aperture to ray height: Smallest $|r_k/y_k|$).
4. Run a chief ray using a $y-u$ trace.
5. Calculate as in #3 above to find the field stop.
6. If there are a few components, use the thin lens equation to find the location and size of the pupils and windows.
7. If there are many components, use a $y-u$ trace forward and backward to find pupils and windows.
8. Sizes of pupils can be determined by taking the slope angle of the chief ray at the aperture stop and dividing by the image and object space slope angles of that ray to get the magnifications of the aperture stop in image space and object space, respectively. Multiplying the aperture stop size by the magnifications gives the exit and entrance pupil sizes. (The basis for this will be found in Section 5.2.1, The Optical Invariant.)

A full report of a system contains:

Image location and magnification; the locations of stops, pupils, and windows; the effective focal length of the system; its *f-number*; and its angular field of view.

camera lens with a 23-mm aperture stop has an optical speed of 2.2 ($=\frac{50}{23}$) or, in a convenient f -notation, a speed of $f/2.2$. Table 2.6 lists a range of lenses and optical speeds and relative costs. Costs, however, are determined by the quantity of lenses required as well as other variables. A single simple lens that is custom made will cost considerably more than a stock multi-element photographic lens of comparable focal length and f -number. For the 35-mm lens in our design the f -number would have to be 35 mm/120 mm or $f/0.29$. This is an unrealistic lens requirement. Rarely is the diameter of a lens greater than its focal length (i.e., the f -number of a lens is almost always greater than 1).

Then what is the solution to our problem? We will have to abandon the two-lens system, as attractive as it may have been, and use a single-lens system. It should be noted that if we had drawn in a chief ray back at the ray-sketching phase, we would have diagnosed the problem then. But that would have been getting ahead of things.

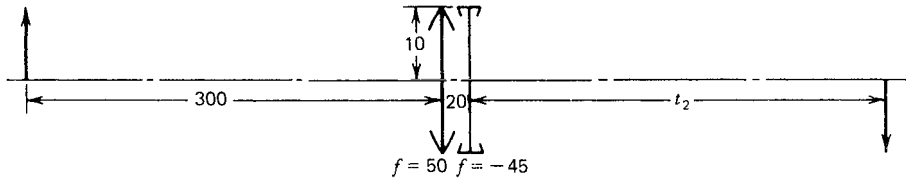
Our design choice is then a single-stage 20:1 reduction lens. The exact focal length to be chosen may be decided by choosing the best overall length of the system, but noting that as f gets smaller, the cost of the system increases. Our next step might be to determine how to fold the system into a convenient package. This will be postponed until Chapter 4.

Because our sample design is a single-lens design, let us examine another design to illustrate the full use of the y - u ray trace and to complement the graphical Example 2.B. This telephoto design consisting of two lenses, one positive followed by one negative, is analyzed in Example 2.C. For an object 300 mm from the first lens a y - u trace gives ray heights of 30 and 20 mm at the front lens and back lens, respectively. If the clear radius of the lens or stop is divided by the ray height at that component $|r_k/y_k|$, we obtain ratios of 0.33 and 0.50 for the front lens and back lens, respectively. The component with the smallest $|r/y|$ ratio will be the aperture stop, as discussed earlier, since this component will first intercept a marginal ray. Thus, the first lens in this system serves as the aperture stop and entrance pupil.

Table 2.6. Lenses, Optical Speeds, and Costs

$f/\#$	Costs
$f/1$	Very expensive, highly corrected
$f/2$	Expensive, well corrected
$f/4$	Moderate, corrected
$f/8$	Moderate, small correction needed
$f/16$	Inexpensive, little correction needed
$f/32$	Inexpensive, nearly a pinhole

EXAMPLE 2.C. Given two lenses as illustrated below, where are the stops of the system? What is the effective focal length of the system?



Solution: Using a y - u trace (shaded values are calculated, unshaded values are either given by the system, by definition of the ray type, or set arbitrarily).

	n	0	1	2	3
System values	t_k	300	20	360	
	ϕ_k	0	0.02	-0.0222	0
	r_k		10	10	
Axial ray	y_k	0	30	20	0
	u_k	0.1	-0.5	-0.05556	
	$ r_k/y_k $		0.33 ^a	0.5	
Chief ray	\bar{y}_k	-30	0	2	54
	\bar{u}_k	0.1	0.1	0.1444	
	$ r_k/\bar{y}_k $		∞	5 ^b	
Parallel ray	y'_k	10	10	6	0 ← At focal plane
	u'_k	0	-0.2	-0.06667	
	$ r_k/y'_k $		1.0	1.67	

^a Aperture stop.

^b Field stop.

Effective focal length of the system is (using parallel ray data)

$$f_{\text{eff}} = \frac{-y'_0}{u'_2} = \frac{10}{0.06667} = 150.$$

Location of the back principal plane may then be found. (Using primes on the t_2 's to distinguish them from the image distance with the object at 300 units).

Closing for t'_2 : set $y_3 = 0$ for parallel ray.

Back Focal Distance: $t'_2 = -y'_2/u'_2 = 6/0.06667 = 90$ units.

Thus the back principal plane is 60 units in front of the second lens and 40 units in front of the first lens. (See Fig. 2.10.)

The exit pupil could be found by ray tracing, but with only one lens between the aperture stop and the image space it is just as easy to employ the thin lens equation with the object distance equal to 20 mm.

Exit pupil:

$$\frac{1}{t_1} = \frac{1}{-45} - \frac{1}{20},$$

$t_1 = -180/13$. The image is virtual, 13.85 mm in front of the back lens.

To find the field stop we need only extend a ray in both directions from the center of the first lens. Obviously, the second lens is the only element to limit the chief ray and is therefore the field stop and exit window. When an axial ray, started at the back lens, is traced through the system, the entrance window is located.

Entrance window:

$$\frac{1}{t_1} = \frac{1}{50} - \frac{1}{20},$$

$t_1 = -100/3$. The image is virtual and appears from object space to be 33 mm behind the first lens.

The results for this system are summarized in Example 2.C. It should be pointed out that these assignments of stops, pupils, and windows are not true for all object distances. For example, if the object is located 30 mm from the first lens, the second lens is the aperture stop and the first lens is the field stop. Here the lens gives a virtual, magnified image [Problem 2.4].

2.6. EFFECTIVE FOCAL LENGTH

The y - u trace can also be used to find the effective focal length of the system. Although applying the thin lens equation twice will give us the back focal point of the combination, a y - u trace will locate not only this point, but also determine the effective focal length of this two-lens system [Problem 2.5].

The lenses of this system can be treated as a single optical component. We can, for example, find the effective focal length for this combination. The *effective focal length* is the focal length of the combination system if it were replaced by a single thin lens. If the initial parallel ($u_0 = 0$) ray and the final ray that crosses the axis are extended toward each other, the point where the rays cross would determine the location for the replacement thin lens, and the distance from the axis crossing (i.e., the focal point) to this position is the effective focal length of the replacement lens (and, therefore, of the system). The position of the thin lens is called the *back principal plane* of the system. It and other parameters of optical systems will be described in Section 5.1.2 on the cardinal points of a lens.

The trace is noted at the bottom of Example 2.C as a “parallel ray.” This is the trace of a ray parallel to the optic axis and is used to find the effective focal length of an optical system. The ray height is somewhat arbitrary, although it is usually less than the entrance pupil. The new object distance may be kept the same as the old object distance since it is of no consequence in this trace. Since a ray parallel to the optic axis is required, the initial ray slope is set equal to zero. The place where this ray crosses the optic axis (or appears to diverge from the optic axis, in the case of a combination of lenses with a negative effective focal length) is the back focal point of the lens. To find the effective focal length where j is the number of surfaces in the system (in Fig. 2.10, $j = 2$), we write

$$\tan u_j = \frac{-y_0}{f_{\text{eff}}}$$

and in the paraxial case, $\tan u_j = u_j$,

$$f_{\text{eff}} = \frac{-y_0}{u_j}, \quad \text{when } u_0 = 0. \quad (2.11)$$

Again, the inclusion of the negative sign is necessary to provide the proper signs on both sides of the equation. For the case illustrated in Example 2.C, the focal length is 150 units. In cases where the ray crosses the axis one or more times before entering image space, even Eq. 2.11 fails. A more general formula for the effective focal length is

$$f_{\text{eff}} = -\frac{y_0}{u_j} \frac{y_j}{|y_j|}, \quad \text{when } u_0 = 0. \quad (2.11')$$

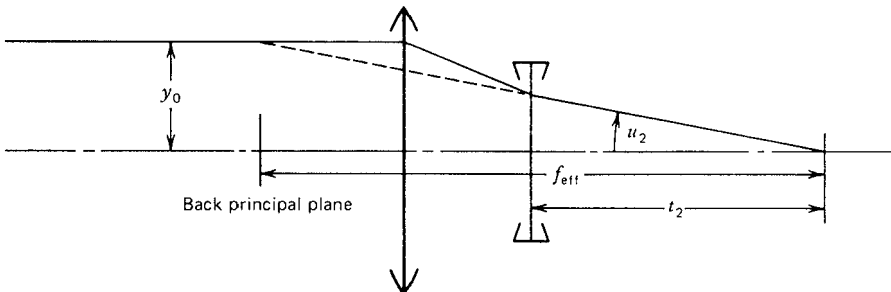


Figure 2.10. Effective focal length of a telephoto lens system.

Since the back focal point is 90 units to the right of lens #2, the principal plane for this system is $150 - 90$ or 60 units to the left of the second lens and $60 - 20$ or 40 units in front of the first lens. See Fig. 2.10. This telephoto combination produces a long-focal-length lens from two short-focal-length lenses located closer to the focal plane than a single lens. Practice in finding the focal lengths and other characteristics of thin lens systems are given in the problems [Problems 2.6–2.15].

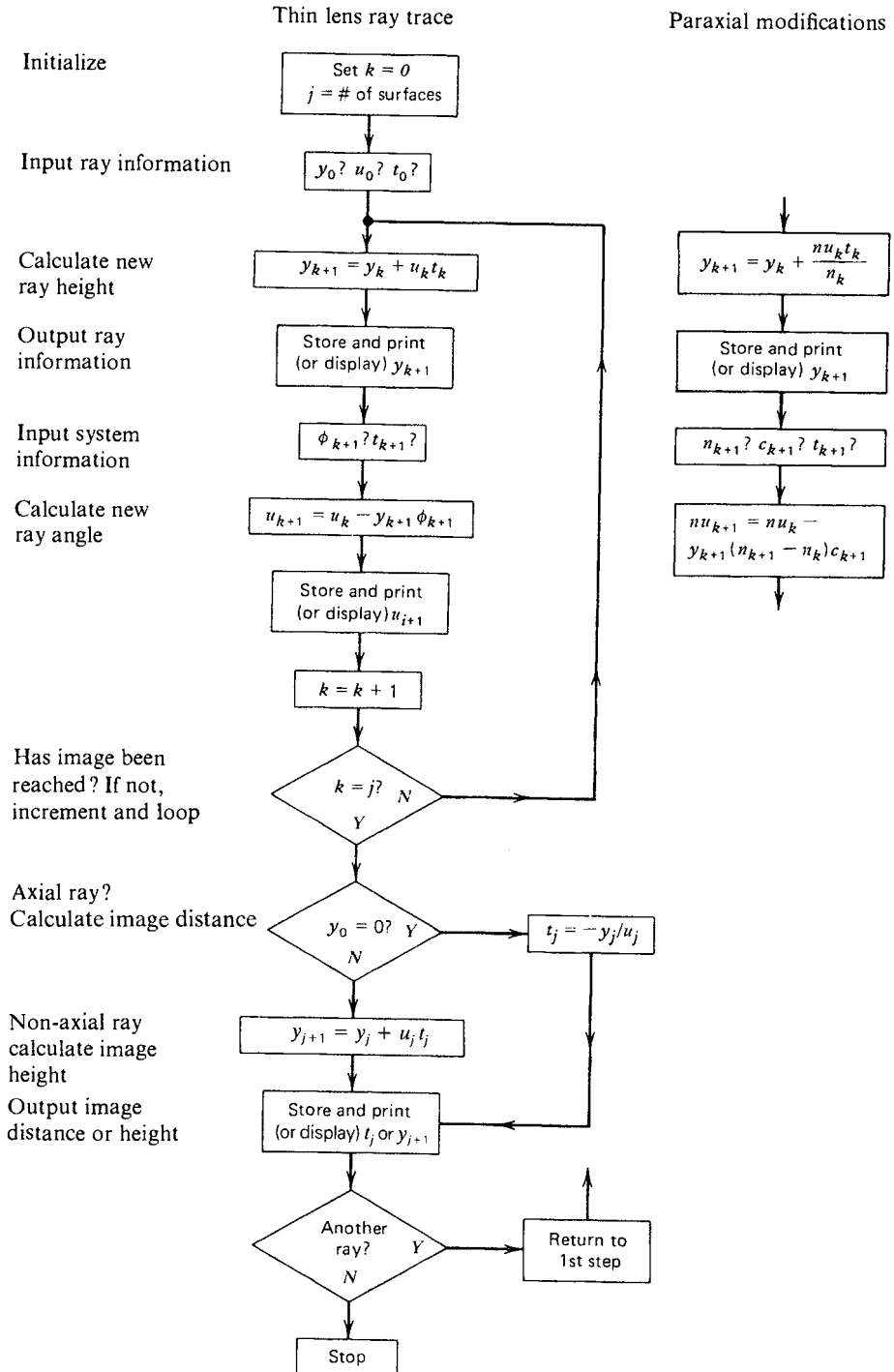
A thin lens ray trace gives us a first assessment of a proposed design system. It provides additional information on the size and position of the stops, pupils, and windows. Note, however, that there must be values assigned to the array before the trace can be made, and this is usually done, as we did here, with simple thin lens calculations. Only after these are in place can a y - u trace be run. Equations 2.5 and 2.8 are all that are needed to trace any ray. As we noted earlier, these equations could be programmed on a small programmable calculator or microcomputer. Although a number of such ray-tracing programs exist for different systems, no programs are listed in this text. Rather than give a specific program for a particular calculator or computer, a flow chart is provided in Example 2.D to enable one to write such a program quickly and efficiently.

EXAMPLE 2.D. Ray Trace Program for Thin Lenses. Rather than giving a specific program for a particular calculator or computer, a flow chart is provided on the following page to enable one to write a program quickly and efficiently.

2.7. CYLINDRICAL LENSES

All of the lenses and mirrors discussed to this point were constructed from spherical surfaces. The unique direction in the system was the optic axis, the z axis in these coordinate systems. In terms of the orientation of an object in a direction at right angles to the optic axis, it made little difference whether the object was oriented along the x axis or y axis or any intermediate axis in the x - y plane, since ray tracing would always give the same result.

If the optical components were made of cylindrical surfaces, instead of spherical surfaces, this would no longer be true, since the optical power of a component would depend on whether a ray was propagating in a plane parallel to the cylinder axis or perpendicular to it. Cylindrical lenses can be used to provide a line focus rather than a point focus. For example, in Fig. 2.11(a), there are two cylindrical lenses that provide two line foci. Analysis of this system is simple: one treats a cylindrical lens in a plane perpendicular to the cylinder axis as a thin lens and in a plane containing the cylinder axis as a slab of glass with no optical power. Thus, at least, two sections of the optical system are needed to



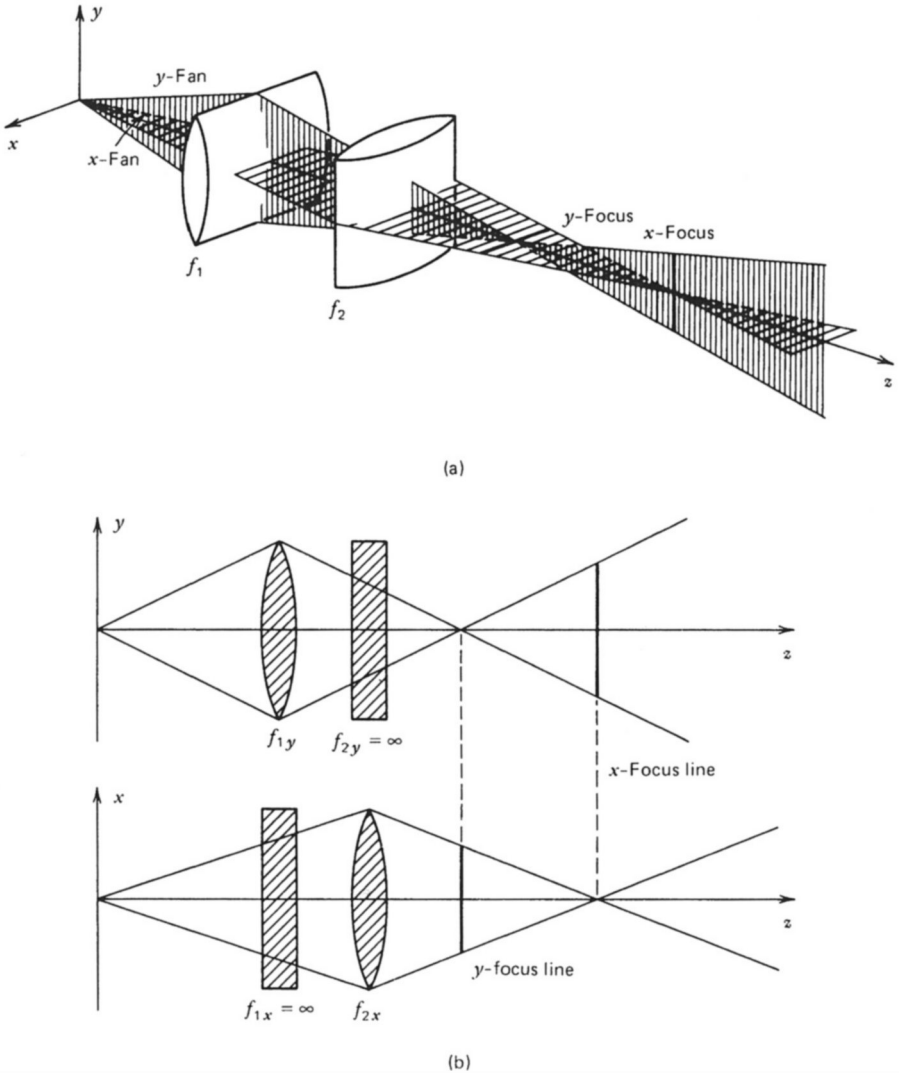


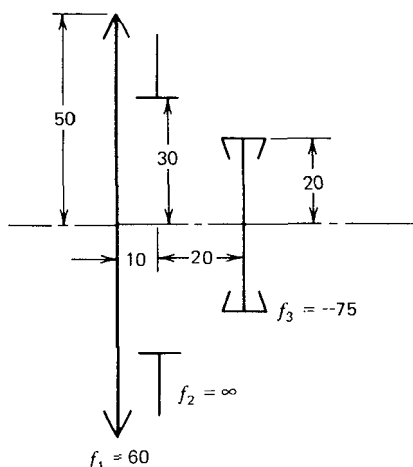
Figure 2.11. Cylindrical lens system. (a) Perspective view showing the focusing of two fans of rays. (b) Ray diagrams in the y - z and x - z planes showing the treatment of cylindrical lenses as thin lenses in one plane and glass slabs in the orthogonal plane.

display the course of rays through a system. The sections for the cylindrical optical system in Fig. 2.11(a) are displayed in Fig. 2.11(b). To describe the system a fan of rays in the y - z plane is shown passing through the system. In this case, only the first lens has optical power. The rays focus to a point (y -focus) as it were the only lens in the system. In the x - z planes, the only lens with optical power is the second lens. The point source is imaged to a different point (x -focus) along the axis. Since the y -fan is diverging at that point, this results in a line

focus. Extensions to more complicated systems follow this same treatment [Problem 2.14].

PROBLEMS

- 2.1. Determine from a computer terminal you are familiar with, the best possible reduction format. You might take the following into consideration: (1) Most microfiche projectors will display an $8\frac{1}{2} \times 11$ -in. page of typescript in a format at 75% of the original size. (2) The optically readable code should fit under or beside the screen. (3) the 20-fold time reduction used in the text may not be optimum. What is the greatest number of frames you can pack onto a standard microfiche and still maintain readability? (There is no correct answer to this problem; there are, however, reasonable versus ridiculous answers.)
- 2.2. The focal lengths of four lenses are $f_1 = 100$ mm, $f_2 = -50$ mm, $f_3 = 80$ mm, and $f_4 = -280$ mm. If the separations between the lenses are $t_1 = 20$ mm, $t_2 = 40$ mm, and $t_3 = 30$ mm. Where is the image located if the object is 400 mm from the first lens? (a) Find the image using the thin lens equation and rule 6 in Section 1.3. (b) Find the image using the y - u trace.
- 2.3. (a) Show that a new *axial* ray with a slope angle $u'_0 = \alpha u_0$, will give new ray values $y'_k = \alpha y_k$ and $u'_k = \alpha u_k$, where y_k and u_k are the ray trace values for the slope angle u_0 . (b) Using the results of part a, show that any ray having both nonzero u_0 and y_0 is just the sum of the axial and chief rays with appropriate multipliers.
- 2.4. If the object distance for Example 2.C is changed to 30 mm, where are the stops, pupils, and windows of the system?
- 2.5. Two lenses of focal lengths 98.59 mm and 84.84 mm are separated by 2.2 mm. Where is the back focal point of this system (a) using the thin lens equation and (b) using the y - u trace? Find the effective focal length and locate the back principal plane of the system. What is the effective focal length if the 2.2-mm separation is ignored?
- 2.6. *Periscope Lens*. Two identical lenses, focal length 104.167 mm, are separated by 4.6 mm with a stop placed symmetrically between the lenses. The stop diameter is 90% of a lens diameter. Where are the stops, pupils, and windows of this system if the object is 75 mm from the first lens? What is the effective focal length of this system and where is the back principal plane located relative to the lenses?
- 2.7. *Telephoto Lens*. A system of two lenses and a stop is shown in the accompanying figure. What are the image distance, locations, and sizes of the stops, pupils, windows, and angular field of view if the object is at 400 mm?



- 2.8. *Telephoto Lens.* A system of two lenses and a stop is shown in the figure for the preceding problem. What are the image distance, locations, and sizes of the stops, pupils, windows, and angular field of view, if the object is at infinity? Where is the principal plane of the system?
- 2.9. *Telescope.* A telescope consists of a 30-mm-diameter objective lens of 147 mm focal length and a 10-mm-diameter single eyepiece of 21 mm focal length separated by the sum of the focal lengths. If the eye pupil, adjusted to the amount of light coming through the telescope, is 5 mm in diameter and 20 mm behind the lens, where are the stops, pupils, and windows of the system? Draw a number of parallel rays at various heights through the system to show its operation. What is the angular field of view of this telescope?
- 2.10. *Microscope.* A microscope consists of a 14.545 mm focal length objective (5-mm diameter) separated from a 28-mm focal length eyepiece ($d = 5$ mm) by 185 mm. If the object is 16 mm from the objective, what is the total magnification of the system? Where are the stops, pupils, and windows of this system?
- 2.11. *Double Gauss.* A symmetrical lens system is composed of four thin lenses. The first component ($f = 45.7$ mm, $d = 10$ mm) is separated from the second component ($f = -83$ mm, $d = 10$ mm) by 4 mm. The second lens is 6 mm from a diaphragm ($d = 7.5$ mm). The diaphragm is at the mirror plane of this symmetrical geometry, so the focal lengths and spacings are reversed in the second half of the lens. What is the effective focal length? With the object at infinity, where are the stops, pupils, and windows for this lens? If an object were at 750 mm, would this change any of these conclusions?

- 2.12. *Cooke Triplet*. Three thin lenses are arranged in the following order: a positive lens ($f = 60$ mm, $d = 8$ mm) separated 13 mm from a negative lens ($f = -40$ mm, $d = 6.4$ mm) and followed by another positive lens ($f = 50$ mm, $d = 7.2$ mm) with a separation of 13 mm from the negative lens. What is the effective focal length of the system? Where are the stops, pupils, and windows for an object at infinity? If an object is 300 mm from the first lens, where are the image and stops for the system?
- 2.13. *Ramsden Eyepiece*.* Two positive lenses spaced 25 mm apart form an eyepiece ($f_1 = 50$ mm, $d = 20$ mm, $f_2 = 36$ mm, and $d = 16$ mm). What is the effective focal length of this combination? If an object is 5 mm in front of the first lens, where is the image located, and where are the stops, pupils, and windows for this system?
- 2.14. *Cylindrical Lens System*. Two cylindrical lenses separated by 20 mm have optical powers of $\phi_x = 103.333$ diopters and $\phi_y = 36.905$ diopters. The subscripts indicate that these lenses have optical power in planes perpendicular to one another. If the object is located 10 mm from the first cylindrical lens, where are the images located? If the object is a square 10 mm \times 10 mm format, what will the image look like?

* An eyepiece is a wide-angle optic for a magnified image.

3

Radiometry

(OR PUTTING LIGHT THROUGH AN OPTICAL SYSTEM)

As anyone who looks at an underexposed snapshot knows, the amount of light converged by an optical system is just as important as the quality of the image. Many of the cameras we use today have elaborate sensing and computing systems that tell us whether we should take the picture. Although many optical systems do not contain such apparatus as part of their design, still the control and distribution of light in a train of optical elements must be determined if useful systems are to be produced. Consider the design problems that are present in residential and commercial illumination, in the visibility of cathode ray tubes in computer terminals, light emitting diodes (LEDs), and liquid crystal displays (LCDs) for calculators, and light propagation in supermarket scanning systems and photographic systems. Each design raises similar questions and demands different answers. Some of the questions that must be answered are:

1. How much radiation is available from the primary source?
2. What are the limitations in the light collection and distribution system?
3. How does the illumination fall off away from the optic axis?
4. How much of the radiation is found in image space?
5. How do you measure what you have?

The following general question should always be uppermost in the mind while designing a system: Is there sufficient radiation available to accomplish the task? Rough calculations of the sort given in this chapter should be made as soon as it is possible to answer this question. It is for this reason that this chapter on radiometry is placed near the beginning of the text.

3.1. THE BASIC CONCEPTS AND TERMS

Before computing any light levels there must be a systematic approach to defining the amount of light in a system and analyzing its propagation through the system. To do this, certain concepts and definitions are required. The purpose of this chapter is to introduce these items and show how they are used

in the simplest cases. One of the most difficult problems in the field of radiometry is coping with the large number of units that are used in the field. In this text we use the set of CIE units accepted by the International Commission on Illumination (Table 3.1). These have also been adopted by the American National Standards Institute (ANSI Z7.1-1967). There are other systems that are equally valid; but as an introduction to the subject it is best to keep the number of units to a minimum. Descriptions of other systems of units and the conversion factors between them can be found in Ref. 3. 1.

A *point source* is a small source of light that radiates equally in all directions. Strictly speaking, a point source is an idealization; however, it can be approximated in the real world by a source of radiation whose dimensions h_x, h_y are small compared to the observation distance R (i.e., $h_x h_y \ll R^2$), as shown in Fig. 3.1. Beyond some sufficiently great distance, even sources of substantial size may be treated as point sources—a useful approximation in many calculations. For example, all of the stars, some of which are larger than the orbit of the earth, are from our vantage point, essentially, point sources.

To measure the portion of space into which a point source is radiating, we use the concept of a solid angle. Recall that a *plane angle*, measured in radians (rad), can be described as the ratio of the arc S of a circle subtended by the angle divided by the radius R of the circle,

$$\phi = \frac{S}{R}, \tag{3.1}$$

Table 3.1. Radiometric and Photometric Units and Symbols

Quantity	Radiometric	Photometric
Energy	Joule (Q_e)	Talbot (Q_v)
Power, flux	Watt (ϕ_e)	lumen (ϕ_v) ^a
Intensity	Watts/sr (I_e)	lumens/sr (I_v)
Irradiance	Watts/m ² (E_e, I)	
Illuminance		lumens/m ² (E_v)
Radiance	Watts/m ² -sr (L_e)	
Luminance (brightness)		lumens/m ² -sr (L_v)

^a 680 lumens = 1 watt at 555 nm. Other wavelengths are scaled to the standard luminosity curve (Fig. 3.16).

Note: The area indicated in the units for irradiance and illuminance is the illuminated surface area, whereas the area indicated in units for radiance and luminance is the area of the source.

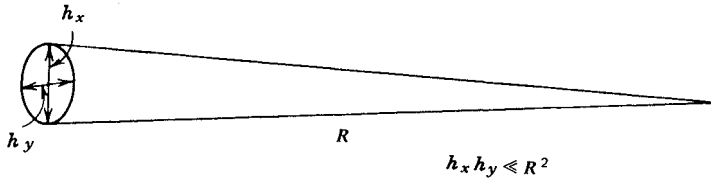


Figure 3.1. Approximation of a source of dimensions h_x by h_y as a point source.

as shown in Fig. 3.2(a). A *solid angle* denoted as Ω in Fig. 3.2(b) is expressed in units of steradians (sr) and defined as the area A on a sphere of radius R subtended by the solid angle divided by the square of the sphere's radius,

$$\Omega = \frac{A}{R^2}. \quad (3.2)$$

For example, a point source radiating in all directions illuminates the area of a sphere, $4\pi R^2$. Therefore, the point source radiates into 4π steradians. (While the area defining the solid angle can be irregular, in optical design it will usually be a circle or an ellipse.) Note that the requirement, $h_x h_y \ll R^2$, for sources of some extent to be considered point sources is equivalent to saying that the source as viewed by an observer must subtend a solid angle very much less than unity.

3.1.1. Radiometry of Point Sources

Starting with the idea of a point source and solid angle, we can do some simple radiometry. The amount of radiative energy per unit time is the radiation flux or *power* (ϕ_e) leaving the source and passing through a reference surface A , as shown in Fig. 3.3. The power is measured in units of watts (joules/sec), as it is for mechanical and electrical systems (Table 3.1). As is discussed later, measurement of the power within certain frequency bands requires a modification of the unit of power. For example, if we measure only that radiation to which the human

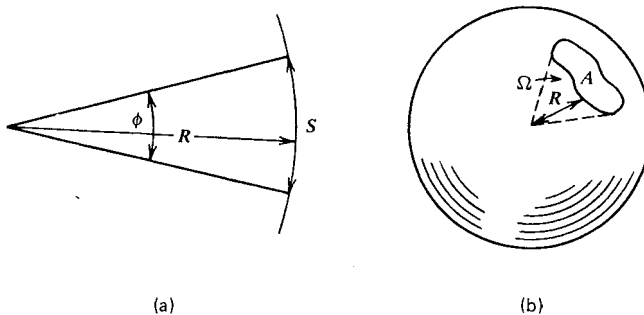


Figure 3.2. Definition of (a) plane and (b) solid angle.

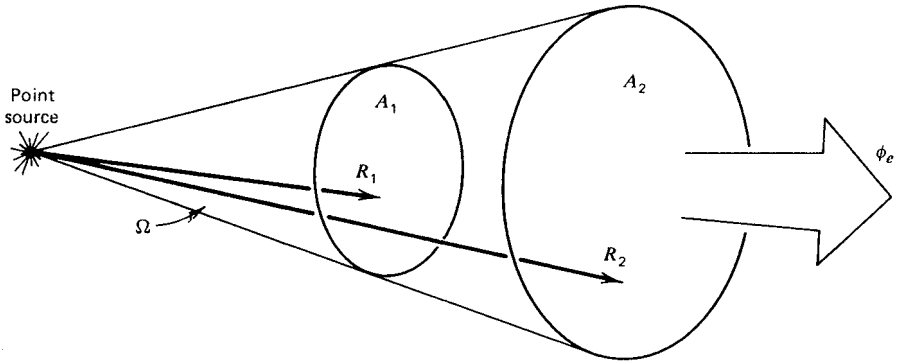


Figure 3.3. Irradiance of surfaces by a point source.

eye is sensitive and scale these measurements to the response of the eye, the unit of power within the subfield of *photometry* is a *lumen*. If we measure the radiation as a function of frequency or wavelength, the unit is watts/Hz or watts/nm. This is the subfield of *spectral radiometry*. (Section 3.5 contains additional discussion of these two subfields.)

The quantity that describes the radiant output of a *point source* is *intensity* (I_e). It is the power ϕ_e emitted per unit solid angle Ω and has the units of watts/steradian, that is

$$I_e \equiv \frac{d\phi_e}{d\Omega}. \quad (3.3)$$

If the source is positioned a distance R_i from a surface, the illumination of that surface is given by the *irradiance* (E_e), the radiant flux density (power per unit area) given in watts/m²,

$$E_e = \frac{\phi_e}{A_i}. \quad (3.4)$$

The irradiance of a surface by a point source can be expressed in terms of its intensity. Since the intensity of a point source does not change with angle, then $I_e = \phi_e/\Omega$. Thus $\phi_e = I_e\Omega$. Now at a distance R_i from the point source the area subtended by the solid angle Ω is $A_i = \Omega R_i^2$, as noted in Fig. 3.3. Inserting these two expressions for ϕ_e and A_i into Eq. 3.4 for E_e , we obtain the *inverse square law*,

$$E_e = \frac{\phi_e}{A_i} = \frac{I_e\Omega}{R_i^2\Omega} = \frac{I_e}{R_i^2}. \quad (3.5)$$

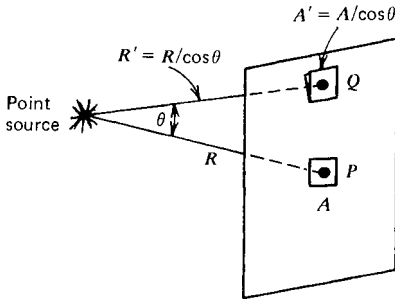


Figure 3.4. Irradiance of a plane surface by a point source.

Thus, the irradiance from a point source on a surface varies as the inverse square of the distance from the point source to the surface [Problem 3.1]. It follows a spherical surface with a point source at its center is a surface of constant irradiance.* But most radiometry problems involve plane surfaces. How does one determine the irradiance on a plane due to a point source located a distance R from the plane, as shown in Fig. 3.4? First, note that compared to the point at the foot of the perpendicular, where the irradiance is I_e/R^2 , the distance to the screen at an angle θ to the perpendicular is greater than R by

$$R' = \frac{R}{\cos \theta}.$$

Also the area A' illuminated from a unit solid angle is slanted, and thus one dimension is larger than an area A normal R' by

$$A' = \frac{A}{\cos \theta}$$

Thus the irradiance on the plane at an angle θ to R is

$$\begin{aligned} E'_e &= \frac{\phi_e}{A'} = \frac{I_e \Omega}{A/\cos \theta} = \frac{I_e A/(R')^2}{A/\cos \theta} = \frac{I_e \cos \theta}{R^2/\cos^2 \theta}, \\ E'_e &= \frac{I_e}{R^2} \cos^3 \theta = E_e \cos^3 \theta. \end{aligned} \quad (3.6)$$

Thus, the irradiance falls off as $\cos^3 \theta$ from the irradiance E_e at the point closest to the source. The simple manipulations of these quantities should give you some idea of the type of calculations that must be made to determine the distribution of light over a surface [Problem 3.2].

* The term "intensity" should only be used with point sources. The amount of radiation falling on a surface (usually termed intensity) is the *irradiance* (W/m^2), independent of the type of source.

3.1.2. Radiometry of Extended Sources

Although it is easy to describe a point source, it is difficult to obtain anything approximating one with any substantial intensity. The reverse is true for an *extended source*—any source of radiation with measurable dimensions. All sources we use are extended sources; but how can we describe them? From projector lamp filaments to the sun, all sources have some degree of variation in the amount and direction of emission with position on the surface unless they are carefully constructed and masked.

A measure of the “brightness” of a source is its *radiance*. The term “brightness” really only applies to visible sources, but if we extend the term to include regions outside the visible, we can use the idea to gain an understanding of radiance. Radiance (L_e) is the amount of power ϕ_e per unit source area as radiated into a unit solid angle Ω ,

$$L_e = \phi_e / A_s \Omega. \quad (3.7)$$

It may help to think of the source as a large collection of point sources distributed uniformly across the source area, so the radiance is a measure of the intensity per unit of source area (and by the definition of intensity, it is a measure of the power per solid angle per unit source area). This concept of an extended source as a distribution of point sources can be used in making some assumptions about the properties of real sources. Suppose that each small area ΔA of an extended source contains a point source, as shown in Fig. 3.5. That is, $I_e = \phi_e / \Omega = \text{constant}$ for all directions. Then the brightness of the source in the direction normal to the surface ($\theta = 0$) is $L_e(0) = \phi_e / \Delta A \Omega$. But if one observes this source at some angle θ to the surface normal, the point sources would appear to be more tightly packed together in the direction of the observer (look at the array in Fig. 3.5 at a glancing angle). Thus the ΔA areas would be reduced to $\Delta A \cos \theta$ and the brightness should be

$$L_e(\theta) = \frac{\phi_e}{\Delta A \cdot \Omega \cdot \cos \theta} = \frac{L(0)}{\cos \theta}.$$



Figure 3.5. Extended source as a collection of point sources.

A source should look brighter at a glancing angle than it does from straight on. But in most cases, this is not the case. If anything, a source either maintains a constant brightness or becomes dimmer as the point of observation moves away from the source normal.

As an approximation to real sources, we assume that the intensity of each area small enough to be considered a point source falls off as $\cos \theta$, instead of radiating equally in all directions, that is

$$I_e(\theta) = \frac{\phi_e}{\Omega} \cos \theta \quad (\text{Lambert's Law}) \quad (3.8)$$

then the brightness of the source is independent of angle,

$$L_e(\theta) = \frac{\phi_e \cos \theta}{\Delta A \cos \theta \Omega} = \frac{\phi_e}{\Delta A \Omega} = L(0).$$

Such a source is referred to as a *Lambertian source*. While most sources are not truly Lambertian, many are so well approximated by the description that the concept is useful in radiometric calculations.

To calculate the irradiance on a screen at some distance from an extended source, we again appeal to the model of the extended source as a collection of point sources. A source of brightness L_e and area A_s is located a distance R from a point P on a screen, as shown in Fig. 3.6. Since all points on S contribute to the irradiance at P , the irradiance due to each point source is $E_e = I_e/R^2$. For an extended source of radiance L_e , the intensity of each point source is $I_e = L_e \Delta A$, where ΔA is a small area on the source.

$$E_e = \frac{L_e \Delta A}{R^2}.$$

If we sum over all the sources, they are equal to the source area, A_s ,

$$E_e = L_e \cdot \frac{A_s}{R^2} = L_e \Omega, \quad (3.9)$$

where Ω is the solid angle of the source from the point P . Note that the actual locations of the source and screen do not enter the equation, only the solid

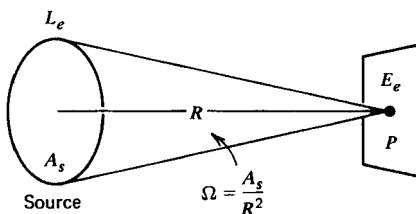


Figure 3.6. Irradiance of a plane surface by an extended source at a point normal to the source plane.

angle. Therefore, if the original source is replaced by an even larger source of the same brightness at some distance from the screen and an aperture is inserted between the source and screen having the same solid angle subtense at the screen as the original source, irradiance will be the same as with the original source! As long as the aperture is filled (or “flushed” as it is termed) and subtends the same angle at the screen as the original source did, the irradiance on the screen will be the same [Problem 3.3].

All of the above applies only to evaluation of the irradiance at a point on the screen closest to the source or the source defining aperture. What is the irradiance at a point elsewhere on the screen? The analysis is similar to that for the $\cos^3\theta$ law for the point source. As may be seen in Fig. 3.7, the area A_s of the source S as observed from point P' is foreshortened by a factor of $\cos\theta$ and the distance is increased by $1/\cos\theta$, so that the solid angle subtended by the source at a point P is

$$\Omega' = \frac{A_s \cos\theta}{(R')^2} = \frac{A_s \cos\theta}{R^2/\cos^2\theta} = \frac{A_s}{R^2} \cos^3\theta = \Omega \cos^3\theta.$$

The irradiance on a surface normal to R' would then be $E'_e = L_e \Omega' = L_e \Omega \cos^3\theta$. But just as in the point source calculation, the area on the screen is greater than that normal to R' by

$$A' = \frac{A}{\cos\theta}$$

so that

$$E''_e = E'_e \cos\theta$$

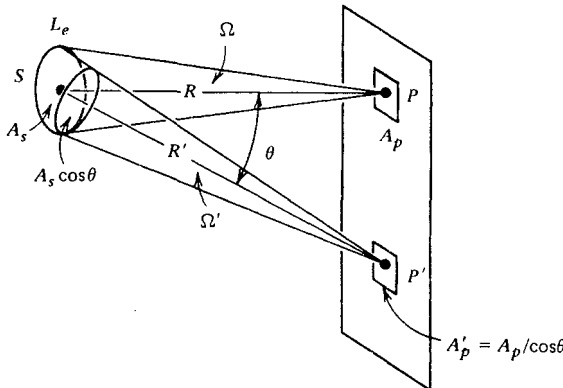


Figure 3.7. Irradiance of a plane surface by an extended source at a general point on the surface.

and

$$E_e'' = L_e \Omega \cos^4 \theta = E_e \cos^4 \theta, \quad (3.10)$$

where E_e is the irradiance at the point closest to the source. This is known as the *cosine fourth law*. Its importance to optical design is that the irradiation on a screen falls off rapidly with angle for a Lambertian source [Problem 3.4].

3.2. RADIOMETRY IN AN OPTICAL SYSTEM

Although the $\cos^4 \theta$ law will work fairly well for everyday sources such as lamps and fluorescent fixtures, in optical system design there are usually lenses or mirrors in a system to provide an image, concentrate the light to a small area, or provide even illumination. In this section the method of calculating the irradiance at various points in optical systems is examined. We start with the point source.

3.2.1. Imaging a Point Source

A point source of intensity I_e is located on the optic axis of a thin lens a distance t_0 from the lens, as shown in Fig. 3.8. The first thing to note is that only a fraction of the light emitted by the source is directed into image space. There are methods of collecting additional light, but the total power ϕ_e in image space will always be less than power emitted by the object [Problem 3.5]. Note also that in object space (to the left of the lens and around its periphery to the right), irradiance would be calculated using the $\cos^3 \theta$ law. Behind the lens in image space, there is, in effect, another point source, the image. The image intensity

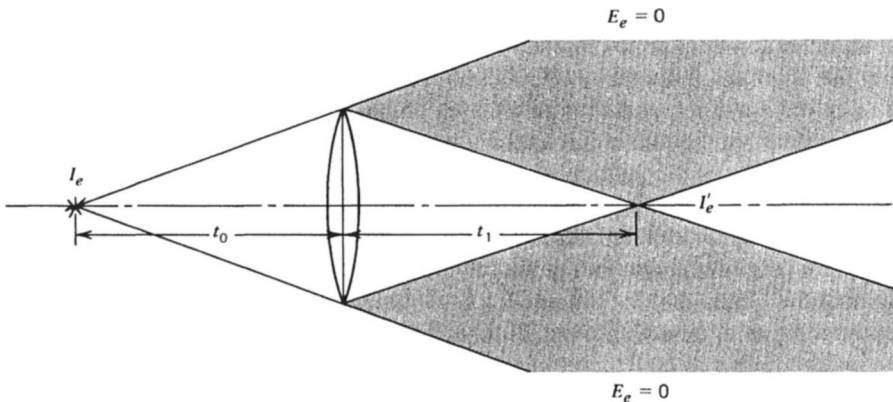


Figure 3.8. Imaging of a point source by a thin lens.

may be calculated by noting that the power ϕ_e collected by the lens within a solid angle $\Omega = A_l/t_0^2$, where A_l is the area of the lens, is refocused to a point in a solid angle $\Omega' = A_l/t_1^2$. Thus

$$\phi_e = I_e \frac{A_l}{t_0^2} = I'_e \frac{A_l}{t_1^2}$$

or

$$I'_e = I_e \cdot \frac{t_1^2}{t_0^2} = M^2 I_e, \quad (3.11)$$

where M is the transverse magnification (Eq. 1.14). Thus the intensity of the image depends on the magnification of the system.

Irradiances in image space may be calculated by treating the image as a point source of intensity I'_e . With the point source intensity I_e the irradiance in image space can then be calculated [Problem 3.6] to be

$$E'_e = I_e M^2 \cos^3 \theta / R^2, \quad (3.12)$$

where R is the distance from the image plane. Note that this holds only inside the solid angle Ω' . *Outside of this region, the irradiance is zero.* This illustrates the major difference between the radiometry of sources and of images. Although the optical system produces secondary sources comparable to the primary sources at a distance from the original, it does so in a restricted region of image space. In addition, the irradiance in image space can be calculated using the distance from the *image* point to the surface in question, even if the screen is between the lens and the image point. This can be seen from the fact that the light intercepted by a plane a distance R from the image point will converge to that point and then diverge to illuminate an equal area at distance R on the other side of the image.

When the optical system is not a thin lens, these calculations must be recast in terms of the pupils of the system. As we have seen, the amount of flux collected by an optical system is determined by the aperture stop. If this is not accessible, then we may determine the solid angle of the light collected by the system as the area of the entrance pupil divided by the square of the distance between the point source and entrance pupil, as shown in Fig. 3.9. In image space, the area that will determine the solid angle is the exit pupil. In image space with your eye anywhere inside the solid angle Ω' you can look through the lens and see the entire exit pupil illuminated. As far as the image of the point source is concerned, the image intensity is due to a uniformly illuminated exit pupil radiating toward the image. Therefore, in radiometric calculations an optical system can be replaced by its entrance and exit pupils. (Transmission losses are accounted for by multiplication of final result by the transmittance T of the system.) While this may seem like a statement of little consequence, it allows considerable simplification of a problem.

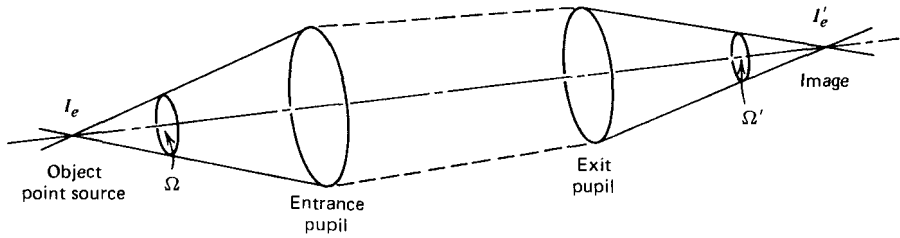
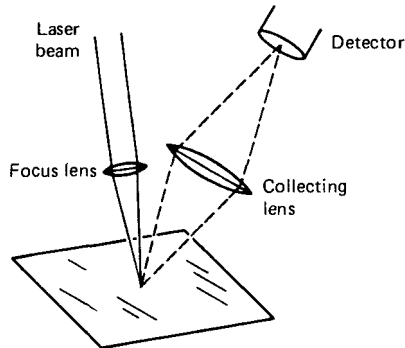


Figure 3.9. Imaging of a point source by an optical system.

EXAMPLE 3.A. A laser beam is focused on a small spot on a slab of material creating a visible point of light on the slab that radiates into a hemisphere (2π sr). If a lens images this spot onto a detector and if the magnification is unity (the lens focal length is 8 cm and its diameter is 3 cm), what fraction of the radiated power will be collected and imaged onto the detector?



Solution: We need to determine the solid angle of the collection system.

$$\Omega_1 = \pi \left(\frac{3}{2} \right)^2 \frac{1}{R^2} \quad \text{If magnification is unity, then } t_0 = -2f. \\ \therefore R = 2f = 16 \text{ cm.}$$

$$\Omega_1 = \frac{\pi}{4} \cdot \frac{9}{(16)^2} = \frac{\pi 9}{4(16)^2} \text{ sr}$$

The total solid angle that the source radiates into is 2π sr. The fraction is then:

$$\frac{\pi 9}{4(16)^2} \cdot \frac{1}{2\pi} = \frac{9}{8(16)^2} = 0.0044.$$

About $\frac{1}{2}\%$ is directed onto the detector.

It should be noted that not all sources that are considered point sources radiate into 4π steradians. If a source underfills a lens, as in the case of a laser beam or the intermediate image of a point source, then (with the exception of transmission losses) all the power is transferred into image space. If a source overfills a lens, then the power transferred is proportional to the ratio of the collection optics solid angle divided by the output solid angle of the source. An example of a calculation involving point sources is given in Example 3.A.

3.2.2. Imaging an Extended Source

As noted above, the irradiance due to an image point source is zero outside the solid angle into which the image is focused and diverged. The same is true for images of extended sources. In image space, the observer will see an image only if the eye is in a region where light from the image is directed. The radiometry of images applies only in regions where the optics has redirected the light.

An optical system made up of passive components (lenses, mirrors, prisms, etc.) can redirect radiant energy, but it cannot increase it. (A light amplifier, a laser, or an image intensifier is needed for that.) Therefore, with the exception of absorption and reflection losses, the power out of the optical system ϕ'_e is less than or equal to the power in, ϕ_e :

$$\phi'_e = T\phi_e, \quad (3.13)$$

where T is the transmittance of the system, $0 < T < 1$.

In Fig. 3.10, light from a source of brightness L_e is collected by a lens to produce an image of brightness L'_e . What is the relation between source and image brightness? First, from Eq. 3.7,

$$\phi_e = L_e \cdot A \cdot \Omega$$

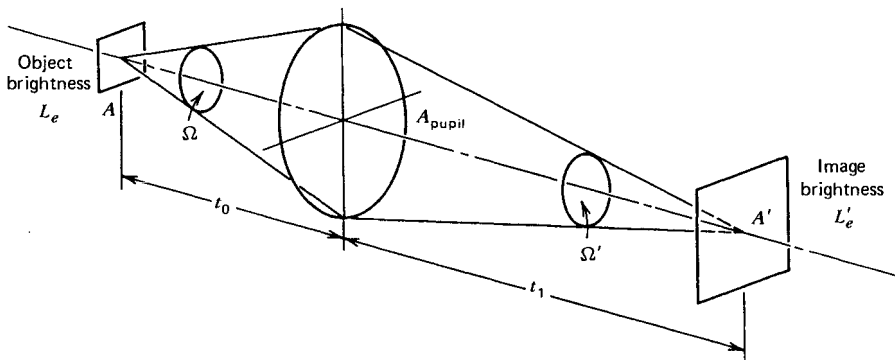


Figure 3.10. Imaging of an extended source by a thin lens.

and

$$\phi'_e = L'_e \cdot A' \cdot \Omega',$$

where L_e is the source brightness, A is the source area, and Ω is the solid angle subtended by the entrance pupil. The primed quantities are the corresponding values for image space. The area of the source is related to the area of the image by the magnification of the system:

$$\frac{A'}{A} = M^2 = \frac{t_1^2}{t_0^2},$$

and the ratio of the solid angles can also be expressed in terms of the square of the magnification:

$$\Omega = \frac{A_{\text{pupil}}}{t_0^2} \quad \Omega' = \frac{A_{\text{pupil}}}{t_1^2}$$

or

$$\frac{\Omega'}{\Omega} = \frac{t_0^2}{t_1^2} = \frac{1}{M^2}.$$

But $A/A' = 1/M^2$, therefore

$$A\Omega = A'\Omega'. \quad (3.14)$$

The product of the source area times the solid angle into which it radiates is referred to as the *throughput* of the system. This last equation tells us that it is equal to the corresponding throughput of the image. It can also be shown [Problem 3.7] that the solid angle subtended by the source or the image, Ω'' at the lens multiplied by the corresponding pupil area A_{pupil} is also equal to the throughput

$$A\Omega = A'\Omega' = A_{\text{pupil}}\Omega''. \quad (3.15)$$

Using Eq. 3.13 and the throughput relation, Eq. 3.14, we find that

$$L'_e \cdot A'\Omega' = TL_e A\Omega,$$

$$L'_e = TL_e,$$

or, since $T \leq 1$

$$L'_e \leq L_e. \quad (3.16)$$

Thus the brightness of an image can never exceed the brightness of its object.

The above statement, referred to as the *Brightness Theorem*, is, in effect, an optical form of the law of conservation of energy. No amount of repositioning of optical elements can make the image brighter than the object. For example, note that: If the lens is moved closer to the object to collect more light (ϕ_e is larger because Ω is larger), the image is magnified so the energy is spread into a larger area A' . If the lens is moved closer to the image to reduce the image size, the collection angle is smaller, reducing the light flux. And if the size of the lens is increased, the power through the system goes into a larger solid angle Ω' , so the brightness remains the same.

Then why does anyone buy larger, more expensive lenses for his or her camera? Because increasing Ω and Ω' means the irradiance on the film E'_e is larger, since $E'_e = L'_e \Omega'$. Since L'_e is constant, as we have just argued, the larger the solid angle, the greater E'_e , and since the exposure of a film is the product of the irradiance E'_e and the exposure time, an increase in E'_e enables one to decrease the exposure time, capturing motion or preventing a blur due to camera movement.

3.3. RADIOMETRY OF EXTENDED SOURCES

While the Brightness Theorem sets limits on image brightness, it does not tell us the irradiance distribution in the image plane or elsewhere in image space. From Fig. 3.10 and Eqs. 3.7. and 3.14, we find that the irradiance in the image plane is

$$E'_e = \frac{\phi'_e}{A'} = \frac{L'_e \cdot A' \cdot \Omega'}{A'} = T \cdot L_e \Omega', \quad (3.17)$$

since $L_e \Omega = E_e$ and $\Omega' = \Omega/M^2$, the image irradiance is

$$E'_e = \frac{T}{M^2} E_e. \quad (3.18)$$

Thus the image irradiance can be increased by demagnification, but the details of the image are harder to see and harder to record.

What happens to the irradiance off axis? The power into the lens from a point at an angle θ between the source normal and the lens axis is

$$\phi_e = L_e \cdot A \cdot \Omega \cos^4 \theta,$$

and the irradiance is

$$E'_e = \frac{\phi'_e}{A'} = \frac{T \phi_e}{A'} = T L_e \left(\frac{A \Omega}{A'} \right) \cos^4 \theta.$$

Using the throughput relation (Eq. 3.14), the quantity in parentheses can be replaced by Ω' , so that

$$E'_e = TL_e \Omega' \cos^4 \theta. \quad (3.19)$$

Note that with the exception of the transmission factor T , this is the same as if the source were located in place of the lens. This holds until the edge of the image is reached and then $E'_e = 0$, assuming no diffraction or lens aberration. In the simple example shown in Fig. 3.10, the lens is the aperture stop and the pupils. The field stop (and entrance window) is the edge of the source, and the exit window is the edge of the image. Thus the only reduction in image irradiance is due to the cosine fourth law.

3.3.1. Vignetting

In more complicated systems the field stop is defined by some aperture in the system. In Example 2.C, given in the previous chapter, the front lens was found to be the aperture stop and the back lens to be the field stop. Given a source of unlimited extent and brightness L_e located 300 mm from the positive lens, we would like to calculate the distribution of light in the image plane. For the point on the screen on axis, only the aperture stop limits the rays (Fig. 3.11(a)). But at a certain angle the back lens limits the rays, as shown in Fig. 3.11(b)). To analyze what is happening, we can replace the two lenses and the source by an exit window $P_3 P_4$ (here the field stop) and an illuminated exit pupil $P_1 P_2$. Because the source has unlimited extent, the exit pupil is always filled with rays at all angles. Thus the calculation of the on-axis and off-axis points is simplified to that shown in Fig. 3.11. Note that in part c of the figure, the exit window begins to cut off the bottom of the ray bundle. This reduction in the effective aperture stop of the system is called *vignetting* (vin-yét-ting). At a very high angle such that a ray from P_1 hits P_4 , there will be no light passing through the optical system. If you positioned your eye on the optic axis and looked back through the system, you would see the illuminated exit pupil centered in the exit window. As you move your eye down to point B , the exit pupil moves relative to the exit window to a point where the bottom edges of the two circles are just touching. Beyond this, moving toward D , progressively more of the exit pupil is vignetted until it is completely cut off. A plot of the irradiance as a function of angle is given to the right of Fig. 3.11.

The value of the irradiance in the image plane can be written as

$$E'_e = TL_e \Omega' V(\theta) \cos^4 \theta, \quad (3.20)$$

where $V(\theta)$ is the fraction of the solid angle *not* vignetted. As the angle θ is increased, the cross-section of the solid angle becomes strongly elliptical, but because the $\cos^4 \theta$ factor already accounts for a foreshortening of a source dimension and therefore accounts for the elliptical shape, the factor $V(\theta)$ is just

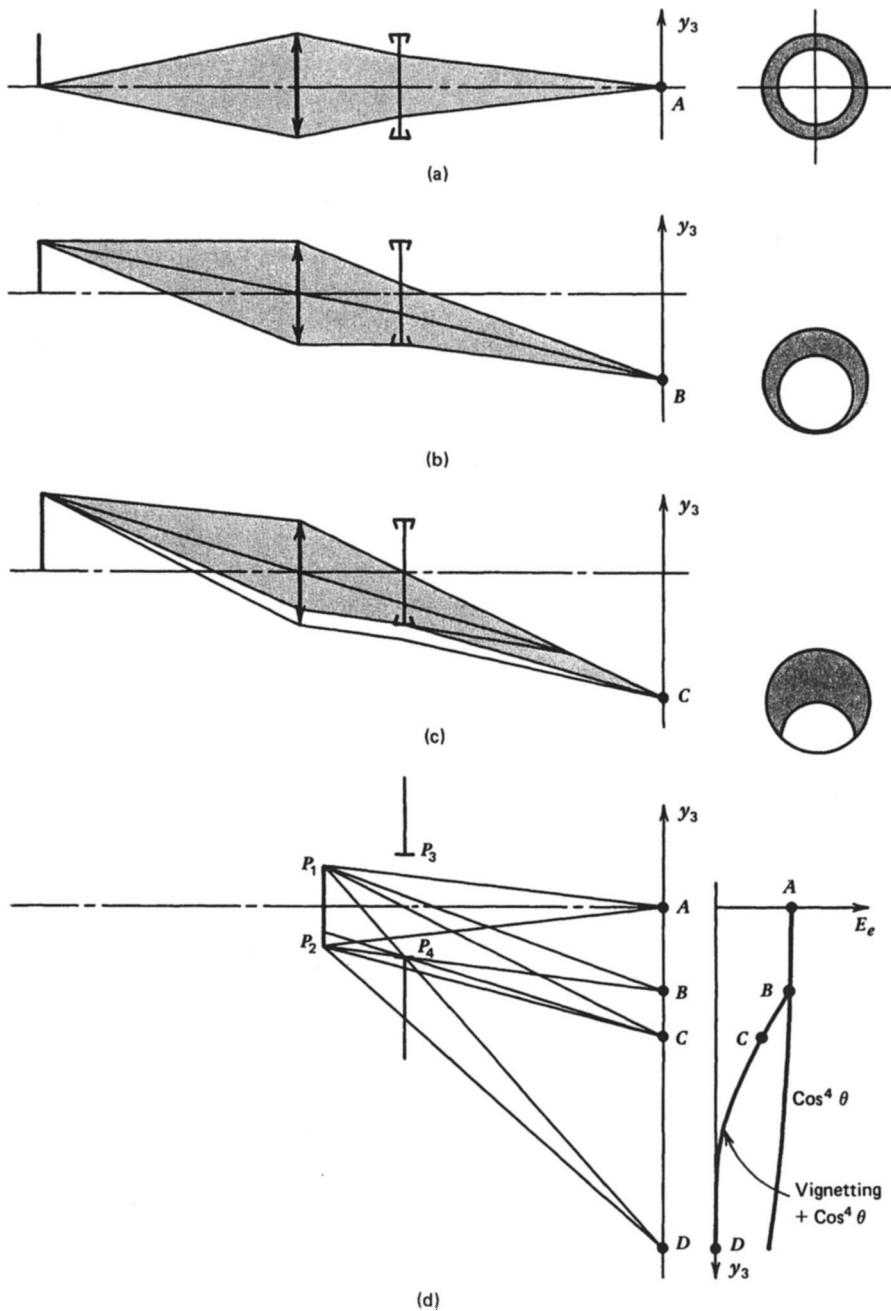


Figure 3.11. Vignetting. (a)–(c) Left side—marginal ray traces of an extended source seen from different points in the image; right side—view of the exit pupil as seen through the exit window from different points in the image plane. (a) on the optic axis, (b) at the point where vignetting begins, (c) at a point where there is substantial vignetting. (d) Left side—ray traces for the above three points plus an additional point where vignetting is complete; right side—plot of irradiance versus distance from the axis corresponding to the points illustrated in the left-side figure.

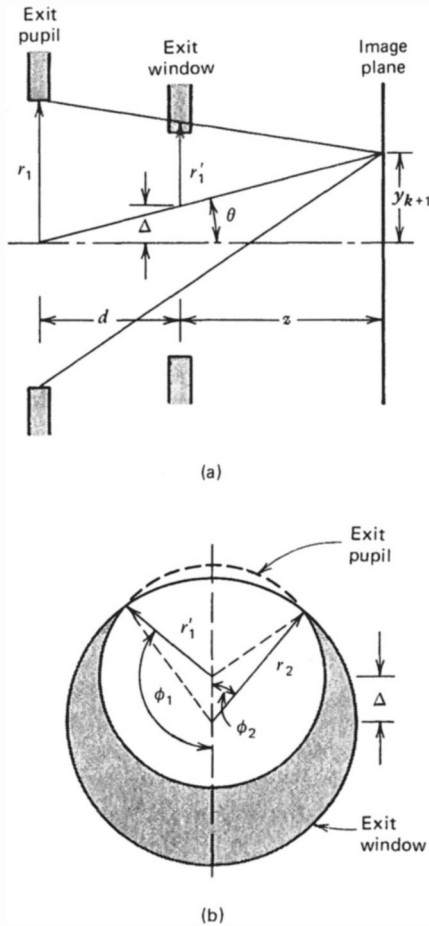


Figure 3.12. Vignetting. (a) Projection of exit pupil onto the exit window. (b) Off-axis view of exit window and pupil. Δ is the separation between the center of the exit window and the center of the projection of the exit pupil.

the overlap of two circles, the exit pupil and window, of different radii and centers. To account for the separation between the circles, the exit pupil is projected onto the exit window plane, as shown in Fig. 3.12(a), so that its radius is reduced from r_1 to

$$r_1' = \frac{z}{d+z} \cdot r_1, \quad (3.21)$$

where z is the exit window to image distance and d is the separation between the exit pupil and the exit window. The angle θ between a ray from the center of the exit pupil to a point y'_{k+1} on the image and the axis of the system is given by

$$\tan \theta = \frac{y'_{k+1}}{d+z}. \quad (3.22)$$

and the separation between the two circles Δ is

$$\Delta = d \tan \theta = \frac{d}{d+z} \cdot y'_{k+1}. \quad (3.23)$$

The area between the chord of a circle and the arc is the segment of the circle. The unvignetted area is the sum of the areas of two segments of the circles as shown in Fig. 3.12(b), whose half angles can be found from

$$\cos \phi_1 = \frac{(r_1'^2 + \Delta^2 - r_2^2)}{2r_1'\Delta}, \quad (3.24a)$$

$$\cos \phi_2 = \frac{(r_2^2 + \Delta^2 - r_1'^2)}{2r_2\Delta}. \quad (3.24b)$$

A preliminary inspection of the overlap of the two circles would seem to require two cases, one for Δ greater than the smaller of the two radii and one for Δ less than this quantity. In fact, the overlap area for both cases can be expressed as [Problem 3.8]

$$A(\theta) = \phi_1 r_1'^2 + \phi_2 r_2^2 - r_1' r_2 \sin(\phi_1 + \phi_2), \quad (3.25)$$

where

$$(|r_1' - r_2| \leq \Delta \leq r_1' + r_2).$$

The vignetting factor in Eq. 3.20 above is then

$$V(\theta) = V(\Delta) = \frac{A(\theta)}{\pi r^2}. \quad (3.26)$$

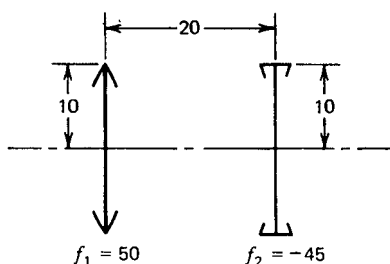
where r is the smaller of r_1' and r_2 . Note that for each value of $V(\theta)$, Δ has to be determined and then ϕ_1 and ϕ_2 found before the overlap area can be calculated. Obviously, this is the type of a problem a small computer or programmable calculator can solve with ease. Starting from the axis the solid angle could be evaluated and r_1' found. The calculation of $V(\theta)$ does not begin until the angle where $\Delta = |r_1' - r_2|$ (edges of the two circles touching). To that point the $\cos^4 \theta$ factor is the only correction. Beyond that and until $\Delta = r_1' + r_2$ (complete cutoff of light), Δ is incremented by small amounts to provide a smooth plot in a minimum amount of time. (Where on the image plane in Fig. 3.12(a) will the arrangement in part (b) be seen?) An application of this technique to Example 2.C is given in Example 3.B. Note that the function $V(\theta)$ could be approximated by a straight line between the points when $\Delta = |r_1' - r_2|$ and $\Delta = r_1' + r_2$. The

greatest error occurs at the smaller values of $V(\theta)$ but in that region the cosine fourth factor reduces it. Thus a linear approximation of $V(\theta)$ by

$$V(\theta) = \frac{(r'_1 + r_2) - \Delta}{(r'_1 + r_2) - |r'_1 + r_2|} \quad (3.27)$$

can provide a good approximation to the more exact calculation in much less time [Problem 3.9]. In the example given above and in Example 3.B the exit window is closer to the image plane than the exit pupil. This is not always the case; sometimes they are reversed [Problem 3.10]. In such cases the same procedure is followed except that the exit window is projected on the exit pupil. It should also be noted that the irradiated plane need not be in the image plane. For an extended source the system can be replaced by its illuminated exit apertures and any plane can be evaluated using vignetting and the cosine fourth law.

EXAMPLE 3.B. What is the degree of vignetting at 25° to the optic axis for the system described in Example 2.C?

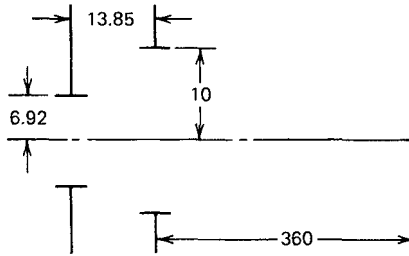


Solution: From Example 2.C, we know that the positive lens is the aperture stop and that the negative lens is therefore the field stop and exit window. The exit pupil is found by finding the image of the positive lens in image space. Using the thin lens formula,

$$t_0 = -20, \quad f = -45 \rightarrow t_1 = -13.85,$$

$$M = 0.692, \quad r_1 = 6.92.$$

STEP 1. Thus, we can replace the lenses with two apertures as shown. The distance from the exit window to the image is $z = t_2 = 360$ mm (see Example 2.C).



Projecting the exit pupil onto the exit window, its radius is reduced (by Eq. 3.21) to:

$$r'_1 = \frac{z}{d+z} \cdot r_1 = \frac{360}{13.85 + 360} \cdot 6.92 = 6.66 \text{ mm.}$$

STEP 2. The size of $\Delta = d \tan \theta$ (Eq. 3.23)

$$\Delta = 13.85 \tan 25^\circ = 6.458.$$

STEP 3. Since r_2 , r'_1 , and Δ are known, ϕ_1 and ϕ_2 can be calculated from Eqs. 3.24 to be

$$\phi_1 = 1.733 \text{ rad } (99.3^\circ),$$

$$\phi_2 = 0.717 \text{ rad } (41^\circ).$$

The overlap area is (using Eq. 3.25)

$$A(25^\circ) = 106.16.$$

The unvignetted area is $\pi r_1'^2 = \pi 6.66^2 = 139.3$. Thus the vignetting factor is

$$V(\theta) = \frac{A(\theta)}{\pi r_1'^2} = \frac{106.16}{139.3} = 0.76, \text{ or } 76\%.$$

If one includes the $\cos^4 \theta$ factor ($\cos^4 \theta(25^\circ) = 0.675$), the reduction in illumination from the on axis point is

$$0.76 \times 0.675 = 0.51.$$

At 25° off-axis the irradiance is half the on-axis value with both $\cos^4 \theta$ and vignetting contributing comparable amounts.

The preceding analysis is somewhat simplified in that only two apertures (or their images) are considered. In some cases, as the observation point moves off axis, the edges of other apertures may interrupt the solid angle subtended by the already vignetted sources. Although the calculation involving the additional sectors increases the difficulty of the computation, the method of calculation remains the same. For rough calculations, a linear extrapolation between two values of Δ that represent the end points of the vignetting region may prove adequate.

3.4. APPLICATIONS

The purpose of this section is to show how radiometry can be applied to some practical situations. Beginning first with the radiometer, we will then look at searchlights and illuminators.

3.4.1. Radiometer

A *radiometer* is a device that measures radiant power. It consists of a light sensitive detector, usually an absorbing black surface and a transducer that produces a signal caused by the heating of the surfaces as described in Section 10.2; a set of concentric apertures or baffles, which define the amount of radiation measured and reduce extraneous radiation from other sources; and a calibrated meter or display to provide a readout, as shown in Fig. 3.13(a). While

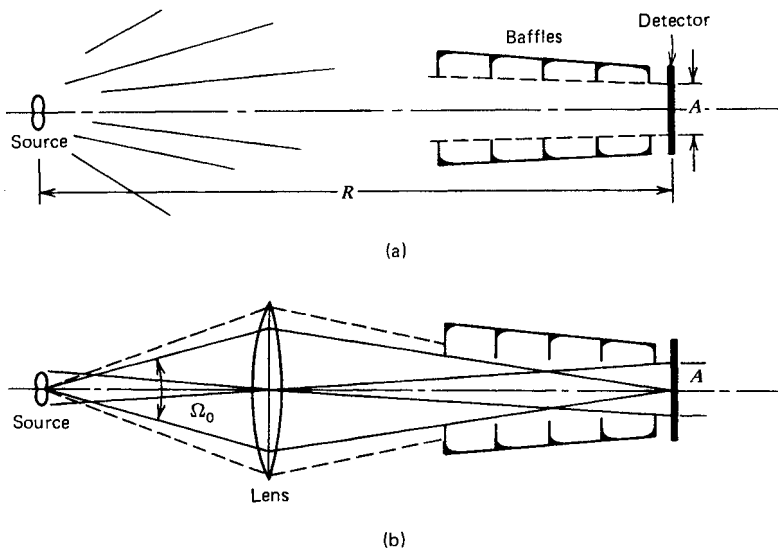


Figure 3.13. Radiometer. (a) Direct measurement of irradiance of an extended source. (b) Measurement of an extended source with improved sensitivity by imaging the source onto the detector.

a radiometer is relatively insensitive to the wavelength of radiation (otherwise it is not a radiometer, but a narrow-band power meter), the region over which this is true should be determined either from specifications or experiment. To be able to use a radiometer intelligently, the acceptance angle, detector surface area, and location of the detector within the radiometer head should also be known.

Usually a radiometer will be used in one of two configurations. In the first, it will be positioned a distance R in front of a source and the power ϕ_e falling on the detector surface (the area, A) will give a value for the irradiance $E_e = \phi_e/A$ at distance R , as shown in Fig. 3.13(a). From the geometry of the source and assumptions as type of source (isotropic, Lambertian, etc.), a value of the total power emitted by the source may be found. If the source is too weak, a second configuration, shown in Fig. 3.13(b), in which a lens is used to image the source onto the detector, is appropriate. In the second case the fraction of source power measured must be determined. Here, however, the entrance pupil and its distance from the object must be found so that the solid angle collected by the radiometer can be found. By determining the fraction of the source radiation that enters the radiometer, the total radiated power of the source may be calculated. Whether the image of the source underfills or overfills the detector will determine whether the source or the image of the detector surface at the source is the entrance window. Thus, in most radiometric calculations, the measurement hinges on the correct assessment of the geometry and application of certain assumptions. In Fig. 3.13(b), the source is smaller than the detector image and, therefore, the source outline is the field stop. An example of calculations for this configuration is given in Example 3.C.

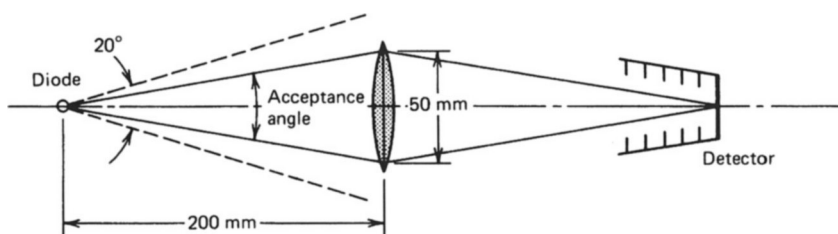
3.4.2. Searchlights

In the realm of geometrical optics, a point source at the focal point of a lens will produce parallel rays showing no divergence. However, without point sources and aberration-free lenses, we are forced to produce collimated light as best we can and analyze the results to understand how brightly we can illuminate a remote area, actors on stage, or the night sky around shopping malls.

A *searchlight* consists of an extended source located at the focal point of a lens or mirror, as shown in Fig. 3.14(a). To simplify matters we assume the source has a circular shape. Again the replacement of source and lens by the exit pupil and exit window makes the analysis simple. In this case, the aperture stop (and therefore, exit pupil) is the lens. The exit window is then the virtual image of the source, which is located at infinity, but whose angular extent is the same as that of the source located one focal length f from the lens, as shown in Fig. 3.14(b). Since the image of the source is at infinity, all rays coming from the top of the source are parallel to each other. The same is true for all other points on the source. If the above argument is less than convincing, consider that if you cover the region to the left of the lens for Fig. 3.14 parts (a) and (b), the ray paths look the same for both figures. Both the original source and the image at infinity subtend a solid angle $\Omega_1 = \pi r^2/f^2$, where r is the radius of the circular source.

EXAMPLE 3.C. A light emitting diode (LED) with an internal collimator design emits in a 20° cone (full angle). The broadband detector with a 10° (full angle) acceptance angle is used to measure the output. Set up the geometry to measure the light output of the diode. (Radiometer data: 1-cm² circular detector, 1-cm diameter entrance aperture.)

Solution: Unless the radiometer is very sensitive, a lens is needed to image the object into the detector. Assume $T = 0.92$ for a glass lens. We choose a 10-cm lens and set it up in a unit magnification geometry



($-t_0 = 2f = t_1$). If the radius of the lens is 25 mm, the solid angle subtended by the lens at the diode is:

$$\Omega = \frac{\pi(25)^2}{(200)^2} = 0.05 \text{ sr.}$$

The diode emits into an angle of 20° . This 20° cone corresponds to a solid angle of $\pi \tan^2 10^\circ = 0.10 \text{ sr}$. Thus the lens will intercept approximately half the diode output. Because the diode is so small and the detector area is so large the entire image will fall on the detector. The output of the diode will be

$$\phi_{\text{out}} = \frac{\phi_{\text{measured}}}{0.92 \times 0.50}.$$

One region of particular interest is near the lens. Note that within region *A* in Fig. 3.14 the entire disk of the source will be seen. Since the source is at infinity, a change of position along the optic axis does not change the size of the solid angle subtended by the source. Thus, anywhere in region *A*, Ω' is constant, and since L_e is constant, the illumination E_e is constant within this area. Beyond a distance d_0 , where the solid angle just fills the exit pupil (i.e., when $d_0^2 = A_{\text{lens}}/\Omega_1$; check this out for yourself), the illumination along the axis will be $L_e \Omega'$, where Ω' is now equal to A_{lens}/d^2 and d is the lens to observation point

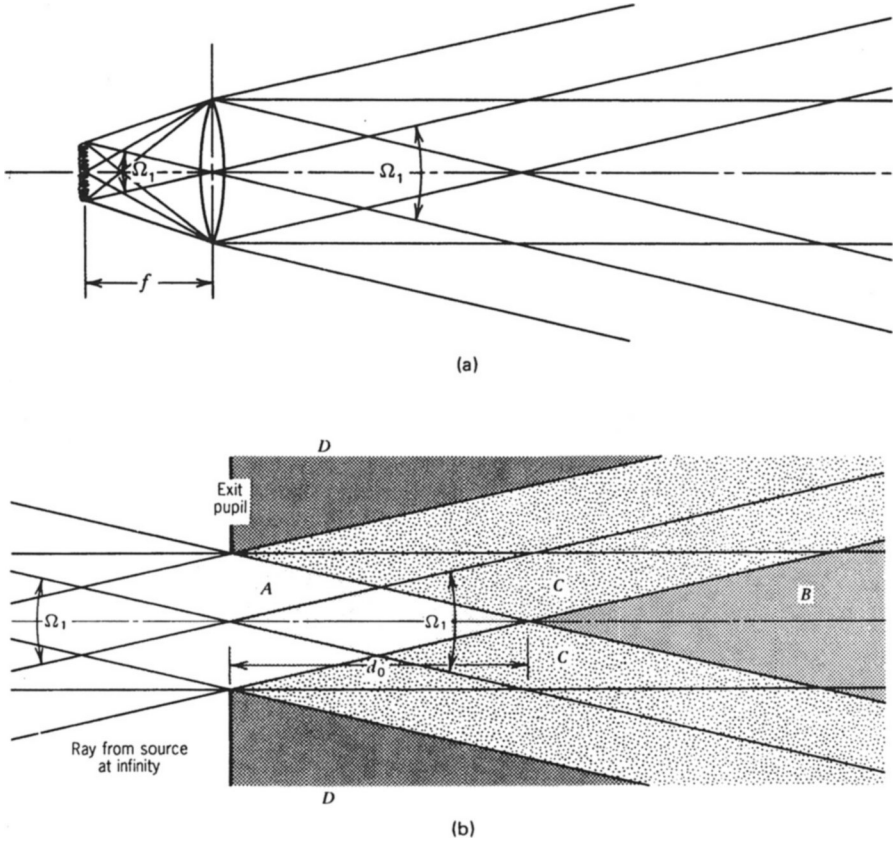


Figure 3.14. Searchlight. (a) Ray tracing of center and marginal rays for an extended source at the focal point of a lens. (b) Replacement of source and lens by a virtual object at minus infinity and an exit pupil. Region *A* is a region of constant irradiance, region *B* is a region where a point source approximation may be used, region *C* is a vignetting region, and *D* is the shadow region.

distance. Therefore, in region *B*, the lens acts like a point source of intensity $L_e A_{\text{lens}}$ [Problem 3.11].

In region *C* vignetting occurs, and the irradiance is determined by the fractional area of the source as seen through the hole that is the exit pupil (the lens). This is another vignetting problem, but the second aperture cannot be assigned a finite distance, only a finite subtense Ω_1 . Thus, at any distance d from the lens, the source subtends a solid angle Ω_1 . Therefore, the projection of the source on the exit pupil plane is a circle of radius $r'_1 = \sqrt{d^2 \Omega_1 / \pi}$. In the case of the searchlight, the separation between the centers of the circles Δ is equal to the distance y' between the point of observation and the lens axis. So the irradiance in region *C* can be treated as a vignetting problem, but with r'_1 and Δ calculated differently [Problem 3.12] and with no cosine fourth dependence.

In summary, it is possible to separate the searchlight irradiation into four regions: the region (*A*) of constant irradiation, the region (*B*) of point source irradiation, the region (*C*) of vignetted irradiation, and the shadow region (*D*). In region *C* and beyond distance d_0 , the solid angle subtended by the lens is smaller than the source as it is in region *B*. Thus in region *C* beyond d_0 the irradiance falls off more rapidly than in region *B* because of vignetting in *C* in addition to the $1/d^2$ dependence. Although this discussion of such a simple system may seem more detailed than necessary, it is done with the purpose of illustrating how an optical system may be replaced by its equivalent pupil and window to assess the effect of vignetting. In this case, the values were not found in an image plane (for there is none), but for the entire irradiated region [Problem 3.13]. Although the light from a searchlight can be well collimated,* it cannot compare with the collimation that can be achieved by using a laser source. That subject is explored in Chapter 7.

3.4.3. Illuminators

In many applications the image we produce does not come from a self-luminous source. Instead, it is illuminated by another source. The sun, roomlights, or a flashlamp provide the light in photography. In optical systems, the most interesting cases are transilluminated objects such as a photographic slide, a film in a projector, or samples in a microscope. The ideal solution would be a bright planar source located next to the transparency, but it is impossible to get a source of uniform brightness as large as a frame of motion picture film or a 35-mm slide. Some optics must therefore be used to increase the irradiance at the object and the irradiance at the screen. One of the most important requirements of a good transilluminated system is a nearly uniform illumination of the image plane. While the border of the transparency or film frame may serve as a possible system stop, the uniformity of illumination one would like to achieve is independent of the content of the transparency. Thus, we need only worry about the uniformity of light distribution on a screen. We will examine several illumination methods and compare the results of these methods.

Although the direct illumination method shown in Fig. 3.15(a) is simple in that it only requires a projection lens, it is inefficient in directing radiation to the screen. If the source is small enough to be considered a point source, then only a cone of light equal to the image solid angle would illuminate the center of the screen, leaving the edges dark, as shown in the figure. The irradiance on the screen is given by Eq. 3.12; $E_e = I_e \cdot M^2 \cos^3 \theta / R^2$ for an area inside Ω' . Outside

* It is interesting to note that carbon arc searchlights that one often sees at shopping centers and movie openings usually emit a *diverging* radiation bundle. This is to compensate for the perspective distortion, and the net effect (if the divergence is properly set) is to see a parallel bundle or pencil of radiation proceeding skyward and simply stopping. If the radiation were collimated, it would come to a point when viewed from near the searchlight, much as two railroad tracks converge due to the perspective distortion.

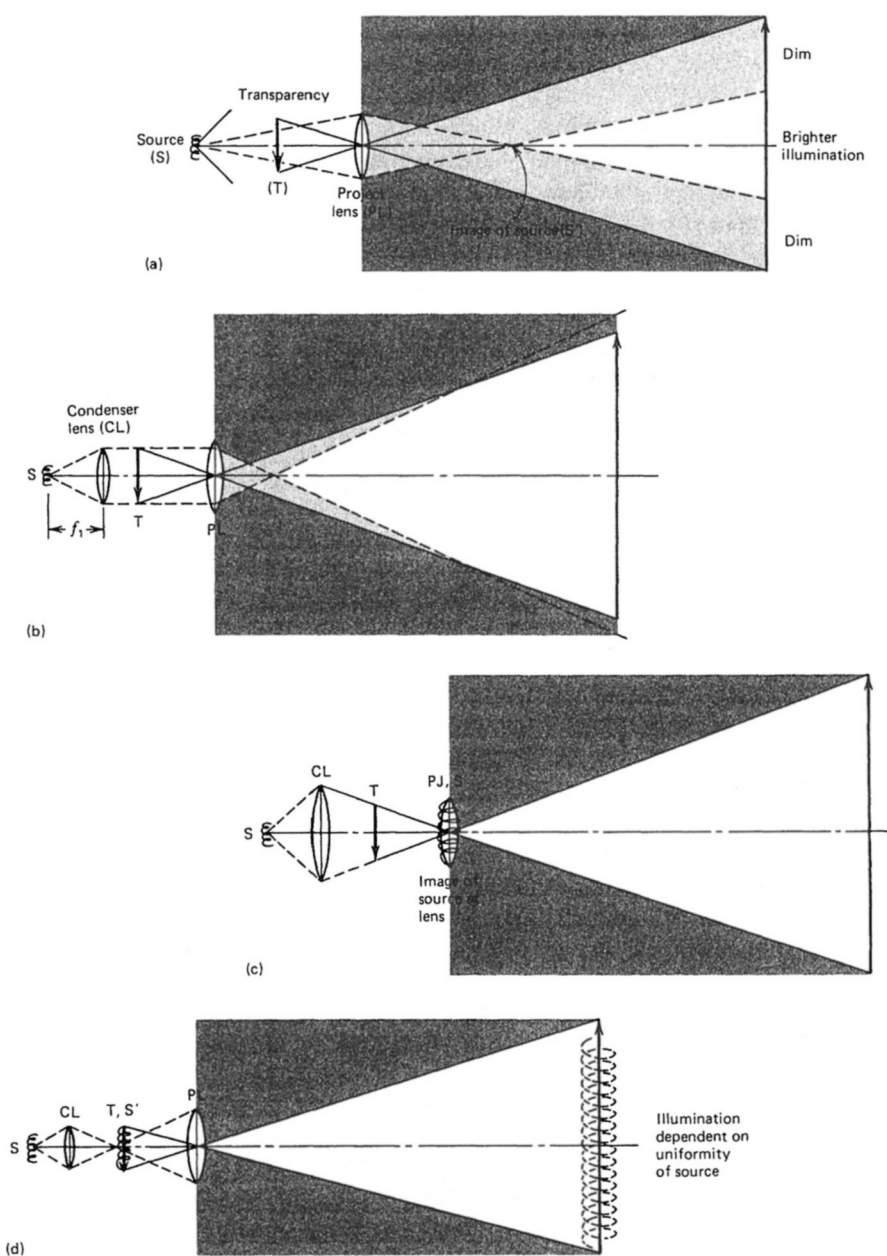


Figure 3.15. Types of illumination. (a) Point source illumination, (b) Collimated illumination, (c) Köhler illumination, and (d) Abbe illumination.

Ω' it is zero. Any attempt to increase E_e by increasing the magnification of the source will cause the image of the source to be located closer to the screen, resulting in a smaller illuminated area. If the surrounding illumination is weak and the magnification of the object modest so that only the narrow central region of the system could be used, such illumination might be used in an inexpensive projector.

A lens may be inserted into the system so that the light from the source is collimated as a mini-searchlight. If the collimation lens is placed close to the object, as shown in Fig. 3.15(b), the object is in the region of constant illumination (region A, discussed in the previous section), and the distribution of light on the screen should be more uniform. However, the illumination is again dependent on the relative sizes and the separation of the exit apertures. From Fig. 3.15(b), it should be easy to see that the film frame *and* its projection on the imaging lens (since the space between them is filled with collimated light) are aperture stops and that the image of the source is the field stop. The focal length of the condenser lens is made as short as possible and the diameter as large as possible to increase its light collection. The focal length of the projection lens is determined by the magnification needed and the projection distance. From a simple ray trace one can see that the light source is magnified as the ratio of the projector focal length over the condenser focal length [Problem 3.14]. Unless the source area is very large, the image area and thus the field stop will normally be small, so that vignetting occurs at relatively small angles.

If the source is moved farther back from the condensing lens, at some point a magnified image of the source is located at the projection lens, as shown in Fig. 3.15(c). This lens is now the field stop, while the aperture stop is the film frame. But the image of the film frame in image space is the screen image, which is also the exit pupil. This is an ideal situation since the image surface and the exit pupil are the same. Provided the film frame is not too large and the projection lens is filled with the source image, there will be little or no vignetting. This configuration, in which the source is imaged into the projection lens, is called *Köhler illumination*. The same configuration can be used in microscopy to image the source into the microscope objective.

There are conditions where Köhler illumination is not feasible. For example, when a small frame of a motion picture film must be imaged onto a large screen in an auditorium or drive-in, the illumination would not be sufficient since the film is placed in the constant irradiance region of the condenser lens. Instead the source can be imaged onto the film frame and then the transparency and source are imaged onto the screen, as shown in Fig. 3.15(d). The problem with this arrangement is that unless the source brightness is extremely uniform, the irregularities in the source will be evident on the screen. This is usually achieved by making the condenser optics slightly irregular to produce a uniformly blurred source image in the film frame. This type of configuration is called *Abbe illumination* and has a counterpart in microscopy. Some microscope illuminators come equipped with a swing-away lens to enable the microscopist to choose between Köhler and Abbe illumination.

One last type of illumination might be referred to as focused illumination. In this case the object is to provide the largest amount of flux into the smallest area. With the sources discussed so far, this is achieved by choosing a lens with a low f -number. The large diameter will collect a large flux and the short focal length lens will produce a small image of the original source [Problem 3.15]. In Chapter 7 we consider another illumination problem, that of focused laser beams. There the approach is somewhat different.

3.5. PHOTOMETRY AND SPECTROPHOTOMETRY

In the preceding sections, the spectral output of the sources was not discussed. The rate of energy output was given in watts; no distinction was made between the radiative power of a source and the power of mechanical or electrical power sources. There are situations, however, when it is necessary to make this distinction, to single out certain wavelength regions and measure radiation only within these regions. For example, the human eye responds to a narrow range of wavelengths, between 400 and 700 nm. If one tried to design a system without taking this fact into consideration, radiometric calculations might show an excellent power flux, but the eye would see nothing.

The subfield of radiometry that deals with the human eye as a detector is called *photometry*. All the concepts that have been discussed thus far still hold. All that will be done is to evaluate the output of a system based on the response of the human eye. This is accomplished by converting radiometric measurements and units to photometric ones or by evaluating a system initially in photometric units. The point of contact between the two fields and systems of units comes through curves shown in Fig. 3.16, that give the relative response of the human eye to different wavelength bands. These curves are called the *spectral luminous efficiency curves*. One of them (the *photopic* curve) peaks at 555 nm in the yellow-green region and falls off toward each end of the spectrum—to 400 nm in the violet and 700 nm in the red. This is the curve for the light-adapted eye. For very dim illumination, a second curve, called the *scotopic* curve, that peaks further to the blue at 510 nm, is used. The connection between radiometric units, valid in all regions of the electromagnetic spectrum, and photometric units, defined only in this narrow region, is that 1 watt of radiometric power at 555 nm is equal to 680 lumens. All other powers are then scaled to the photopic spectral luminous efficiency curve. For example, a 50-mW helium-neon laser, which has an output at 633 nm, has a photometric power of about $50 \text{ mW} \times 680 \text{ lumens/watt} \times 0.25$, or 8.5 lumens. The photometric output of a gallium arsenide diode that operates at 900 nm is virtually zero. (It is not quite zero. The human eye is still slightly sensitive at that wavelength—about 10^{-8} that of the peak value at 555 nm.) Once the connection between photometric and radiometric quantities is made, there is little difficulty transferring back and forth between the two systems, providing the spectral distribution of the source and the spectral variation in the transmittance is known. A list of corresponding radiometric and photometric quantities is given in Table 3.1.

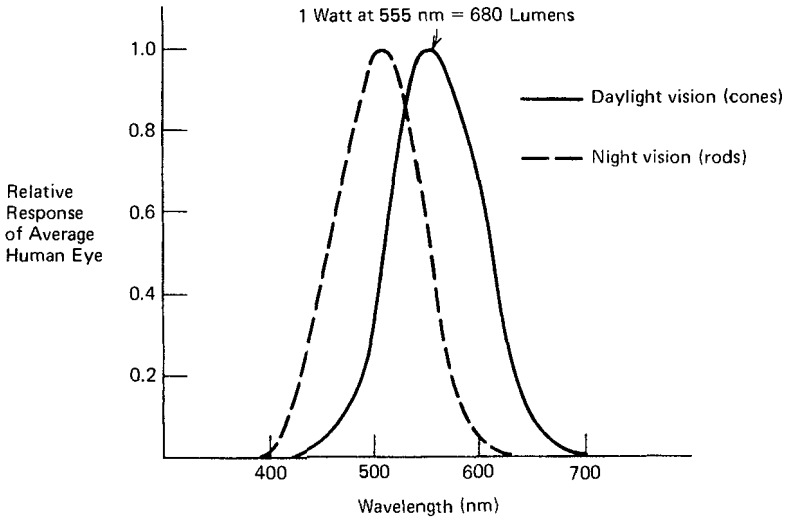


Figure 3.16. Spectral luminous efficiency curves. The solid curve is a plot of relative response of a daylight-adapted eye to visible wavelengths (photopic curve); the dashed curve plots the response of the dark-adapted eye (scotopic curve).

Note that the radiometric and photometric quantities are distinguished by the radiant and luminous modifiers and roots and by the subscripts $_e$ for radiometric terms and $_v$ for visible (photometric) terms. One of these quantities, the luminous intensity I_v , when expressed in lumens/ster, is equal to one candela (formerly a “candle”) and is very close to the luminous intensity of an ordinary candle. At one time the candle was the standard for all photometric measurements. It has now been redefined more precisely as $\frac{1}{60} \text{ cm}^2$ of a blackbody radiator at the freezing point of platinum, 2042 K.

The calculation of a source whose output is constant per unit wavelength band across the visible range would be done by adding up the contributions for each small wavelength band on the luminous efficiency curve to get the total photometric output. It is proportional to the area under the curve [Problem 3.16]. For sources that do not provide a constant power per unit wavelength band, a multiplication of these two curves and summation would be needed. For example, if $s(\lambda)$ represents the spectral luminous efficiency curve shown in Fig. 3.16, then the photometric power of source, ϕ_v is equal to

$$\phi_v = \frac{680 \text{ lumens}}{\text{watt}} \int_{\lambda_2}^{\lambda_1} \phi_e(\lambda) s(\lambda) d\lambda, \quad (3.28)$$

where λ_1 and λ_2 are 400 nm and 700 nm (provided there are no extremely large contributions to ϕ_e in the near ultraviolet or infrared) and ϕ_e is the power (in watts) per unit band width. In Example 3.D, the method for carrying out a calculation for a rather simple case is shown.

EXAMPLE 3.D. The strongest spectral lines in the output of an argon ion laser are at 488 nm and 514.5 nm. If a laser has an output consisting of 1.5 watts at the 488 nm line and 2 watts at 514.5 nm, what is the photometric power of this laser?

Solution: The relative efficiencies for these two lines, based on Fig. 3.16, are 0.19 and 0.60. Thus, the total power is

$$680 \frac{\text{lumens}}{\text{watt}} [(1.5 \times 0.19) + (2 \times 0.60)] = 1010 \text{ lumens.}$$

In the case of more complicated spectrum, there may be a need to analyze the spectrum, measure the total power output and apply a point-by-point weighting to arrive at a photometric value. Experimentally, it is possible to construct a filter that approximates the standard luminous efficiency curve. When the filter is placed in front of a broadband radiometer and calibrated in lumens, the radiometer will serve as a photometer.

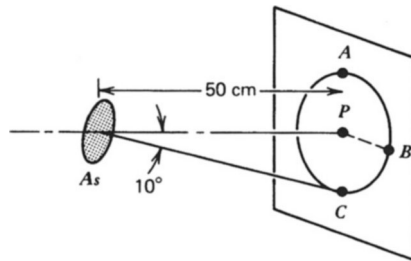
One additional quantity introduced in this discussion is that of a spectral output. In a radiometric measurement the power may be spread across many wavelengths, but only the total power is measured. In many cases, however, the actual distribution of power across the spectrum must be determined. The spectral variation of a source is described using the same quantities defined earlier in either radiometric or photometric terms but modified by the word "spectral." This adjective then specifies that the quantity is measured over a small unit bandwidth (nm, μm , etc., in wavelength; Hz in frequency). The symbols given in Table 3.1 are also modified by the addition of another subscript (λ for wavelength variation; ν for frequency variation) to the radiometric/photometric subscript.

Spectroradiometry, the measurement of spectral variation of sources and system outputs, plays a large part in the design and details of many optical systems. The methods for determining these variations, as well as the devices that must be incorporated into optical systems to disperse and analyze the radiation transmitted by a system, is covered in Chapter 9.

PROBLEMS

- 3.1. A point source of light is buried deep in an experimental optical system. Unfortunately, the plans for the optics are missing and the location of the source can only be roughly estimated. However, measurements of the irradiance from the point source show that at some distance from the opening from which the radiation is emitted d_1 , the irradiance is

- 6.25 W/m², and at a distance d_2 , which is 40 cm further from the first measurement point, the irradiance is 4 W/m². How far from the opening is the source located, and what is its intensity?
- 3.2. At what angle to the normal is the irradiance of a screen illuminated by a point source equal to 75% of the peak value?
 - 3.3. A light source whose surface is approximated by a circle with a radius of 3 cm is located 2 m from a screen. If an aperture with a 1-cm radius is placed between the source and the screen, how far from the screen can the aperture be placed without affecting the irradiance in the screen at the center of the patch of light passed by the aperture? If the aperture is moved twice as far from the screen, what will be the irradiance at the center of the patch as a fraction of the unobstructed irradiance?
 - 3.4. A Lambertian source of surface area 5 cm² is located 50 cm from a screen. The surface is tilted at an angle of 10° with respect to the screen (i.e., the normals of the screen and source make an angle of 10° to each other). What is the irradiance at point P if the source radiance is 100 W/m²-sr? What is the irradiance of points A , B , and C on a circle about point P ?



- 3.5. Devise some methods to collect more of the radiation from the source than the system in Fig. 3.8. Use ray sketches. Estimate the maximum fraction of the source energy that it is possible to collect.
- 3.6. A point source whose intensity is 1 W/sr is located 25 cm from a 10-cm focal length lens with a 5-cm diameter. What is the image intensity? What is the irradiance at the center of a screen (a) 10 cm from the lens in object space, (b) 25 cm from the lens in image space, and (c) 10 cm from the lens in image space?
- 3.7. Show that the solid angle at the lens subtended by the source or its image, Ω'' , multiplied by the pupil area, A_{pupil} , is also equal to the throughput, $A\Omega$.
- 3.8. Show that the overlap area $A(\theta)$ can be expressed as Eq. 3.25 and that this holds whether Δ is greater than or less than the smaller of the two radii. Sketch these two cases.
- 3.9. Plot the exact expression for the vignetting factor, $V(\theta)$, and its linear approximation, Eq. (3.27) from $\Delta = |r'_1 - r_2|$ to $\Delta = r'_1 + r_2$. (A small

programmable calculator should be used for this problem.) For simplicity, assume $r_1 = r_2$.

- 3.10. A 2-cm-diameter Lambertian source is located 15 cm in front of a 20-cm focal length lens (diameter, 4 cm). On a screen 1 m beyond the lens, what is the radius of a circle inside of which the irradiance is greater than 90% of the peak irradiance? (This problem cannot be solved in a closed form. You will have to pick a radius and calculate the value. Another method that is somewhat lengthy is to plot the irradiance versus the radial distance.)
- 3.11. A searchlight is constructed from a 5-cm-diameter source of brightness $L_e = 50$ watts/m²-sr placed at the focal point of a 25-cm lens with a diameter of 10 cm. Plot the irradiance along the optic axis from the lens to a distance of 250 cm.
- 3.12. Using the values for the searchlight given in the previous problem, plot the values of the irradiance from the optic axis out to the shadow region on a plane located at a distance $d_0 = \sqrt{A_{\text{lens}}/\Omega_1}$ from the lens.
- 3.13. Do Problem 3.12 for a plane located at $d_0/2$.
- 3.14. Using a ray trace, show that the source image is magnified by a factor $f_{\text{projector}}/f_{\text{condenser}}$ for collimated illumination (Fig. 3.15(b)).
- 3.15. A tungsten filament whose brightness is 100 W/m²-sr and whose filament area is 4 mm² is enclosed in a glass bulb of 2-cm radius is to be used to irradiate a small sample.
 - (a) What is the maximum irradiance that can be achieved without the aid of a lens?
 - (b) If an $f/2$ lens with a focal length of 5 cm is used to focus the light, what is the maximum achievable irradiance?
- 3.16. What would be the total photometric output of a source whose power output is constant across the visible region and is equal to 300 watts in that region? First, provide a rough estimate and then a more accurate value.

REFERENCES

- 3.1. *Handbook of Optics*, edited by W. G. Driscoll and W. Vaughn (McGraw-Hill, New York, 1978). Technical handbook. First chapter on radiometry and photometry includes tables of translation between different symbols.
- 3.2. W. J. Smith, *Modern Optical Engineering* (McGraw-Hill, New York, 1966). Chapter 8 contains material similar to that in this chapter. Symbols are different from those used here.
- 3.3. H. A. E. Keitz, *Light Calculations and Measurements* (Macmillan, London, 2nd edition, 1971). Comprehensive text that treats a large number of cases that are progressively more difficult. Uses units other than those defined in this text.

4

Mirrors and Prisms

(HOW TO GET EVERYTHING IN THE BOX)

Although it would be nice if all optical systems could be constructed as they are originally designed, as one long optical chain, many systems must be folded to fit a certain length or certain volume. In some systems, several images must be combined or an image must be moved or a beam scanned as a part of the unit's operation. Although it is sometimes possible to scan with elements having optical power, lenses or curved mirrors, in most instances these operations are assigned to elements with no optical power, that is, plane mirrors and prisms. The purpose of this chapter is to show how mirrors and prisms can be used to fold and combine optical trains in order to provide the most compact and versatile system possible. The orientation of mirror images will be discussed first. Then different types of mirrors and prisms will be described. Finally, the use of mirrors and prisms in the design of optical systems will be presented.

4.1. MIRROR IMAGES

When you turn on the light in the bathroom in the morning, some of that light scattered by your skin is directed to the plane mirror over the sink and redirected back to your eye, where you can see what you look like at the beginning of the day. The image you see has its hair parted on the “wrong side,” and it appears to use the wrong hand to comb its hair.

Our interpretation of our mirror image is based on the idea that there is another person facing us. Therefore, when we shake hands with someone, we expect their right hand to move toward us, but when we try this with our mirror image it moves its “left hand” towards us. This would seem to argue that a mirror reverses right and left. If we look at writing in a mirror, that certainly appears to be so. Look, for example, at the writing in the mirror to the left of the object in Fig. 4.1. But, consider the writing in the mirror above the object in the same figure. There is no left and right reversal. Instead up and down are reversed. How is one to interpret and understand mirror images? A ray trace of an object in front of a plane mirror (Fig. 4.2) shows that the *only coordinate reversed* is the direction normal to the surface of the mirror. The result of this front-back reversal as we perceive it depends on where the mirror is located

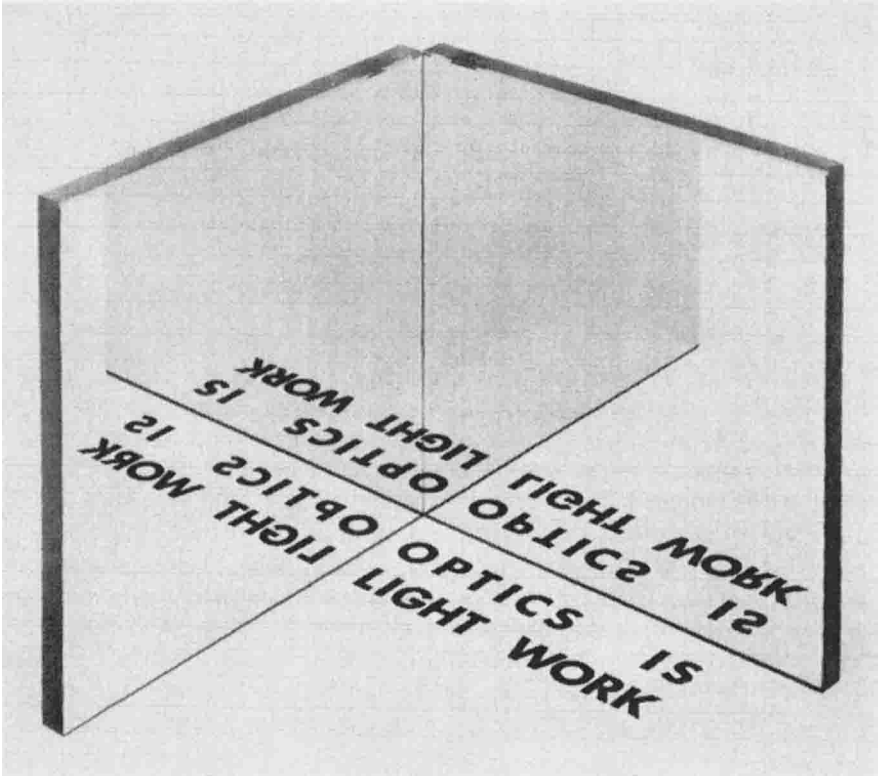


Figure 4.1. Mirror symmetry. A mirror reverses the front-to-back direction of an image (direction normal) to the mirror surface. Depending on the placement of the mirror, this can cause right-to-left reversal (mirror on left of figure) or a top-to-bottom reversal (mirror on right of figure). The result of two reflections is an inversion as shown by the upside down image in the center. (Photo by Vincent Mallette.)

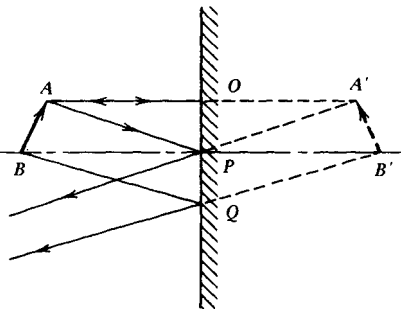


Figure 4.2. Reflection of an object by a plane mirror illustrating front-to-back reversal.

relative to the object. In the cases shown in Fig. 4.1, there is a left–right reversal for the mirror to one side of the writing and an up–down reversal for the mirror located above the writing. In the case of our own image in front of a mirror, the front–back reversal turns a right hand into a left hand. Any *single* reflection by a plane reflecting surface results in a reversal of the image. This single *reversal* changes the handedness of the image and causes printing to be unreadable no matter how the head is oriented. Two reversals due to reflections from two mirrors restore the image to the same handedness and the printing to readability. This is easily seen with an object between parallel mirrors. All odd number reflections are reversed; all even number reflections are not reversed, and their printing can be read without difficulty.

In Fig. 4.1 there is a third image, which, although upside down, is readable. These two reflections from surfaces at right angles to each other produce an unreversed image rotated by 180° . We refer to this image as *inverted*. In this case inversion means a rotation of 180° of the image about the line of intersection between mirrors. With mirrors at other angles, two reflections yield unreversed images rotated through some angle with respect to the object. There is some confusion in the literature as to the image orientation described by a reversal and an inversion. In this text, as in many others, a reversal is a reversing of one coordinate which will make printing unreadable; inversion is a rotation of the image by 180° . This convention is in accord with the description of real images in the case of lens and mirror systems.

For simple calculations of image orientation, it is useful to think of the effect of reflections on an image as though a pencil, an arrow, or a matchstick were bounced off the mirror surface. In some cases, to distinguish both inversion and reversal, a figure or character (the letter “R” is a good example) is chosen to follow an image orientation through an optical system. Examples of this method are given later in this chapter. Please note that if you are following an image from object space, that the rule of reversal (an odd number of reflections causes a reversal) is correct only if you view the image from the same side as you viewed the object. If an image is viewed in the opposite direction of light propagation from that of the object, an additional image reversal occurs. For example, an image on a rear screen projection has an additional reversal compared to an image on a conventional projection screen. To be absolutely certain what one is seeing, one should use an unsymmetric three-dimensional object (for example, the letter “R”) whose front is white and whose back is black [Problem 4.1]. The calculus of image translation and rotation by plane mirrors can become quite complex, particularly if the mirror orientations change as they do in a scanning system.

Two special mirror configurations should be described. The first is an answer to Robert Burn’s prayer: “God give us the gift to see ourselves as others see us.” If you arrange two mirrors at right angles to one another and look in perpendicular to their intersection, as shown in Fig. 4.3, you will see yourself as others see you. Your hair is parted correctly and your image understands the etiquette of hand shaking. This, of course, occurs because each ray leaving your

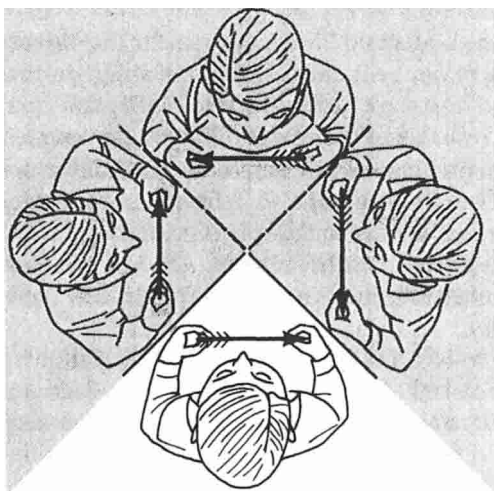


Figure 4.3. A two-mirror reflection. Someone looking in such a mirror would see themselves as others do.

head is doubly reflected before reaching your eye. If you try this with two hand-held mirrors, notice how little tolerance there is in the angle between the mirrors. The eye immediately detects any small deviations from the right-angle geometry. This can be of value in the alignment and testing of optical systems. This nonreversing mirror arrangement does, indeed, let us see our face and figure as others see us. (Robert Burns had a somewhat more profound gift in mind.)

This first example is really a special case of the general class of *constant deviation mirrors*. If two mirrors intersect at an angle θ (Fig. 4.4), light perpendicular to the line of intersection of the two mirrors is deviated by

$$\phi = 2\theta \quad (4.1)$$

independent of the angle of incidence (providing the ray hits both mirrors) [Problem 4.2]. Obviously, a two-mirror configuration is slightly more complicated than a single mirror to deviate a beam, but when a system designer cannot specify the exact position of an entering ray, a constant-deviation mirror pair may be appropriate.

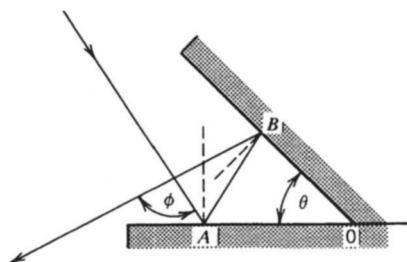


Figure 4.4. Constant deviation mirrors. Two mirrors at an angle of θ to each other deviate a ray through twice that angle ($\phi = 2\theta$), if the ray reflects off of each mirror once.

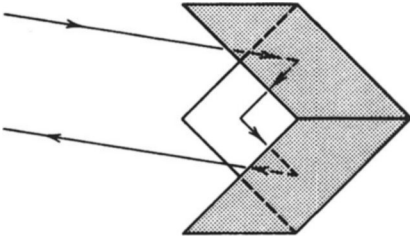


Figure 4.5. A cube corner (or corner cube) reflector. The sides of the reflector are three sides of a cube. A ray reflected from this device returns parallel to the incident ray.

The second example consists of three mirrors arranged as the corner of a cube and is called either a *cube corner*, or a *corner cube* (Fig. 4.5). The former is more descriptive; the latter is easier to say. The singular property of a corner cube reflector is that a light ray incident upon it is returned parallel to itself after three reflections. Because of this, a corner cube reflector is also called a *retroreflector*. Arrays of such reflectors were left on the moon during three of the Apollo missions so that light pulses from earth could be timed to more precisely determine the distance from earth to the moon (Ref. 4.1). Because corner cube reflectors were used, the angular orientation of the array as it was placed on the moon was not critical. The image in a retroreflector is both inverted and reversed.

4.2. PLANE MIRRORS

It would seem that there would be little variety in the types of plane mirrors, but this is not so. Mirrors in optical design may be called upon to do a lot more than just translate an image. While many mirrors are made by metallizing a surface, a large number of optical mirrors are created by optical design from alternating layers of dielectric materials. Because of their importance to optical design, a portion of this section is devoted to an understanding of the details of their construction.

4.2.1. Metal Mirrors

The usual bathroom mirror is an aluminized or silvered second-surface mirror. It is termed a *second-surface mirror* because the metallic coating is on the back surface of the mirror (Fig. 4.6(a)). Light passes through the first surface of plate glass and is reflected off the second surface, which has been coated with a thin layer of metal. A layer of paint is applied over the metal to protect this fragile coating. While this arrangement works well for bathroom and vanity mirrors, it is unacceptable for many optical designs, since reflections off the first surface will contribute unwanted “ghosts” to an image in the optical system. Most optical systems are designed with *first-surface mirrors* to eliminate any ghosts (Fig. 4.6(b)). However, the mirror surface is exposed to the environment. To provide

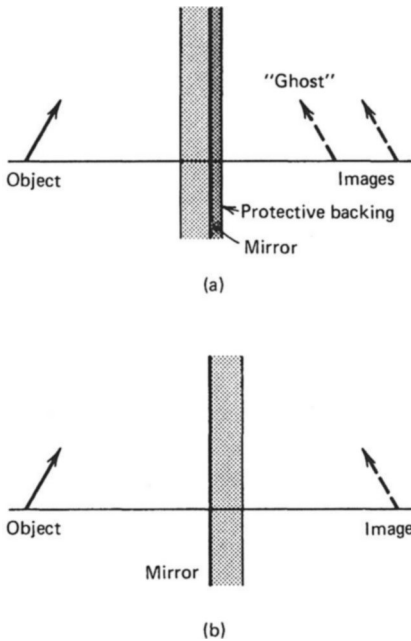


Figure 4.6. Mirrors. (a) A second-surface mirror can be protected by a tough paint, but it introduces a ghost image as well as the intended image. (b) A first-surface mirror introduces no ghost image, but is subject to scratches and abrasion.

some protection against oxidation and abrasion, the metal may be overcoated with a thin layer of dielectric material such as silicon monoxide (SiO) or magnesium fluoride (MgF_2).

4.2.2. Dielectric Mirrors

Many metallic materials, such as aluminum, silver, and gold, are highly reflective compared to other materials, but they also absorb a fraction of the incident radiation. For some applications, particularly those involving high-power lasers, this “small” absorption of radiation contains sufficient energy to vaporize the material, rendering it useless as a mirror. In these and other applications, the optical designer must use mirrors made entirely of dielectric materials that absorb little or no radiation in the region of interest. These *dielectric mirrors* consist of alternating layers of materials of high and low refractive index, each layer having an optical thickness, nd , equal to a small fraction (usually one quarter or one half) of the wavelength of the light for which the reflector is designed. The method by which this is achieved can be seen in the construction of an anti-reflection coating, the simplest example of a dielectric coating.

The laws of geometrical optics listed in Chapter 1 tell us *where* light will go after hitting an interface, but they do not tell us *how much* light is reflected and *how much* is transmitted. That can be found from a series of expressions called *Fresnel's Equations* (Ref. 4.2). If E_0 is the electric field incident on a surface and

E_r is the electric field reflected from that surface, the fraction of the incident field reflected is given by the *reflectivity*, R , given here for two polarizations:

$$R_{\parallel} = \frac{E_{r\parallel}}{E_{0\parallel}} = \frac{n_2 \cos \theta_1 - n_1 \cos \theta_2}{n_2 \cos \theta_1 + n_1 \cos \theta_2}. \quad (4.2)$$

$$R_{\perp} = \frac{E_{r\perp}}{E_{0\perp}} = \frac{n_1 \cos \theta_1 - n_2 \cos \theta_2}{n_1 \cos \theta_1 + n_2 \cos \theta_2}, \quad (4.3)$$

where n_1 and n_2 are the refractive indices of the incident and refracted media, and θ_1 and θ_2 are the incident and refracted angles, related by Snell's Law (Eq. 1.7). The subscripts \parallel and \perp indicate light polarized with the electric field parallel (\parallel) and perpendicular (\perp) to the plane of incidence. There are similar equations for the amount of transmitted radiation. In the case of normal incidence ($\theta_1 = \theta_2 = 0$), only R_{\perp} applies since all light is polarized perpendicular to the plane of incidence, and the equation simplifies to

$$R = R_{\perp} = \frac{n_1 - n_2}{n_1 + n_2}. \quad (4.4)$$

Note that if light in a low refractive index medium is incident upon a higher refractive index medium ($n_2 > n_1$) the sign of R is negative, indicating a 180° phase change of the electric field upon reflection. If $n_1 > n_2$, no such phase reversal occurs (Fig. 4.7).

The amount of light reflected from an interface is proportional to the *square* of the electric field amplitudes, therefore, the fraction of light reflected from an

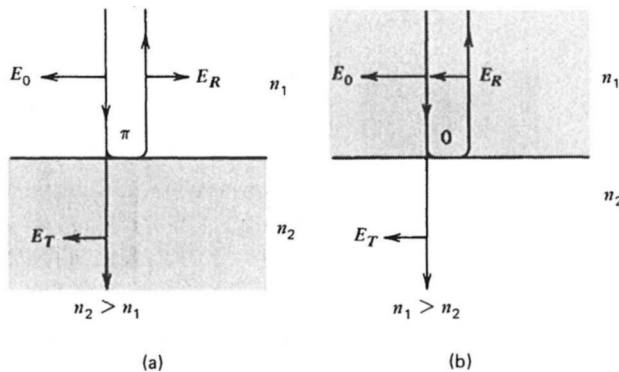


Figure 4.7. Reflection of light at an interface. (a) Incident medium refractive index (n_1) less than transmitting medium index (n_2) with a 180° (π) phase change between the incident and reflected rays. (b) n_1 greater than n_2 with no phase difference between incident and reflected rays.

interface, the *reflectance*, is the square of the reflectivity. In the case of normal incidence,

$$r = |R|^2 = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 \quad (4.5)$$

Therefore, the fraction of light reflected from a glass surface ($n_g = n_2 = 1.5$) at an air–glass interface is

$$\left(\frac{1 - 1.5}{1 + 1.5} \right)^2 = \left(\frac{-0.5}{2.5} \right)^2 = (0.2)^2 = 0.04. \quad (4.6)$$

That is, 4% of the light from a glass surface is reflected. At a pane of window glass, about 8% is reflected since there are two surfaces. While this may seem to be a small amount, in a five-element lens, where each surface will transmit about 96% of the energy, the amount of light transmitted through ten surfaces would be $(0.96)^{10}$ or 66%. This substantial reduction or transmission of light due to reflection losses is why anti-reflection coatings are used on many camera lenses.

An *anti-reflection coating* often consists of a low-refractive-index material, such as MgF_2 ($n = 1.38$), whose optical thickness is one quarter wavelength ($nd = \lambda_0/4$), deposited on the glass (Fig. 4.8(a)). The incident light reflecting off the air– MgF_2 interface undergoes a 180° phase change. The light reflecting off the MgF_2 –glass interface also undergoes a 180° phase change. However, the light reflecting off the glass must also travel an optical path-length distance equal to $\lambda_0/2$ (one quarter wavelength into the material and the same distance back out to the air), which retards the wave by another 180° . Thus one wave suffers a 180° phase change while the other suffers 360° phase change. Since the two waves are out of phase by 180° they will destructively interfere. The cancellation between the two reflected waves can be exact only if R for the air–

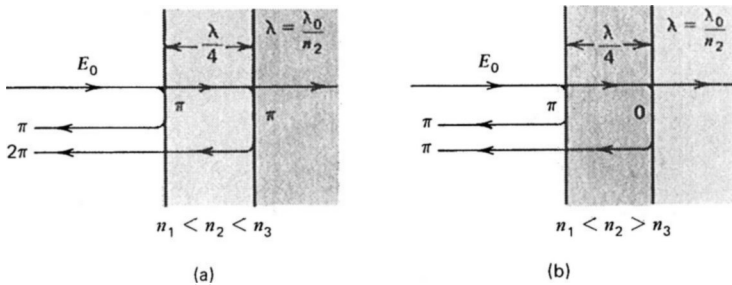


Figure 4.8. Quarter-wave layer coatings. (a) An anti-reflection coating. The rays reflected from both interfaces are 180° (π) out of phase with each other. (b) A reflection-enhancing coating. The rays reflected from both interfaces are in phase providing constructive interference for the reflecting rays.

MgF_2 interface equaled that for the MgF_2 -glass interface. This would be true only if $n_2 = \sqrt{n_1 n_3}$ [Problem 4.4]. However, the deposition of this quarter-wave layer reduces the reflected light to about 1% of the incident radiation [Problem 4.5]. With additional layers, reflected light can be reduced virtually to zero.

The above discussion would seem to be the wrong topic in a chapter about mirrors, but consider the consequences of putting a quarter-wave thickness of material with a refractive index *higher* than that of the glass (Fig. 4.8(b)). The reflection at the air-high-index interface again causes a 180° phase change, but the reflection at the second surface does not. Thus the additional phase contribution to the second wave from traveling through the material results in both reflections adding constructively, since they are in phase. By adding a high-refractive-index layer, the reflectance of the interface is increased almost ninefold over the uncoated substrate, as shown in Example 4.A. Additional quarter-wave optical thicknesses of high- and low-refractive-index materials increase the reflectance further. In Fig. 4.9 is shown the construction of a five-layer reflector. After more than 12 layers, the reflectance of a dielectric mirror approaches 100% with virtually no absorption or scattering. This reflectance is at a maximum only in a region about the design wavelength. The size of the region depends on the design of the stack of multiple dielectric coatings. Away from this region the reflectance is reduced. For example, at one half the design wavelength it falls to that of the uncoated substrate. The dielectric mirror is also designed to be used at a specific angle of incident radiation. At other angles, the performance is reduced, and the wavelength of maximum reflectance is shifted (Ref. 4.3).

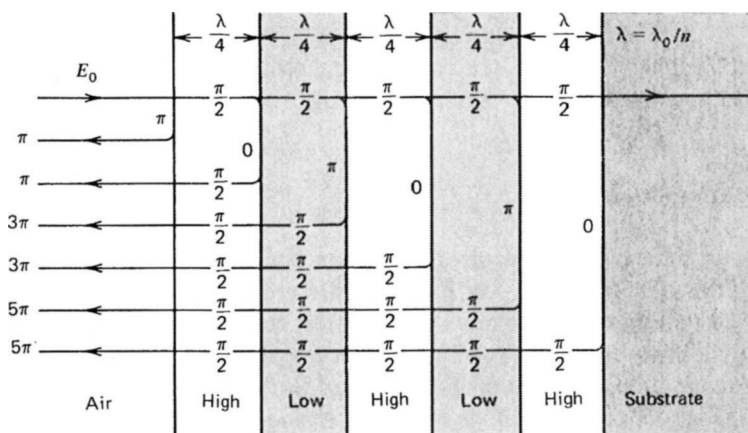


Figure 4.9. A 5-layer reflector. Alternating layers of higher and lower refractive index materials provide multiple reflections all in phase. This produces constructive interference at the design wavelength and an efficient nonabsorbing reflector.

EXAMPLE 4.A. If a quarter-wave layer of zinc sulfide ($n = 2.3$) is deposited on a crown glass surface ($n = 1.52$), what fraction of the light is reflected from the surfaces at normal incidence?

Solution: Before the irradiance can be found, the reflectivities of the surfaces must be found. These are given by Eq. 4.4.

$$R_{12} = \frac{n_1 - n_2}{n_2 + n_1} = \frac{1 - 2.3}{2.3 + 1} = \frac{-1.3}{3.3} = -0.39,$$

$$R_{23} = \frac{n_2 - n_3}{n_2 + n_3} = \frac{2.3 - 1.52}{2.3 + 1.52} = 0.204.$$

These are then added together (R_{23} must be multiplied by -1 because light reflected off the ZnS–glass interface suffers a 180° phase change due to the additional path length of one-half wavelength).

$$R = R_{12} + R_{23} = -0.39 - 0.204 = -0.598,$$

$$r = R^2 = 0.35.$$

Thus the reflectance for a single surface has been increased from 4% to 35%.

4.2.3. Liquid Mirrors

One other mirror with a very special orientation is that formed by the surface of a liquid. A pool of mercury, for example, has been used with a laser beam to provide a local vertical and serve as the critical element of an optical plumbline. Less-toxic and less-reflective materials can also be used to perform the same task.

4.2.4. Beamsplitters

At times, an optical design requires more than one image or beam. This can be achieved through the use of a partially silvered mirror, or *beamsplitter*, a component that partially transmits and partially reflects the light incident on it. Besides dividing a beam into two beams, a beamsplitter can be used to combine two beams or images into one. Examples of both uses of a beamsplitter are shown in Fig. 4.10. The methods for constructing beamsplitters are similar to those for mirrors and use either metallic or dielectric coatings. The dielectric coatings, however, must be designed to give the proper reflectance and transmittance at the angle for which the beamsplitter is designed. Since the

EXAMPLE 4.B. What is the minimum reflectance for a dielectric film of ZnS ($n = 2.3$) sandwiched between two right-angle prisms ($n = 1.52$) to produce a polarizing beamsplitter?

Solution: The reflectivities are the same at both surfaces since n_1 , n_2 , θ_1 , and θ_2 in Eq. 4.4 are interchanged. The thickness of the ZnS layer will determine whether the reflectivity is maximum or minimum. The problem can be divided into two parts:

PART 1. What is the reflectivity at an interface between $n_2 = 2.3$ and $n_1 = 1.52$ for $\theta_1 = 45^\circ$?

Snell's Law gives

$$\sin \theta_2 = \frac{1.52}{2.3} \times \sin 45^\circ = \frac{1.52}{2.3} \times \frac{1}{\sqrt{2}} = 0.4673;$$

$$\theta_2 = 27.86^\circ; \quad \cos \theta_1 = 0.707; \quad \cos \theta_2 = 0.884.$$

Using Eqs. 4.2 and 4.3

$$R_{\parallel} = \frac{2.3 \times 0.707 - 1.52 \times 0.884}{2.3 \times 0.707 + 1.52 \times 0.884} = \frac{0.2825}{2.9702} = 0.095119,$$

$$R_{\perp} = \frac{1.52 \times 0.707 - 2.3 \times 0.884}{1.52 \times 0.707 + 2.3 \times 0.884} = \frac{-0.9586}{3.1082} = -0.3084.$$

There is no phase change at the first interface. At the second interface, the subscripts 1 and 2 are reversed. Thus,

$$R'_{\parallel} = -R_{\parallel}; \quad R'_{\perp} = -R_{\perp}.$$

PART 2. What is the thickness needed to obtain the maximum reflectance?

The optical thickness of the light path inside the dielectric film must be either $\lambda_0/4$ or $\lambda_0/2$. If it is $\lambda_0/4$, then R'_{\parallel} and R'_{\perp} must be multiplied by -1 . This will then give reflectances that are equal and in phase so that the reflecting waves add constructively. Thus, R_{\parallel} for both waves will be twice that for a single surface, or 0.19, and $r_{\parallel} = 0.036$ and R_{\perp} will be -0.62 , and $r_{\perp} = 0.38$. Therefore, this simple beamsplitter will reflect about 38% of the light out of the beam that is polarized perpendicular to the plane of incidence, while transmitting nearly all of the light polarized parallel to the plane of incidence. The actual thickness d of the layer can be determined from the requirement that $d/L = \cos 27.86^\circ$ and $nL = \lambda_0/4$, where L is the geometrical path inside the layer. Therefore, $d = (\lambda_0/(4 \times 2.3)) \cos 27.86^\circ$. At 550 nm, $d = 52.9$ nm.

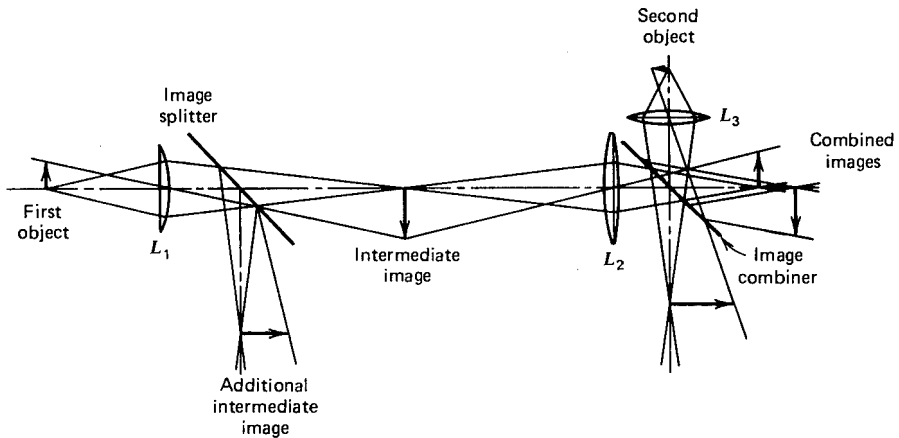


Figure 4.10. Beamsplitters in an optical system. Beamsplitters can be used to provide additional images or to combine two images. In the case shown here, the images are not located in the same plane in order to clarify the ray traces. In many instances the image planes would be adjusted to overlap.

incident light is not perpendicular to the dielectric layers, the full Fresnel equations must be used (Example 4.B). At large angles of incidence, the reflectances may be quite different for parallel and perpendicular polarizations. Therefore, some dielectric beamsplitters divide the light according to its polarization content. Thus, by altering the polarization of a beam before it enters a polarizing beamsplitter, one can control the ratio of transmitted-to-reflected irradiance beyond the component [Problem 4.5].

Just as with second-surface mirrors, beamsplitters by their very nature have a second surface, the uncoated surface of the substrate. Again the possibility of ghost images occurs. Anti-reflection coatings must be applied to this surface taking into account the nonzero angle of incidence of the light. Beamsplitters can also be made from metallized, stretched plastic membranes called *pellicles*. The surfaces are so thin that the first and second surface images are essentially identical. Used in white light illumination at a nonzero angle of incidence pellicles work well. In monochromatic light, for near normal incidence, interference fringes may be visible and can pose a problem in some designs. Because of the fragility of the membranes, pellicles are not used in an exposed environment or in a system that cannot be isolated from vibrations.

4.3. PRISMS

In Chapter 1 it was noted that light sent through the corner of a glass prism would be separated into its colors because of the wavelength dependence or dispersion of the refractive index of the glass. We return to this concept of

dispersion in Chapter 9, where the analysis of light will be treated. In this section only nondispersing types of prisms will be treated, those prisms that can serve in a compact arrangement of multiple reflecting surfaces. The attractiveness of prisms over several mirrors is that once they are constructed they retain the design orientations and require no adjustment in the final assembly except for the prism as a whole.

For example, consider the right-angle prism, as shown in Fig. 4.11(a). In cross section this simple prism is a 45° - 90° - 45° triangle. Light entering one of the small faces is reflected off the hypotenuse face and exits through the other small face. Provided the refractive index of the prism is greater than $\sqrt{2}$ (~ 1.414), the light will be reflected by total internal reflection (discussed in Section 1.2). This is another reason that prisms are so useful as mirrors. Since there is no need for metallic or dielectric coatings on the reflecting surface, it serves as a nearly lossless reflector. Only scattering due to small surface imperfections and Fresnel reflections at the entrance and exit faces, which can be anti-reflection coated, cause any losses. This same right-angle prism, oriented so that the hypotenuse face is the entrance and exit face (Fig. 4.11(b)), is the prism version of the nonreversing mirror described at the end of Section 4.1. Again the permanent alignment of the reflecting faces and the total internal reflections suggest the use of such a prism in many applications. With such compelling advantages, why not use a prism in all instances where a mirror is needed? There are two disadvantages to prisms: cost and weight. If the cross-sectional area of the entrance or exit opening in the prism is much greater than 5 cm^2 , the weight of the prism becomes considerable, as does the cost of fabricating a large chunk of glass into a precisely ground and polished prism. Beyond a 5-cm^2 cross section, a mirror has the advantage. Thus, total internal reflection prisms and plane mirrors should be thought of as complementary rather than interchangeable devices in most optical designs.

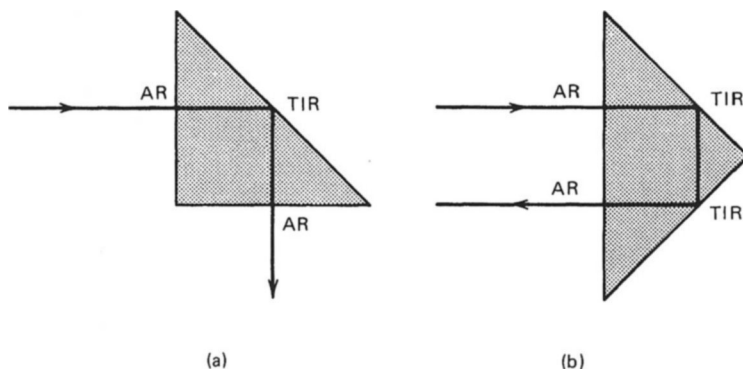


Figure 4.11. Prism reflectors. (a) A prism used as a mirror reflector. (b) A prism used as a constant deviation mirror similar to Fig. 4.3. AR: Anti-reflection coating, TIR: total internal reflection.

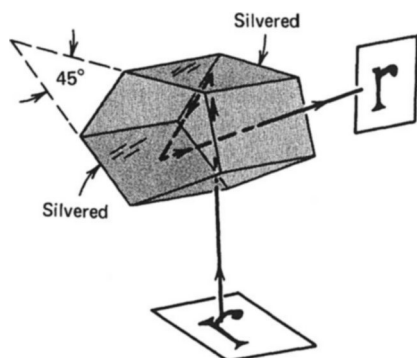


Figure 4.12. Pentaprism. The two reflecting surfaces must be silvered since total internal reflection does not occur for light entering through the correct faces.

The constant deviation prisms and corner cube reflection prisms serve as prism counterparts to the mirror arrangements discussed earlier. Listed below are a number of additional prism types and some of their benefits.

1. *Pentaprism.* As can be seen from Fig. 4.12, a pentaprism is a constant deviation prism that deviates an incoming ray by 90° . Because of this property, it serves as an "optical square," the optical equivalent of a carpenter's square. It is useful in alignment and layout systems that depend on ray paths that intersect at 90° [Problem 4.7]. Because of the small angle of incidence of the first internal reflection, total internal reflection does not apply here, and the reflecting surfaces of a pentaprism must be coated with a reflecting film.

2. *Porro prisms.* These are two right-angle prisms used in the 180° constant deviation arrangement (Fig. 4.13). While both prisms produce a reversal, the combination produces an inversion. These two prisms fold the path of an optical system and also displace an image both horizontally and vertically by half the hypotenuse length in each direction. Porro prisms can be used to reduce the length of a Keplerian telescope, providing at the same time the additional inversion needed to erect the inverted telescope image. For this reason Porro prisms are found in many binoculars.

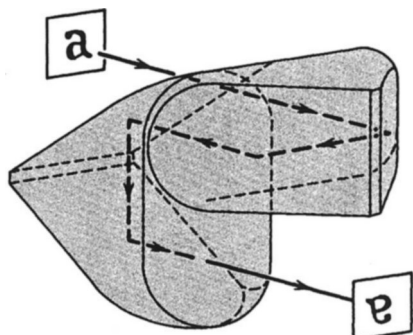


Figure 4.13. Porro prisms. Two right-angle prisms used as shown in Fig. 4.11(b). In addition to folding the optic axis, the two reflections provide an inversion of the image.

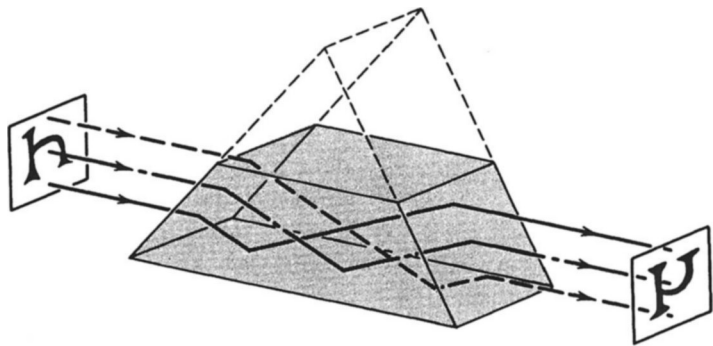


Figure 4.14. Dove prism. Light is refracted toward the prism base and is then totally internally reflected off the base of the prism. Because light near the apex of the prism never reflects off the base before the second refraction, the apex of the prism may be eliminated to save weight.

3. *Dove prism.* Consider light entering a prism parallel to its base, as shown in Fig. 4.14. The light will be refracted toward the base, where it is totally internally reflected and then refracted at the opposite side to give parallel light again. Since the apex portion of the prism has no effect on the rays reflected off the base surface, it is usually removed. What remains is a Dove prism. While it may at first appear that the non-normal angle of incidence would give rise to dispersion, traversal of light rays through a Dove prism is equivalent to passage

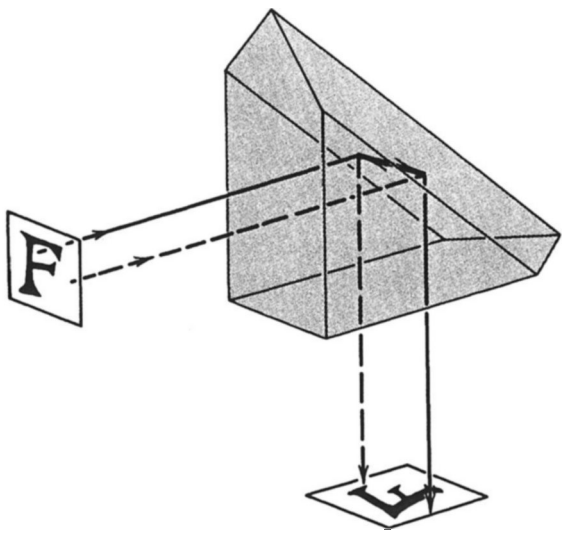


Figure 4.15. Amici prism. This is a right-angle prism used as a reflector as in Figure 4.11(a), but a second “right-angle roof” has been added to the hypotenuse. The additional reflection surface provides an inversion in place of the reversal introduced by the simple reflector.

through a slab of glass, as we will see in the next section. One of the most useful properties of the Dove prism is that a rotation of the prism about an axis along the light propagation direction outside the prism results in a rotation of the reversed image through twice the rotation angle (Problem 4.8).

There are numerous other prism combinations (Ref. 4.4); many of these are explored in the problems to this chapter. It is worthwhile to test your understanding of image orientation by mentally sending rays through each of the prisms to determine the orientation of the final image. In many cases a single mirror surface is replaced by a *roof* (two mirrors at right angles), as in the case of the Amici prism shown in Fig. 4.15. This eliminates (or adds, if you prefer) one reversal in the optical train.

4.4. FITTING IT IN OR SPREADING IT OUT

The above information on image orientation and reflector types can be combined with ray tracing to enable one to analyze and design optical systems. (There is a great deal to be learned about optical design by analyzing existing systems.)

4.4.1. Plane Parallel Plates—Beam or Image Displacers

Besides prisms and mirrors, one of the simplest methods of redirecting light, particularly for small displacements, is to pass it through a plate of glass with parallel surfaces, as shown in Fig. 4.16. If the light ray is normal to the surface, the only effect is to increase the optical path length by an amount $(n - 1)d$, where n is the refractive index of the plate and d is its thickness. Successive applications of Snell's Law shows that even if the ray enters the plate at an angle to the normal, the ray will emerge parallel to its original direction; however, the

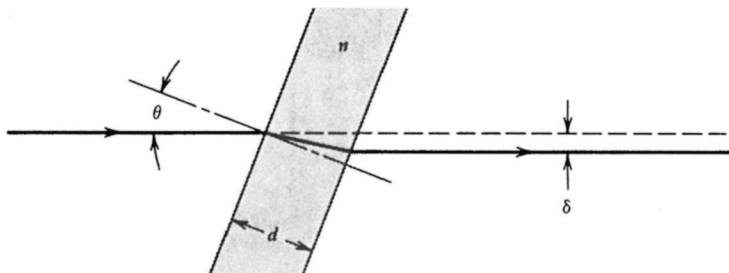


Figure 4.16. Lateral displacement of a ray by a slab of material with parallel surfaces.

beam will be displaced laterally by a distance, δ . Using Snell's Law and geometry (Fig. 4.16), this lateral displacement can be shown to be [Problem 4.10]

$$\delta = d \sin \theta \left[1 - \frac{\cos \theta}{\sqrt{n^2 - \sin^2 \theta}} \right]. \quad (4.7)$$

For a small angle of incidence, θ , ($\sin \theta = \theta$; $\cos \theta = 1$), this simplifies to

$$\delta = d \frac{(n - 1)}{n} \theta. \quad (4.8)$$

Thus, for small angles the displacement is proportional to the angle of incidence. A thick glass plate can be used to displace a beam or an image by a small amount by rotating the plate about an axis perpendicular to the displacement direction. It will also shift an image along the optic axis. This axial image shift is discussed in the next chapter (Section 5.1.4).

4.4.2. Analyzing and Designing a Folded System

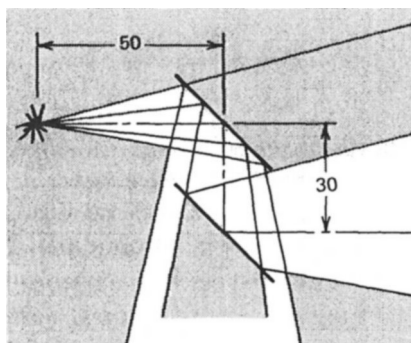
Any system that includes a number of mirrors can best be analyzed by unfolding the optics. This is done by removing the reflecting plane mirror, straightening out the optic axis, and, if it serves as either an aperture stop or field stop, replacing the mirror with an equivalent *slanted* aperture of the same size, as shown in Example 4.C. This concept of mirrors as filled apertures is useful in many analyses such as laser beam scanning. With beamsplitters it may be necessary to draw two separate systems, or, if the layout is not too crowded, one system on top of the other.

In the case of nondispersing prisms, an unfolding shows, as in Fig. 4.17(a), that a prism is equivalent to a parallel plate. For right-angle prisms, the thickness of the parallel plate is equal to the length of the input side of the triangle. In Fig. 4.17(b), the effect of unfolding a Dove prism is seen. Here the plate is oriented at an angle to the beam. As we will see in a later chapter, in convergent or divergent light a slab of glass can introduce considerable aberrations. Thus a Dove prism is always used in parallel light. It is also evident in the example that the height of the input face is limited by the length of the base, making the Dove prism a long, narrow component. Its total aperture may be improved by placing another Dove prism base-to-base with the first.

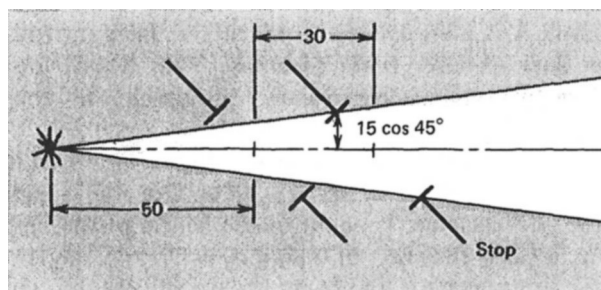
When designing a system that must include folding, it may be useful to make a ray sketch on tracing paper of the unfolded geometry that can be cut out and folded to determine the best possible placement of mirrors. There are a number of guidelines that are useful in designing folded systems.

EXAMPLE 4.C. A point source of 5 watts/sr is located 50 cm from the first of a pair of circular mirrors, 30 cm in diameter. The first mirror deviates the light by 90° to the second mirror 30 cm beyond, which directs the light parallel to the original direction. How much flux is directed in this system?

Solution: Draw the system.



Then redraw the system in an unfolded form, replacing the mirrors with apertures.



The second mirror is the aperture stop for this system. Because of the divergent rays, the top of the second mirror intercepts the rays before the bottom. The aperture is nearly elliptical with semi-minor axes of approximately $15 \text{ mm} \times \cos 45^\circ$ or 10.6 mm. The semi-major axis is the radius of the mirror, 15 mm. The area of the ellipse is $\pi \times 10.6 \times 15 \text{ mm}^2 = 500 \text{ mm}^2$. The distance to the mirror is 80 mm. Therefore, the solid angle subtended by the mirror is $500/(80)^2 = 0.0781 \text{ sr}$. Thus, the flux redirected by the mirrors is $5 \times 0.0781 = 0.39 \text{ watts}$.

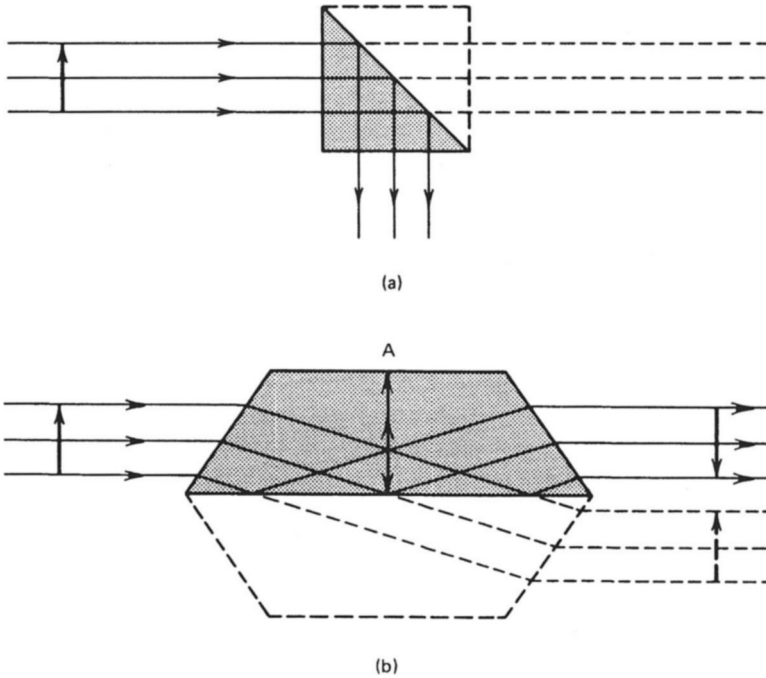


Figure 4.17. Unfolding of (a) a right-angle prism and (b) a Dove prism, to show that both may be treated as slabs of material with parallel surfaces.

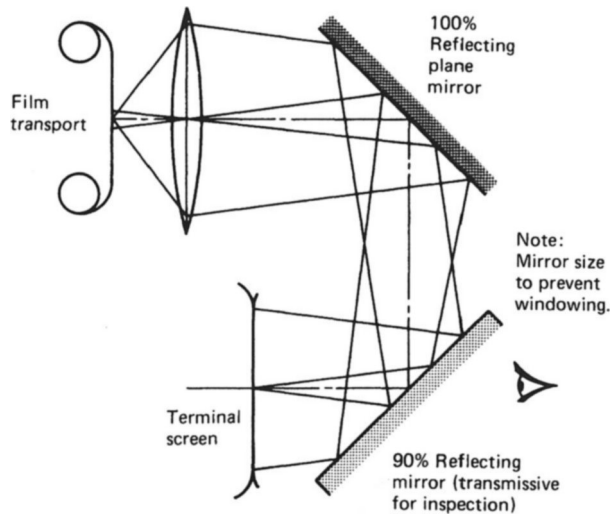
1. Place the folds near, but not at, a component. Mounting components in groups including the folding mirrors can simplify the mechanical design of a system.
2. Fold where the light is collimated, if possible. This is particularly true if a prism is being used.
3. Do not put a folding surface (mirror, beamsplitter, or prism surfaces including their transmitting surfaces) at an intermediate image point, if possible. Any imperfection in the surface will show up as part of the final image.
4. Think in three dimensions. Folding may be done in an x - y - z geometry. Some systems must be folded and stacked to fit a volume.
5. In some stacked geometries and some folded geometries, a component can be used twice.
6. There should be a very compelling reason for *not* folding with 90° and 180° deviations.

In Example 4.D we present one idea for folding our sample design from Chapter 2.

EXAMPLE 4.D. Design a folded system for the design problem in Chapter 2.

Solution: Since this system in an unfolded geometry would stick straight out from the terminal screen, the obvious thing to do is to fold the system at least once to reduce the overall depth of the system. One fold, however, would leave a tall tower in front of the terminal. The approach, not a unique one, taken in this case is to fold the system twice to reduce its length and place the imaging lens and the film transport mechanism atop the terminal.

One method of determining the approximate locations of folds in the system is to cut out a diagram of the optical system drawn to scale and make trial folds in the cutout. Tracing Fig. 2.8 and making trial folds in a cutout, the geometry, shown in the figure, was produced. Notice that the chief rays from both extremes of the screen were drawn to assure that neither of the mirrors acted as field stops.



The provision for visual inspection of the screen was incorporated in the system by making the first mirror 10% transmitting to allow direct inspection. Because there is no intermediate image, the border coding must enter either at the object (the screen) or at the image (the microfiche film frame). While a number of considerations would have to be explored further, some sort of simple bar code generated in a transmissive liquid crystal device might be used in contact with the microfilm to produce a shadowgraph of the computer readable code. Such questions as the information content in a simple code, its resolution, and the amount of light needed to produce such a code with the exposure requirements as the screen image must be explored.

4.4.3. Examples of Systems with Prisms and Mirrors

In film editing systems with very simple film drives (in some cases nothing more than a hand crank), the image of each frame must be stopped momentarily, although the film passes the viewing optics without stopping. This can be achieved by inserting a rotating octagonal prism in front of the film to produce a deflection of the image to the same position for different film positions, as shown in Fig. 4.18. In other designs, a rotating prism can be used to create motion. For example, a laser beam transmitted through a rotating octagonal prism can be scanned over a limited region without the beam pausing at the endpoints of the scan as would occur if an oscillating mirror were used.

A second example is the *range finder*, an afocal system, made from a beamsplitter and a plane mirror on a calibrated circle, as shown in Fig. 4.19. The observer looks through the beamsplitter and observes a distant object. The observer also sees the same object reflected off the fully reflecting mirror and the beamsplitter, but shifted with respect to the one transmitted through the beamsplitter. By rotating the mirror, the two objects may be superimposed. From the geometry given in Fig. 4.19, we see that if the angle θ is measured and the distance d is known accurately, the range to the distant object may be calculated from

$$r = d \tan \theta. \quad (4.9)$$

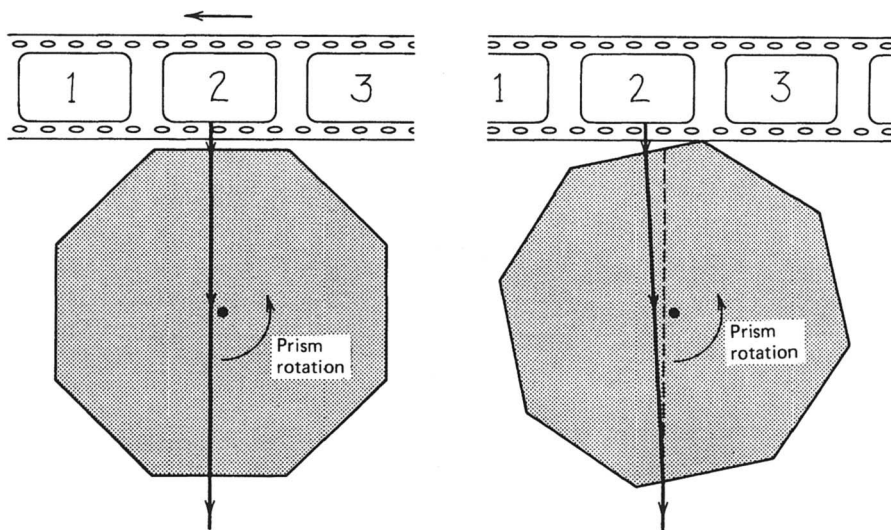


Figure 4.18. A motion-stopping prism. Rotating the octagonal prism at the correct rate relative to the film speed results in stopping the motion of the frame. The prism as depicted here for clarity is too large for the frame size. Frame 4 should be seen through the next pair of facets in the prism.

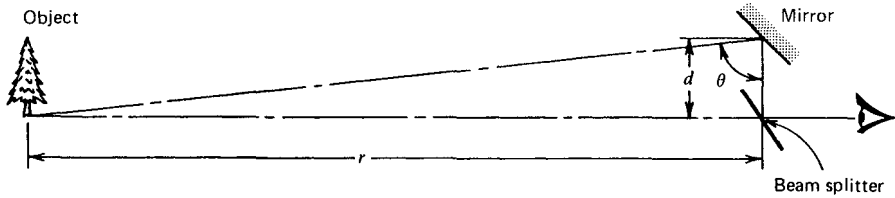


Figure 4.19. Range finder. The mirror is oriented to superimpose the virtual image on top of the object viewed through the beamsplitter.

The accuracy of this device is limited by the accuracy by which we can position and measure the mirror angle and by the accuracy with which we know d [Problem 4.11].

Sometimes the objective of an optical design requires not a highly compact system set in a minimum volume, but rather a system with a maximum length. This might be the case if a system were designed to measure the absorption of light due to some trace constituent in a gas—some atmospheric pollutant, for example. In such a system, one might direct a source to a distant corner cube reflector and collect the light returning from it, as shown in Fig. 4.20. Because of the divergence of the beam and the diffraction by the corner cube aperture, the collection lens is considerably larger than the output lens. A detector behind the collection lens records a fraction of the amount of light returning from the corner cube. Monitored over some duration, the time variation of the trace concentration could be determined.

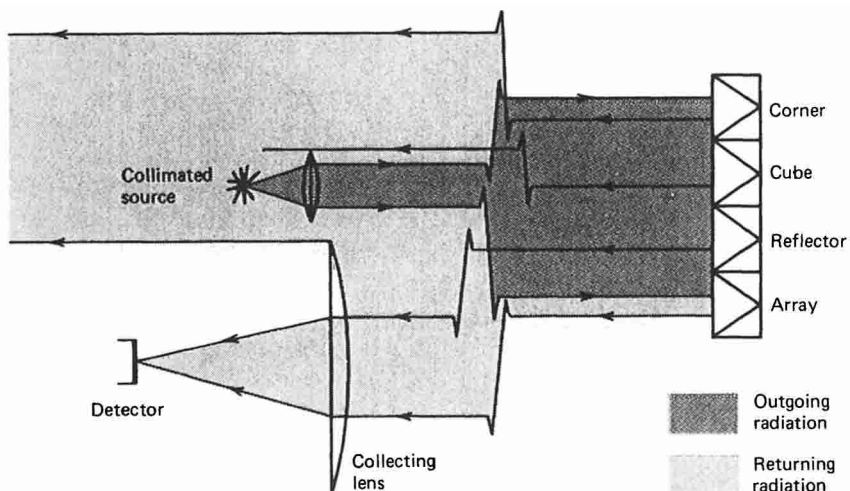
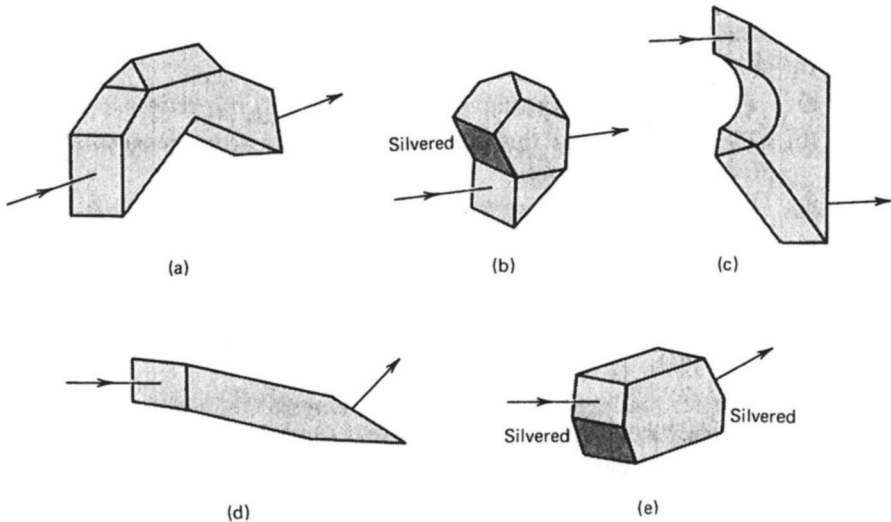


Figure 4.20. A long-path mirror system. Cube corners in an array are used to return a collimated beam to a detector in the vicinity of the source.

PROBLEMS

- 4.1. A television screen is to be magnified by a lens to produce a large screen display. What must be done to the television to get a correct unreversed, uninverted image in the following situations? (One of the things that can be done is to change the electronic circuitry to sweep from right to left, reversing the image. Thus one can turn the TV upside down or reverse the circuitry or both.) Draw each configuration.
- (a) The light passing through the lens hits a rear projection screen.
 - (b) The light passing through the lens hits a front projection (i.e., reflecting) screen.
 - (c) The light after passing through the lens is reflected back to a rear projection screen just above the TV.
 - (d) The light passing through the lens is reflected back onto a front projection screen just above the TV.
- 4.2. Using Fig. 4.4, prove that Eq. 4.1 is true.
- 4.3. The refractive index of diamond at the center of the visible part of the spectrum is 2.42. What is the reflectivity of light normal to the surface? What is the critical angle for this material? What does this have to do with the value of diamond as a precious stone?
- 4.4. Prove that the refractive index n_1 of a deposited quarter-wave layer is the geometric mean of the product of the air n_0 and substrate n_2 refractive indices (i.e., $n_1 = \sqrt{n_0 n_2}$) for complete cancellation of the reflected wave. Hint: for complete cancellation, the amount of light reflected by both interfaces must be equal.
- 4.5. For an air–glass ($n_2 = 1.52$) interface the best anti-reflection coating material is magnesium fluoride (MgF_2). If the refractive index of MgF_2 is 1.38, what percent of the incident light will be reflected off a surface coated with a quarter-wave layer of MgF_2 ? For a five-element-lens, what will be the fraction of light transmitted if they are coated in this manner?
- 4.6. How much variation in the power of a linearly polarized laser beam could be introduced by passing it through the polarizing beamsplitter described in Example 4.B and rotating the beamsplitter about the laser beam direction?
- 4.7. Using a pentaprism, a laser, and a mercury mirror, design a system that would serve as an “optical plumb” to define both a local vertical and a line perpendicular to the local vertical.
- 4.8. Using diagrams of the aperture of a Dove prism, show that rotation of a Dove prism about its base edge appears to rotate an image through twice the angle of rotation. (A closer examination will show that one is just reorienting the plane of reversal.)

- 4.9. Analyze the following prisms. All rays enter at normal incidence to the first surface. What is the final orientation of an image? Which surface might have to be silvered? What could it be used for?



- 4.10. Derive Eq. 4.7 using Fig. 4.16.
- 4.11. Assuming a 50-mm mirror separation and a 1-minute accuracy in angle measurement for a simple range finder, plot the accuracy of such an instrument as a function of range from 1 m to 10 m. At a measurement angle of $89^{\circ}42'$, what is the range? How accurate is the instrument?
- 4.12. An optical system whose total length is 2 m must be folded into a box whose sides can be no longer than 37.5 cm. The optic axis at the input point must be parallel, but not necessarily coincident, with the optic axis at the exit point. Because of losses in the optical system no more than 4 mirrors or reflections may be used. The reflectors must be large enough to reflect a $50\text{ mm} \times 50\text{ mm}$ beam. Design a system to meet these requirements. Give the locations of the mirrors and the direction of their normals to determine their orientation.

REFERENCES

- 4.1. J. E. Fuller and E. S. Wampler, "The Lunar Laser Reflector," *Scientific American*, Vol. 222, No. 3, pp. 38–49, 1970. Description of the experiment using the cube corner reflectors left by the Apollo astronauts on the moon to measure the distance from the earth to the moon.

- 4.2. E. Hecht and A. Zajac, *Optics* (Addison-Wesley, Reading, Mass., 1974), Section 4.3, pp. 71–89.
- 4.3. *Military Handbook 141, Optical Design* (U.S. Department of Defense, Washington, D.C., 1962). Contains several chapters on the design and construction of dielectric mirrors. (Available from the Naval Documents and Forms Center, Tabor Avenue, Philadelphia, Pa., 19120.)
- 4.4. L. Levi, *Applied Optics* (Wiley, New York, 1968), Section 8.3, “Prisms” and especially Appendix 8.2, “Major Types of Reflecting Prisms,” a collection of prism designs by F. Gardner.

5

Paraxial Ray Tracing

(HOW TO GET A LOT OF INFORMATION WITH VERY LITTLE EFFORT)

It would be convenient to be able to design and build all systems as if they were composed of thin lenses, but at some point we must take into account the thickness of a lens or other optical elements. This can be done using paraxial ray tracing, a method similar to the thin lens ray tracing described in Chapter 2. It serves as the first evaluation of a complex optical system and as the reference trace to which all other traces can be compared. After describing a method for calculating a paraxial ray trace in detail, we will use it to evaluate a set of aberrations that occur in many optical systems, chromatic aberration.

5.1. THE y - nu TRACE

Instead of attributing the bending of a ray to the focal power of a thin lens, we must determine the effect of a number of curved surfaces as a ray propagates through a system. Between interfaces, from surface k to surface $k + 1$, the law of transmission still holds:

$$y_{k+1} = y_k + u_k t_k, \quad (5.1)$$

where y is the ray height at the surface, u is the slope angle, and t is the separation between the two surfaces.

5.1.1. Derivation of Paraxial Equations

An equation must be derived to calculate the change in slope angle due to refraction at the interface. The conventions for angles, curvatures, and heights are those given in Table 2.1. Note that a clockwise rotation of the ray into the optic axis gives a positive angle. *The same rotation of the ray into a surface normal is also positive* (Fig. 5.1). To simplify the mathematics we use the *paraxial approximation* by choosing rays propagating very close to the optic axis. In this approximation the sine of an angle is equal to its angle ($\sin i_k = i_k$), and the Law of Refraction (Eq. 1.7) becomes

$$n_k i_k = n_{k+1} i_{k+1}. \quad (5.2)$$

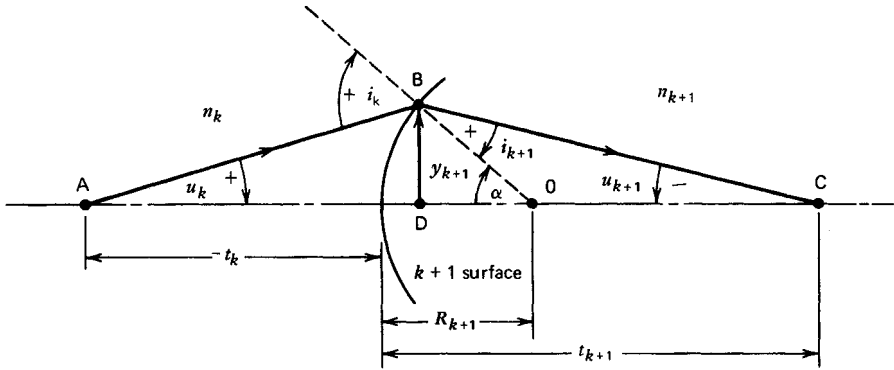


Figure 5.1. Ray trace at an interface. The sign of each of the ray angles is noted along with the symbol. In the paraxial approximation, the curved interface is replaced by a tangent plane at the surface vertex, the point where the interface intersects the optical axis.

The geometry for the refraction of rays at an interface is given in Fig. 5.1. To keep the notation general, we will look at the $k + 1$ st surface in a system where the slope angle of the incident ray u_k is changed to u_{k+1} by refraction at the surface. At the point where the ray intercepts the surface it has a height of y_{k+1} , calculated from Eq. 5.1. Also at that point the ray is incident at an angle i_k and refracted to an angle i_{k+1} and the local normal to the surface makes an angle α with the optic axis. In the same small angle approximation, the angles α , u_k , and u_{k+1} can be expressed as

$$\alpha \cong \tan \alpha = \frac{y_{k+1}}{R_{k+1}}, \quad (5.3a)$$

$$u_k \cong \tan u_k = \frac{y_{k+1}}{-t_k}, \quad (5.3b)$$

$$u_{k+1} \cong \tan u_{k+1} = -\frac{y_{k+1}}{t_{k+1}}, \quad (5.3c)$$

where R_{k+1} is the radius of curvature of the $k + 1$ st surface. The negative sign in Eq. 5.3c is due to the fact that u_{k+1} is a negative angle for this geometry. From the geometry shown in Fig. 5.1, we can equate i_k , the exterior angle to triangle ABO, to the sum of the interior angles

$$i_k = \alpha + u_k. \quad (5.4)$$

In triangle OCB the exterior angle, α , is equal to the sums of the interior angles,

$$\alpha = i_{k+1} - u_{k+1}. \quad (5.5)$$

The negative sign is inserted because u_{k+1} is a negative angle in the figure. Substituting for α in both Eqs. 5.4 and 5.5, using Eq. 5.3a and solving for i_k and i_{k+1} , we obtain:

$$i_k = \frac{y_{k+1}}{R_{k+1}} + u_k; \quad i_{k+1} = \frac{y_{k+1}}{R_{k+1}} + u_{k+1}. \quad (5.6)$$

Multiplying the first equation by n_k and the second by n_{k+1} and equating, using Snell's Law (Eq. 5.2), yields

$$n_k \frac{y_{k+1}}{R_{k+1}} + n_k u_k = n_k i_k = n_{k+1} i_{k+1} = n_{k+1} \frac{y_{k+1}}{R_{k+1}} + n_{k+1} u_{k+1}. \quad (5.7)$$

Finally, solving for the slope angle of the refracted ray times its refractive index,

$$n_{k+1} u_{k+1} = n_k u_k - \frac{(n_{k+1} - n_k)}{R_{k+1}} y_{k+1}. \quad (5.8)$$

This can be expressed more compactly by collecting subscripts and defining new (or maybe, nu) quantities,

$$nu_{k+1} \equiv n_{k+1} u_{k+1}; \quad nu_k \equiv n_k u_k \quad (5.9)$$

and

$$c_{k+1} \equiv \frac{1}{R_{k+1}}, \quad (5.10)$$

where c_{k+1} is the curvature of the $k + 1$ st surface, with the result

$$nu_{k+1} = nu_k - y_{k+1}(n_{k+1} - n_k)c_{k+1}. \quad (5.11)$$

The factor $(n_{k+1} - n_k)c_{k+1}$, the power of the surface, has the same effect that the power of the lens had in the thin lens equation.

$$\phi_{k+1} = \frac{(n_{k+1} - n_k)}{R_{k+1}} y_{k+1} = (n_{k+1} - n_k)c_{k+1} y_{k+1}. \quad (5.12)$$

Note that there are an infinite combination of curvatures and refractive index differences that can produce the same refracting power. This property is a valuable asset when it comes to correcting aberrations.

Just as for thin lenses, there are two equations, 5.1 and 5.8, that are linear in y and u . In some cases, especially for programming calculations, it may

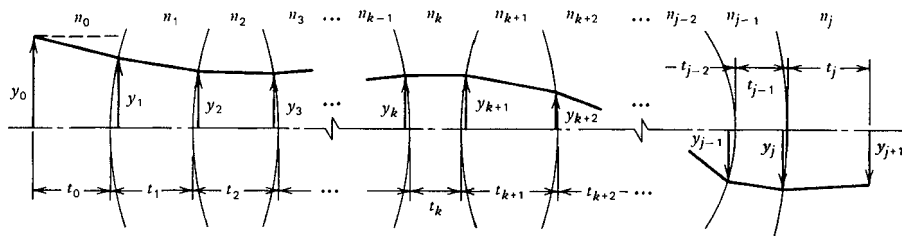


Figure 5.2. Subscript notation for ray tracing through multiple surfaces. (The vertical arrows represent ray heights and not the usual symbol for an object.)

be useful to express both equations in terms of the refractive-index-angle product, nu :

$$y_{k+1} = y_k + \frac{nu_k t_k}{n_k} \quad (5.13a)$$

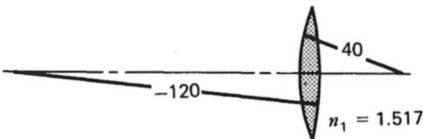
and

$$nu_{k+1} = nu_k - y_{k+1}(n_{k+1} - n_k)c_{k+1}. \quad (5.13b)$$

The method used for ray tracing through a series of surfaces is similar to that used for thin lenses. The differences are that the lens power is replaced by the product of the refractive index difference and the surface curvature (i.e., by the surface power), and the angle variable u in the thin lens ray trace is replaced by nu variable. The method is called the y - nu trace. An example of a paraxial trace of a single thick lens is given in Example 5.A. As before, we can arrange the system parameters (R_{k+1} or c_{k+1} , n_k , t_k) in an array and calculate the ray values through the system. Figure 5.2 gives an overall idea of the subscript notation for a y - nu trace. Note that the subscript changes *at* the interface. All values immediately to the left of the interface have a subscript one less than those at or immediately to the right of the interface. The zero subscript values are assigned to object space and, assuming j surfaces and interfaces, the j th subscript is assigned to image space. The one exception is y_{j+1} , which is the image height. A stop is treated as any other surface except that the refractive index does not change across the stop surface and the surface has zero curvature ($c_k = 0$).

To *open*, one assigns an object height and object distance and initial angle, unless the object distance is infinite and one is tracing a parallel ray, in which case $u_0 = 0$. To *close*, one can specify a number of different parameters. Each of these is called a *solve*. In many cases, all that is required is to find the image distance in an axial ray trace. Thus, to close one need only set $y_{j+1} = 0$ and solve Eq. 5.1 for t_j , as was done in the thin lens method. In other cases, one may

EXAMPLE 5.A. Calculate the effective focal length for a lens of crown glass (glass # 517645) of a thickness of 4 mm, with radii of curvatures, $R_1 = 40$ mm and $R_2 = -120$ mm. Assume $n = n_d = 1.517$.



Solution: Set up a system array:

t_k	∞	4	56.53	closing
n_k	1.0	1.517	1.0	
c_k	0	1/40	-1/120	
r_k	—	—	—	— no radii given

Parallel ray				
y_k	10	10	9.659	0 to close
nu_k	0	-0.1292	-0.17086	
u_k	0	-0.0852	-0.17086	

$f_{\text{eff}} = -y_o/u_2 = 10/0.17086 = 58.526$ mm

Since there is only one aperture, it will serve as the aperture stop. The field stop is either in object or image space.

Note: If you need to find an angle in a material when you are doing a y - nu trace, remember to divide out the index.

wish to maintain a certain light-collecting ability, or to position an image plane or a pupil plane. In these instances, as will be discussed later, other solves are needed.

5.1.2. Cardinal Points of a Lens

The introduction of element thickness into our design not only forces us to revise the method of calculation, but it also requires that some new definitions be provided to describe the properties of a thick lens. When the thin lens was discussed there was no doubt where the lens was located and what the focal length was. But, from what point is a focal distance or any other quantity measured in a thick lens or a system composed of multiple elements? The

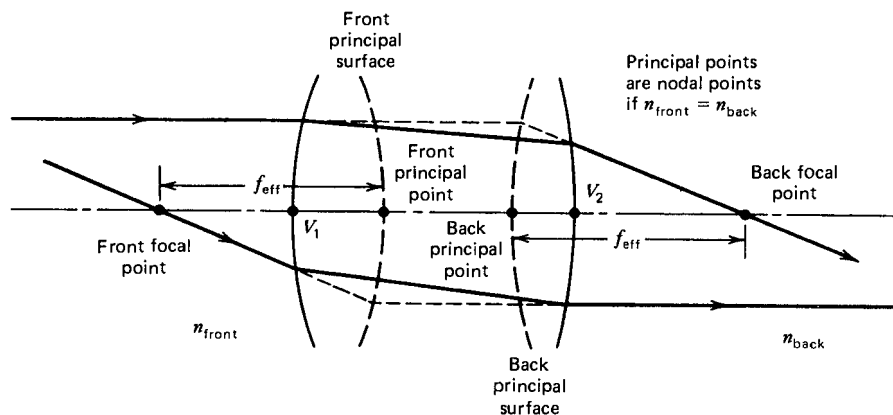


Figure 5.3. Cardinal points and principal surfaces of an optical system. In addition to the cardinal points, the vertices of the system (V_1 and V_2) are shown. If the refractive indices in the front and back of the system are the same, the principal points and nodal points of the system are coincident.

procedure for designing systems can be simplified since the lens, or even a series of lenses, can be replaced with six quantities, known as the *cardinal points* of a lens. These points, lying on the optic axis, are illustrated in Fig. 5.3. To describe them it is necessary, however, to set up some directions. When referring to lenses (and optical systems), the *front* of the lens is the side facing the object, and the *back* of the lens faces into the space where light is transmitted by the lens. Two points in the figure not defined by ray tracing and not included as cardinal points are the *front* and *back vertices* of the lens, the points where the optic axis intersects the front and back surfaces of the lens.

The first set of cardinal points are the front and back focal points. Starting from the concept of a focal point as the point on the optic axis through which all parallel input rays are directed by the lens, the focal length of the lens could be defined as the distance between the back vertex and the back focal point. This distance is the *back focal length* (for Example 5.A it is equal to 56.53 mm). Parallel light, incident from the back of the lens, will focus at the front focal point (57.86 mm from the front vertex for Example 5.A). This distance is the *front focal length*. But these focal lengths usually do not correspond to the effective focal length of a lens that is used in most calculations. The effective focal length of the lens is defined in relation to the so-called principal surfaces of the lens.

As with our example of the telephoto lens in Chapter 2, a parallel ray is extended as a straight line, and the ray segment through the focal point is extended back along the ray to cross the parallel ray; their intersection is a point on the *principal surface* of the lens. When this is done for rays of differing heights to the axis, the principal surface is defined. The intersections of the two principal surfaces with the optic axis are the *principal points* of the lens, the second set of cardinal points (Fig. 5.3). Just as in Chapter 2, the *effective focal length* of a lens or an optical system is defined as the distance from the focal point to the

principal point of the lens and is the same for the front and back of the lens, provided the surrounding media are the same on both sides. In the paraxial approximation, the curved principal surfaces can be replaced by planes located at the principal points. These *principal planes* are used to simplify the layout of optical designs.

The final set of cardinal points are the nodal points. If parallel rays along the optic axis are incident on a lens, they will focus at the focal point. If the lens is rotated at an angle to these rays, the focus of the rays will, in general, move as the lens is rotated. There is one point on the optic axis where the focus remains fixed when the lens is rotated (Fig. 5.4). This is one of the *nodal points* of the system. Thus, by scanning the rotation point along the optic axis until the image movement stops, the location of the nodal point in a lens system can be found. Obviously, the lens has another nodal point, which is found by turning the lens around so that parallel light enters the other side. If the refractive indices of the media in the front and back of the lens are the same, the nodal points are coincident with the principal points.

The advantage of these concepts is that a thick lens or an entire complicated optical system can be replaced by its cardinal points and principal planes in any analysis. Similar rules for ray sketching that applied to thin lenses can be applied to these replacements:

1. A parallel ray entering the system will emerge from the back principal plane directed toward the back focal point.
2. A ray through the front focal point will emerge from the front principal plane parallel to the optic axis.
3. A ray through the front nodal point emerges from the back nodal point with the same slope angle.

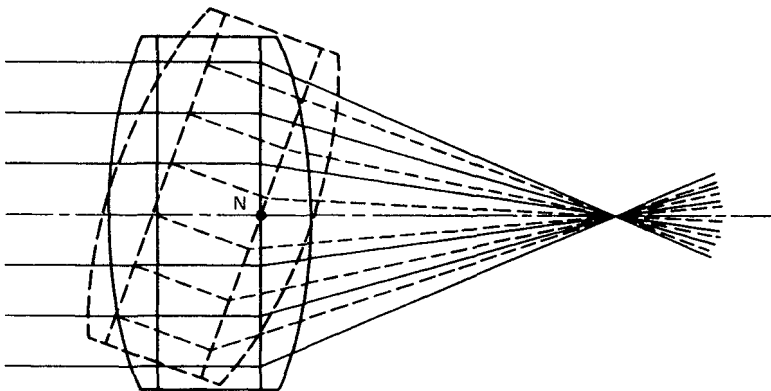


Figure 5.4. The nodal point of a lens is defined as the point on the optical axis of a lens around which the lens may be rotated without producing motion of the image. Here, the lens is rotated about the back nodal point, *N*. (Shown in dashed lines.)

4. All other rays travel *between the two principal planes* as if parallel to the optic axis, then continue their transmission from the back principal plane as shown in Fig. 5.5.

In the calculations in Example 5.A, we see that the angle of the ray through the focus (u_2) in the paraxial approximation is equal to the input ray height (y_0) divided by the distance from the principal point to the back focal point. This distance, the effective focal length, may be calculated from:

$$f_{\text{eff}} = - \frac{\text{parallel input ray height}}{\text{output ray angle}} = - \frac{y_0}{u_j} \cdot \frac{y_j}{|y_j|} \quad (5.14)$$

The effective focal length of the lens in Example 5.A is 58.526 mm. Using the values for a parallel ray traced backward through the system to the front focal point, one gets the same value for the effective focal length [Problem 5.2]. In this case, however, the intersection of the input and output rays is at the front principal plane. These concepts can be applied not only to single thick lenses, but also to elaborate multi-element imaging systems [Problem 5.3]. Thus six points—two focal points, two principal points, and two nodal points—plus entrance and exit pupils can replace an extremely complex optical system in many design studies.

5.1.3. The Lensmaker's Equation

Although it is useful for only two surfaces, a y - nu trace of a single thick lens with symbols instead of specific values provides a simple derivation for the power of a thick lens [Problem 5.5]:

$$\phi = \frac{1}{f} = (n - 1) \left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{t(n - 1)}{nR_1R_2} \right]. \quad (5.15)$$

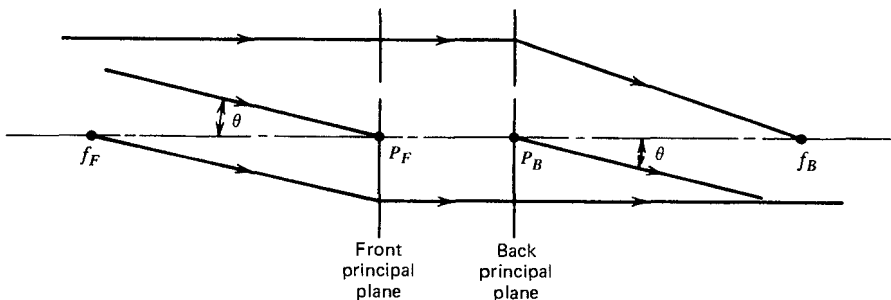


Figure 5.5. Ray trace of a system in terms of its principal planes.

In the limit of a thin lens ($t \rightarrow 0$), Eq. 5.15 becomes the lensmaker's equation, which was stated in Chapter 1 (Eq. 1.18):

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right). \quad (5.16)$$

If one of the surfaces of a thick lens is flat (a plano lens), the power of the lens is independent of thickness. (It has the same value as a thin lens with the same radius and index.) In general, however, an infinite number of combinations of construction parameters (radii, index, and thickness) can be used to arrive at a particular focal length. Two simple cases are shown in Example 5.B. Choosing a refractive index of 1.5 (so that $n - 1 = \frac{1}{2}$) and using two simple cases, the range of radii of curvature is from about $f/2$ to f , unless a meniscus-type lens is used. So, to start from scratch, a glass and a focal length are chosen, and if there are no other guidelines, a plano or symmetrical lens is used. The curvatures are calculated. Later, when other requirements are imposed, these assumptions will have to be modified.

EXAMPLE 5.B. What should the radii of curvature be for a thin lens with a focal length of 45 mm?

Solution: Not enough information yet, but we can look at two simple cases.

- (a) **Plano-Convex Lens:** $c_2 = 1/R_2 = 0$.
Then $f = 45 \text{ mm} = R_1/(n - 1)$.
Thus if we chose a crown glass $n = 1.5$,

$$R_1 = (n - 1)f = f/2 = 22.5 \text{ mm}.$$

- (b) **Biconvex Lens:** $R_2 = -R_1$ ($c_2 = -c_1$).
Then $1/f = (n - 1) \cdot (2/R_1)$ and for crown glass $(n - 1) = \frac{1}{2}$.

$$f = \frac{R_1}{2(n - 1)} = R_1 = 45 \text{ mm}.$$

So to start from scratch, choose either the glass or the radius and, using either type of lens, calculate the other to get some initial values for the lens.

5.1.4. Parallel Plates in an Imaging System

Besides causing a lateral (sideways) displacement of a ray, a parallel plate inserted between an element and an image will displace the image position

along the optical axis. The paraxial ray trace is useful in determining the effect of a parallel plate in an optical system. It establishes whether or not the plate serves as a stop, and it enables the designer to specify the smallest possible components without restricting either the light throughput or the extent of the image. As shown in Fig. 5.6, the incident and emerging ray angles u_{k-1} and u_{k+1} are equal. By Snell's Law, for paraxial rays the ray angle inside the plate is $u_k = u_{k+1}/n$, where n is the refractive index of the plate. Without the plate, the distance from the image to the point where the ray height is equal to y_k is

$$t_k = -\frac{y_k}{u_{k-1}}. \quad (5.17)$$

(The negative sign here and in the derivation below is consistent with the fact that all of the ray slope angles in the example are negative.) With the plate in place, one additional surface is added so that the ray height at the other side of the plate is

$$y_{k+1} = y_k + u_k d = y_k + \frac{u_{k+1}}{n} \cdot d. \quad (5.18)$$

The distance from the front of the plate to the image is the thickness of the plate plus $(-y_{k+1}/u_{k+1})$ or

$$\begin{aligned} t'_k &= d - \frac{y_{k+1}}{u_{k+1}} = d - \frac{1}{u_{k+1}} \left(y_k + \frac{u_{k+1}}{n} \cdot d \right), \\ &= d - \frac{y_k}{u_{k+1}} - \frac{d}{n} = d - \frac{d}{n} + t_k, \end{aligned}$$

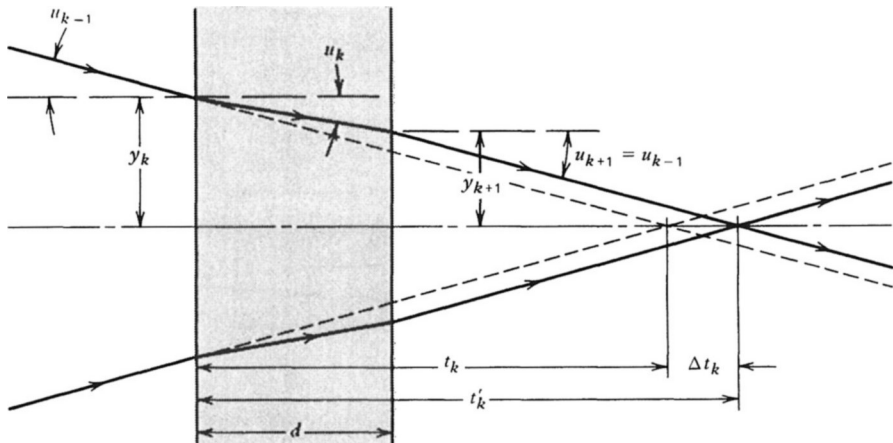


Figure 5.6. Axial displacement of an image by a plate of glass with parallel surfaces.

where we have used Eqs. 5.18 and 5.17. The total distance to the image point is $t_k = -y_k/u_{k-1}$ without the plate. Thus the difference in the image locations is

$$\Delta t_k = t'_k - t_k = d - \frac{d}{n} = \frac{(n-1)d}{n}. \quad (5.19)$$

As an example, we apply these equations to determine the effect of a set of Porro prisms in an optical system. The two prisms must be unfolded and replaced by the equivalent parallel plates, as shown in Fig. 5.7. We see that a prism whose hypotenuse is equal to L is equivalent to a plate of glass of thickness L . Thus, each prism adds $(n-1)L/n$ to the distance. In drawing prisms in optical systems it is sometimes useful to reduce the width of the equivalent parallel plate by the image displacement distance Δt_k . This equivalent width,

$$d_{eq} = 2L - \Delta t_k = 2L - 2L \frac{(n-1)}{n} = \frac{2L}{n}, \quad (5.20)$$

allows the rays through the prism to be drawn as straight lines, as if the prisms were not there. It is referred to as the *equivalent air thickness or reduced thickness* [Problem 5.4]. The height of the slab is still L , and it must be taken into account when determining the aperture and field stops of a system. In Example 5.C, the reduction in length of a telescope and the minimum size of the prisms are determined using the above principles.

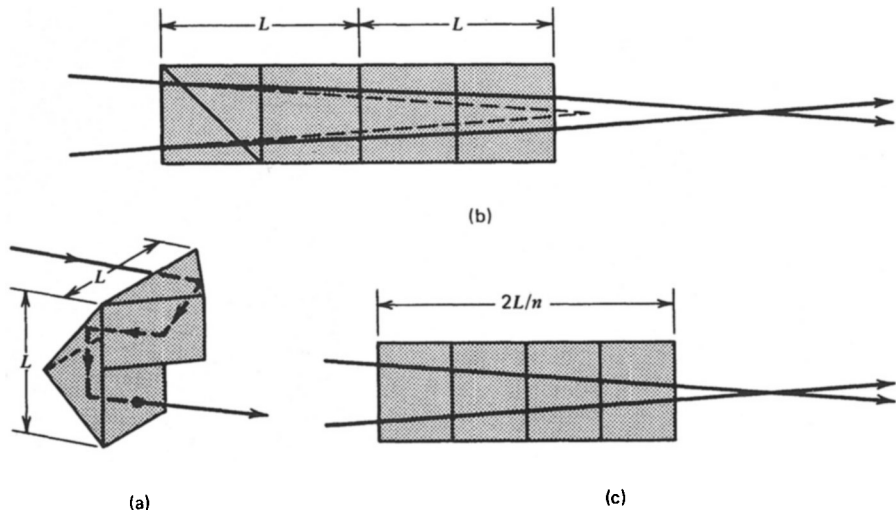
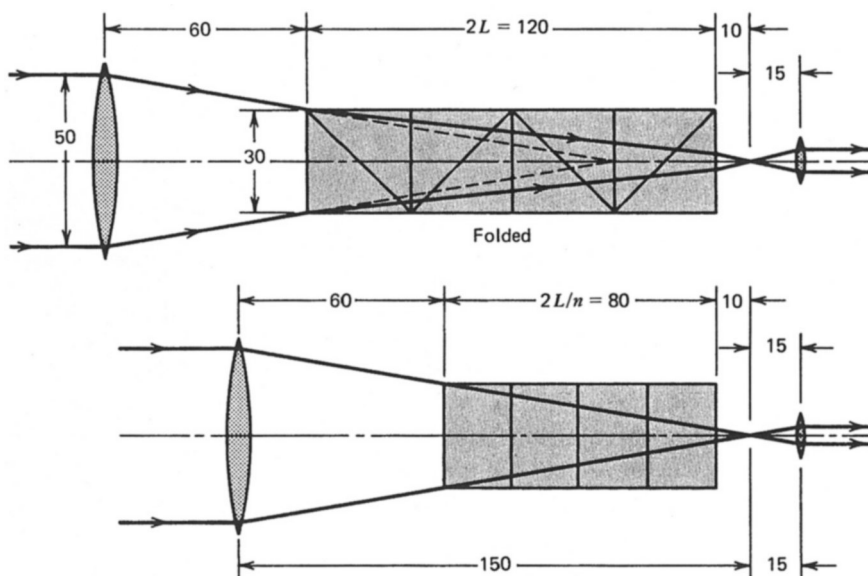


Figure 5.7. Effect of Porro prisms on imaging. (a) A perspective drawing of an axial ray. (b) The unfolded diagram of two rays traversing the prism. (c) The reduced thickness drawing in which rays may be drawn undeviated.

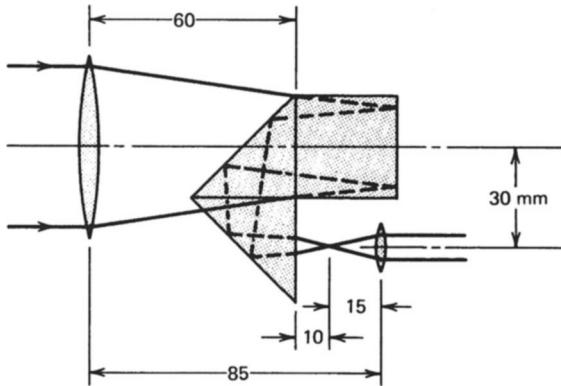
EXAMPLE 5.C. What is the reduction in the length of a $10\times$ telescope designed with Porro prisms (see Section 4.3 and Fig. 4.13) so that the prisms do not reduce the optical speed of the system (i.e., none will become an aperture stop). The refractive index of the prism is 1.5.

Solution: Assuming a nominal eye pupil of 5 mm, a $10\times$ telescope must have an objective with a clear aperture that is 10 times as large, or else the eye becomes the aperture stop. The only room for the prism is between the objective lens and the intermediate image. Smaller (i.e., less expensive) prisms could be used if they were placed close to the intermediate image. However, a prism surface should not be placed at the intermediate image since any defect or dust on the prism will be in focus with the image.

Assuming a 150-mm objective with a 50-mm clear aperture (large!), the prism apertures can be somewhat less. If the first prism is placed 60 mm from the objective lens, the hypotenuse of the prism must be 60 mm since the edge rays entering one-half the prism are 30 mm apart. A scaled diagram with unfolded prisms ($n = 1.5$) of reduced length, as illustrated in Fig. 5.7, is shown below.



Note that the effect of sending the rays through a slab of glass of thickness $2L$ is to increase the pathlength by $2L - 2L/n$ or $2(n - 1)L/n$. But the input plane for the prisms is also the output plane, reducing the length by $2L$. The net reduction then is $2L - (2(n - 1)L/n)$ or simply $2L/n$. The overall length (folded) will be $150 + 15 - 80$ mm or 85 mm.



The optical axis is placed downward by half the hypotenuse of the prism, or 30 mm.

5.2. THE GAUSSIAN CONSTANTS

Because of the linearity of the paraxial equations in y_k and u_k , a number of useful expressions and relations can be derived. Although the paraxial equations are approximations, they can be used to derive the basic parameters of an optical system, known as the *Gaussian constants*. These include the various focal lengths already discussed, the stops, pupils, windows, and f -number of the system, the image distance and magnification, and the optical invariant.

5.2.1. The Optical Invariant

Using the paraxial ray trace equations, we can show that a combination of axial and oblique ray parameters are constant throughout an optical system. In Fig. 5.8 an axial ray and an oblique (nonaxial, nonparallel) ray are traced through an optical system defined by refractive indices n , curvatures c , and thicknesses t . At the $k + 1$ st surface, the refracting equations give

$$\text{(Axial)} \quad nu_{k+1} = nu_k + y_{k+1}(n_{k+1} - n_k)c_{k+1},$$

$$\text{(Oblique)} \quad \bar{n}u_{k+1} = \bar{n}u_k + \bar{y}_{k+1}(n_{k+1} - n_k)c_{k+1},$$

where the barred quantities refer to the oblique ray. Solving both of these equations for the power of the $k + 1$ st surface, $(n_{k+1} - n_k)c_{k+1}$, and equating yields

$$\frac{nu_{k+1} - nu_k}{y_{k+1}} = \frac{\bar{n}u_{k+1} - \bar{n}u_k}{\bar{y}_{k+1}}.$$

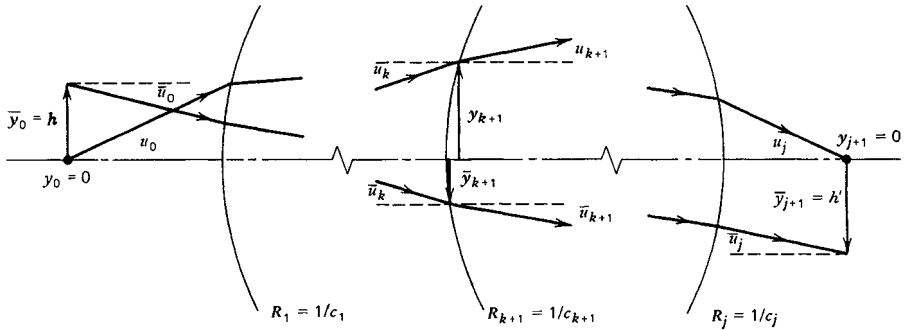


Figure 5.8. Diagram for derivation of the optical invariant.

Multiplying and rearranging, we obtain

$$\bar{y}_{k+1} n u_{k+1} - y_{k+1} \bar{n} \bar{u}_{k+1} = \bar{y}_{k+1} n u_k - y_{k+1} \bar{n} \bar{u}_k.$$

Using the transfer equation [Problem 5.6], the right side of this equation can be written in terms of variables with only k subscripts:

$$\bar{y}_{k+1} n u_{k+1} - y_{k+1} \bar{n} \bar{u}_{k+1} = \bar{y}_k n u_k - y_k \bar{n} \bar{u}_k.$$

Note that the quantities on each side of the equation have only one subscript, and since this derivation was completely general, the quantity

$$I = \bar{y}_k n_k u_k - y_k \bar{n}_k \bar{u}_k \quad (5.21)$$

is constant from one medium to another and is an invariant of the system, called the *optical* or *Lagrange invariant*. Although we called the unbarred ray, the axial ray, and the barred ray an off-axis ray, the derivation is true for any two rays as long as one is not a scaled version of the other. That is, the second ray is not the first ray with all the ray heights and slope angles multiplied by a constant.

One particularly simple and useful case where the optical invariant may be applied is in determining the image magnification from an axial ray. First the optical invariant is evaluated in both object space and image space. In object space,

$$y_0 = 0; \bar{y}_0 = h, \text{ object height,}$$

and

$$I = h \cdot n u_0 - 0 \cdot \bar{n} \bar{u}_0 = h n u_0 = h n_0 u_0.$$

In image space, after j surfaces:

$$y_{j+1} = 0; \bar{y}_{j+1} = h', \text{ the image height,}$$

and

$$I = h' \cdot nu_j - 0 \cdot \bar{n}u_j = h' \cdot nu_j = h'n_ju_j.$$

Equating the values for the invariant in object and image space we obtain

$$hn_0u_0 = h'n_ju_j.$$

The transverse magnification of the system is therefore

$$M = \frac{h'}{h} = \frac{n_0u_0}{n_ju_j} \quad \text{or} \quad M = \frac{u_0}{u_j} \quad \text{if} \quad n_0 = n_j. \quad (5.22)$$

Since both of the ray angles are those for the axial ray, this means that an axial ray trace through a series of lenses will not only determine the image position but also its magnification. This obviously holds for a thin lens ray trace too.

5.2.2. Finding the Stops—Scaling

The linearity of Eqs. 5.13a and 5.13b also allows one to write down a whole family of rays by simply multiplying the values found in an axial trace by a constant, a procedure known as *scaling*. For example, the paraxial approximation to the marginal ray can be determined by dividing the clear radius of the k th element by y_k found by an axial ray trace. The smallest absolute value of these ratios determines the aperture stop. If the smallest ratio is at the k th element, then

$$\gamma = \left| \frac{r_k}{y_k} \right|$$

and

$$y'_k = \gamma y_k \quad \text{and} \quad u'_k = \gamma u_k$$

give the values of the marginal ray in the system, since $y'_k = \gamma y_k = r_k$. The ray height at the aperture stop is equal to the clear aperture radius. The next problem to be attacked is to find the entrance and exit pupils. This is done by tracing a chief ray through the system starting at the aperture stop and tracing into image and object spaces. The images of the aperture stop locate the entrance and exit pupils. Since the chief ray, by definition, crosses the optic axis

at the aperture stop, the locations of the entrance and exit pupils are found by setting $\bar{y}_0 = 0$ and $\bar{y}_{j+1} = 0$, respectively. For the entrance pupil

$$\bar{y}_1 = \bar{y}_0 + \bar{u}_0 \bar{t}_0 = +\bar{u}_0 \bar{t}_0$$

or

$$\bar{t}_0 = \frac{\bar{y}_1}{\bar{u}_0} \quad (\text{entrance pupil location}),$$

and after j surfaces:

$$\bar{y}_{j+1} = \bar{y}_j + \bar{u}_j \bar{t}_j = 0,$$

$$\bar{t}_j = -\frac{\bar{y}_j}{\bar{u}_j} \quad (\text{exit pupil location}).$$

By using the optical invariant, the size of the pupils can be determined from the magnification as shown in Example 5.D.

Again, the limiting aperture for the chief ray can be found, giving the field stop. Dividing the chief ray heights at each element by the clear aperture radius, the field stop occurs at the element with the smallest $|r_k/\bar{y}_k|$ ratio. It might seem that another ray trace would be needed to find the entrance and exit windows of the system, but we already have sufficient information to determine their location and size. As was noted in Chapter 2, any new paraxial ray having nonzero y_0 and u_0 is just the sum of an axial ray and chief ray with appropriate multipliers. In the equations below let the unprimed variables be those for an axial ray, the barred variables, a chief ray, and primed variables, the new paraxial ray. If we wish to find the image of surface m in image and object space, then we set $y'_m = 0$. But y'_m can be written as the sum of an axial ray and a chief ray, so that at surface m ,

$$y'_m = \alpha y_m + \beta \bar{y}_m = 0. \quad (5.23)$$

Because y'_m can be scaled as can any other ray, we arbitrarily get $\alpha = 1$, then

$$\beta = -\frac{y_m}{\bar{y}_m}. \quad (5.24)$$

In object space

$$y'_1 = y_1 + \beta \bar{y}_1,$$

$$u'_0 = u_0 + \beta \bar{u}_0.$$

The image of the m th surface is found by setting $y'_0 = y'_1 - u'_0 t'_0 = 0$. (As was noted in Section 2.5, when a ray is traced backward, the sign convention for distances is reversed. If the entrance pupil location is to the right of surface 1 then it will be negative.) The location of the image of this surface in object space is at

$$t'_0 = \frac{y'_1}{u'_0} = \frac{y_1 + \beta \bar{y}_1}{u_0 + \beta \bar{u}_0} = \frac{y_1 - y_m \bar{y}_1 / \bar{y}_m}{u_0 - y_m \bar{u}_0 / \bar{y}_m}, \quad (5.25)$$

using values found from the two earlier traces. Similarly, in image space

$$y'_j = y_j + \beta \bar{y}_j,$$

$$u'_j = u_j + \beta \bar{u}_j,$$

and the image of the m th surface is at

$$t'_j = -\frac{y'_j}{u'_j} = -\frac{y_j - y_m \bar{y}_j / \bar{y}_m}{u_j - y_m \bar{u}_j / \bar{y}_m}. \quad (5.26)$$

The value of this method is that once the axial and chief rays are traced, no additional traces need be done, and no calculations are needed at intermediate surfaces. If the m th surface were the field stop of the system, then t'_0 and t'_j locate the entrance and exit windows, respectively. By finding the magnification of the field stop, the sizes of the windows can also be calculated [Problem 5.7–5.12]. In Example 5.D, an example of this method is given. Obviously, such a general method can be programmed into a calculator or small computer to provide these values in the design.

EXAMPLE 5.D. Determine the size of the entrance and exit pupils of a two-lens system using the optical invariant. The object distance is 11 units.

Solution: Set up the array and trace axial and chief rays.

t_k	11	0.3	4.0	0.3	1.452	
n_k	1.0	1.517	1.0	1.517	1.0	
c_k	0	0	-0.5	0.0888	-0.0888	0
r_k	—	.20	.20	.10	.10	—

Axial ray						
y_k	0	1.1	1.11978	0.36194	0.3212	0
nu_k	0.1	0.1	-0.18946	-0.2061	-0.22085	
u_k	0.1	0.06592	-0.18946	-0.13586	-0.22085	
r_k/y_k		.1818	.1786°	.2763	.3113	

	Chief ray								
\bar{y}_k	-1.1198	-0.01978	0	0.4	0.4161	0.5070			
$\bar{m}\bar{u}_k$	0.1	0.1	0.1	0.0816	0.0625				
\bar{u}_k	0.1	0.06592	0.1	0.0538	0.0625				
r_k/\bar{y}_k	10.11	—	.25	.2405 ^b					

^a Aperture stop.^b Field stop.

To determine the locations of the pupils: the aperture stop at surface 2 is the “object” point and the pupils are its images in object and image space. The “object” slope angle is the angle at the chief ray \bar{u}_2 and the “image” slope angles are the chief ray angles in the object space for the entrance pupil and in the image space for the exit pupil.

$$M_{\text{ent pup}} = \frac{\bar{u}_2}{\bar{u}_0} = \frac{0.1}{0.1} = 1 \quad r_{\text{ent pup}} = 0.20$$

(As though looking through a pane of glass, since the first surface is plano.)

$$M_{\text{ext pup}} = \frac{\bar{u}_2}{\bar{u}_4} = \frac{0.1}{0.0625} = 1.6 \quad r_{\text{ext pup}} = 0.20 \times 1.6 = 32.$$

Using the chief ray values, the entrance pupil can be located at

$$t'_0 = \frac{\bar{y}_1}{\bar{u}_0} = \frac{-0.01978}{0.1} = -0.1978,$$

which lies to the right of the first surface because of the sign convention for rays traced backward. The exit pupil is located at

$$t'_4 = \frac{-\bar{y}_4}{\bar{u}_4} = -\frac{0.4161}{0.0625} = -6.6576$$

to the left of the last surface. Since the field stop is located in image space, the exit window is also there. The entrance window may be found with another ray trace or using the β -ratio. Here $m = 4$ and

$$\beta = \frac{-y_4}{\bar{y}_4} = -\frac{0.3212}{0.4161} = -0.7719,$$

$$t'_0 \approx \frac{1.1 + (-0.772)(-0.01978)}{0.1 + (-0.772)(0.1)} = \frac{1.11527}{0.0228}$$

$$\approx 48.9$$

to the left of the first surface based on the backward ray trace sign convention.

The optical speed, or f -number, of the system, which we will denote as F , is defined for a parallel ray entering the entrance pupil of the system at a height of y as

$$F = \frac{\text{effective focal length}}{\text{entrance pupil diameter}}. \quad (5.27)$$

But from the definition of f_{eff} (Eq. 5.14) this is equal to

$$F = \frac{f_{\text{eff}}}{2y_1} = \frac{1}{2|u_j|}. \quad (5.28)$$

Since the f -number measures the light-collecting ability, f_{eff} is positive and the f -number is also positive. If y_1 is the entrance pupil radius, u_j is the exit slope angle of a marginal ray with an object at infinity. If the image space refractive index is not unity, then

$$F = \frac{1}{2n_j|u_j|}. \quad (5.29)$$

The procedure of scaling can be applied not only to rays in optical systems, but also to the optical system itself. A system can be analyzed using any convenient set of units. When an analysis is finished and a specific quantity, say, the effective focal length, is required, one need only multiply all radii, separations, and apertures by a constant equal to the new quantity divided by the old quantity. This scaling procedure does not affect the angular field of view or the system f -number since they are the ratios of two scaled distances.

5.3. CHROMATIC ABERRATION

Until now we have assumed that a single refractive index could characterize the glass in an optical system. But, as was pointed out in Chapter 1, the refractive index of all materials varies with wavelength. This dispersion in optical materials is responsible for chromatic aberration. The effects of chromatic aberration can be investigated by using the paraxial ray trace. After determining the amount of chromatic aberration present in a system, it is possible, through the use of glasses of different dispersions, to reduce the amount of aberration with the aid of some simple thin lens formulae.

5.3.1. Chromatic Aberration of a Simple Lens

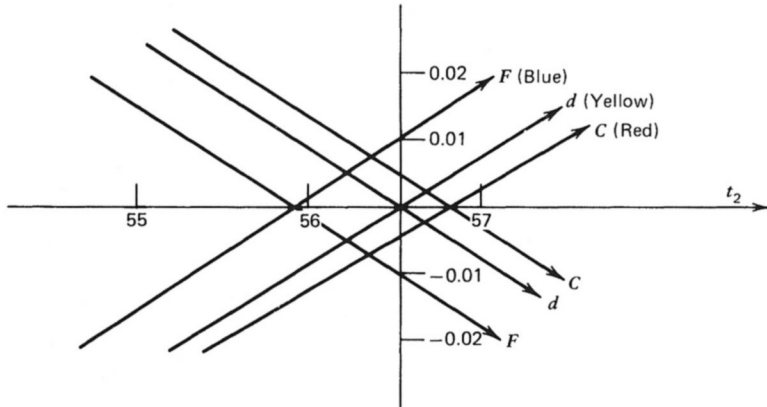
In Example 5.E, the effective focal length of a crown glass (Glass # 517645) lens is determined for three wavelengths. In this case we choose the C , d , and F lines at 656.3, 587.6, and 486.1 nm. (The concepts and notation for the glass number, the dispersion in the form of the V -number, and the notation and wavelengths of

the Fraunhofer lines were given in Section 1.1.) To serve as a reference point, the trace is closed on the central d -line. (The point at which the rays cross the axis are also shown as t_2 in the example.) Note that the blue rays are focused inside the paraxial focus for the d -line, while the red rays are focused beyond it. This defect in the lens is called *axial color*, although it occurs at off-axis points, also.

EXAMPLE 5.E. Calculate the effective focal length for a lens of crown glass (glass #517645) of thickness 4 mm with radii of curvatures, $R_1 = 40$ mm and $R_2 = -120$ mm for wavelengths d , F , and C lines.

				C: 56.7983	The array has been set up to parallel all three.
				d: 56.5311	
				F: 55.9126	
t	∞		4		
n_C			1.51461		
n_d	1.0		1.51700	1.0	
n_F			1.52262		
c	0	$\frac{1}{40}$	$-\frac{1}{120}$		
<hr/>					
y	C		0.9660236	0.004544	(to close,
d	1.0	1.0	0.9659196	0.0	← determines
F			0.9656763	-0.010683	reference
	C	0	-0.0084941	-0.017008	image plane)
nu	d	0	-0.0085201	-0.017086	
	F	0	-0.0085809	-0.0172712	

f_{eff} $C: 1/0.017008 = 58.796$
 $d: 1/0.017086 = 58.526$
 $F: 1/0.0172712 = 57.900$



Axial color can be measured in terms of the separation between the C and F image points, the *longitudinal axial chromatic aberration*,

$$L_{\text{ch}} = t_F - t_C, \quad (5.30)$$

where t_F is the distance beyond the final element where the F -ray crosses the axis and t_C is the corresponding distance for the C -ray. Thus, in Example 5.E, $L_{\text{ch}} = 55.9126 - 56.7983 = -0.8857$. It is sometimes useful, for comparison, to normalize this error by the focal length, so that $\bar{L}_{\text{ch}} = L_{\text{ch}}/f_d = -0.8857/58.526 = -0.01513$.

Axial color can also be measured in terms of the separations between the F - and C -rays in the focal plane of the d -ray. This is the *transverse axial chromatic aberration*

$$T_{\text{ch}} = y_F - y_C, \quad (5.31)$$

where y_F and y_C are the ray heights of the F - and C -rays in the plane of the d -ray image. In Example 5.E, $T_{\text{ch}} = -0.01068 - 0.00454 = -0.01522$. A lens, that has *negative* chromatic aberration, such as the one in the example, is said to be “*undercorrected*.”

If, instead of examining the distribution of rays around a focus, one examines an image at a point off the optical axis for a lens uncorrected for chromatic aberration, the edges in the image would not be sharp but would be a blur of colors. Because of dispersion, the magnification of the lens is different for different wavelengths, leading to a succession of centered images of different size for each color (Fig. 5.9). This chromatic aberration, although arising from the same source as axial color, is referred to as *lateral color*. It is defined in terms of the chief ray of the system and therefore varies, as does the chief ray, with the stop position. Depending upon the application, it may be more important to reduce the lateral color more than the axial color. While the correction for one chromatic aberration usually reduces the other, the amount of correction may not be the same for both.

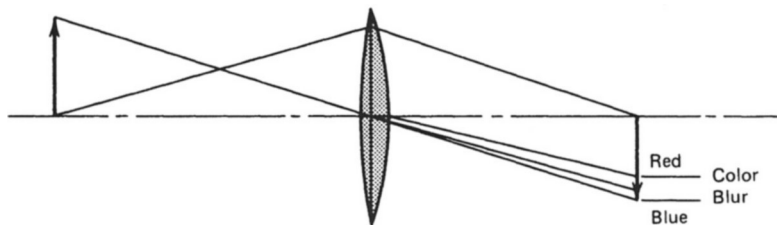


Figure 5.9. Lateral color. Because the magnification of a system can vary with the refractive index of a lens, dispersion of glass produces images whose sizes vary with wavelength. This results in an overall image that is blurred.

5.3.2. Achromatization—Cancellation of Dispersive Power

The lensmaker's equation (Eq. 5.16) can be used to examine the change in the power of a thin lens with refractive index. Given two wavelengths λ_C and λ_F , the powers of a thin lens with curvatures $c_1 = 1/R_1$ and $c_2 = 1/R_2$ are

$$\phi_C = \frac{1}{f_C} = (n_C - 1)(c_1 - c_2) = (n_C - 1)c$$

and

$$\phi_F = \frac{1}{f_F} = (n_F - 1)c, \quad (5.32)$$

where $c = (c_1 - c_2) = (1/R_1 - 1/R_2)$. The difference in powers of this lens at the two wavelengths is

$$\Delta\phi = \phi_F - \phi_C = (n_F - n_C)c = \Delta n \cdot c, \quad (5.33)$$

where $\Delta n = n_F - n_C$. If we define the d -line focal length as the average focal length,

$$\phi_d = \frac{1}{f_d} = (n_d - 1)c, \quad (5.34)$$

then

$$\frac{\Delta\phi}{\phi_d} = \frac{-\Delta f}{f} = \frac{n_F - n_C}{n_d - 1} = \frac{1}{V_d}, \quad (5.35)$$

where V_d is the dispersion given in Eq. 1.5 of Section 1.1. Thus the variation in power can be expressed as

$$\Delta\phi = \frac{\phi}{V_d} \quad (5.36)$$

Now it can be seen why the dispersion is expressed as it is.

By combining two lenses made up of different glasses, it is possible to cancel the $\Delta\phi$'s while maintaining the overall power. That is, given two lenses A and B , we can use the lensmaker's equation to construct two lenses such that

$$\phi_A + \phi_B = \phi_d \quad (5.37)$$

and

$$\Delta\phi_A + \Delta\phi_B = 0. \quad (5.38)$$

Here we are assuming the two lenses are in contact, since Eq. 5.37 is just another expression of the equation derived in Example 1.F. If ϕ_d is again defined as the power of the lens at the d -line and n_d is the refractive index for that line, then, the difference in the curvatures can be expressed, using Eq. 5.34, as

$$c_A = \frac{\phi_A}{n_A - 1}; \quad c_B = \frac{\phi_B}{n_B - 1},$$

where we have used the notation $n_A = (n_d)_A$; $n_B = (n_d)_B$. In Eq. 5.33, we also subscript the difference in powers and refractive indices:

$$\Delta\phi_A = \Delta n_A c_A; \quad \Delta\phi_B = \Delta n_B c_B.$$

Then the condition for the cancellation of powers (Eq. 5.38) becomes

$$\Delta n_A c_A + \Delta n_B c_B = 0$$

or

$$\Delta n_A \frac{\phi_A}{n_A - 1} + \Delta n_B \frac{\phi_B}{n_B - 1} = 0. \quad (5.39)$$

But

$$\frac{\Delta n_A}{n_A - 1} = \frac{(n_F - n_C)_A}{(n_d - 1)_A} = \frac{1}{V_A}$$

and similarly for V_B , so that

$$\frac{\phi_A}{V_A} + \frac{\phi_B}{V_B} = 0 \quad (5.40)$$

or

$$f_A V_A + f_B V_B = 0. \quad (5.41)$$

Since all transparent glasses and materials have a normal dispersion in the visible—that is with $n_F > n_d > n_C$ —the dispersion V_d is always positive. This means that

$$-\frac{f_A}{f_B} = \frac{V_B}{V_A} > 0 \quad (5.42)$$

and f_A and f_B must therefore have different signs. Thus, an achromatic combination must be made of a positive and a negative lens and their focal length–dispersion product (fV) must be equal amounts with opposite signs.

Using this thin lens achromatization method, let us do a preliminary design of an achromat to replace the lens in Examples 5.A and 5.E. Therefore, $f_d = 58.526$; $\phi_d = 0.017087$. Using Eqs. 5.37 and 5.40, we can solve ϕ_A and ϕ_B in terms of ϕ_d , V_A , and V_B .

$$\phi_A = \phi_d \frac{V_A}{V_A - V_B} \quad \text{and} \quad \phi_B = \phi_d \frac{V_B}{V_B - V_A} \quad (5.43)$$

Note that these lenses will require considerable power unless $V_A - V_B$ is made as large as practicable. Usually a value of ΔV of better than 15 is needed. The glass map shown in Chapter 1 is a guide to ranges of glasses available. The ones shown in that map represent a selection of about 20% of the glasses offered commercially (Ref. 5.1). We choose two glasses, a crown (K5, 522595) and a dense flint (F5, 603380) (Ref. 5.2), giving $V_A - V_B = 21.5$. Using the above equations we find that with a crown in the front lens

$$\phi_A = \frac{0.017069 \times 59.5}{21.5} = 0.0472; \quad f_A = 21.170 \text{ mm},$$

$$\phi_B = \frac{0.017069 \times 38.0}{-21.5} = -0.03017; \quad f_B = -33.147 \text{ mm}.$$

The radii of curvature of these lenses may be calculated, if we are willing to make some assumptions. If we use a biconvex lens for the positive lens (a reasonable choice, since it saves on tooling when fabricated), $c_2 = -c_1$, and

$$c = 2c_1 = \frac{\phi_A}{n_{dA} - 1} = \frac{0.0472}{0.522} = 0.09042,$$

$$c_1 = 0.04521,$$

$$R_1 = -R_2 = \frac{1}{c_1} = 22.12 \text{ mm}.$$

If this is made as a cemented doublet (saving one lens mount), then $c_2 = c_3$ and only the last surface curvature needs to be calculated:

$$c_3 - c_4 = \frac{\phi_B}{n_{dB} - 1} = \frac{-0.03017}{0.6034} = -0.05,$$

$$c_4 = 0.05 + (-0.04521) = 0.00479,$$

$$R_4 = 208.77 \text{ mm}.$$

EXAMPLE 5.F. Ray trace the achromat designed using thin lenses. Tentatively assign thicknesses of $t_1 = 5$, $t_2 = 3$.

t	∞	5	3	
				C 49.7882
				d 49.72288
				F 49.6367
n_C		1.51982	1.59874	
n_d	1.0	1.52249	1.60342	1.0
n_F		1.52860	1.61461	
c	0	0.04531	-0.04531	+0.0047
<hr/>				
C			0.922514	0.884507
y	1.0	1.0	0.922252	0.884286
F			0.921658	0.883830
C	0	-0.015497	-0.012669	-0.017765
u	0	-0.015550	-0.012656	-0.017784
F	0	-0.015668	-0.012609	-0.017806

$$C: 1/0.0176715 = 56.2905$$

$$f_{\text{eff}} \quad d: 1/0.0176901 = 56.2303$$

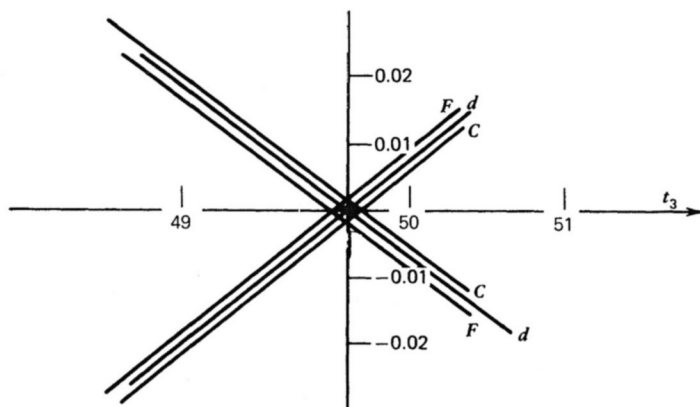
$$F: 1/0.0177107 = 56.1608$$

Plotting on the same scale as before.

$$L_{\text{ch}} = 49.6367 - 49.7882 = -0.1515$$

$$\bar{L}_{\text{ch}} = -0.00269 \text{ (or } \frac{1}{6} \text{ of } \bar{L}_{\text{ch}} \text{ in Example 5.E)}$$

$$T_{\text{ch}} = -0.00153 - 0.00117 = -0.00270$$



Because we inserted some thicknesses, the calculations were thrown off somewhat, and the f_{eff} has been changed by about 4%. This could be remedied by rescaling the system values.

In Example 5.F, the achromat that has just been designed is analyzed using a paraxial ray trace. Thicknesses of 5 mm and 3 mm were tentatively assigned to the two lenses and the ray traces computed for the three wavelengths. The rays are plotted on the same scale as before. The longitudinal axial chromatic aberration is now $L_{ch} = 49.6367 - 49.7882 = -0.1515$ and $\bar{L}_{ch} = -0.00269$, or about one sixth of the aberration in Example 5.E. Here, $T_{ch} = -0.00270$. Because we inserted the thickness, the lens values were thrown off somewhat. The effective focal length is about 4% less than before. This can be remedied by scaling the system values to the target effective focal length. For additional information see Ref. 5.3.

5.4. PARAXIAL RAY TRACING, CALCULATORS AND COMPUTERS

The y - nu method presented here is only one of a number of ways of calculating the Gaussian constants for an optical system. There are methods, such as a matrix method preferred by some optical designers (Refs. 5.4–5.6). Although the matrix method has certain benefits, particularly when there is a periodic array of identical lenses and stops, it suffers from a more complex set up and increased computation time when working problems on paper. The particular method shown here has the advantage of being easy to set up. It enables one to determine the ray heights along an optical system so that stops, pupils, and windows are rapidly found.

Programs for paraxial ray tracing have been written for a large number of calculators and computers. You may want to use one yourself. Rather than add to the slew of programs already in existence, the flow chart for a paraxial ray trace was given as a modification to the thin lens ray trace in Example 2.D. With a small amount of effort, you can program your own paraxial ray-trace program that fits your needs or use one of the “spreadsheet” programs that are available on all microcomputers (see box). Once this is done, check your routines with the calculations in this or other texts. Be sure that you understand the sign and input conventions of other programs, if you choose to ray trace with an unfamiliar program. Again, use simple problems to check your understanding of the program.

PARAXIAL RAY TRACING USING SPREADSHEETS

Since paraxial ray tracing requires nothing more than a series of linear equations, a good deal of computation can be done on a microcomputer using spreadsheet programs, such as Visicalc and 1-2-3*. These are programs that enable one to define arithmetic and logical relations between the contents of one position, or *cell* in a large array and all others

* Visicalc is a trademark of Visicorp; 1-2-3 is a trademark of Lotus Development Corp.

cells in the array. Any subsequent changes in the contents of cells result in recomputations of all relations to give new values. Once the system values block (t_k, n_k, c_k, r_k) has been defined for the surfaces in the system, an axial ray trace block may be built out of the two paraxial ray trace equations. As soon as initial ray values are entered, the values for the subsequent ray heights and slope angles are computed. The axial ray block may be replicated for the chief ray block. Computations of S_k, i_k, \bar{i}_k enable one to calculate the third-order coefficients (see Section 6.2) surface by surface.

If the spreadsheet program has @IF commands, one can close a ray trace to find an image point. One useful procedure, which is easy to perform, is bending a lens. By writing the optical power of second surface as the difference of the optical power of the lens minus the optical power of the first surface, the effective focal length of a system can be maintained without resorting to additional scaling and redefinition of the size of the entrance pupil, as is necessary with simple ray-trace programs.

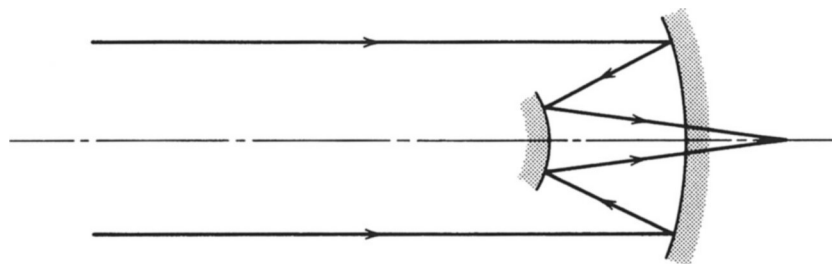
Spreadsheets for exact ray traces are not as useful because the computational time becomes substantial, although this may change with the introduction of more powerful programs of this type. But, for an initial rapid evaluation of a relatively complex system, spreadsheets provide enormous flexibility once you develop the knack of defining relationships between cells.

Another question that should be asked at this point is: Why bother with this *y-nu* method at all? If all of this can be programmed into a calculator, why not just teach someone how to run a ray-trace program? The reason is that an understanding of paraxial ray tracing is required before you can intelligently run a ray-trace program. Both as an introduction to the concepts and as a model for much more sophisticated routines, the paraxial ray trace is ideal. We urge you to work a few problems before continuing.

5.4.1. Ray Tracing with Reflecting Surfaces

In the examples given so far, we have used only refracting surfaces. How is a reflecting surface handled in a paraxial ray trace? Looking at Example 5.G, we see that the propagation direction and thus the optic axis has been reversed. To account for a large number of possible sign changes, all conventions in terms of height and angles can be preserved if only two things are changed (as noted in Table 2.1): After a reflection, the signs of all refractive indices (n_k) are reversed and the signs of all spacings and thicknesses (t_k) are reversed. It can be shown that this will give the mirror equations, Eqs. 1.13 and 1.14. Thus, after one reflection, all distances and surface separations t_k , are negative (see rule # 7 in Table 2.1). An example of a simple paraxial ray trace containing mirrors is given in Example 5.G. Note that the second mirror surface cancels the effect of the first surface so that all the conventions are back to the original state.

EXAMPLE 5.G. Find the focal point for a combination of two mirrors: a concave mirror ($R = 100$ units, $D = 40$ units) reflecting to a convex mirror ($R = 25$ units, $D = 10$ units) at a separation of 40 units. What is the effective focal length and f -number of the system?



Solution: The diagram and array are given below. Please note the number of signs that are present. You must be very careful to keep track of them.

t_k		∞		-40		50
n_k		1.0		-1.0^a		1.0^a
c_k	0		-0.01		-0.04	
r_k	$—$		20		5	
<hr/>						
y_k	20		20		4	0
nu_k		0		-0.4		-0.08
u_k		0		0.4		-0.08
r_k/y_k	$—$		1^b		1.25	

^a Changed sign on reflection.

^b Aperture stop.

Calculations for some of the quantities are:

$$\begin{aligned} nu_1 &= nu_0 - y_1(n_1 - n_0)c_1 \\ &= 0 - 20(-1 - 1)(-0.01) = -20(-2)(-0.01) = -0.4, \end{aligned}$$

$$u_1 = nu_1/n_1 = -0.4/-1 = 0.4,$$

$$\begin{aligned} y_2 &= y_1 + u_1 t_1 \\ &= 20 + (0.4)(-40) = 20 - 16 = 4, \end{aligned}$$

$$\begin{aligned} nu_2 &= nu_1 - y_2(n_2 - n_1)c_2 \\ &= -0.4 - 4(1 - (-1))(-0.04) = -0.4 - 4(2)(-0.04) \\ &= -0.4 + 0.32 = -0.08, \end{aligned}$$

$$u_2 = nu_2 = -0.08.$$

Location of the focal point:

$t_2 = -y_2/u_2 = 4/0.08 = 50$ units from second mirror, 10 units from the first. The effective focal length is

$$f_{\text{eff}} = -y_0/u_2 = -20/-0.08 = 250 \text{ units.}$$

The f -number is $F = 1/2|u_2| = 1/0.16 = 6.25$.

5.4.2. Solves and Pickups

There are certain situations in optical design where the ray trace is carried out to determine a certain lens characteristic and not simply to find the image location by “closing” the trace as we did earlier. For example, the f -number of a system is the reciprocal of twice the final index-marginal ray-angle product (Eq. 5.29). If a system is required to have a specific f -number and the ray height at the final surface has been computed, then instead of closing by setting $y_{j+1} = 0$, one solves for the radius of curvature in the final surface that will give the correct f -number. Since $1/nu_j = 2F$,

$$nu_j = \frac{1}{2F} = nu_{j-1} - y_j(n_j - n_{j-1})c_j$$

and the required final curvature is

$$c_j = -\frac{(1/2F) - nu_{j-1}}{y_j(n_j - n_{j-1})} \quad (5.44)$$

and image distance is

$$t_j = \frac{y_j}{2F} \quad \text{if } n_j = 1. \quad (5.45)$$

Although the example given here applied to ending the ray trace, the same method could be used anywhere in a ray trace if a particular ray angle is required in a specific region of the optical system. This is sometimes called an *angle solve*, although it is a surface curvature that is solved for.

Similarly, if a ray was required to hit an element at a certain height above the axis, as might be the case if there were a limiting stop somewhere in a design, one could vary the space between the elements. In this case, both y_{k+1} , y_k and u_k are known and the translation equation provides a *height solve* for t_k .

$$t_k = \frac{y_{k+1} - y_k}{u_k}. \quad (5.46)$$

Obviously, both of the paraxial equations may be solved for a particular system parameter to preserve a particular ray property. Although this also could include refractive indices, and the range of values is small, the leverage to change other values is not as great as with other solves.

Sometimes in an optical design it is useful to reduce the number of system parameters for structural or economic reasons. For example, equating of the neighboring radii of curvature in our calculation of the achromatic doublet is an example of a *pickup*, an assignment of one system parameter value to another system parameter. Usually curvatures are picked up because this allows the lenses to be cemented together, enhancing the mechanical integrity of the lens and reducing mounting costs.

As was pointed out at the beginning of this chapter, the paraxial ray trace serves as a reference trace against which more exact calculations can be compared. Considering that the image distance, the locations of the cardinal points of the system, as well as its stops, pupils, and windows, its magnification and f -number can be obtained through such a trace, it certainly provides a lot of information with very little effort.

PROBLEMS

- 5.1. Calculate the maximum possible clear radius of the lens described in Example 5.A. Note this is a problem somewhat similar to the determination of segments in the vignetting problems of Chapter 3.
- 5.2. Given a negative lens with radii of curvature $R_1 = -76$ mm and $R_2 = 24$ mm, and a thickness of 4.5 mm, what is the effective focal length of this lens if it is made from glass number 573425? Use y - nu trace rather than the thick lens formula.
- 5.3. Three thin lenses form a triplet. Their focal lengths are 100 mm, -50 mm, and 25 mm, and they are spaced 20 mm apart. What is the effective focal length of the combination? Where are the principal planes of the triplet located?
- 5.4. A Dove prism has a base length of $B = 50$ mm and a base angle of 60° .
 - (a) What is the equivalent air thickness of this prism if its refractive index is 1.621?
 - (b) What is the maximum clear aperture height (A in Fig. 4.17) of the prism?
- 5.5. Using the y - nu trace, derive the power of a thick lens as given in Eq. 5.15.

- 5.6. Show that Eq. 5.21 can be derived from the equation just before it in the text.
- 5.7. You have two large batches of glass on hand, a 541599 and a 689309. Can you make an achromat with a 30-mm focal length from these? If so, what are the respective focal lengths of the individual lenses? What would be the radii of curvature of the two lenses if the contact surface of a cemented doublet is a plane?
- 5.8–5.12. The following problems are thick lens versions of the thin lens ray traces of Chapter 2. In each of the following problems find:
- The stops, magnification, effective focal length, image location and f -number of the system.
 - The pupils and windows of the system.

5.8. *Periscopic Lens*

Object distance: 75 mm

Element #1: $R_1 = \text{flat}$, $t = 2.1 \text{ mm}$, $R_2 = -54 \text{ mm}$, $d = 10 \text{ mm}$,
540597 glass

Separation: 2.3 mm

Element #2: Stop; $d = 9 \text{ mm}$

Separation: 2.3 mm

Element #3: Element #1 (reversed) $R_1 = 54 \text{ mm}$, $t = 2.1 \text{ mm}$,
 $R_2 = \text{flat}$, $d = 10 \text{ mm}$, 540597 glass

5.9. *Double Gauss*

Object distance: infinity

Element #1: $R_1 = 31.5 \text{ mm}$, $t = 4 \text{ mm}$, $R_2 = \text{infinity}$, $d = 10 \text{ mm}$,
689495 glass

Separation: 0.2 mm

Element #2: $R_1 = 19.47 \text{ mm}$, $t = 7.0 \text{ mm}$, $R_2 = 12.47 \text{ mm}$,
 $d = 9 \text{ mm}$, 744448 glass

Separation: 6.0 mm

Element #3: Stop; $d = 7.5 \text{ mm}$

Separation: 6.0 mm

Element #4: Element #2 (reversed) $R_1 = -12.47 \text{ mm}$,
 $t = 7.0 \text{ mm}$, $R_2 = -19.47 \text{ mm}$, 744448 glass

Separation: 0.2 mm

Element #5: Element #1 (reversed) $R_1 = \text{infinity}$; $t = 4 \text{ mm}$;
 $R_2 = -31.5 \text{ mm}$, 689495 glass

For an example of some approaches to design, see Mandler's discussion, "Design of Basic Double Gauss Lenses," (Ref. 5.7).

5.10. *Triplet*

Object distance: 400 mm

Element #1: $R_1 = 100$ mm, $t = 5.7$ mm, $R_2 = -69.18$ mm,
 $d = 8$ mm, 691547 glass

Separation: 13 mm

Element #2: $R_1 = -66.67$ mm, $t = 3.1$ mm, $R_2 = 21.81$ mm,
 $d = 6.5$ mm, 487704 glass

Separation: 13 mm

Element #3: $R_1 = \text{flat}$, $t = 4.2$ mm, $R_2 = \text{unknown}$, $d = 7$ mm,
702410 glass

Solve for R_2 to obtain an effective focal length of 90.46 mm. (Hint: What is the final slope angle needed to obtain this focal length for the particular input ray height you have chosen?)

5.11. *Ramsden Eyepiece*

Object distance: 5 mm

Element #1: $R_1 = 85$ mm, $t = 4.6$ mm, $R_2 = -44.17$ mm,
 $d = 20$ mm, 589613 glass

Separation: 25 mm

Element #2: $R_1 = 44$ mm, $t = 5.0$ mm, $R_2 = -48$ mm,
 $d = 16$ mm, 652585 glass

5.12. *Mangin Mirror*

Object distance: infinity

Element #1: $R_1 = -120$ mm, $t = 6$ mm, $R_2 = -126$ mm,
 $d = 50$ mm, 532516 glass

Separation: 0 (mirror coated on second surface)

Element #2: $R_2 = -120$ mm, mirror surface, $t = -6$ mm,
 $R_3 = -120$ mm, $d = 50$ mm

REFERENCES

- 5.1. B. H. Walker, "The 20% rule: Selecting optical glass", *Optical Spectra* (now *Photonics Spectra*), December 1973, pp. 42–45. While this selection of desirable glass types was one person's attempt to reduce the glass list needed for most optical designs, it has acquired an unofficial status as a standard list. At least one international lens design competition was based on the list as a starting point for determining which glasses could be used in the design. (Note: the list in the article omits LF7 (575415) and SSKN8 (618498), although they appear on the glass map.)
- 5.2. "Optical Glass", published by Schott Optical Glass Inc., Duryea, Pa., 18642 (undated). Glass catalog of 238 glasses. Included is a comprehensive glass map of all the glasses.

- 5.3. R. Kingslake, *Lens Design Fundamentals* (Academic, New York, 1978). Chapter 4 on chromatic aberrations extends the introduction given in the text. This text is the result of 45 years of teaching lens design.
- 5.4. J. W. Blaker, *Geometric Optics, The Matrix Theory* (Dekker, New York, 1971).
- 5.5. W. Brouwer, *Matrix Methods in Optical Instrument Design* (Benjamin, New York, 1964).
- 5.6. A. Nussbaum, *Geometric Optics, An Introduction* (Addison-Wesley, Reading, Mass., 1968). Matrix approach.
- 5.7. W. Mandler, "Design of basic double Gauss lenses," in *Proceedings of SPIE* (1980 International Lens Design Conference), **237**, 222–232 (1980).

6

Exact Ray Traces

(DISCOVERING WHERE THE PROBLEMS ARE)

Thus far small angle approximations have been used to find the Gaussian constants that describe an optical system to first order; paraxial ray traces were used to find the baseline or reference values for a given system. These values can be compared with exact calculations for rays near the edge of the pupils and for object points that are not on the optic axis. The differences between the exact values and the paraxial values are a measure of the performance of the optical system. These differences may be described by the aberrations in the optical system. In this chapter, after a discussion of some exact ray-trace methods, the five monochromatic third-order Seidel aberrations are described, and some methods for reducing these aberrations are given. Because of the power of today's computers and calculators, emphasis is placed on understanding these aberrations rather than deriving the necessary equations. Discussions of the advantages and difficulties of aspheric optics and the concept of the modulation transfer function conclude the chapter.

6.1. RAY TRACING TECHNIQUES

The simplest set of exact ray traces is the meridional ray trace with spherical surfaces, where rays are restricted to a plane containing the optic axis. The meridional ray trace is a special case of the more general skew ray trace, which has no restriction except that the ray pass through the entrance pupil of the system. A ray-trace scheme for skew rays will not be described. The necessary equations for calculating a skew ray are found in the references [6.1–6.2].

The conventions for a meridional ray trace are the same as those for the paraxial ray trace and thin lens ray trace given in Table 2.1. Figure 6.1 shows a ray trace across an interface. Because of the modern sign convention, the slope angles in the figure are negative and the angles of incidence and refraction are positive. While this might lead to a certain amount of confusion in the present derivation, especially in the geometry, the point to bear in mind is that the quantities $-U_k$ and $-U'_k$ are positive.

A number of different sets of equations for a meridional ray trace can be used; the particular method is largely a matter of personal preference. The set of equations to follow are especially useful for programmable calculators and

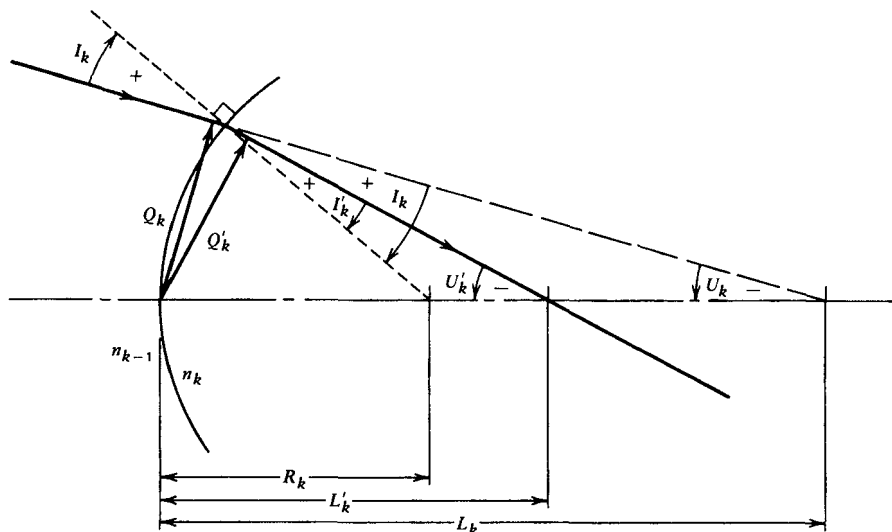


Figure 6.1. Ray trace of a ray across an interface. The angles U_k and U'_k are negative.

computers because all computations involving angles are done through their trigonometric functions, and these are easily called by most calculator and computer programs.

Referring to Fig. 6.1, note that two new variables have been introduced, Q_k and Q'_k , the *ray-vertex perpendiculars*. Although the Y_k 's still exist as vertical coordinates of the ray intersection at the surface, their computational role has been taken by the Q 's, so this method may be called a *Q-U trace*. To distinguish between an exact and a paraxial ray trace, all exact quantities are labeled with capital letters, whereas all paraxial quantities are expressed in small letters, as they have been in earlier chapters. We will again use j as an index denoting the number of surfaces with the object surface index being zero and the image surface index being $j + 1$.

To open the trace, at least two of the following three variables must be set (Fig. 6.2):

1. The sine of the initial slope angle, $\sin U_0$
2. The distance to the first vertex, $L = t_0$
3. The ray-vertex perpendicular *before* the first surface,

$$Q_1 = H \cos U_0 + t_0 \sin U_0, \quad (6.1)$$

where H is the object height. In the case of an axial ray, $H = 0$, and Eq. 6.1 becomes $Q_1 = t_0 \sin U_0$ (Fig. 6.2(a)).

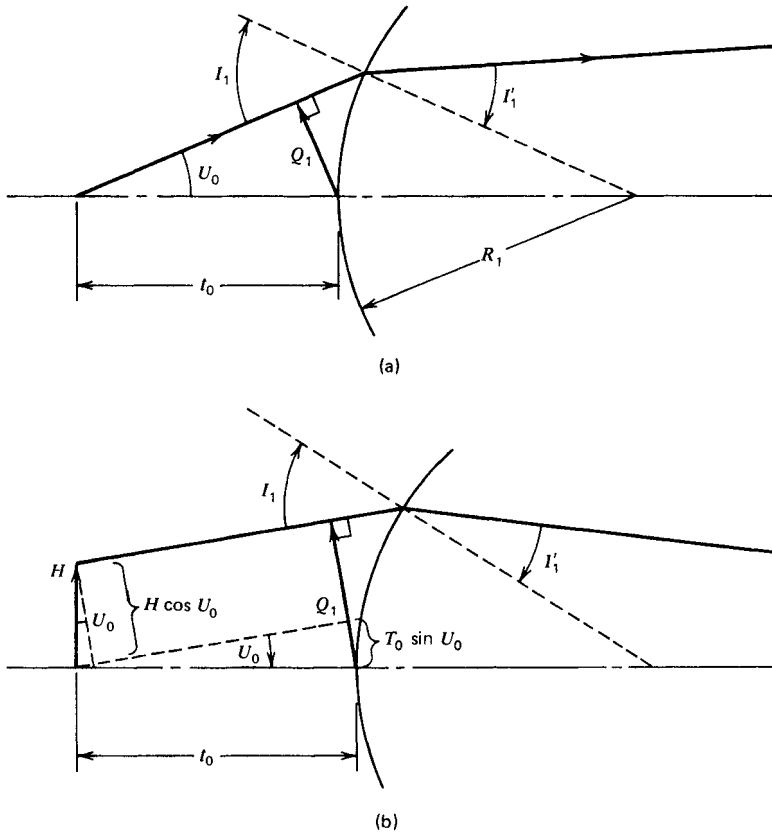


Figure 6.2. Opening a ray trace. (a) An axial ray. (b) A ray from an object of height H with a slope angle U_0 .

In contrast to the paraxial ray trace, where there is a single ray height at a surface, in an exact ray trace there are two ray-vertex perpendiculars, one on each side of the surface. To distinguish between variables before and after refraction, values before refraction will be unprimed, those after refraction will be primed. Thus the paraxial ray height y_k at the k th surface is replaced by Q_k and Q'_k .

For refraction, the angle of incidence at a surface may be calculated by the geometry at the k th surface (Fig. 6.3):

$$\sin I_k = Q_k c_k + \sin U_k, \quad (6.2)$$

where $c_k = 1/R_k$ and $U_k = U'_{k-1}$ [Problem 6.1]. In the case of object space, $U_1 = U'_0 = U_0$. The cosine is then found using the trigonometric identity

$$\cos I_k = \sqrt{1 - \sin^2 I_k}.$$

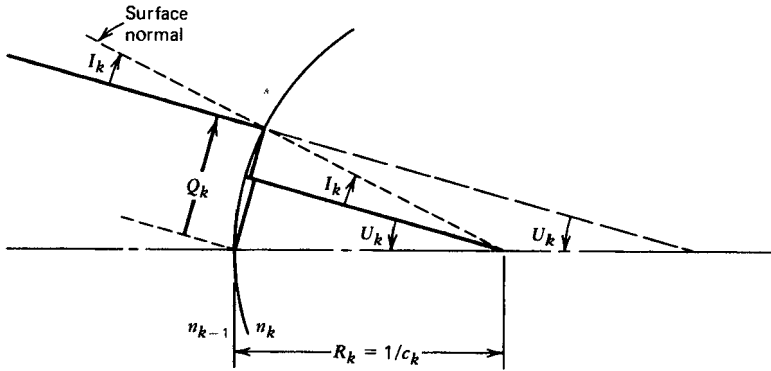


Figure 6.3. Refraction of a ray at an interface. Determination of the angle of incidence in terms of the initial slope angle U_k . The same geometry will hold for the refracted ray also.

By geometry (using Fig. 6.1 and noting that U_k and U'_k are negative) it can be shown [Problem 6.2] that

$$I'_k - U'_k = I_k - U_k, \quad (6.3)$$

where U'_k and I'_k are the refracted slope angle and the angle of refraction after the k th surface. The sine and cosine of the combination $I_k - U_k$ can be calculated using the trigonometric identities

$$\sin (I_k - U_k) = \sin I_k \cos U_k - \cos I_k \sin U_k, \quad (6.4a)$$

$$\cos (I_k - U_k) = \cos I_k \cos U_k + \sin U_k \sin I_k. \quad (6.4b)$$

The sine of the angle of refraction is obtained from Snell's Law and the cosine calculated from the sine:

$$\sin I'_k = \sin I_k \frac{n_{k-1}}{n_k}, \quad (6.5)$$

$$\cos I'_k = \sqrt{1 - \sin^2 I'_k}.$$

Using trigonometric identities again, with $U'_k = I'_k - (I_k - U_k)$, we can determine the new slope angle trigonometric functions from previously computed values:

$$\sin U'_k = \cos (I_k - U_k) \sin I'_k - \sin (I_k - U_k) \cos I'_k, \quad (6.6a)$$

$$\cos U'_k = \cos (I_k - U_k) \cos I'_k + \sin (I_k - U_k) \sin I'_k. \quad (6.6b)$$

The new ray-vertex distance can now be shown to be [Problem 6.3]:

$$Q'_k = Q_k \frac{(\cos U'_k + \cos I'_k)}{(\cos U_k + \cos I_k)}. \quad (6.7)$$

With the calculation of Q'_k and U'_k a new ray direction has been determined. The transfer to the next surface is similar to the paraxial case, i.e.,

$$U_{k+1} = U'_k \quad (6.8)$$

$$Q_{k+1} = Q'_k + t_k \sin U'_k. \quad (6.9)$$

The ray values at the $k + 1$ st interface are computed and the next refraction is determined using Eq. 6.2.

The meridional trace can be closed in two ways: (1) An axial ray, in which case the image distance is found by setting Q_{j+1} equal to zero.

$$L' = T_j = -\frac{Q'_j}{\sin U'_j}; \quad (6.10)$$

EXAMPLE 6.A. Run a meridional ray trace on the system in Example 5.A and compare to the paraxial values.

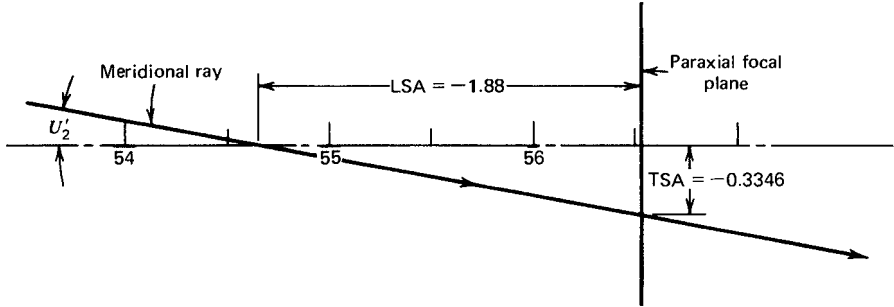
Solution: Run both y - nu and exact ray traces.

System values	t		∞		4	56.53	(Paraxial values)
	n		1		1.517	54.65	(Meridional values)
	c	0		1/40	-1/120	1	
Paraxial values	y_k	10		10		9.6592	
	u_k	0		-0.0852		-0.1709	
	Q_k	10		10		9.7245	0 (6.1, 6.9)
	$\sin U_k$	0		0		-0.0872	-0.1752 Transfer
	$\cos U_k$	1		1		0.9962	0.9845 Trig.
	$\sin I_k$	0		0.2500		-0.1681	(6.2)
	$\cos I_k$	1		0.9682		0.9858	Trig.
	$\sin (I_k - U_k)$	0		0.2500		-0.0816	(6.4a)
	$\cos (I_k - U_k)$	1		0.9682		0.9967	(6.4b)
	$\sin I'_k$		0		0.1648		-0.2549 (6.5)
	$\cos I'_k$		1		0.9863		0.9670 Trig.
	$\sin U'_k$		0		-0.0870		-0.1752 (6.6a)
	$\cos U'_k$		1		0.9962		0.9845 (6.6b)
	Q'_k		10		10.0726		9.5750 (6.7)

Closing,

$$L' = -Q'_2 / \sin U'_2 = -9.5750 / -0.1752 = 54.65$$

$$f_{\text{eff}} = -Q_0 / \tan U'_2 = -10 / -0.1779 = 56.21$$



Comparing to the paraxial ray trace, $t_2 = 56.53$.

$$\text{LSA} = L' - t_2 = 54.65 - 56.53 = -1.88,$$

$$\text{TSA} = -\text{LSA} \times \tan U'_2 = 1.88 \times (-0.1779) = -0.3346.$$

This is an illustration of the approximate nature of the paraxial trace.

(2) An oblique ray, where it is assumed L' has already been found using the counterpart of Eq. 6.1 in image space. The image height is then given by

$$H' = \frac{Q'_j + L' \sin U'_j}{\cos U'_j} \quad (6.11)$$

Just as with the paraxial ray trace, a meridional ray trace can be set up in an array format and each of the ray parameters calculated. An example of this procedure is given in Example 6.A, using the same lens parameters as for the paraxial example, Example 5.A.

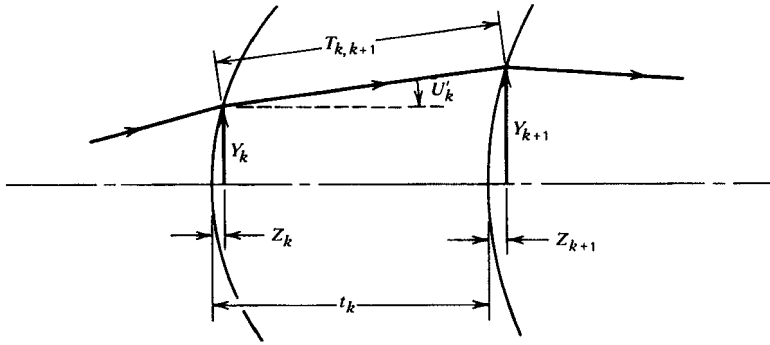


Figure 6.4. Meridional ray-trace quantities, Y_k , Z_k , coordinates of the point of intersection and $T_{k,k+1}$, the distance along a ray between surfaces.

Other quantities of interest that can be obtained from a meridional ray trace are the coordinates (Y_k, Z_k) of the point of intersection of the ray with a surface and the distance $(T_{k,k+1})$ between ray-surface intersection points (Fig. 6.4):

$$Y_k = \frac{Q_k(1 + \cos(I_k - U_k))}{\cos U_k + \cos I_k} = \frac{Q'_k(1 + \cos(U'_k - I'_k))}{\cos U'_k + \cos I'_k}, \quad (6.12a)$$

$$Z_k = \frac{Q_k \sin(I_k - U_k)}{\cos U_k + \cos I_k} = \frac{Q'_k \sin(I'_k - U'_k)}{\cos U'_k + \cos I'_k}, \quad (6.12b)$$

$$T_{k,k+1} = \frac{t_k - Z_k + Z_{k+1}}{\cos U'_k}. \quad (6.12c)$$

Meridional rays, confined to a plane containing the optic axis, constitute a small fraction of the rays emanating from an off-axis object point. Most rays are skew rays, which can only be described by vectors through their direction cosines and their coordinates at surfaces of interest. The equations for tracing a skew ray are not presented here; discussions and derivations are given in Refs. 6.1 and 6.2. Although the computations are quite involved, skew ray traces are needed to assess most designs. For many years it was necessary to trace some skew rays by hand, now it is done as part of most ray-trace programs. In most programs, all exact ray traces are computed as skew rays.

One set of calculations that is quite useful in evaluating a lens or an optical system consists of imposing a rectangular grid on the entrance pupil and then tracing rays from an object point to all the intersections of grid lines in the pupil and plotting the location of these rays on a plane in image space, usually (but not always) at the paraxial image plane (Fig. 6.5). If the system were perfect, all of the rays would intersect at one point, the image point. But, because of

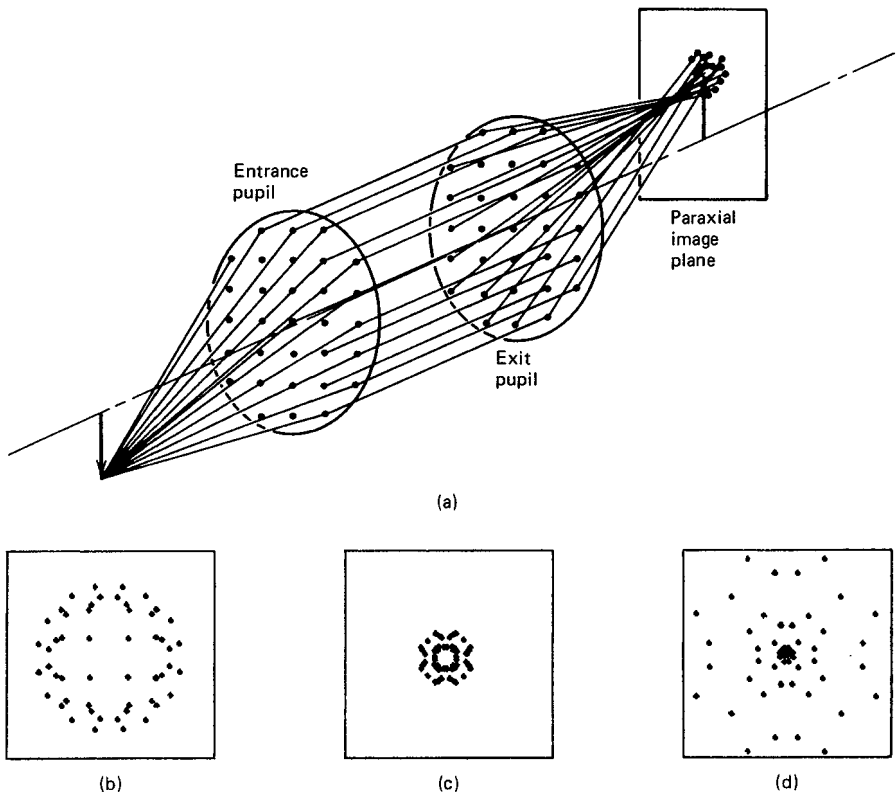


Figure 6.5. Spot diagram. (a) Generation of a spot diagram. Three spot diagrams from Example 6.A: (b) between the lens and the best focus, (c) at best focus, and (d) paraxial image plane. This shows graphically the spherical aberration in this lens.

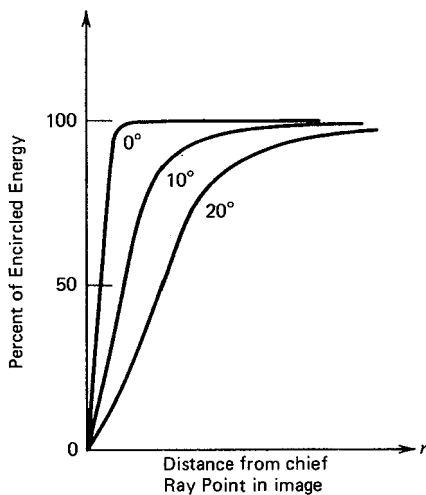


Figure 6.6. Encircled energy plots. The steepness of the rise in the curve indicates the quality of performance for different chief ray angles to the optic axis.

aberrations, the rays are spread out in a region in the vicinity of the image point. This generates a series of *spot diagrams* (Fig. 6.5(b)–(d)) that provide insight into the performance of the optical system. If one assumes that each ray in the spot diagram carries the same fraction of the total energy, then a plot of the amount of energy found within some small circle as a function of the increasing circle radius about the image point provides additional information on the system performance. With *plots of encircled energy* for different distances off axis as shown in Fig. 6.6, a designer gains some idea of the off-axis performance of a lens. In the figure, 0° represents an axial object, while the 10° and 20° plots represent ray bundles whose chief ray is at the specified angle to the optic axis. The more steeply the plot rises to 100% the closer the performance is to that of the ideal lens.

6.2. ABERRATIONS

Once a series of exact ray traces are made, one can discover where the problems are in a design. It is at this point that the knowledge of optics must be combined with practical experience to progress further. Faced with the evidence of aberrations, a fledgling designer has two questions: (1) How well does my design perform the assigned task? and (2) Once I know how bad the aberrations are, what can I do about it? The complete answers are not in this text. That would be akin to publishing a map of an entire country that would enable a traveler to find any particular house. What can be done is to provide some general methods for evaluating and reducing aberrations. More detailed texts, colleagues, and other designers can provide additional help. What follows is akin to a map of the country's major highway system; others will have to provide local maps.

The differences between the paraxial image height y_{j+1} and the skew ray at the paraxial image plane is a measure of the aberrations of a system. In Fig. 6.7

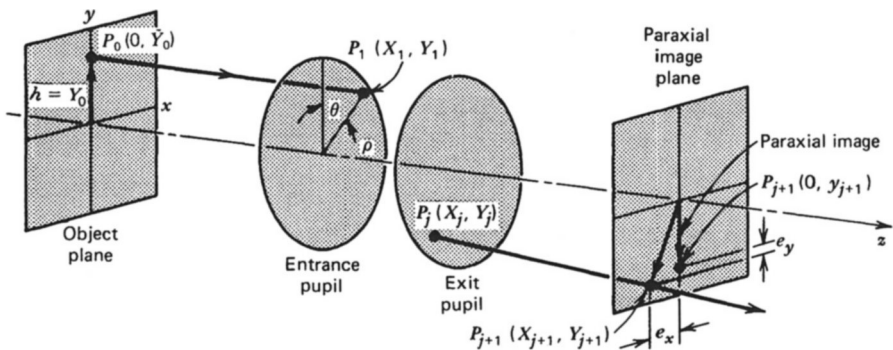


Figure 6.7. Diagram of skew ray trace quantities.

a skew ray is launched from an object point $P_0(0, \bar{Y}_0)$ toward a point in the entrance pupil $P_1(X_1, Y_1)$. After traversing the optical system the ray pierces the paraxial image plane at $P_{j+1}(X_{j+1}, Y_{j+1})$. The differences between the y -coordinates of the exact and paraxial traces, $Y_{j+1} - y_{j+1}$, and the nonzero X_{j+1} coordinate ($x_{j+1} = 0$, by definition for the paraxial ray), are a measure of the aberrations of the system and can be expressed as a mixed third-order polynomial in the variables Y_1, X_1 , and \bar{Y}_0 :

$$Y_{j+1} - y_{j+1} = -\frac{1}{2n_j u_j} \left[\sum_{k=1}^j B_k (Y_1^2 + X_1^2) \frac{Y_1}{y_1^3} + \sum_{k=1}^j F_k \frac{(3Y_1^2 + X_1^2)}{y_1^2} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right) + \sum_{k=1}^j (3C_k + P_k I^2) \frac{Y_1}{y_1} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^2 + \sum_{k=1}^j E_k \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^3 \right] + O(5) \quad (6.13a)$$

and

$$X_{j+1} = -\frac{1}{2n_j u_j} \left[\sum_{k=1}^j B_k (Y_1^2 + X_1^2) \frac{X_1}{y_1^3} + \sum_{k=1}^j F_k \frac{(2Y_1 X_1)}{y_1^2} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right) + \sum_{k=1}^j (C_k + P_k I^2) \frac{X_1}{y_1} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^2 \right] + O(5), \quad (6.13b)$$

where $O(5)$ represents fifth-order and higher terms, $I = \bar{y}_k n_k u_k - y_k n_k \bar{u}_k$, the optical invariant (Eq. 5.21), and the barred quantities are paraxial chief ray values. Each term in the polynomial can be identified with a particular type of aberration, called a *Seidel aberration*, and the coefficients of these terms represent the contribution of that aberration to the overall performance of the system. One of the most attractive features of this polynomial expansion is that each of the five third-order Seidel aberrations—spherical aberration (whose coefficient is B_k in the above expansion), coma (F_k), astigmatism (C_k), field curvature (P_k), and distortion (E_k)—may be computed using values derived from two paraxial ray traces, an axial trace, and a chief ray trace. Another powerful feature of this expansion is that the effect of each of the surfaces can be evaluated separately. This enables the designer to determine which elements are contributing the greatest amount of aberration and to correct a design by balancing out the contributions for the various surfaces.

6.2.1. Spherical Aberration

In Example 6.A, an exact trace of a ray coming from an object at infinity was made and compared with the paraxial trace for the same lens. It was found in the meridional trace that the ray crossed the optic axis 54.65 mm from the second surface. In Example 5.A, the ray crossed the axis 56.53 mm from the second surface. Thus, a marginal parallel ray crosses the axis closer to this positive lens than its paraxial counterpart. The failure of the lens to direct axial

rays from an object point to the corresponding point in image space is called *spherical aberration* (Fig. 6.8). It is the first of the five lowest order, or Seidel, aberrations, and usually one must correct this aberration before attempting to reduce the others. From an examination of the ray trace we see that there are two possible measures of this aberration: (1) the distance between the points where the paraxial and exact rays cross the axis (*longitudinal spherical aberration*, or LSA),

$$\text{LSA} = T_j - t_j \quad (6.14)$$

and (2) the distance between the points where the rays cross the paraxial focal plane (*transverse spherical aberration*, or TSA),

$$\begin{aligned} \text{TSA} &= (T_j - t_j) \tan U'_j \\ &= -\text{LSA} \tan U'_j \end{aligned} \quad (6.15)$$

All of the calculations given above serve to quantify the degree of aberration at the extreme points of the field, but they do not provide a full understanding of what is happening. This can be remedied by using a series of graphs of ray data plotted in different ways.

An *axial intercept plot* is constructed by plotting the longitudinal spherical aberration as a function of input ray height. In Example 6.A, it was shown that the LSA for an input ray height of 10 mm was -1.88 mm, but there is no information for rays at intermediate heights. If a set of meridional traces is run for values of, say, 20%, 40%, 60%, and 80% of the maximum ray height, a set of LSA values may be plotted against y_0 , as shown in Fig. 6.9. The forward bending of this curve indicates *undercorrected spherical aberration*. If the marginal ray crossed the axis *beyond* the paraxial ray, the spherical aberration is said to be *overcorrected*, and the axial intercept plot would bend backward.

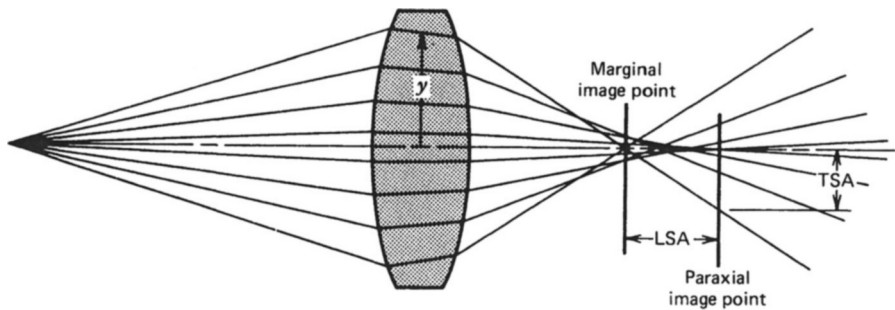


Figure 6.8. Spherical aberration. The distance between the points where the paraxial and marginal rays cross is the longitudinal spherical aberration (LSA). The distance between the optic axis and the marginal ray in the paraxial image plane is the transverse spherical aberration (TSA).

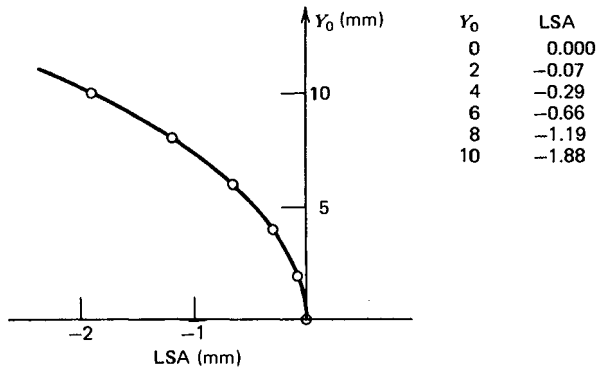


Figure 6.9. Plot of longitudinal spherical aberration (LSA) versus input ray height for Example 6.A.

A more general method of plotting aberrations is the *ray intercept plot*. Here the ray height at a reference plane (usually the paraxial image plane) is plotted against the negative of the tangent of the ray slope angle U'_k . Thus if a lens has no spherical aberration, a fan of axial rays would cross the axis at the same point and $Y_{k+1} = 0$ for all rays, as shown in Fig. 6.10(a). The plot is a straight line along the x axis (Fig. 6.10(b)). If the reference plane were not at the paraxial image plane but shifted by an amount Z' , the plot would still be a straight line since the ray height at the reference plane would be equal to $-Z' \tan U'_k$ and the slope of the line would be $-Y'/\tan U'_k = Z'$, a constant. Therefore, except for a change of slope, the choice of reference plane does not affect the ray intercept plot. From the data listed in Fig. 6.9, the amount of TSA can be calculated using Eq. 6.14 and plotted against $-\tan U'_k$ for these rays. This gives the characteristic ray intercept plot for undercorrected spherical aberration (Fig. 6.10(c)). The deviation from a straight line indicates the presence and amount of spherical aberration. (What would a plot of overcorrected spherical aberration look like? [Problem 6.5].)

A plot of the transverse spherical aberration versus the ray height at the entrance pupil Y_1 would produce a plot similar to the ray intercept plot. Both curves can be approximated by odd-order polynomials. For an image of a point on the optic axis, the curve can be fitted to a power series

$$Y_k - y_k = b_1 Y_1 + b_3 Y_1^3 + b_5 Y_1^5 + O(7),$$

where b_3, b_5 are spherical aberration coefficients. Since the curves are plotted in the paraxial focal plane, b_1 , the slope of the curve near $Y_1 = 0$, must be zero as discussed earlier. The third-order contribution b_3 can be calculated using one paraxial ray trace. Evaluating Eq. 6.13, $\bar{X}_0 = 0$ and $\bar{Y}_0 = 0$ (object point on the axis), we see that $Y_{j+1} = -(1/2n_j u_j) \sum_{k=1}^j B_k Y_1^3 / y_1^3$. If the ray height is the same at the entrance pupil for the exact ray trace as for the paraxial trace ($Y_1 = y_1$),

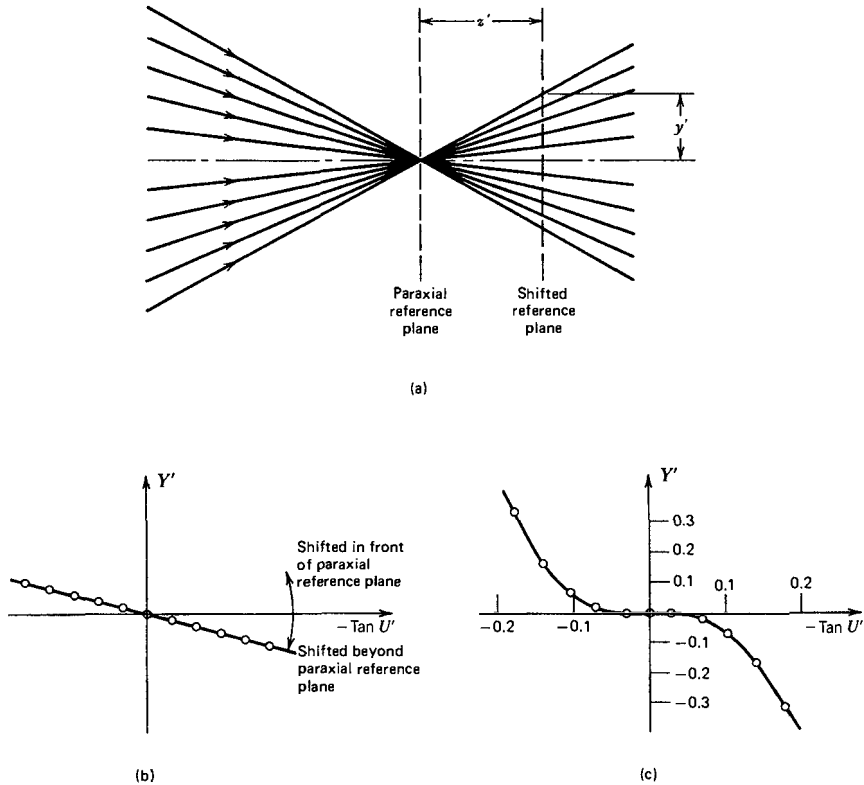


Figure 6.10. Ray intercept plot. (a) Fan of rays exhibiting no spherical aberration. (b) Ray intercept plot for a reference plane shifted beyond the paraxial reference plane. (c) Ray intercept plot of the lens in Example 6.A.

then the transverse spherical aberration is just the summation of all surface contributions multiplied by $(-1/2n_j u_j)$. If the object is at infinity this multiplier is the f -number of the system. In some texts the sum $\sum B_k$ is represented as B [Ref. 6.1] or SC [Ref. 6.2] or LA [Ref. 6.3], the Seidel transverse spherical aberration coefficient. In this text B_k refers to the coefficient, while TSA and LSA refer to the *values* of the aberration as defined in Eq. 6.14 and 6.15. The expression for B_k for the k th surface can be determined from paraxial ray values from:

$$B_k = S_k i_k^2 \quad (6.16)$$

and

$$S_k = y_k n_{k-1} \left(\frac{n_{k-1}}{n_k} - 1 \right) (u_k + i_k), \quad (6.17)$$

where i_k is the paraxial angle of incidence before the k th surface,

$$i_k = y_k c_k + u_{k-1}, \tag{6.18}$$

and u_{k-1} is the paraxial slope angle before the k th surface. In Example 6.B, the transverse aberration coefficient is calculated for the lens in Examples 5.A and 6.A.

EXAMPLE 6.B. Calculate the Seidel third-order spherical aberration coefficient for the lens in Example 6.A and compare the result to the spherical aberration computed from the exact and paraxial ray traces.

Solution: An array format similar to the paraxial ray trace format is used to order the quantities to be used in the calculation. Some of these are taken from Example 6.A.

System	t_k	∞	4		56.53
	n_k	1	1.517		1.0
	c_k	0	$\frac{1}{40}$	$-\frac{1}{120}$	
Axial ray	y_k	10	10	9.6592	0
	nu_k	0	-0.1292	-0.1709	
	u_k	0	-0.0852	-0.1709	
	i_k		0.2500	-0.1657	
Seidel coefficients	S_k		-0.5616	-2.5496	
	B_k		-0.0351	-0.07000	

$$B = SC = B_1 + B_2 = -0.1051$$

$$Y_3 = \text{TSA} = -\frac{B}{2n_2u_2} = -\frac{1}{2(-0.1709)}(-0.1051)$$

$$= -0.3075 \text{ mm}$$

$$\text{LSA} = -\text{TSA}/u_2 = 0.3075/-0.1709 = -1.80$$

This can be compared to a value of -0.3346 mm for the transverse spherical aberration calculated in Example 6.A. Higher orders of the power series contribute an additional 10% to the actual transverse spherical aberration.

6.2.2. Coma and the Offense Against Sine Condition

Assuming the system has been corrected for spherical aberration, the imaging of off-axis points may still show aberration. A bundle of rays about the chief ray will focus in the paraxial image plane. But marginal rays in the meridional plane, as shown in Fig. 6.11(a), do not focus on the chief ray but come to a focus closer to the optic axis. The reason is that the local optical powers at the top and bottom of the lens are different for off-axis rays. However, these are only three points on the lens. The full aberration pattern is more complicated. It is useful to label certain rays in the system in order to follow their progress through the system and to make the analysis of the system more precise. *Meridional rays*, because they are confined to a plane tangent to the optic axis, are also referred to as *tangential rays* and the plane of confinement as the *tangential plane*. Skew rays propagating in a plane and containing the chief ray at right angles to the tangential plane are called *sagittal rays*. The corresponding plane is called the *sagittal plane*.

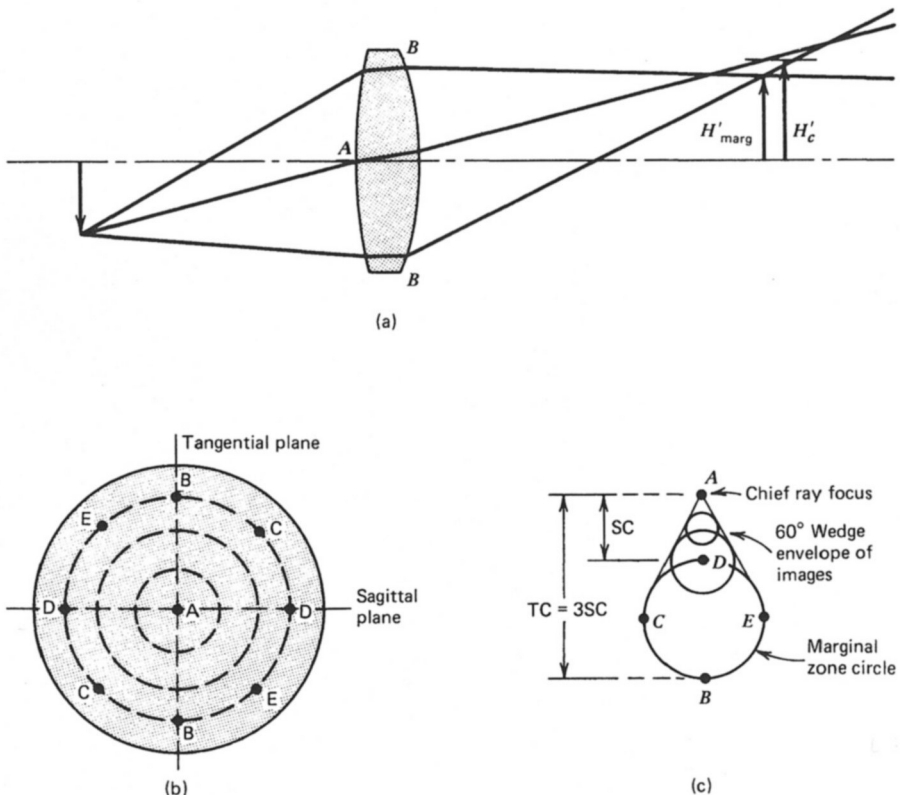


Figure 6.11. Comatic aberration or coma. (a) Marginal and chief rays illustrating of coma. (b) Entrance pupil pattern used to map the comatic image in (c).

If the entrance pupil is divided into a series of concentric zones centered about the chief ray (Fig. 6.11(b)), then rays from the object point through that circular zone do not focus to a point. Instead they form a circular image in the paraxial image plane (Fig. 6.11(c)). The contributions from the marginal zone produce the largest circle; intermediate zones produce smaller circles. The labels on the marginal zone in Figure 6.11(b) provide a means of seeing how the circular image is produced. The point (*D*) on the circle closest to the focus point (*A*) of the chief ray is due to the sagittal rays; the point (*B*) due to the tangential rays is three times farther away. The other rays in the zone (e.g., *C* and *E*) fill in the rest of the circle. The radius of the circle is equal to *AD*, distance from the chief ray focus to the sagittal focus. Together the superposition of these circles of increasing size and their displacement from the chief ray focus point produce a cometlike image, or *coma*, as indicated in Fig. 6.11(c).

A measure of tangential coma is the distance between the intersection of the two marginal tangential rays and the chief ray in the same plane:

$$\text{COM}_t = H' - H'_c, \quad (6.19)$$

where *H'* is the distance of the marginal image point from the optic axis and *H'_c* is the chief ray height in the same plane. In Eq. 6.13 the comatic contribution is denoted by the term with coefficients *F_k*. From the variables multiplying the coefficient, it can be seen that coma varies as the square of the entrance pupil diameter, but only linearly with the object height. A relation similar to that for the optical invariant for paraxial rays can be derived from sagittal rays. As in the case of the optical invariant, each side of the relation consists of parameters in one region, and it may be equated to the parameters in the other regions in the system. If in object space $\bar{Y}_0 = H$, the object height, and in image space (after *j* surfaces) $\bar{Y}_{j+1} = H'$, the sagittal image height, then

$$H n_0 \sin U_0 = H' n_j \sin U'_j \quad (6.20)$$

and the magnification for the skew ray image is

$$M = \frac{H'}{H} = \frac{n_0 \sin U_0}{n_j \sin U'_j}. \quad (6.21)$$

In the paraxial limit, $\sin U_0 = u_0$ and $\sin U'_j = u_j$ and we get the paraxial magnification equation

$$m = \frac{h'}{h} = \frac{n_0 u_0}{n_j u_j}. \quad (5.22)$$

As noted earlier coma can be considered a variation in the power of the lens (and therefore the magnification) from one zone of the lens to another. If a

system has been corrected for spherical aberration, then for a system to be free of coma also the paraxial magnification m and the meridional magnification M should be equal:

$$M = \frac{n_0 \sin U_0}{n_j \sin U_j} = \frac{n_0 u_0}{n_j u_j} = m. \quad (6.22)$$

The relationship for coma-free imaging is called the *Abbe sine condition* after the investigator who first derived it:

$$\frac{\sin U_0}{\sin U_j'} = \frac{u_0}{u_j}. \quad (6.23)$$

A measure of coma in a system free of spherical aberration is a deviation from this equality or “*offense against sine condition*” (OSC). The amount of sagittal coma is expressed in a manner corresponding to Eq. 6.19 for tangential coma,

$$\text{COM}_s = H_s' - H_c' \quad (6.24)$$

and the ratio of this to the chief ray height H_c' is defined as the OSC:

$$\text{OSC} = \frac{H_s' - H_c'}{H_c'} = \frac{H_s'}{H_c'} - 1. \quad (6.25)$$

If there is no longitudinal spherical aberration, the marginal ray intersection plane will lie in the paraxial plane and $H_c' = h'$, the paraxial ray height. If the first term is multiplied by unity in the form of h/h , then

$$\text{OSC} = \frac{H_s'}{h} \cdot \frac{h}{h'} - 1 = \frac{M}{m} - 1. \quad (6.26)$$

Since any nonzero value of OSC indicates a violation of the Abbe sine condition, this measure of sagittal coma can be justly termed an offense against the sine condition. This parameter is used for evaluation even when spherical aberration is present and a contribution from spherical aberration is included in the OSC.

The third-order comatic contribution in Eq. 6.13 is given by

$$Y_{j+1} - y_{j+1} = -\frac{1}{2n_j u_j} \sum_{k=1}^J F_k \frac{(3Y_1^2 + X_1^2) \bar{Y}_0}{y_1^2 \bar{y}_0}$$

and

$$X_{j+1} = -\frac{1}{2n_j u_j} \sum_{k=1}^J F_k \frac{(2Y_1 X_1) \bar{Y}_0}{y_1^2 \bar{y}_0}.$$

The factor of three ratio between the tangential and sagittal aberrations is evident from these equations. The expression for the Seidel coefficient for coma is similar to that for spherical aberration:

$$F_k = S_k i_k \bar{i}_k \quad (6.27)$$

$$S_k = y_k n_{k-1} \left(\frac{n_{k-1}}{n_k} - 1 \right) (u_k + i_k) \quad (6.17)$$

where S_k is the same as for the spherical aberration case and $i_k = y_k c_k + u_{k-1}$ is the paraxial angle of incidence before the k th surface and $\bar{i}_k = \bar{y}_k c_k + \bar{u}_{k-1}$ is the chief ray angle of incidence before the k th surface. A calculation of the Seidel coefficient for coma and a comparison to exact ray trace data is given in Example 6.C.

6.2.3. Astigmatism

Pure coma is hard to display since it is usually combined with another off-axis aberration, astigmatism. Consider the diagram in Fig. 6.12. Suppose we examine a “fan” of rays, rays emanating from an object point that are all confined to the same plane. A fan of rays containing only meridional rays is called a *tangential fan* since the ray plane is the tangential plane. Another set of rays of importance are another fan of rays in the sagittal plane called the *sagittal fan*. After passing through the lens, the fans come to a point focus at different points along the chief ray. In most positive lenses the sagittal image falls beyond the tangential focus, but in front of the paraxial plane. This results in two “line” images, the two lines at right angles to each other with a circular region between them that represents a compromise image, a circular blur, called the *circle of least confusion*. This aberration is called *astigmatism* and is the result of the

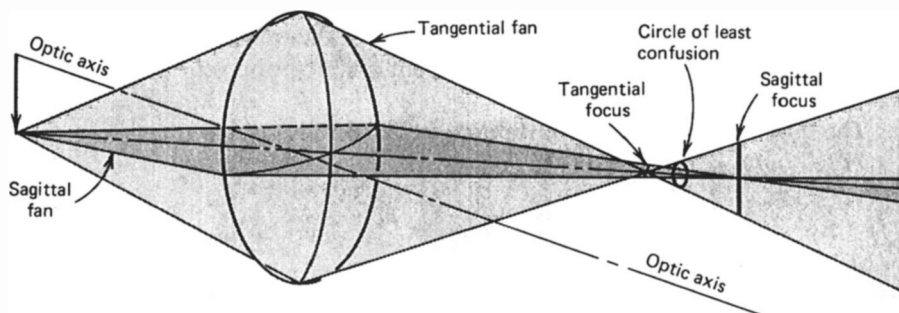


Figure 6.12. Astigmatism.

variation in the power of the lens for two off-axis fans of rays. In Eq. 6.13 the astigmatic contribution is denoted by the term with coefficients C_k . From the variables multiplying the coefficient, it can be seen that astigmatism varies as the square of the object height \bar{Y}_0 and linearly with the entrance pupil radius Y_1 .

Although it is possible to determine the amount of astigmatism by tracing meridional and skew rays, a simpler method of calculating the astigmatic image points is to trace a chief ray and then trace two additional rays, one meridional

EXAMPLE 6.C. Calculate the Seidel third-order tangential coma for the lens in Example 6.A and compare to the results from an exact ray trace. Assume the chief ray trace makes an angle of 10° with the optic axis and that the entrance pupil is at the first surface.

Solution: Use an array slightly expanded for Example 6.A.

System	t_k	∞	4	56.53	
	n_k	1	1.517	1.0	
	c_k	0	$\frac{1}{40}$	$-\frac{1}{120}$	
Axial ray	y_k	10	10	9.6542	0.0
	nu_k	0	-0.1292	-0.1709	
	u_k	0	-0.0852	-0.1709	
	i_k		0.2500 ^a	-0.1657	
	Chief ray	\bar{y}_k		0	0.4602
\bar{nu}_k		0.1745	0.1745	0.1726	
\bar{u}_k		0.1745	0.1151	0.1726	
\bar{i}_k			0.1745	0.1127	
Seidel coefficients		S_k		-0.5616	-2.5496
	F_k		-0.0245	0.0470	$\sum F_k = 0.0231$

^a Aperture stop.

For tangential marginal rays, $Y_1 = y_1$; $X_1 = 0$

$$\bar{Y}_3 - \bar{y}_3 = \frac{3 \sum F_k}{-2n_2u_2} = \frac{3(0.0231)}{-2(-0.1709)} = 0.2028 \text{ mm}$$

An exact ray trace gives a tangential focus 9.677 mm above the optic axis. The principal ray in that plane is 9.450 mm. Thus the coma by a direct ray trace is 0.227.

and one skew, very close to the chief ray. It is essentially a paraxial ray trace with a chief ray as the axis. Using the angles of incidence and refraction I_c and I'_c for the chief ray, one can show that the ray trace equation is

$$\frac{n' \cos^2 I'_c}{t'} - \frac{n \cos^2 I_c}{t} = \frac{n' \cos I'_c - n \cos I_c}{R} \quad (6.28)$$

for the tangential ray, where t' and t are the distances from the refracting surface to the points where the refracted and incident rays, respectively, cross the chief ray. (The equation is a more general version of the simple paraxial relation Eq. 6.38 used in the next section.) For the sagittal ray,

$$\frac{n'}{s'} - \frac{n}{s} = \frac{n' \cos I'_c - n \cos I_c}{R}. \quad (6.29)$$

The transfer equations are

$$t = t' - D \quad \text{and} \quad s = s' - D, \quad (6.30)$$

where D is the distance along the chief ray between the surface just evaluated and the next one as given in Eq. 6.12(c). Together these equations are referred to as a *Coddington*, or *para-principal*, *trace*. Because an exact meridional trace of a chief ray would be needed before a Coddington trace could be performed, applications of these equations are left for use in a computer program.

In some optical systems, including the eye, even on-axis rays suffer astigmatism because the lens surface is not spherical and, therefore, has different powers in different directions. It is as if a cylindrical lens were placed in contact with a spherical lens. In the case of visual systems, an ophthalmologist corrects for astigmatism of the eye by prescribing glasses that compensate for the additional cylindrical power in one place by reducing the power of the lens in that direction. At other times astigmatism can be useful and is intentionally designed into systems to produce an indication of a detector position relative to a focus, as we will see in some of our examples of laser beam focus designs. Those systems that incorporate cylindrical power are termed *anamorphic* systems.

Using the Seidel coefficients, one can determine the amount of astigmatism in a system:

$$Y_{j+1} - y_{j+1} = -\frac{1}{2n_j u_j} \sum_{k=1}^j 3C_k \frac{Y_1}{y_1} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^2,$$

$$X_{j+1} = -\frac{1}{2n_j u_j} \sum_{k=1}^j C_k \frac{X_1}{y_1} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^2.$$

The actual distances plotted in most curves are the longitudinal contributions, which are derived from the transverse values given above by dividing by the final slope angle of the axial ray. Thus the longitudinal astigmatisms are:

$$Z_s = -\frac{X_{j+1}}{u_j}. \quad (6.31)$$

$$Z_t = 3Z_s \quad (6.32)$$

The distance between the sagittal and tangential foci is

$$\begin{aligned} \Delta Z &= Z_t - Z_s = 2Z_s = -\frac{2X_{j+1}}{u_j} \\ &= \sum_{k=1}^j \frac{C_k X_1}{n_j u_j^2 y_1} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^2. \end{aligned} \quad (6.33)$$

If the coefficients for spherical aberration and coma have already been computed, the astigmatism contributions are easily found. The expression for the Seidel coefficient for astigmatism is similar to that for spherical aberration and coma:

$$C_k = S_k \bar{i}_k^2, \quad (6.34)$$

where S_k and \bar{i}_k are defined as before. It can be shown that $C_k = F_k^2/B_k$ for each surface [Problem 6.6]. Thus the astigmatic contributions for the lens in Example 6.A are $(0.0245)^2/(-0.0351) = -0.0171$ for surface 1 and $(-0.0476)^2/(-0.0700) = -0.0324$ for surface 2. The total contribution is $C = C_1 + C_2 = -0.0495$.

Just as with spherical aberrations, one can plot a ray intercept plot for off-axis aberrations. However, instead of plotting the object height versus the ray slope angle (since all of the points are clustered about an off-axis point), the values of the ray intercepts, relative to the chief ray, are plotted against the difference between the tangent of the ray slope angle and that of the chief ray. Since for coma the points for both the marginal rays above and below the chief ray are closer to the axis than the chief ray, the comatic ray intercept curve is an umbrella-shaped curve. In more complex systems, coma may not be as simple as this and the corresponding ray intercept curve can be more complex.

The ray intercept curve can only be plotted for meridional rays, since only in the meridional plane can one be assured that the rays will cross the chief ray. To describe astigmatism graphically, the z -coordinate of the tangential and sagittal image points are plotted for a range of object heights. The result is a pair of paraboloidal surfaces centered on the optic axis. A measure of astigmatism is then the distance between the tangential and sagittal image points for a particular chief ray. A plot of these surfaces for an uncorrected lens is shown in Fig. 6.13.

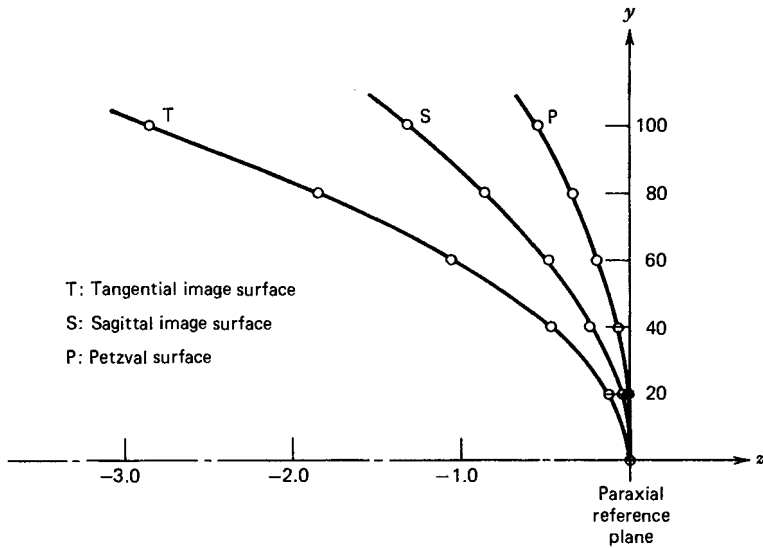


Figure 6.13. Plot of the image surfaces for a lens with astigmatism.

6.2.4. Field Curvature

A lens corrected for spherical aberration and coma is called an *aplanat*. If it is also corrected for astigmatism it is termed an *anastigmat*. Correction against these three aberrations does not assure that the system is aberration free. In an anastigmat the tangential and sagittal surfaces collapse to a single image surface called the Petzval surface, which unfortunately does not resemble the shape of the object surface (usually flat), but rather is a forward curving surface, as shown in Fig. 6.13, in most instances.

Field curvature, as this aberration is called, can be seen even in a thin lens. In Fig. 6.14 is shown a thin lens with an object at t_0 ($t_0 \neq \infty$, here). As was noted in Chapter 1

$$\frac{1}{t_1} = \frac{1}{f} + \frac{1}{t_0} \quad (1.16)$$

At a point on the object whose height is h from the axis, a chief ray may be traced. Because the distance t'_0 to the lens center is greater than t_0 , the off-axis image will be found at distance

$$\frac{1}{t'_1} = \frac{1}{f} + \frac{1}{t'_0} = \frac{1}{f} + \frac{\cos \theta}{t_0} \quad (6.35)$$

For a positive lens, a real image will be found on a surface curving toward the lens. It can be shown [Problem 6.7], that in the case of small angles the axial

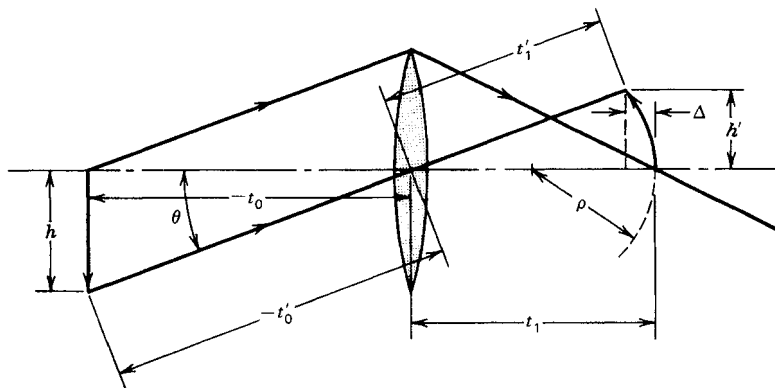


Figure 6.14. Illustration of field curvature with a thin lens.

shift of the off-axis image point from the on-axis point, called the “sagitta” or “sag,” is

$$\Delta = t_1 - t'_1 \cos \theta \doteq \frac{t_1^2 \theta^2}{2f}. \quad (6.36)$$

The curve traced out by the points on the object approximates a circle, the radius of which can be determined from the sag of the circle, the distance between the arc and a half-chord of length h' [Problem 6.7]:

$$\rho \doteq \frac{h'^2}{2\Delta} = \frac{h'^2}{2} \cdot \frac{2f}{t_1^2 \theta^2} \quad (6.37)$$

or since $h' = t_1 \theta$ for small angles

$$\rho \doteq f.$$

This radius of curvature of the image plane is called the *Petzval radius*, and the aberration is called *Petzval curvature* or *field curvature*.

In the paraxial treatment it is possible to derive an even more useful formula, which expresses the contributions to field curvature in terms of the optical power of each surface. This can be derived with the help of Fig. 6.15 from the refraction equation for a paraxial ray trace. We start with an object surface that is spherical, with radius of curvature r_0 that has the same center of curvature as the refracting surface:

$$nu_1 = nu_0 + y_1 \frac{(n_1 - n_0)}{R_1}.$$

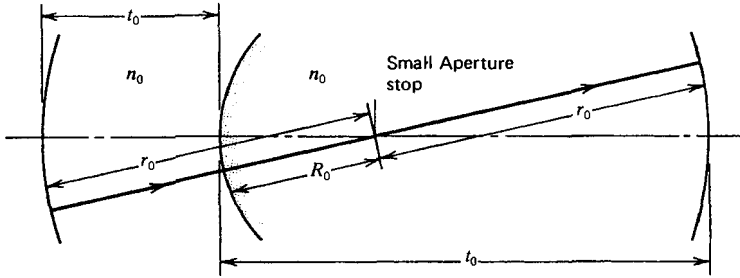


Figure 6.15. Derivation of field curvature by paraxial ray trace.

Replacing u_1 with y_1/t_1 and u_0 with y_1/t_0 , and dividing out of the common factor y_1 , we get

$$\frac{n_1}{t_1} - \frac{n_0}{t_0} = \frac{n_1 - n_0}{R_1}. \quad (6.38)$$

One can see that the image surface will also be a sphere of radius r_1 centered at the same point as r_0 and R_1 . Substituting for $t_1 = R_1 - r_1$ and $t_0 = R_1 - r_0$ and multiplying through all denominators and cancelling terms, we obtain

$$\frac{1}{n_0 r_0} - \frac{1}{n_1 r_1} = \frac{n_1 - n_0}{n_1 n_0 R_1}.$$

If there is a succession of surfaces, then a second surface would give

$$\frac{1}{n_1 r_1} - \frac{1}{n_2 r_2} = \frac{n_2 - n_1}{n_2 n_1 R_2}.$$

Substituting for $1/n_1 r_1$, we obtain

$$\frac{1}{n_0 r_0} - \frac{1}{n_2 r_2} = \frac{n_1 - n_0}{n_1 n_0 R_1} + \frac{n_2 - n_1}{n_2 n_1 R_2},$$

and for $j + 1$ surfaces

$$\frac{1}{n_0 r_0} - \frac{1}{n_{j+1} r_{j+1}} = \sum_{k=0}^j \frac{(n_{k+1} - n_k) c_{k+1}}{n_{k+1} n_k},$$

where the radius of curvature R_k has been expressed using the surface curvature, $c_k = 1/R_k$. If the object surface is a plane, then $1/n_0 r_0 = 0$ and c_{j+1} is the curvature of the final image, its reciprocal of the Petzval radius, r_{j+1} , is

$$\frac{1}{r_{j+1}} = -n_{j+1} \sum_{k=0}^j \frac{(n_{k+1} - n_k) c_{k+1}}{n_{k+1} n_k} \quad (6.39)$$

The above is an expression of the *Petzval Theorem* and the summation term is known as the *Petzval sum*. Note that the field curvature is dependent on the power of the surface divided by the index products, but independent of the placement of the surfaces. Normally the amount of field curvature is given by the difference between the edge of the field image plane and the paraxial image plane, again the “sag” of the curve. For small angles, this is given by

$$\Delta_{pc} = Z_{\text{edge}} - Z_{\text{parax}} = \frac{h^2}{2r_{j+1}}, \quad (6.40)$$

or, using Eq. 6.39 and assuming $n_{j+1} = n_j$, by

$$\Delta_{pc} = -\frac{h^2}{2} n_j \sum_{k=1}^j \frac{(n_{k+1} - n_k)c_{k+1}}{n_k n_{k+1}}. \quad (6.41)$$

This represents the longitudinal value of the Petzval curvature. Multiplying Δ_{pc} by the final slope angle u_j and substituting for h^2 , the object height using the optical invariant $I^2 = h^2 n_j^2 u_j^2$, the transverse Petzval sum is given by

$$PC = \Delta_{pc} \cdot u_j = \frac{I^2}{-2n_j u_j} \sum_{k=1}^j \frac{n_{k+1} - n_k}{n_{k+1} n_k} c_{k+1}.$$

This can be identified with the term

$$\frac{\sum P_k I^2}{-2n_j u_j} \frac{Y_1}{y_1} \left(\frac{\bar{Y}_0}{\bar{y}_0} \right)^2$$

in Equation 6.13(a), where Seidel coefficient is found to be

$$P_k = \frac{n_k - n_{k-1}}{n_{k-1} n_k} c_k = \frac{\phi_k}{n_{k-1} n_k} \quad (6.42)$$

and we have assumed that $Y_1 = y_1$ and $\bar{Y}_0 = \bar{y}_0$. Note that the value of the Seidel coefficient P_k must be multiplied by the square of the optical invariant for this term to be comparable to the other aberration coefficients. Evaluation of Eq. 6.39 for the lens in Example 6.A gives a Petzval radius of 88.03 mm. For a chief ray angle of 10° , the paraxial image height is 10.215 mm. At that point in the image plane the sag in the Petzval surface is 0.593 mm.

Using Eq. 6.39 it can be shown that for a thin lens of refractive index n and focal length f , the Petzval radius is $p = -nf$ [Problem 6.8]. This is similar to the result derived above for a thin lens. In a system of thin lenses the Petzval curvature can be written as

$$c_p = \frac{1}{r_p} = -\sum_{k=1}^j \frac{\phi_k}{n_k}. \quad (6.43)$$

The relationship between the astigmatic image surfaces and the Petzval surface consists of more than a limiting surface as astigmatism is reduced. As was noted, the tangential image surface is three times farther from the Petzval surface than the sagittal surface. To third order the surfaces have the same sign and therefore are on the same side of the Petzval surface. With the introduction of fifth and higher-order corrections it is possible to cause the sagittal surfaces to curve back and cross the Petzval surface at the same point, as shown in Fig. 6.16. The lens has been corrected at zone Z and is always correct on axis at A . For intermediate image points between Z and A , the astigmatism is not too bad, but for image points outside the zone the correction for marginal rays deteriorates rapidly.

6.2.5. Distortion

If field curvature represents a longitudinal bending of the image field along the optic axis, *distortion* represents its transverse counterpart. If a rectilinear grid is used as an object, its image in a system with distortion will be a grid of curved lines. This is because the transverse magnification in a system with distortion changes as a function of the distance from the optic axis. If the magnification decreases as the distance from the axis increases, a “barrel-shaped” image occurs; if the magnification increases with distance, a “pincushion” image is formed, as shown in Fig. 6.17. The amount of distortion is the difference between the actual ray height y_c and the paraxial value y_p , normalized to the paraxial value: $100\%(y_p - y_c)/y_p$. A distortion plot is a graph of the ray height versus the amount of distortion, as shown in Fig. 6.17. Visually, most people do not notice distortions of less than 1%. However, in cartography and other precision imaging applications, such as photomask imaging for large-scale integrated circuit fabrication, even small amounts cannot be tolerated.

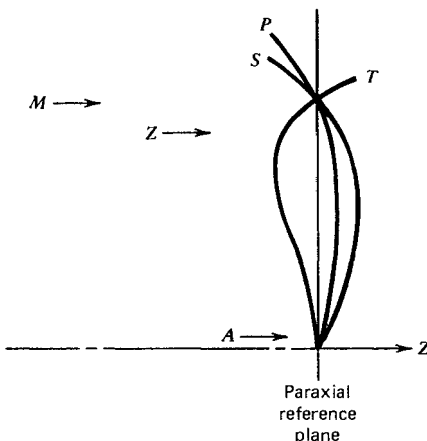


Figure 6.16. Plot of image surfaces for an anastigmat.

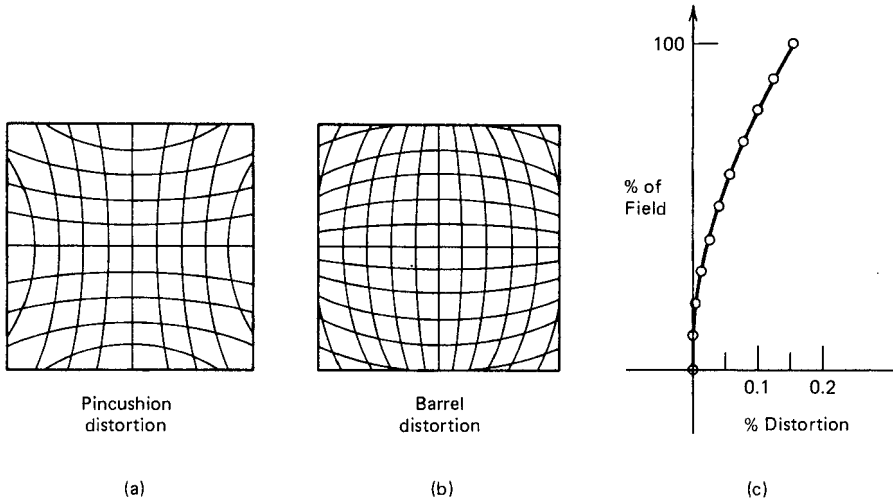


Figure 6.17. Distortion of a rectangular grid object. (a) Pincushion distortion. (b) Barrel distortion. (c) Plot of distortion as a function of object height.

The Seidel third-order contribution, the term E_k in Eq. 6.13, varies as the cube of the object height. The fractional distortion $(\bar{Y}_{j+1} - \bar{y}_{j+1})/\bar{y}_{j+1}$ can be expressed as the square of the object height. The expression for the Seidel coefficient of distortion is similar to the expressions for the other aberrations, but with an additional term

$$E_k = \bar{S}_k \bar{i}_k \bar{i}_k + I(\bar{u}_{k-1}^2 - \bar{u}_k^2), \quad (6.44)$$

where

$$\bar{S}_k = \bar{y}_k n_{k-1} \left(\frac{n_{k-1}}{n_k} - 1 \right) (\bar{u}_k + \bar{i}_k), \quad (6.45)$$

the same expression as S_k (Eq. 6.17) except with chief ray values. I is the optical invariant.

As is evident from the examples in this chapter, the Seidel third order aberration coefficients are a useful means of evaluating a design. However, it should be pointed out that these are *only* third order contributions to the aberrations of the system. When these are reduced to relatively small values the higher order contributions begin to dominate. At that point other methods of evaluating the performance of a system need to be employed.

6.3. REDUCING ABERRATIONS

The purpose of this section is to give some idea of the methods used to correct aberrations in an optical system. We shall look at a small number of methods that can be used to eliminate, or at least reduce, the aberrations of a system to an acceptable level. The adjective “acceptable” should be emphasized. No lens system can be corrected *exactly* for *all* aberrations over an entire field. Instead a lens designer must set limits on the aberrations to produce the performance required with a finite amount of money and a small number of elements that can be fabricated in a reasonable amount of time. This part of design cannot be taught in a textbook of this level. In fact, it is doubtful if it can be taught successfully in the classroom at all. Rather, an ability to set limits on aberrations grows out of real problems presented by the real world. It is possible to describe each aberration and talk about its correction at that point in the text. However, it would give a false sense of how aberrations are corrected. A designer is usually faced with a multi-dimensional crossword puzzle which in many cases cannot be solved totally, but must be solved in some small region where it will do the most good.

6.3.1. Bending a Lens

It was pointed out earlier that a thin lens with a particular optical power $\phi = 1/f$ can be fabricated an infinite number of ways. All that is necessary is that values of n , R_1 , and R_2 in the lensmaker's equation (Eq. 5.16) give the required focal length $f = 1/\phi$. Therefore, these degrees of freedom can be used to correct aberrations in the lens. For example, the shape of a lens affects the amount of spherical aberration and coma. In Fig. 6.18, the spherical aberration and coma Seidel coefficients have been plotted for a 58.53-mm lens with a 20-mm entrance pupil diameter and the aperture stop at the front surface (the same values as in Example 6.A) as a function of the front surface curvature c_1 normalized to the total curvature of the lens, c . Note that the spherical aberration is a parabolic function of c_1/c with a minimum near $c_1/c = 0.85$ and that the coma crosses zero in the same region. There is no shape for a single lens with a refractive index of 1.517 that is free of spherical aberration.

The plot also shows that the orientation of a lens has an effect on the aberrations. In the figure the aberrations at the value of c_1/c equal to that of Example 6.A are considerably smaller than those for the same lens turned around ($R_1 = 120$ mm; $R_2 = -40$ mm) so that the longer radius of curvature surface faces the object. The spherical aberrations are 2.5 times larger and the coma more than 4 times greater when the lens is oriented with the shorter radius of curvature surface facing the object than when the lens is reversed. One can also see from Fig. 6.18 that in the case of an object at infinity, a plano-convex lens with its curved surface facing the object will show less aberration than if the lens is reversed (plano surface facing the object) [Problem 6.10].

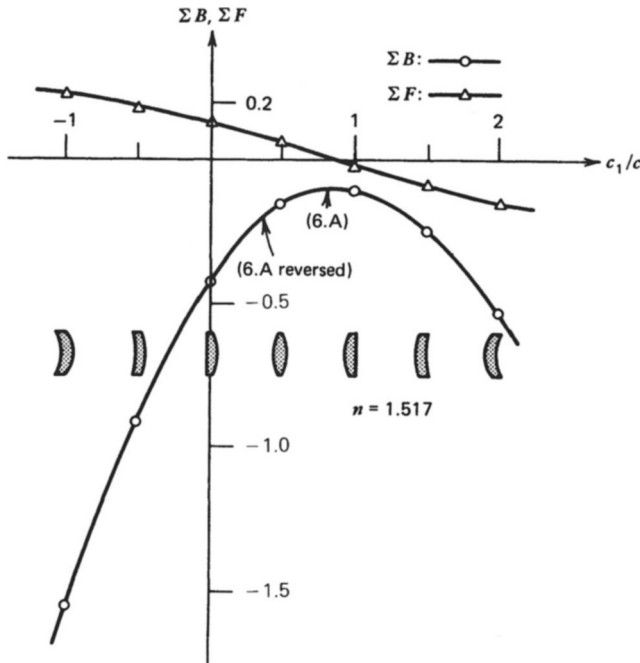


Figure 6.18. Lens bending I. Plot of the spherical aberration (B) and coma (F) of a lens as a function of surface curvature. The locations of the values for the lens in Example 6.A and the same lens with its surfaces reversed are shown.

Bending also has an effect on the other Seidel aberrations. Both astigmatism and distortion change as the lens' shape is changed, as shown in Fig. 6.19. Because the power of the lens is not changed, the field curvature remains constant for all bendings. With combinations of lenses it is possible to correct several aberrations at the same time by bending lenses in such a manner that the overall aberrations of the combination are less than the values for the separate lenses [Problem 6.11].

6.3.2. Moving the Aperture Stop—Stop Shift Equations

An examination of the expressions for the Seidel coefficients shows that three of the aberrations—coma, astigmatism, and distortion—depend on quantities derived from a chief ray trace. Since the chief ray is determined by the location of the aperture stop, relocation of the stop position will change the aberrations of the system. For Example 6.A, the aberrations were calculated with the stop at the first surface of the lens. If the stop is moved out in front of the lens the amount of coma is reduced. If coma is plotted as a function of stop position zero coma will occur when the stop is located about 12.6 mm in front of the lens. To provide some idea on how the movement of a stop can affect coma, a schematic

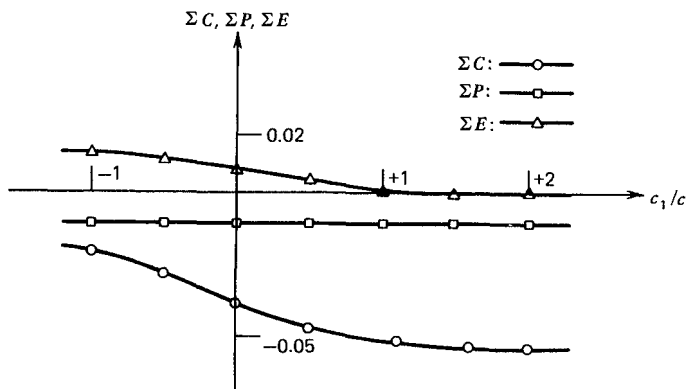


Figure 6.19. Lens bending II. Plot of astigmatism (C), field curvature (P) and distortion (E) as a function of surface curvature. Note, that since the optical power of the lens is constant, the field curvature (P) does not change.

trace of four rays through a thick lens is shown in Fig. 6.20. The three lower rays correspond to the chief and marginal rays for a stop located at the lens. The separation between the chief ray ② and the intersection of the marginal rays ① and ③ shows that coma is present. When the stop is moved to position B , the top marginal ray of the previous setup becomes the chief ray. The intersection of all three of the rays, ②, ③, and ④, at the same point means the lens with the stop at this position is free of coma.

Although it would be possible to move a stop, recalculate the aberrations, and plot the results, there is a simpler method for determining the effect of repositioning a stop. If a stop is moved some distance, the new chief ray from an

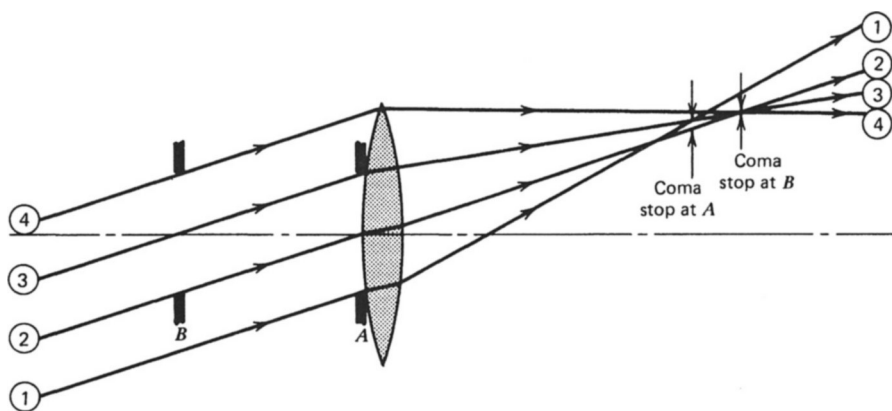


Figure 6.20. Stop shifting. Repositioning of a stop from A to B redefines the chief ray of the system. With the stop at A , ray ② is the chief ray and coma is evident. By moving the stop to B , ray ③ is the chief ray and coma is eliminated.

object of height \bar{Y}_0 crosses the plane of the old aperture stop at a height y' , and if the axial ray in the plane of the old aperture stop is y , the ratio of these two ray heights $Q = y'/y$ is the only parameter needed to compute the aberrations of the new system from those of the old. Since the entrance pupil is the image of the aperture stop in object space, the ray heights in the pupil bear the same ratio as at the aperture stop, $Q = y'_1/y_1$ (assuming the entrance pupil is defined as the first surface in the system, a common practice). The effect of shifting the stop on the aberrations of a system may be determined from the following equations, the *stop-shift equations* (the indices on the summation sign have been omitted).

$$\sum B_k^* = \sum B_k, \quad (6.46a)$$

$$\sum F_k^* = \sum F_k + Q \sum B_k, \quad (6.46b)$$

$$\sum C_k^* = \sum C_k + 2Q \sum F_k + Q^2 \sum B_k, \quad (6.46c)$$

$$\sum P_k^* = \sum P_k, \quad (6.46d)$$

$$\sum E_k^* = \sum E_k + Q \sum (3C_k + P_k I^2) + 3Q^2 \sum F_k + Q^3 \sum B_k. \quad (6.46e)$$

One of the things to be noted in the equations is the method by which coma in the lens of Example 6.A was reduced. The reduction was based on the fact that the lens possessed some spherical aberration. (If $\sum B_k = 0$, then from Eq. 6.46b $\sum F_k^* = \sum F_k$). In fact, since B and F for the stop at the lens have been calculated, it is easy to find the location of the coma-free stop position from the second of the stop-shift equations as shown in Example 6.D. When the stop is shifted, the principal ray defined by the stop is moved closer to the intersection of the marginal rays thus reducing the coma. If there were no spherical aberration, however, this shift would not occur and the coma would be the same for both stop positions.

Besides changes in the curvature of lens surfaces (bending), changes in the thickness and/or separations of individual elements and their refractive indices can be used to modify the aberration coefficients. Problems 6.17 through 6.21 provide some practice and insight into the effects of these changes.

EXAMPLE 6.D. Given the values of the third-order aberration coefficients for spherical and comatic aberration B and F , respectively, for the lens in Example 6.A, where is the coma-free stop position located?

Solution: From Example 6.B, $B = -0.1051$.

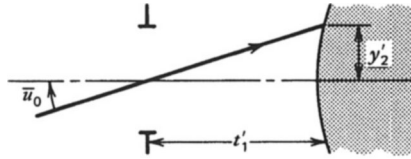
From Example 6.C, $F = 0.0231$.

The stop-shift equation (6.43b) is $F^* = F + QB$.

In this case we require $F^* = 0$. That is,

$$F + QB = 0 \quad \text{or} \quad Q = -\frac{F}{B} = -\left(\frac{0.0231}{-0.1051}\right) = 0.2198.$$

Assume a stop position in front of the first surface.



The stop was originally at the first surface and therefore the entrance pupil and the first surface were identical. Upon moving the surface, the position of the *new* chief ray on the *old* entrance pupil is $y'_2 = \bar{u}_0 t'_1$. The new chief ray angle is the same as the old one ($\bar{u}_0 = 0.17453$) since the object is at infinity. The distance t'_1 is to be determined. Since the axial ray was a parallel ray 10 mm from the axis, it crosses the old entrance pupil at 10 mm ($y_2 = 10$ mm).

$$Q = \frac{y'_2}{y_2} = \frac{t'_1 \cdot \bar{u}_0}{y_2} = -\frac{F}{B}.$$

Only t'_1 is unknown, so inserting values for the other quantities in the equation

$$t'_1 = \frac{10 \text{ mm}}{0.17453} \times 0.2198 = 12.59 \text{ mm}.$$

Thus, moving the system stop 12.6 mm in front of the lens will eliminate coma.

Note that for this calculation, the same chief ray angle \bar{u}_0 for the object at infinity was used for computing the coma-free stop location as was used to compute F . In the case of finite object distances, the computation of t'_1 is somewhat more involved, but easily done. It can be shown that coma will be zero if the new entrance pupil is shifted a distance $t'_1 = Q y_2 / (\bar{u}_0 + Q u_0)$, where $Q = -F/B$, u_0 and \bar{u}_0 are the original axial and chief ray slope angles and y_2 is the axial ray height in the original entrance pupil.

6.3.3. Other Correction Techniques

The other parameters of the optical system, refractive index, component thickness, and separation, may be varied to reduce the aberrations of the system. Although the choice of refractive index is most useful in reducing chromatic aberrations, the sequence of refractive indices in a system and their differences can provide an additional degree of freedom when correcting a system [Problem 6.12].

Since most optical systems consist of more than one component, the separation between components can be modified as well as the thickness of the individual components. For example, in the telephoto lens in Chapter 2, the separation between the front and back lenses could be changed. This would, of course, change the effective focal length of the system. Because of the linear properties of the paraxial equations that determine the Gaussian constants, the focal length of the revised system is obtained by simply multiplying all curvatures and thicknesses by a constant equal to the ratio of the old effective focal length to the new effective focal length. In Example 6.E a ray trace of the telephoto lens in Example 2.C is run with thin lenses ($t_1, t_3 = 0$), and the aberration coefficients computed. Then the separation between lenses is increased, the entire system scaled to the original focal length, and aberrations recomputed. For the comparisons to be valid, the f -number of the system must be maintained by changing the entrance pupil diameter.

EXAMPLE 6.E. For an object at infinity, what are the aberrations for Example 2.C if the separation between lenses is 20 mm? If the separation is 40 mm? If the lenses are rescaled to the original 150-mm effective focal length with the original f -number ($f/7.5$), what are the aberrations?

Solution: Rather than work out all values, the system parameters and aberrations will be quoted for each setup. In all cases, a paraxial ray trace was done, but the thickness of each lens was set at zero. The lenses were assumed to be symmetrical lenses:

Case 1

t	∞	0	20	0	90
n	1	1.5	1	1.5	1
c	0	1/50	-1/50	-1/45	1/45

Effective focal length = 150 mm, Entrance pupil diameter = 20 mm,
 f -number = $150/20 = 7.5$, Chief ray angle = 10 degrees.

Aberration coefficients:

$$B = -0.1730, \quad F = -0.0648, \quad C = -0.0578, \\ PI^2 = 0.0045, \quad E = -0.0788.$$

Case 2

t	∞	0	40	0	12.86
n	1	1.5	1	1.5	1
c	0	1/50	-1/50	-1/45	1/45

Effective focal length = 64.29, Entrance pupil diameter = 20 mm,
 f -number = $64.3/20 = 3.2$, Chief ray angle = 10 degrees.

Aberration coefficients:

$$B = -0.2541, \quad F = -0.1002, \quad C = -0.0968, \\ PI^2 = 0.0045, \quad E = -0.0567.$$

Note that the f -number of the system is now smaller due to the shorter focal length of the lens. Thus the aberrations are not truly comparable.

If the entrance pupil diameter is reduced to 8.57 mm, the f -number is again 7.5. Then the aberration coefficients are:

$$B = -0.0086, \quad E = -0.0079, \quad C = -0.0177, \\ PI^2 = -0.0008, \quad E = -0.0243.$$

Case 3

t	∞	0	93.33	0	30
n	1	1.5	1	1.5	1
c	0	1/116.67	-1/116.67	-1/105	1/105

Effective focal length = 150, Entrance pupil diameter = 20 mm,
 f -number = 7.5, Chief ray angle = 10 degrees.

Aberration coefficients:

$$B = 0.0200, \quad F = -0.0184, \quad C = -0.0415, \\ PI^2 = 0.0019, \quad E = -0.0567.$$

In the last case the system was scaled from a 64.29-mm up to a 150-mm effective focal length and 20-mm entrance pupil diameter. Note in both the unscaled and scaled versions, the aberration coefficients for a lens system with the same f -number change dramatically with a change in the separation of the two lenses.

A lens corrected for both spherical aberration and coma, as noted earlier is called an aplanatic lens, or aplanat. Since it is corrected for coma it obeys the Abbe sine condition. Although one can correct for spherical aberration and chromatic aberration with a cemented doublet, to produce an aplanat an airspace must be added between the lenses and the interior curvatures c_2 and c_3 made unequal. (Kingslake, Ref. 6.3, has an entire chapter on aplanatic objectives.)

There are a few special geometries in an optical design that can simplify the calculations and reduce aberrations. As a very simple example, an object at the vertex of a surface (Fig. 6.21(a)) has an image at the same point, so there are no aberrations. More important, a point object at the radius of curvature of a surface (Fig. 6.21(b)) will suffer no spherical aberration since all rays are perpendicular to the surface. Most important is the case where the angle of

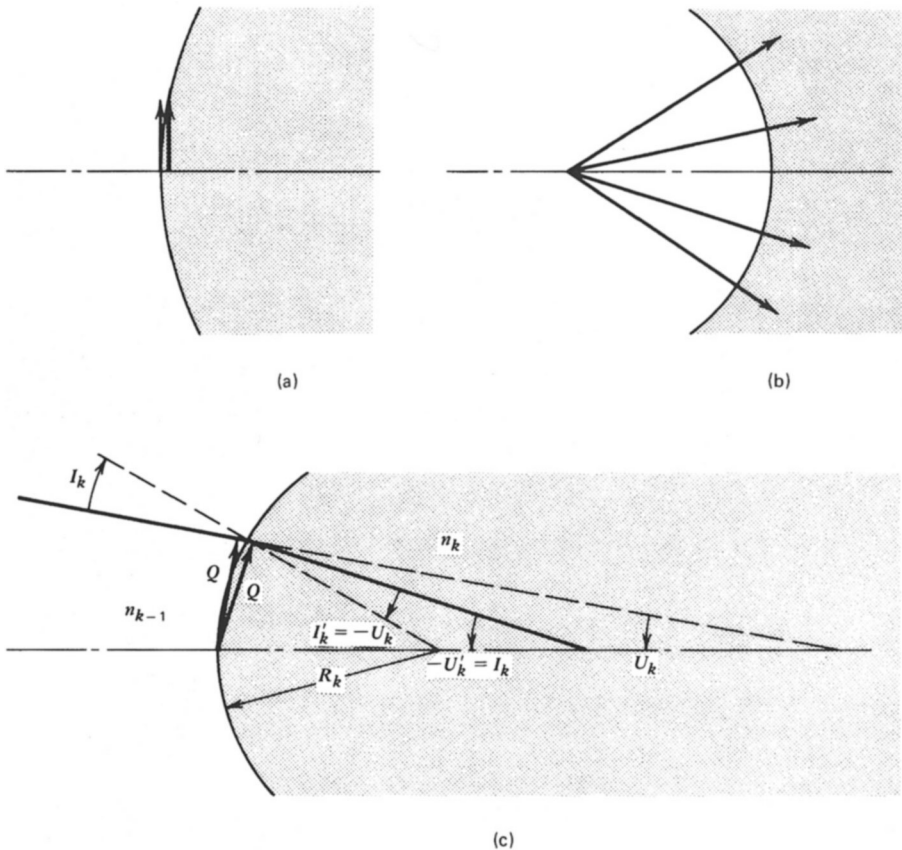


Figure 6.21. Aberration-free geometries. (a) Object at a surface. (b) Object at the center of curvature of a surface. (c) Aplanatic surface, where the angle of refraction I'_k equals the initial slope angle $-U_k$.

refraction I'_k is equal to the original slope angle for a ray U_k , Fig. 6.21(c). Using Eq. 6.3, it is obvious that if $I'_k = -U_k$, then $U'_k = -I_k$. Referring back to the meridional ray trace equations

$$L' = T_k = \frac{-Q'_k}{\sin U'_k}, \quad (6.10)$$

$$Q'_k = Q_k \frac{\cos U'_k + \cos I'_k}{\cos U_k + \cos I_k}, \quad (6.7)$$

it is clear that

$$Q'_k = Q_k$$

Using Eq. 6.2 and Snell's Law, we can write the image distance

$$\begin{aligned} L' &= \frac{-Q'_k}{\sin U'_k} = -R_k \frac{\sin I'_k - \sin U'_k}{\sin U'_k} \\ &= R_k \frac{\sin I_k + \sin I'_k}{\sin I_k} = R_k \left(1 + \frac{\sin I'_k}{\sin I_k} \right) \\ &= R_k \left(1 + \frac{n_{k-1}}{n_k} \right) \end{aligned}$$

or

$$L' = R_k \frac{(n_k + n_{k-1})}{n_k}. \quad (6.47)$$

Note that the image distance is independent of the height of the ray, Q_k . Therefore, for this special surface geometry known as *aplanatic surface* there will be no spherical aberration due to this surface. The axial intercept distance of a ray prior to refraction can be shown to be [Problem 6.13]:

$$L = R_k \frac{n_k + n_{k-1}}{n_{k-1}}.$$

Thus, if the axial intercept distance before refraction by an aplanatic surface and the refractive indices are known, the required curvature can be calculated. Moreover, this geometry is coma free in that it fulfills the Abbe sine condition, since

$$\frac{\sin U_k}{\sin U'_k} = \frac{\sin I'_k}{\sin I_k} = \frac{n_k}{n_{k-1}} = \frac{u_k}{u'_k}$$

in the limit of small angles, $u_k = -i'_k$, and $u'_k = -i_k$, and Snell's Law becomes $n_k i'_k = n_{k-1} i_k$ or $n_k u_k = n_{k-1} u'_k$. The pair of object and image points are known as the *aplanatic* points of this surface since there is neither spherical aberration nor coma for this geometry. ($S_k = 0$ (Eq. 6.17) in such cases.) As we shall see later, a combination of special geometries (e.g., the aplanatic conjugates) and symmetrical design can remove or reduce some aberrations before any corrections are applied to a system.

The symmetry argument is particularly true for astigmatism. If a lens is symmetrical about its aperture stop, then the optic axis is not unique and any chief ray could serve as an optic axis. Since all rays are meridional rays, there is no astigmatism. In cases where symmetry cannot be preserved, it may take a combination of stop movements, bendings, refractive index changes, and movement of elements to reduce astigmatism.

6.3.4. Correcting Field Aberrations

From the expression for the Petzval curvature (Eq. 6.42), it is evident that the reduced power of each surface, $\phi_{k+1}/(n_{k+1}n_k) = (n_{k+1} - n_k)/(R_{k+1}n_{k+1}n_k) = (n_{k+1} - n_k)c_{k+1}/n_{k+1}n_k$ contributes to the field curvature. Thus a combination of curvatures and refractive indices must be devised to provide the correct focal length and still reduce the curvature of field.

As an example, take a simple lens that is split into a positive and negative doublet, as was done when correcting chromatic aberration. In this case the controlling equations are

$$\phi = \phi_1 + \phi_2 = 1/f \quad (5.37)$$

and by Eq. 6.43, $c_p = 0$,

$$\frac{\phi_1}{n_1} + \frac{\phi_2}{n_2} = 0. \quad (6.48)$$

If ϕ_1 is the power of the positive lens, then $|\phi_1| > |\phi_2|$, so that there will be a net positive power. However, the two terms in Eq. 6.48 must be equal with opposite signs, and this can only be achieved if the refractive index n_1 is greater than n_2 . The positive lens in the doublet would therefore probably be of flint glass and the negative lens of crown, as in Example 6.E. If the lens must also be an achromat then the condition placed on the glasses is that [Problem 6.14]

$$\frac{n_1}{V_1} = \frac{n_2}{V_2}. \quad (6.49)$$

EXAMPLE 6.F. Design a lens whose focal length is to be 80 units to be free of chromatic aberration and have a zero Petzval sum using thin lens considerations.

Solution: Choose two glasses based on Eq. 6.47. Let the negative lens be a crown, LaK10(720504): $n_1 = 1.720$, $V_1 = 50.4$, $n_1/V_1 = -0.03412$. Searching the glass list (Appendix A), one finds a light flint LLF1(548458) with a ratio $n_2/V_2 = 1.548/45.8 = 0.03380$.

Since $\phi = \phi_1 + \phi_2 = \frac{1}{80}$

$$\frac{\phi_1}{\phi_2} = -\frac{n_1}{n_2} = \frac{1.720}{1.548} = -1.11\bar{1},$$

$$\phi_1 = -1.11\bar{1}\phi_2,$$

$$\frac{1}{80} = -1.11\bar{1}\phi_2 + \phi_2 = -0.11\bar{1}\phi_2,$$

$$\phi_2 = -0.1125 \text{ (} f_2 = -8.888 \text{ units),}$$

$$\phi_1 = -1.11\bar{1}\phi_2 = 0.125 \text{ (} f_1 = 8 \text{ units).}$$

Further variation of these two lenses should allow reduction of spherical aberration also.

In the case of photographic lenses, where the imaging plane is almost always a piece of flat photographic film, field curvature is a controlling aberration. Thus designers of photographic lenses pay a great deal of attention to the Petzval sum in the initial configuration of a design. Since the spacing of the components has no effect on the sum, one method of correcting field curvature is to design a lens without the correction and then add a negative element (assuming a positive lens is being designed) close to the final image surface to compensate for the curvature from the rest of the system. The concentric shell, the second element in Fig. 6.22 is an example of this type of correction. If placed near the final image, the element, known as a *field flattener*, has only a small effect on the image location, yet has its full effect in the Petzval sum. In a single element lens, the field curvature is zero if the radius of curvature, and therefore the power, is the same for both surfaces. Any focusing comes from the separation of the two surfaces, and it can be shown [Problem 6.15] that the power of a thick section spherical shell is proportional to the thickness.

As in the case of an object located at the center of curvature of some optical surface or the design with the equal radii of curvature of the thick meniscus lens in the case just noted, the use of the symmetry is useful for reducing distortion. If a lens is designed that is symmetrical about the aperture stop, it will be not only

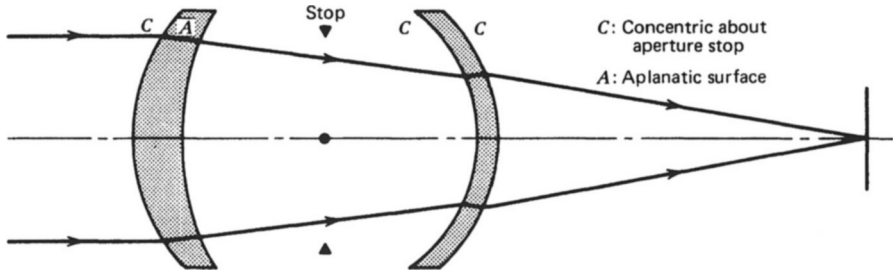


Figure 6.22. Field flattener. Introduction of an element with negative power near the image plane provides a small contribution to the image location, but is effective in reducing the Petzval sum, thereby flattening the image plane.

free of distortion but also free of lateral color and coma for an object at unit magnification. This can be seen simply by the fact that rays traced through an optical system are reversible; thus whatever occurs to one ray on one side of the stop will also occur in mirror symmetry on the other side. Thus a grid object on one side will be imaged *without distortion* (other aberrations may get worse) into a grid on the other side of the stop. When the magnification is not unity the front and back halves of the system must be scaled to produce the distortion free condition. One simple method of reducing distortion in a reproducing system such as a camera-projection system is to project an image with the same lens that was used to take the picture, since the distortion due to reduction will be equal and opposite to that for enlargement. In some laser scanning systems, there is an inherent distortion in the scanning geometry. If the focusing lenses following the scanner are designed with a compensating distortion (for example, a built-in barrel distortion to compensate for pincushion distortion in the scanner), the final scan pattern will be distortion-free.

6.3.6. Aspheric Surfaces

With the exception of a few cylindrical lenses, all of the designs considered so far have used spherical surfaces. This is because ray tracing through spherical surfaces is simple and fabrication is easy compared to other surface shapes. However, there are cases where, because of size, simplicity, or optical speed, the best design must incorporate an aspheric surface. The simplest example of an aspheric surface is the corrected reflecting telescope mirror. With an infinite conjugate a spherical mirror displays spherical aberration. But, by flattening the edges of the curve of the mirror, turning the sphere into a paraboloid, the mirror can be corrected for infinite object distance. For other object and image distance combinations, other conic solids of revolution, obtained by rotating about a curved symmetry axis, will eliminate spherical aberration.

Because there is a need for some reference surface, an asphere is usually defined by its departure from some reference sphere. This is expressed as the

difference between the sphere and the asphere at different heights above the optic axis, as shown in Fig. 6.23. First the distance between the plane at the sphere vertex and the sphere is determined. This is referred to as the sagitta or "sag" of the surface at different distances from the optic axis. For a sphere the sag may be written as

$$z_s = \frac{c\rho^2}{1 + \sqrt{1 - c^2\rho^2}}, \quad (6.50)$$

where $c = 1/R$, the curvature of the surface, and $\rho = \sqrt{x^2 + y^2}$, the distance from the optic axis. If c^2 in the denominator of Eq. 6.50 is replaced by $(1 + \kappa)c^2$, the equation gives the sag for an asphere, which is a conic section of revolution, κ is the conic constant ($\kappa = 0$ for a sphere, $\kappa = -1$ for a parabola, $-1 < \kappa < 0$ for ellipsoid, and $\kappa < -1$ for a hyperbola). Depending on the conjugate distances and the presence of other elements in the system, different conic sections are used to construct systems with no spherical aberration. Additional corrections for off-axis aberrations can be made by introducing surfaces that can be represented as higher order polynomials of $c^2\rho^2$ (i.e., $(x^2 + y^2)/R^2$). The added degrees of freedom provided by allowing surfaces to be aspheric must be balanced against the difficulty and increased cost of producing such surfaces.

An example of an aspheric surface in an optical system is the Schmidt corrector plate used for systems with large light-gathering power, such as TV projection systems, missile tracking cameras, and wide-field telescopes. The plate has a fourth-power curve of the form $z_a = \alpha\rho^2 + \beta\rho^4$.

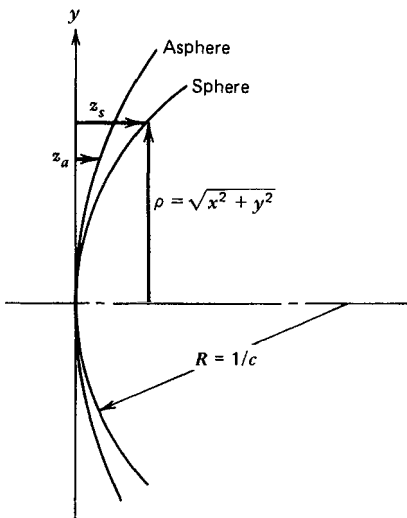


Figure 6.23. Aspheric surface. Definition of an asphere as a departure from a spherical surface.

For large optics it is less costly to generate a spherical surface and then correct it by inserting a thin plate with one aspheric surface. The expansion of the sag of a sphere is given by

$$\Delta = c\rho^2 + \frac{c\rho^4}{4} + \dots$$

For distant objects the reflecting surface that is free of spherical aberration is a paraboloid. This surface has a sag of $c\rho^2$. Thus it is the second term in the spherical expansion $c\rho^4/4$ that must be canceled to approximate a paraboloid. This would require a surface such as surface 1, as shown in Fig. 6.24(a). If a

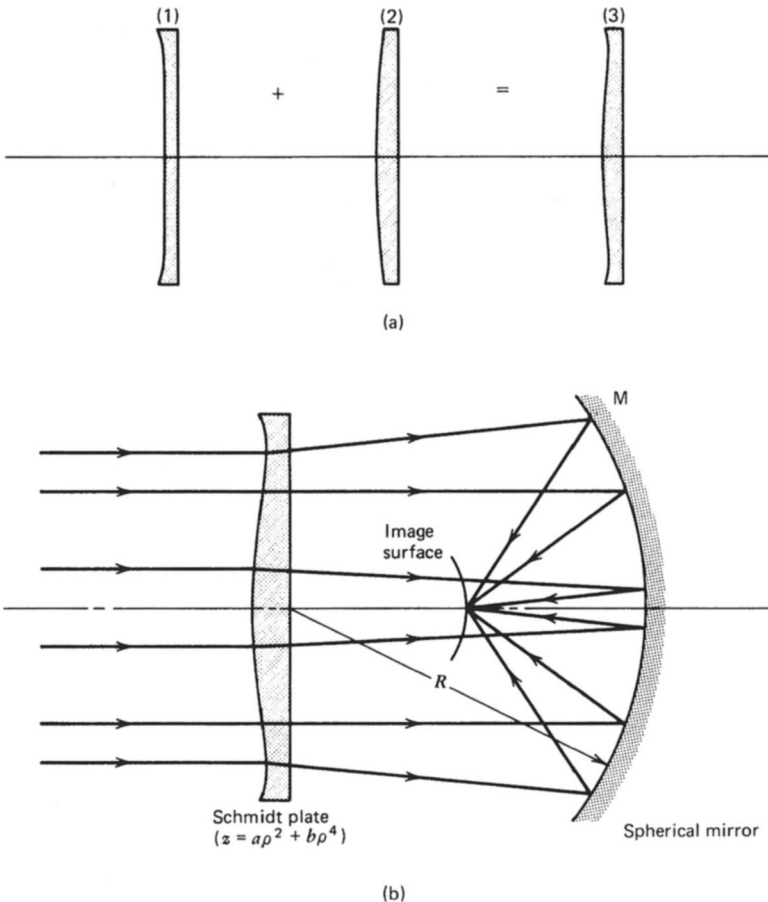


Figure 6.24. Schmidt-type aspheric correction. (a) The correction of a spherical surface to eliminate spherical aberration can be done with a correction plate with a fourth power curve (1). If some spherical power (2) is added, a more producible shape (3) can be used. (b) A Schmidt telescope.

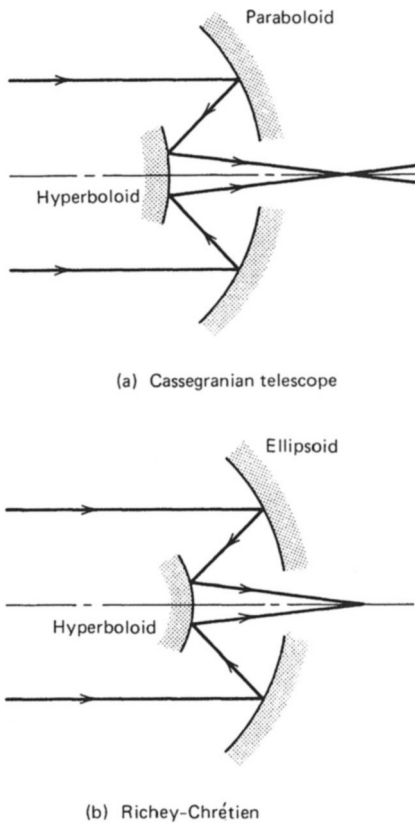


Figure 6.25. Two aspheric telescopes. (a) Cassegrainian. (b) Ritchey-Chrétien.

weak convex sphere is added (surface 2), a more producible shape (surface 3) is generated, since the local radii of curvature are much longer than those near the edge of surface 1 [Problem 6.16]. When the plate is placed at the radius of curvature of a spherical mirror as shown in Fig. 6.24(b), it takes the place of parabolizing the mirror surface and also serves as the system stop. Note the center zone causes rays near the axis to focus short, the intermediate zone provides no correction, and the marginal zone causes the rays to focus long compared to an uncorrected spherical mirror. Using the symmetrical principal, one can see that there will be no coma or astigmatism since all chief rays are normal to the mirror and there is, therefore, no distinction between tangential and sagittal planes. Other mirror systems that use aspherics are the Cassegrain telescope and its aplanatic offspring, the Ritchey-Chrétien telescope (Figs. 6.25(a) and (b)). In many lens systems it is usually simpler to add additional spherical lenses than it is to achieve the same correction with an aspheric surface. However, if the lens system is to be made from plastic, only the molds must be individually fabricated so that it would be economical to use aspherics. One new technique which has been used to fabricate aspheric mirrors is *replication*.

This process transfers a precision optical surface onto a surface that has the approximate shape of the precision surface and is inexpensive to produce. A thin layer of epoxy fills in the difference between the rough surface and the precision surface. After the epoxy sets, the rough surface with its epoxy coats is separated from the precision surface resulting in a copy or replication of the desired surface. Beyond the difficulties presented by fabricating aspheric surfaces, there are the additional difficulties of testing the surfaces or the entire system and specifying the curves so that they can be fabricated. A good review on aspherics by Shannon can be found in Ref. 6.4.

6.4. OPTICAL COMPUTATION

The examples presented in this chapter and the previous one were used to show the type of computation that goes into the preliminary design and evaluation of optical systems. Those calculations are based on a small number of equations and concepts that are easily understood yet powerful in their application. For years such ray tracing was done with aid of tables of logarithms. First adding machines and then mechanical calculators and finally electronic calculators were used to reduce the tedium. Most recently computers have been used to perform ray traces for a rapid evaluation of optical systems. It might seem in the face of this progress that optical design could be handled by anyone with access to a computer. However, without some intelligent choices as to the number of elements, their glasses, powers, and spacings, the computer programs can do nothing but evaluate a poor optical system. And, without intelligent choices as to the parameters to be modified when reducing the aberrations, the improvement of optical performance may take a long time. To paraphrase one of the catch phrases of the computer world: Optical garbage in, optical garbage out.

6.4.1. Predesign

The first stage of lens design might be called the predesign stage. At this point the focal length, f -number, and other Gaussian constants are decided on. This initial design can be refined by evaluating the third-order aberration coefficients and revising the design. Most of this can be carried out on a programmable calculator or a microcomputer.

Until recently, computation at this stage was done using thin lenses because the amount of computation could be reduced to a minimum yet the important factors of the design could still be evaluated. Discussions of thin lens predesign can be found in Refs. 6.1–6.3. With the advent of the microcomputer paraxial calculations are easily performed on actual systems and it is not necessary to start from the thin lens formulations, although the design principles explained in the references are valid for most thick lens problems. Usually if a lens has been designed (and perhaps patented) to work in an application similar to the one at hand, the values for this lens can be used as a starting point for modification and

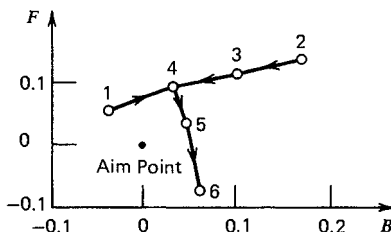
refinement. A paraxial ray trace with rescaling to the size and f -number needed can be followed by an evaluation of the aberrations to provide a baseline against which future attempts at improvement can be compared. One useful method for understanding parameter changes is the “double plot” (see box). Remember the third order aberrations serve only as the lowest-order gauge of performance.

THE DOUBLE PLOT

The evaluation of lens designs has been simplified by the computer. Through the use of merit functions and iteration programs a great deal of data can be generated on a large number of lens variations. The problem is usually that there is too much information and this leads to a tendency to run another iteration to see if things get better. It takes very little time to find out that a mindless use of these techniques usually ends in incredibly weird designs—most of which are impractical or unwieldy. It is only through the use of experience and insight that practical lenses can be produced.

One means of gaining insight into the effect of the variations of lens parameters on a design is to look at some designs in which only a few parameters are changed and the results are plotted. For example, a plot of coma and spherical aberrations as a function of bending the first surface of a lens would give insight into how this affects the overall design. Then freezing that parameter at some value (not necessarily at the minimum points for both aberrations), the effect of changing the first thickness or second curvature, could be explored. A double plot, as it is called, is shown for a lens in which c_1 and c_2 are changed.

1. $c_1 = 0.0500, c_2 = -0.150$
2. $c_1 = 0.0550, c_2 = -0.150$
3. $c_1 = 0.0530, c_2 = -0.150$
4. $c_1 = 0.0515, c_2 = -0.150$
5. $c_1 = 0.0515, c_2 = -0.180$
6. $c_1 = 0.0515, c_2 = -0.2$



In some instances, an “aim point” is used as an indication of departure from the ideal result. It is not always the case that the aim point is at the coordinate point (0, 0).

The choices made regarding the size of the field, the acceptable amounts of aberration, and other values to be used in evaluating a lens are arrived at by experience. It is sometimes useful to examine patented lenses by performing

your own evaluation to see what the designer was able to achieve at various field angles and apertures. An extensive discussion of the design of an achromatic objective, a Cooke triplet, and a number of other systems can be found in Smith (Ref. 6.2).

6.4.2. Design With Computer Programs

Because of the incredible computational speed and memory of large computers, a first-rate lens design program can provide an enormous amount of information about the performance of a lens. There can be so much data that it is difficult to evaluate it. To provide some method to compare different lens designs most programs provide some sort of merit function. A simple example of a merit function might be the sum of the squares of the Seidel coefficients:

$$MF = B^2 + F^2 + C^2 + P^2 + E^2. \quad (6.51)$$

The idea behind the merit function of this kind is to find a set of system parameters that minimizes its value. The reason the coefficients are squared is to prevent the possibility of obtaining a small merit function with large aberration coefficients of opposite sign.

The simple merit function given above treats all aberrations equally. For many applications some aberrations are more important than others. At times, a certain amount of aberration may be desired in a system (see, for example, the discussion on distortion) and a target value should be assigned. An improved merit function would be

$$MF = a_1(B - B_0)^2 + a_2(F - F_0)^2 + a_3(C - C_0)^2 + a_4(P - P_0)^2 + a_5(E - E_0)^2 + a_6(F' - F'_0)^2, \quad (6.52)$$

where B_0 , F_0 , C_0 , P_0 , E_0 , and F'_0 are the target values for the five Seidel coefficients and f -number of the system, and a_1 through a_6 are the weights assigned to each of the terms in MF. The f -number was included as a target value because one of the methods of reducing aberrations is to reduce the size of the entrance pupil ("stopping down the system"). To guard against artificially reduced merit functions, targets such as the one for the f -numbers are included in the merit function [Problems 6.17–6.21].

But optical design programs can do more than trace a series of rays and then calculate system performance in the form of a merit function. Most programs are written so that the system parameters (index, thickness, curvature) are varied in a systematic way to search for the smallest merit function. This series of searches is referred to as an *iteration*, since as the system is evaluated the parameters change slightly and the evaluation is iterated. Sometimes a designer will request a sequence of iterations; the program finds a system with a minimum merit function and then uses the newly found system parameters as a starting point for a new iteration.

The target values for most merit functions can be formidable. There can be of the order of 30 to 40 targets for a system. While most can be set to zero, some must be carefully defined (the f -number, for example). Before an iteration, the designer decides which surfaces, glasses, and thicknesses he or she will allow to vary and which will be held fixed. It is usually wise not to vary too many parameters; otherwise it is difficult to determine why the design is changing and what would be necessary to produce additional improvement. Most programs are capable of generating ray-intercept plots, spot diagrams, and other graphic output that can provide valuable insight into a lens' performance.

6.4.3. Tilts and Decentration

All optical systems are first designed on paper (or a computer terminal, these days) to provide image quality with the smallest amount of aberration. The designs are based on the premise that the axis of each component lies on the optic axis of the system. But either fabrication or mounting can cause the axis of a given element to be displaced by an amount Δy (*decentration*) or tipped at an angle $\Delta\theta$ (*tilt*), as shown in Fig. 6.26. Both of these errors can seriously degrade the design performance to a point where a simple design might provide a better final performance, even if the more complicated design were, on paper, superior to the simple one.

Thus, beyond the design analysis for aberrations, additional design must be done on the system to test it for sensitivity to tilts and decentrations, which reduce the image quality. Beyond the considerations of image correction, a designer must be aware if small errors in mounting components produce large image errors. Therefore, most lens design programs include the ability to specify tilts and decentrations as part of a tolerance analysis. Through the use of statistical programs that assign random errors, realistic projections of manufacturing yields can be made.

Beyond the sophisticated design process that can produce elegant new arrangements of lens elements, there is another design process that is needed to specify the barrel size, the spacers, retainer rings, and the lens cells. These must keep the decentration and tilt errors to a minimum, produce and maintain the design spacing, and hold the components in a strain-free manner. For a review of lens mounting and centering see Ref. 6.4.

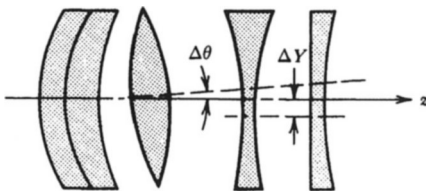


Figure 6.26. Illustration of tilt ($\Delta\theta$) and decentration (ΔY).

6.5. IMAGE EVALUATION

Although the Seidel aberrations give the designer an idea of the performance of the system, the overall effects of these aberrations on the final image is hard to visualize. By examining the effect that an optical system has on point objects, line objects, and periodic objects (gratings), a better evaluation of system performance can be made.

6.5.1. The Point Spread Function

A lens without spherical aberration will image parallel rays to a single spot on the axis. It is imaging a point object located on the axis at infinity to a point image. Ray tracing does not take into account diffraction by the edges of the lens, which will produce a diffraction pattern of the point source (Section 1.7.3). In most instances the aberrations account for a larger spot than does diffraction. For example, when spherical aberration is present in a lens, the size of the spot in the paraxial focal plane would extend to the point where the marginal rays hit the plane. For the case of Example 6.A, the radius of the blur caused by spherical aberration is 0.335 mm. However, if you examine Fig. 6.8 you will note that there is a point between the paraxial and marginal foci where the rays form a smaller blur spot. This point of best focus occurs at 75% of the longitudinal spherical aberration distance (about 1.41 mm in Example 6.A) and has a spot radius equal to one fourth of the transverse spherical aberration (0.084 mm for Example 6.A) (Ref. 6.2).

This method does not indicate how the light is distributed in the spot, but only where the edges of the spot are. To get information on the light distribution a spot diagram (Section 6.1) must be generated. If a large enough number of rays are traced, the center of the spot (x_0, y_0) can be found by averaging the x and y coordinates of all the rays traced and the size of the spot can be found by calculating the root mean square average for all rays intersecting a particular plane,

$$\bar{x} = \sqrt{\sum (x_j - x_0)^2},$$

where the sum is over all rays traced. The rms radius is then $\bar{r} = \sqrt{\bar{x}^2 + \bar{y}^2}$. This *rms spot size* is a measure of system performance. As one tries to improve a system by reducing the third order aberration coefficients the spot size will also be reduced. However, as the minimum values are approached, it may be found that the rms spot size at the best focus for a system with small third-order aberration coefficients is not as good as a system with larger third-order aberrations. This is because as the correction of the system is improved, the higher order aberration terms become important. Which, then, is a better measure of the performance of the system? Since the spot diagram is generated from exact ray traces and the rms spot size is generated from those traces, the rms spot size is a better measure of the system performance.

Although a single number is an easy method of determining the blur spot dimensions, it provides no information on how the light is distributed in the system. However, a spot diagram generated with a large number of rays will give a better idea of how the light is distributed through the spot. If it is assumed, as in the case of the encircled energy plots, that each ray carries the same fraction of the total energy from the object point, a distribution of the energy in the blur spot can be determined and plotted as a three dimensional plot of energy versus coordinates in the plane where the spot is evaluated (Fig. 6.27). This function, which represents the spread in light from a point object caused by an imperfect imaging system, is called the *point spread function*. A measure of the blur spot would be the diameter of the circle when the amount of energy is 50% of that at the center of the spot. This can be very useful for applications where small separated point objects (e.g., stars in astronomy) are present.

An experimental test of optical systems is the *star test*. Although the object could be a star, in most cases it is simply a small source of light. To test systems with infinite conjugates the source is located at the focal point of a well corrected lens to collimate the light going into the system. It is possible with experience to interpret the on-axis and off-axis image patterns formed by the optical system and to determine something about the aberrations present [Ref. 6.1].

6.5.2. The Line Spread Function

The point spread function is a two dimensional distribution, one that contains more information than can be handled easily. In place of the point spread

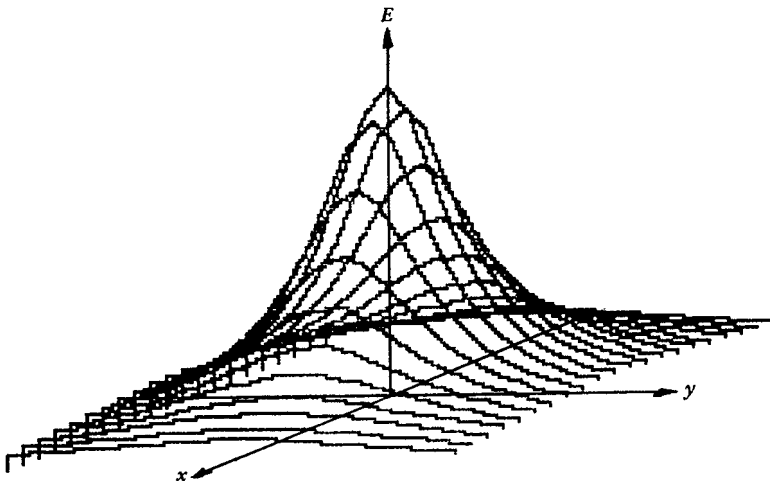


Figure 6.27. Point spread function. A three-dimensional plot of energy in the image of a point source as a function of image plane coordinates.

function, one can use a line object and integrate the point spread function along that line. This gives rise to a broadened line image whose cross-section perpendicular to the length of the line is a one-dimensional function, called, appropriately, the *line spread function*. If there is astigmatism present in a system it is necessary to determine two line spread functions for off-axis rays, one for a line in the tangential plane and another for a line in the sagittal plane.

Suppose instead of a point or a line, the object was a set of three black bars separated by white bars, all with the same width, as shown in Fig. 6.28. What would the image of such an object look like? Consider the white bars as made up of many thin white lines. Each white line will be spread according to the line spread function. Near the edges of the white bars, some of the image will be spread over into the region where the black bar is imaged. The effect will be a "rounding off" the edges of the image and that will produce three dark grey and two light grey bars.

There is an experimental test that corresponds to this arrangement also. A group of alternating black and white bars with progressively finer spacing (the spacing usually decreases as $\sqrt[6]{2}$ so that every sixth set of bars the spacing has been reduced by a factor of 2) is printed as a chart somewhat like the chart used by an eye doctor. A black bar and white bar constitutes one *line pair*. A group is labeled by the number of line pairs per unit length (inches or millimeters). The method is much the same. Examining the image transmitted by the optical system under test either with the unaided eye or a microscope, the resolution is determined from the group of bars with the smallest spacing that can just be resolved. As a simple test of an optical system the resolution chart method is useful. Its use depends, however, on the subjective judgement of the tester. There is a way to convert this test from a subjective test to an objective one, a test that provides additional information about the optical system.

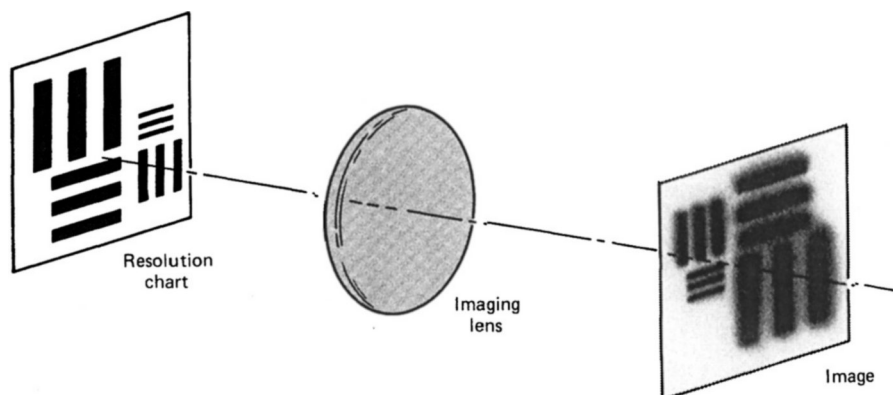


Figure 6.28. Resolution of a lens as measured by imaging a series of sets of black and white bars of equal width.

6.5.3. The Modulation Transfer Function

The quality of an optical system imaging a multiple-bar object can be determined from the image contrast. If I_{\max} is the highest irradiance reading in the bar image and I_{\min} is the lowest, the *contrast* of the image C is defined by

$$C \equiv \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \times 100\%. \quad (6.53)$$

If the system images a black bar as truly black ($I_{\min} = 0$) and a white bar as white ($I_{\max} = 1$) then C will be 100%. As the bars get closer together, they begin to fill in and the contrast drops (Fig. 6.29).

The easiest way to express the variation of bar separations is in terms of *spatial frequency*. A spatial frequency represents in optics much the same concept that the temporal frequency of a monochromatic wave does in wave theory. It is the number of cycles per unit distance. In a way, a spatial frequency is similar to the number of line pairs (one black bar and one white bar) per millimeter on the resolution chart, but a pattern having a single spatial frequency has only a sinusoidal variation. In everyday life, brick walls, railroad ties, and tiled floors are examples of objects with strong, easily perceived spatial frequencies.

A low spatial frequency object, with frequency f_1 , will be imaged by the optical system with a small reduction in contrast, as shown in Fig. 6.30. The same optical system will image an object whose spatial frequency is twice as great, $f_2 = 2f_1$, with considerably smaller contrast. By plotting the image contrast as a function of spatial frequency, the response of an optical system can be determined. For example, our lens might have a response like that shown in Fig. 6.30. We have plotted our two sample points on the curve for lens A. This curve, called the *modulation transfer function*, provides the optical engineer or designer with an objective measure of the performance of an optical system in a manner similar to the audio frequency response curve for a high-fidelity music system. It tells the designer the maximum spatial frequencies the system can

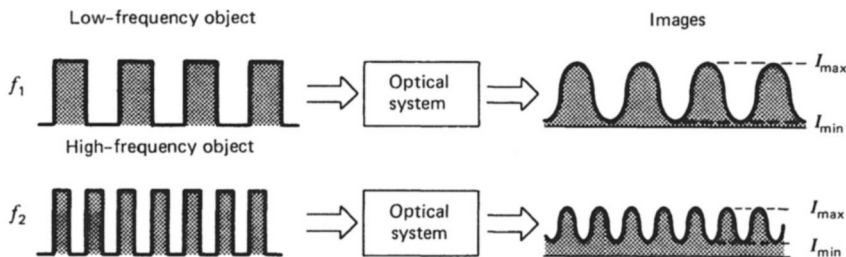


Figure 6.29. Image contrast as a function of the spatial frequency of the object.

resolve, just as its audio counterpart tells the high fidelity buyer the range of high and low frequencies that the sound system will reproduce. The modulation transfer function takes into account all aberrations and the diffraction limits of the lens. Although the high-frequency cutoff may be one criterion for excellent performance, it is not the only one. In some cases, the eye prefers images from systems in which the contrast remains higher at low frequencies at the expense of a high frequency cutoff. In Fig. 6.30 lens B might be preferred to lens A, although the latter has the better high-frequency response.

The modulation transfer function need not be determined just in the image plane. As the plane is defocused, the contrast drops rapidly for high frequencies and more slowly for low frequencies. One of the methods of determining the quality of a lens is to measure the contrast at a particular spatial frequency and then follow the contrast as a function of distance either side of the image plane. This is called "through-focus transfer function". It is a measure of the "depth of focus" of the lens.

There are experimental techniques for determining the modulation transfer function of a system. Some techniques image a black and white grating into an analysis plane and determine how well the grating is imaged onto a slit opening, to produce a measurement of the contrast. A somewhat simpler optical arrangement relies on a measurement of the line spread function. A knife edge slowly cuts the image of a slit. The amount of light passing the slit is detected and stored as a function of knife-edge motion. The distance derivative of this function is the line spread function, and its Fourier Transform is the modulation transfer function. Many firms that use the modulation transfer function for testing large numbers of lenses employ a simpler technique that measures the transfer function at a specific spatial frequency. If the contrast for that part is above a specific level it is accepted, below it, it is rejected.

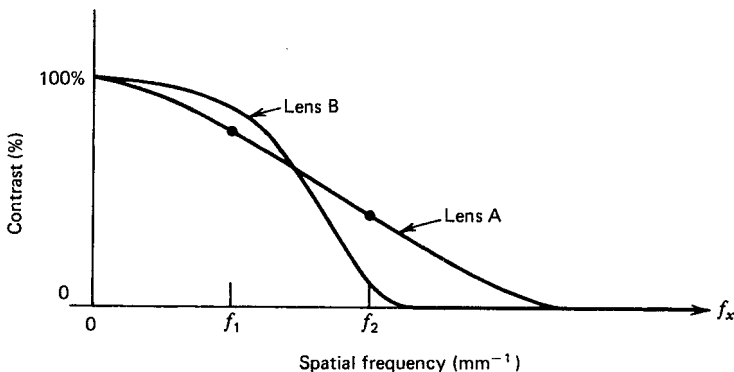


Figure 6.30. Modulation transfer function for two lenses. The frequencies f_1 and f_2 represent the two situations illustrated in Fig. 6.29.

PROBLEMS

- 6.1. Derive Eq. 6.2 using Fig. 6.3. Show it also true for the primed quantities in the other side of the surface.
- 6.2. Derive Eq. 6.3 using Fig. 6.1.
- 6.3. Derive Eq. 6.7. (You will need the results of Problem 6.1.)
- 6.4. Calculate the longitudinal spherical aberration for the lens in Example 6.A if the object were placed 114 units from the first surface. (a) Use an exact ray trace, (b) Use the Seidel coefficients. $Q_1 = 114 \sin 0.1$.
- 6.5. Draw a ray intercept plot for a system with overcorrected spherical aberration.
- 6.6. Show that $C_k = F_k^2/B_k$.
- 6.7. Derive Eq. 6.37 and show that the sagittal distance Δ , the distance between the arc of a circle of radius ρ and its corresponding chord, $2h'$, is given by $\Delta = h'^2/2\rho$, if $h' \ll \rho$.
- 6.8. Using Eq. 6.39 show that the Petzval radius for a thin lens is $-n \div \phi$, where n is the refractive index of the lens.
- 6.9. Calculate the longitudinal spherical aberration for the lens in Examples 6.A and 6.B with the surfaces reversed. ($R_1 = 120$, $R_2 = -40$.) (a) Use an exact ray trace, (b) Use the Seidel coefficients.
- 6.10. For parallel light, a plano-convex lens with the curved surface facing the object is a good approximation of a "best form" ($R_1 = -5R_2$) lens. Calculate the fractional increase in LSA in the plano-convex lens with the curved surface facing the image over the "best form" lens for $n = 1.5$.
- 6.11. Calculate the curvatures needed to correct the lens in Example 5.E for both axial color and spherical aberration using the thin lens technique illustrated in Sections 5.3.2 and a series of bendings on the two components. (This is not a trivial problem. Use a calculator or short computer program.)
- 6.12. Determine the effect that a change of refractive index from 1.517 to 1.626 has on the Seidel aberrations for the lens in Example 6.A.
- 6.13. Show that the axial intercept distance of a ray before refraction for an aplanatic surface is $L = R_k(n_k + n_{k-1})/n_{k-1}$.
- 6.14. Derive the relationship for glasses (Eq. 6.49) that must provide both achromatization and a flat field for a cemented doublet.
- 6.15. Show that the optical power of a thick spherical shell is proportional to its thickness.

- 6.16. If a computer program is available that can handle conic surfaces, determine the amount of conic correction for a corrector plate to a 30-cm *radius* spherical mirror at $f/1$.
- 6.17–6.21. The balance of the problems in this chapter consists of exact ray traces of Problems 5.8–5.12. They should be done with the aid of a calculator or computer. To the extent that the programming contains aberration information the aberrations should be determined. Further, to the extent that the program has the ability to compute a merit function, variations of the lens parameters should be carried out to reduce aberrations in the system. While it is time consuming, it is more instructive to carry out changes in the system in a systematic manner, manually changing the system parameters, than it is to iterate the changes automatically. (Note that closing down the aperture stop is not a valid means of reducing aberrations!)

REFERENCES

- 6.1. *Military Handbook 141, Optical Design* (U.S. Department of Defense, Washington D.C., 1962). Chapters 5 through 10 cover the fundamentals of ray tracing, aberration analysis, and methods of lens design. Numerous examples can be used to check ray-trace programs.* (Available from the Naval Publications and Forms Center, Tabor Avenue, Philadelphia, Pa. 19120.)
- 6.2. W. J. Smith, *Modern Optical Engineering* (McGraw-Hill, New York, 1966). Chapter 10 covers optical computations similar to those in this chapter.**
- 6.3. R. Kingslake, *Lens Design Fundamentals* (Academic Press, New York, 1978). Contains a number of examples of lens design examples.**
- 6.4. *Applied Optics and Optical Engineering*, Vol. VIII, edited by R. Shannon and J. C. Wyant (Academic Press, New York, 1980). Chapter 2, “Lens Centering and Mounting,” is by R. Hopkins and Chapter 3 on Aspherics is by R. Shannon.

* The sign convention is the same as that in the text.

** The sign convention is the “old” convention (i.e., clockwise rotation of a ray into an axis is negative).

7

Gaussian Beams

(MAKING A LASER BEAM DO WHAT YOU WANT IT TO)

The output of a laser is very different than most other light sources. Those sources that have been used in the examples up to this point in the text have been incandescent or fluorescent lights, or in the case of Example 3.C, a light emitting diode. The description of collecting light from such sources and sending it through an optical system was discussed in Chapter 3, Radiometry. Because lasers are used in many optical systems, a chapter is devoted to a description of the propagation of laser beams and their modification by optical systems.

The design of systems incorporating a laser beam should be looked at more as a problem akin to radiometry than as a problem in imaging. Unless the beam is interrupted by a slit, aperture, or transparency, there is no image. The ray optics presented in the first six chapters is insufficient to explain the behavior of laser beams, since the effects of diffraction must be included right from the beginning. Because of this, intuitions based on geometrical optics arguments can be misleading.

After a description of the simplest type of beam, the TEM_{00} mode Gaussian beam and its parameters, we look at means of modifying the beam using equivalent thin lens and paraxial equations for Gaussian beams. The concept of conjugate distances will be introduced and the effect of higher-order modes on the calculations will be described. Finally, the variation in beam irradiance throughout the beam will be described and some methods for reshaping the distribution to give useful non-Gaussian shapes will be given.

7.1. CHARACTERISTICS OF A GAUSSIAN BEAM

The term *Gaussian* describes the variation in the irradiance along a line perpendicular to the direction of propagation and through the center of the beam, as shown in Fig. 7.1. The irradiance is symmetric about the beam axis and varies radially outward from this axis with the form

$$I(r) = I_0 e^{-2r^2/r_1^2} \quad (7.1)$$

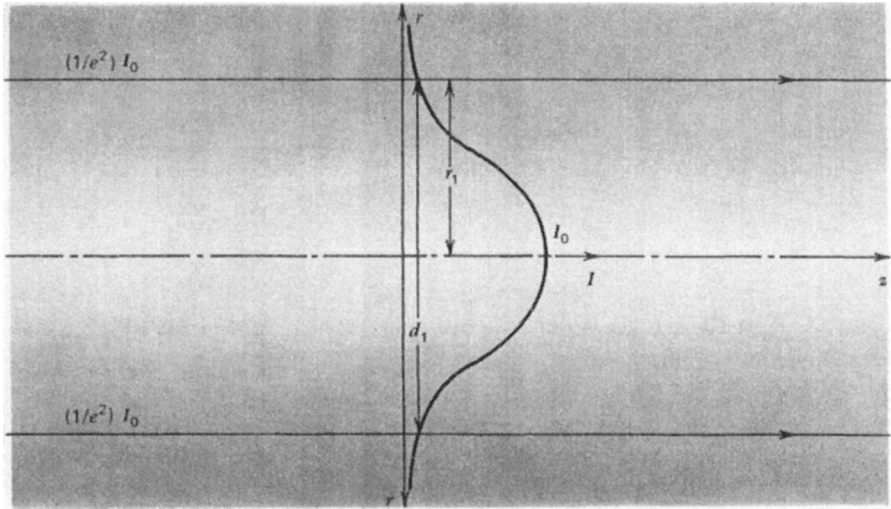


Figure 7.1. Gaussian beam profile. Plot of irradiance versus radial distance from the beam axis.

or in terms of a beam diameter

$$I(d) = I_0 e^{-2d^2/d_1^2}, \quad (7.2)$$

where r_1 and d_1 are the quantities that define the radial extent of the beam. These values are, by definition, the radius and diameter of the beam where the irradiance is $1/e^2$ of the value on the beam axis, I_0 .

7.1.1. Beam Waist and Beam Divergence

Figure 7.1 shows a beam of parallel rays. In reality, a Gaussian beam either diverges from a region where the beam is smallest, called the *beam waist*, or converges to one, as shown in Fig. 7.2. The amount of divergence or convergence is measured by the *full angle beam divergence* θ , which is the angle subtended by the $1/e^2$ diameter points for distances far from the beam waist as shown in Fig. 7.2. In some laser texts and articles, the angle is measured from the beam axis to the $1/e^2$ asymptote, a *half angle*. However, it is the full angle divergence, as defined here, that is usually given in the specification sheets for most lasers. Because of symmetry on either side of the beam waist, the convergence angle is equal to the divergence angle. We will refer to the latter in both cases.

Under the laws of geometrical optics a Gaussian beam converging at an angle of θ should collapse to a point. Because of diffraction, this, of course, does not occur. However, at the intersection of the asymptotes that define θ , the

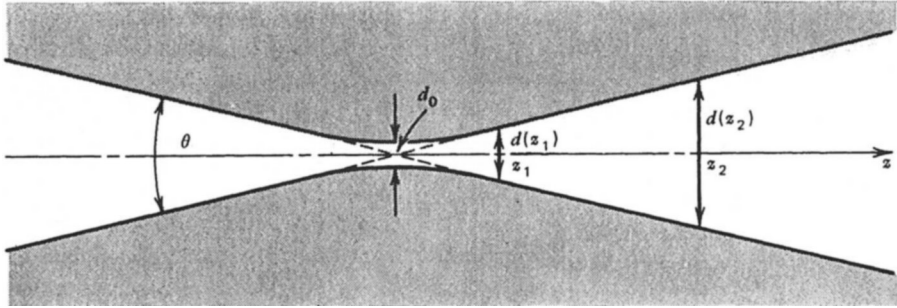


Figure 7.2. Variation of Gaussian beam diameter in the vicinity of the beam waist. The size of the beam at its smallest point is d_0 ; the full angle beam divergence, defined by the asymptotes for the $1/e^2$ points at a large distance from the waist is θ .

beam does reach a minimum value d_0 , the *beam waist diameter*. It can be shown (Ref. 7.1) that for a TEM_{00} mode d_0 depends on the beam divergence angle as:

$$d_0 = \frac{4\lambda}{\pi\theta}, \quad (7.3)$$

where λ is the wavelength of the radiation. Note that for a Gaussian beam of a particular wavelength, the product $d_0\theta$ is a constant. This means that for a very small beam waist the divergence must be large, for a highly collimated beam (small θ), the beam waist must be large.

The variation of the beam diameter in the vicinity of the beam waist is shown in Fig. 7.2 and given as

$$d^2 = d_0^2 + \theta^2 z^2, \quad (7.4)$$

where d is the diameter at a distance $\pm z$ from the waist along the beam axis. Note that far from the beam waist ($\theta z \gg d_0$), $d = \theta z$. Thus, at points z_1 and z_2 satisfying $z_1, z_2 \gg d_0/\theta$, $d_1 = \theta z_1$, and $d_2 = \theta z_2$. Subtracting yields $d_2 - d_1 = \theta(z_2 - z_1)$, and if we set $z_2 - z_1 = t$, an axial distance, we can write the equation as

$$d_2 = d_1 + \theta t. \quad (7.5)$$

This is similar to the transfer equation in geometrical optics. Indeed, well away from the beam waist one can use geometrical optics.

7.1.2. The Rayleigh Range

It is useful to characterize the extent of the beam waist region with a parameter called the *Rayleigh range*. (In other descriptions of Gaussian beams this extent is

sometimes characterized by the *confocal beam parameter* and is equal to twice the Rayleigh range.) Rewriting Eq. 7.4 as

$$d = d_0 \sqrt{1 + \left(\frac{\theta z}{d_0}\right)^2}. \quad (7.6)$$

we define the Rayleigh range as the distance from the beam waist where the diameter has increased to $\sqrt{2} d_0$. Obviously this occurs when the second term under the radical is unity, that is, when

$$z = z_R \equiv \frac{d_0}{\theta}.$$

Although the definition of z_R might seem rather arbitrary, this particular choice offers more than just convenience. Figure 7.3 shows a plot of the radius of curvature of the wavefronts in a Gaussian beam as a function of z . For large distances from the beam waist the wavefronts are nearly planar, giving values tending toward infinity. At the beam waist the wavefronts are also planar, and, therefore, the absolute value of the radius of curvature of the wavefronts must

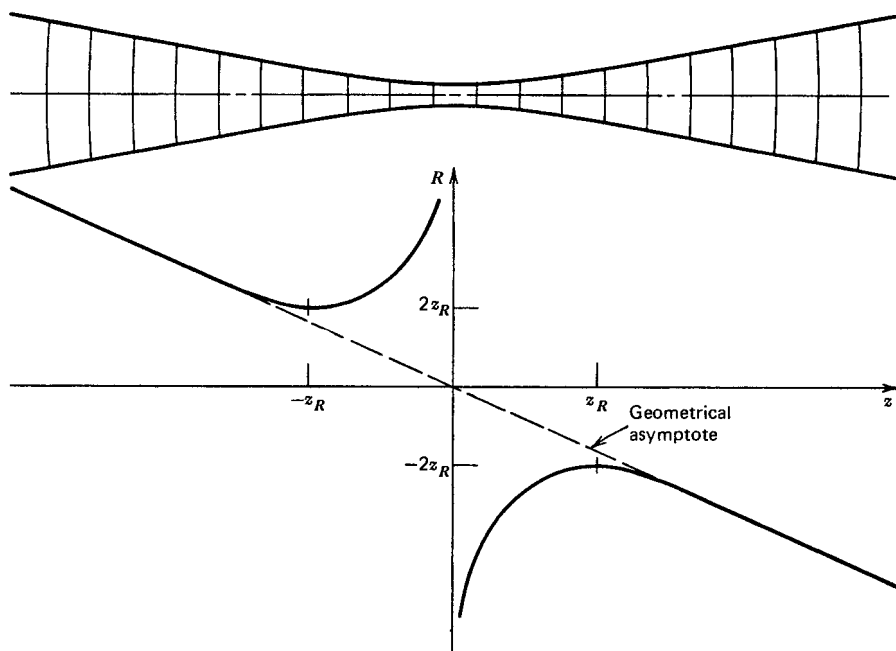


Figure 7.3. Plot of radius of curvature (R) versus distance (z) from the beam waist. The absolute value of the radius is a minimum at the Rayleigh range point, z_R . In the limit of geometrical optics, the radius of curvature of the wavefronts follows the dashed line.

go from infinity at large distances through a minimum and return to infinity at the beam waist. This is also true on the other side of the beam waist but with the opposite sign. It can be shown that the minimum in the absolute value of the radius of curvature occurs at $z = \pm z_R$, that is, at a distance one Rayleigh range either side of the beam waist. From Fig. 7.3, the “collimated” region of Gaussian beam waist can be taken as $2z_R$.

The Rayleigh range can be expressed in a number of ways:

$$z_R = \frac{d_0}{\theta} = \frac{4\lambda}{\pi\theta^2} = \frac{\pi d_0^2}{4\lambda}. \quad (7.7)$$

From this we see that all three characteristics of a Gaussian beam are dependent on each other. Given any of the three quantities, d_0 , θ , z_R , and the wavelength of the radiation, the behavior of the beam is completely described [Problem 7.1]. For example, if a helium–neon laser ($\lambda = 633 \text{ nm}$) has a specified TEM_{00} beam diameter of 1 mm, then $\theta = 4\lambda/\pi d_0 = (1.27 \times 6.33 \times 10^{-7} \text{ m})/(1 \times 10^{-3} \text{ m}) = 0.8 \text{ mrad}$ and $z_R = d_0/\theta = (1 \times 10^{-3} \text{ m})/(0.8 \times 10^{-3} \text{ rad}) = 1.25 \text{ m}$. The Rayleigh range of a typical helium–neon laser is considerable.

The initial Rayleigh range (and, therefore, beam waist and divergence) is determined by the radii of curvatures of the laser mirrors R_1 and R_2 and by their separation L . If the same sign convention used earlier in the text is retained, the Rayleigh range of a laser is given by

$$z_R^2 = \frac{L(R_2 + L)(R_1 - L)(R_2 - R_1 + L)}{(R_2 - R_1 + 2L)^2} \quad (7.8)$$

(The expression is the same as in Ref. 7.2, except that the sign convention for the radii of curvature is the opposite of standard ray-tracing conventions.) It should be noted that there are some geometries for which the value of z_R^2 would be negative. These mirror arrangements are referred to as *unstable resonators*. If we define $g_1 \equiv (1 - L/R_1)$ and $g_2 \equiv (1 + L/R_2)$, it can be shown that [Problem 7.2] the condition for stable resonators based upon the requirement that z_R^2 be positive is that $0 < g_1 g_2 < 1$. This is known as the *stability condition* for laser resonators (Ref. 7.3).

7.2. MODIFYING A GAUSSIAN BEAM

There are many design situations in which the “raw” beam as it comes out of the laser cannot be used. The beam must be modified in some way—expanded, focused, recollimated. This modification can be performed by lenses, mirrors, prisms, or even slabs of glass. For our example, we will use a lens, but the focal length or surface power of any optical element can be substituted into the equations in the appropriate place.

The thin lens equation was expressed in Chapter 1 in both the Gaussian form

$$\frac{1}{t_1} - \frac{1}{t_0} = \frac{1}{f}. \quad (1.16)$$

where t_0 and t_1 are the object and image distances, respectively, and in the Newtonian form [Problem 1.13]

$$-xx' = f^2 \quad (7.9)$$

where x is the distance from the front focal point to the object and x' is the distance from the back focal point to the image. For Gaussian beams we can use the Newtonian form with some minor redefinitions and the addition of a single term to account for the effects of diffraction.

7.2.1. Thin Lens Equation For Gaussian Beams

In Fig. 7.4 the relevant distances are shown. Note that in contrast to the neatly subscripted variables used to trace a ray through a system, a series of unprimed and primed variables is used. This is done because the Gaussian beam is most easily described by locating the origin of the coordinate system at the beam waist. In contrast, the ray-trace coordinate origins are located at the vertices of each element or in the plane of interest. The beam waist always has a zero subscript. Other beam diameters have other subscripts, usually d_1 is to the left of d_0 and d_2 to the right. If there is more than one lens, each additional region between lenses is given another prime. Also unlike conventional ray tracing, beam waist distances relative to the lens position are positive if on the input side of the lens the beam narrows down to a waist before striking the lens, and negative if the beam is intercepted by the lens before the beam waist occurs.

If the positions of the input beam waist and output beam waist are replaced by the object and image distances in Eq. 7.9, the thin lens equation in Newtonian form becomes ($-x \rightarrow z_2 - f$; $x' \rightarrow z'_1 - f$)

$$(z_2 - f)(z'_1 - f) = f^2 - f_0^2, \quad (7.10)$$

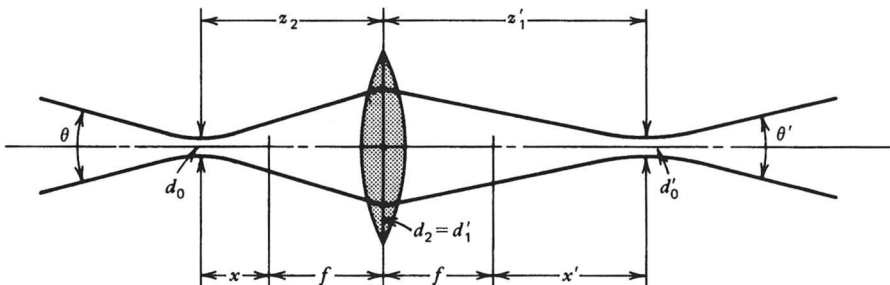


Figure 7.4. Transformation of a Gaussian beam by a thin lens.

where f_0^2 is the term introduced to account for diffraction. There are a number of ways of writing this term since it is an expression of the beam properties on both sides of the lens. The most symmetric form is

$$f_0^2 = z_R z'_R, \quad (7.11)$$

but other useful forms are

$$\begin{aligned} f_0 &= \frac{d'_0}{\theta} \\ &= \frac{d_0}{\theta'} \\ &= \frac{\pi d_0 d'_0}{4\lambda} \\ &= \frac{4\lambda}{\pi \theta' \theta} \end{aligned} \quad (7.12)$$

Another useful relation between the beam waist diameters and their locations, derived in Ref. 7.1, is

$$\frac{(z_2 - f)}{(z'_1 - f)} = \frac{d_0^2}{d'^2_0}. \quad (7.13)$$

Using Eqs. 7.10 and 7.13, a set of simple relations can be derived between the positions of the beam waists and the beam parameters. Solving for $(z'_1 - f)$ in Eq. 7.10, inserting it into Eq. 7.13 and rearranging terms one has

$$\frac{(z_2 - f)^2}{f^2 - d'^2_0/\theta^2} = \frac{d_0^2}{d'^2_0}.$$

Multiplying by the left-side denominator,

$$(z_2 - f)^2 = f^2 \frac{d_0^2}{d'^2_0} - \frac{d_0^2}{\theta^2} = f^2 \frac{d_0^2}{d'^2_0} - z_R^2,$$

and then by d'^2_0

$$d'^2_0 [(z_2 - f)^2 + z_R^2] = f^2 d_0^2,$$

One obtains the new beam waist d'_0 in terms of the old beam waist d_0 :

$$d'^2_0 = \frac{f^2 d_0^2}{(z_2 - f)^2 + z_R^2} = \alpha^2 d_0^2, \quad (7.14)$$

The parameter α , given by

$$\alpha = \frac{|f|}{\sqrt{(z_2 - f)^2 + z_R^2}}, \quad (7.15)$$

is a function of axial quantities related to the input beam (z_R), the lens (f) and the beam-lens position ($z_2 - f$). Once α is evaluated for the arrangement at hand, the output parameters are quickly determined since

$$d'_0 = \alpha d_0, \quad (7.16)$$

$$\theta' = \frac{4\lambda}{\pi d'_0} = \frac{4\lambda}{\pi \alpha d_0} = \frac{\theta}{\alpha}, \quad (7.17)$$

$$z'_R = \frac{d'_0}{\theta'} = \frac{\alpha d_0}{\theta/\alpha} = \alpha^2 z_R. \quad (7.18)$$

Using Eqs. 7.13 and 7.16, the position of the new beam waist z'_1 can be expressed as

$$\begin{aligned} (z'_1 - f) &= \frac{d'^2_0}{d_0^2} (z_2 - f) = \alpha^2 (z_2 - f), \\ z'_1 &= f + \alpha^2 (z_2 - f) = f + \frac{f^2 (z_2 - f)}{(z_2 - f)^2 + z_R^2}. \end{aligned} \quad (7.19)$$

In the thin lens approximation, all of these equations are quite general. Now we will look at some approximations.

7.2.2. Focusing of a Collimated Beam to a Small Spot

Since α is directly proportional to f , a small spot requires a short focal length lens. Since the beam is collimated z_2 and z_R are large compared to f , $\alpha = f/\sqrt{z_2^2 + z_R^2}$. There are two limiting cases:

1. The lens is well inside z_R ($z_2 \ll z_R$). Then $\alpha = f/z_R = f\theta/d_0$. This gives

$$d'_0 = \alpha d_0 = f\theta, \quad (7.20)$$

a simple equation for finding d'_0 . Also

$$\theta' = \frac{\theta}{\alpha} = \frac{\theta d_0}{f\theta} = \frac{d_0}{f}, \quad (7.21)$$

$$z'_R = \alpha^2 z_R = \frac{f^2}{z_R^2} \cdot z_R = \frac{f^2}{z_R}, \quad (7.22)$$

$$z'_1 = f + \alpha^2(z_2 - f) = f + \frac{z_2 \cdot f^2}{z_R^2} \cong f. \quad (7.23)$$

2. The lens is well outside z_R ($z_2 \gg z_R$; beware, the lens may be overfilled if it is too far away). Then $\alpha = f/z_2$ and

$$d'_0 = \frac{f d_0}{z_2}, \quad (7.24)$$

$$\theta'_0 = \frac{z_2 \theta}{f} \quad (7.25)$$

(note $z_2 \theta = d$ at the lens),

$$z'_R = \alpha^2 z_R = \frac{f^2 z_R}{z_2^2} \quad (7.26)$$

$$z'_1 = f + \frac{(z_2 - f) \cdot f^2}{z_2^2} = f + \frac{f^2}{z_2} \cong f. \quad (7.27)$$

In Example 7.A three situations are illustrated. Note that the placement of the lens relative to the size of the Rayleigh range has a dramatic effect on the spot size.

It should be understood that we do not always know where the beam waist is in a laser. In many cases the beam waist is at or near the output mirror, and for an initial attempt to design a system this is a reasonable initial assumption. However, later in the design, the sensitivity of the design to the initial beam waist location should be checked. If the sensitivity is high, a more precise calculation can be made. Given the radius of curvature of the two laser mirrors R_1 and R_2 and their separation L , the distance between the beam waist and the first mirror is

$$z_1 = \frac{L(R_2 + L)}{R_1 - R_2 + 2L}. \quad (7.28)$$

EXAMPLE 7.A. A helium–neon laser beam with a divergence of 0.8 mrad is focused with a 100-mm focal length lens. What are the beam diameters in the cases

- (a) $z_2 \ll z_R$
- (b) $z_2 = z_R$
- (c) $z_2 \gg z_R$?

Solution: Since $\theta = 0.8 \times 10^{-3}$ rad, $d_0 = \frac{4\lambda}{\pi\theta} = 1$ mm. Therefore $z_R = d_0/\theta = 1 \times 10^{-3}/0.8 \times 10^{-3} = 1.25$ m.

- (a) $z_2 \ll z_R$, $d'_0 = f \cdot \theta = 10^{-1} \text{ m} \times 0.8 \times 10^{-3} = 80 \mu\text{m}$.
- (b) $z_2 = z_R$. The complete expression must be used.

$$d'_0 = \frac{fd_0}{\sqrt{(z_2 - f)^2 + z_R^2}} = \frac{10^{-1} \cdot 10^{-3}}{\sqrt{(1.25 - 0.1)^2 + (1.25)^2}} = 60 \mu\text{m}.$$

- (c) $z_2 \gg z_R$. Let us assume $z_2 = 12.5$ m. The beam diameter there is about 10 mm

$$d'_0 = \frac{fd_0}{z_2} = \frac{0.1}{12.5} \times 1 \times 10^{-3} \text{ m} = 8 \mu\text{m}.$$

Note the distance between the beam waist and the lens is extremely large.

The distance between the beam waist and the second mirror is $z_2 = L + z_1$ ($z_1 < 0$) [Problem 7.3]. If the outside surface of the output mirror is not flat, then the effect of the optical power of this surface must also be calculated using the paraxial method given in Section 7.2.6.

7.2.3. Refocusing a Beam Waist to a New Beam Waist

This is a case of magnification or demagnification of the beam waist. Since $d'_0 = \alpha d_0$, α is the magnification constant for the waist. At the same time the beam divergence will be demagnified by θ/α , and the Rayleigh range will be multiplied by α^2 . (This parallels the relation between the transverse magnification and the longitudinal magnification for imaging: $M_l = M_t^2$.)

Once the value of α has been decided on, there are still two degrees of freedom in the system: the focal length of the lens (f) and location of the lens with respect to the beam waist (z_2). While there is one independent input parameter and one independent output parameter, there is an infinite combination of the focusing geometries that will provide the required beam waist. There

are some restrictions, however, on the choice of the focal length. Assuming the value of α has been chosen, then we can solve (7.15) for z_2 :

$$z_2 = f \pm \sqrt{\frac{f^2}{\alpha^2} - z_R^2} = f \pm \frac{1}{\alpha} \sqrt{f^2 - f_0^2} \quad (7.29)$$

Thus, f may be chosen to be any focal length practical (or handy) subject to the limitation that $f^2 > f_0^2 = z_R z'_R$ [Problems 7.4–7.6]. In Example 7.B, the use of Eq. 7.29 is illustrated by a practical problem.

EXAMPLE 7.B. Reduce the beam waist of a helium–neon laser ($d_0 = 1$ mm; $\theta = 0.8$ mrad; $z_R = 1.25$ m) to 184 μm . Assume waist is at the output mirror.

Solution:

$$\frac{d'_0}{d_0} = \frac{184 \mu\text{m}}{1 \text{ mm}} = 0.184 = \alpha,$$

$$z_R = 1.25 \text{ m} \quad z'_R = \alpha^2 z_R = 42 \text{ mm},$$

$$f_0^2 = z_R z'_R = 1250 \times 42 \text{ mm}^2,$$

$$f_0 = 230 \text{ mm} \quad f > 230 \text{ mm}.$$

(a) Choose $f = 300$ mm, then

$$z_2 = 300 + \frac{1}{0.184} \sqrt{(300)^2 - (230)^2} = 1347 \text{ mm}.$$

The positive root was chosen since the laser would get in the way for the negative root. This is an extremely long distance.

(b) Then choose $f = 235$ mm and

$$z_2 = 235 + \frac{1}{0.184} \sqrt{(235)^2 - (230)^2} = 497 \text{ mm},$$

$$z'_1 = 235 + (0.184)^2 (497 - 235) = 244 \text{ mm}.$$

Total distance from laser to focus: $z_2 + z'_1 = 741$ mm.

(c) For the shortest distance, choose $f = f_0 = 230$ mm. Then $z_2 = z'_1 = f$. Total distance: $z_2 + z'_1 = 2f = 460$ mm. Thus by choosing the correct lens a practical laser-to-new-focus distance can be provided.

A special case of the transfer of focus is that of relaying the waist without magnification ($\alpha = 1$). Starting with the definition of α (7.15),

$$\alpha^2 = \frac{f^2}{(z_2 - f)^2 + z_R^2} = 1,$$

and solving for z_2 , one obtains

$$z_2 = f \pm \sqrt{f^2 - z_R^2}, \quad (f \geq z_R) \quad (7.30)$$

or, in terms of f ,

$$f = \frac{z_2^2 + z_R^2}{2z_2}.$$

The distance between original beam waist and the lens is found to be equal to the distance between the lens and the new beam waist as would be expected for unit magnification [Problem 7.7].

$$z'_1 = f + \alpha^2(z_2 - f) = f + z_2 - f = z_2.$$

Thus the total distance between waists in terms of f and z_R is

$$z_{12} = z_2 + z'_1 = 2f \pm 2\sqrt{f^2 - z_R^2}. \quad (7.31)$$

In addition to this general case there are two special cases (Fig. 7.5)

1. $f \gg z_R$. Then $z_{12} = 4f$. This is similar to the relaying of objects and images in geometrical optics, where the unit magnification occurs at $t_1 = -t_0 = 2f$. Note for this to occur for any reasonable focal length, z_R must be $\ll f$ and therefore the beam must be strongly divergent.
2. $f = z_R$. Then $z_2 = z'_1 = f$ and $z_{12} = 2f$. This is obviously different than the previous case. The beam waists are separated by only $2f$. Since it is not always possible to match f to z_R , some prior beam conditioning is usually done. This second case is sometimes called a *Gaussian relay*.

At one time this type of relay was considered for information transmission over some distances. Since the advent of low loss optical fibers this is no longer considered to be a realistic method of transmitting laser radiation.

7.2.4. Collimating a Gaussian Beam

If one cannot achieve truly collimated light, $\theta = 0$ (since d_0 would have to be infinite in extent), then what should be the definition of collimated light? There

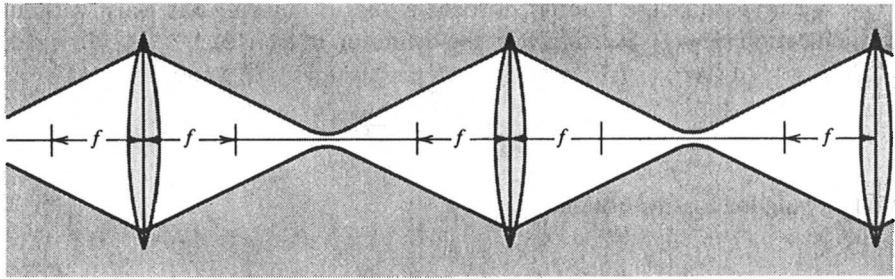
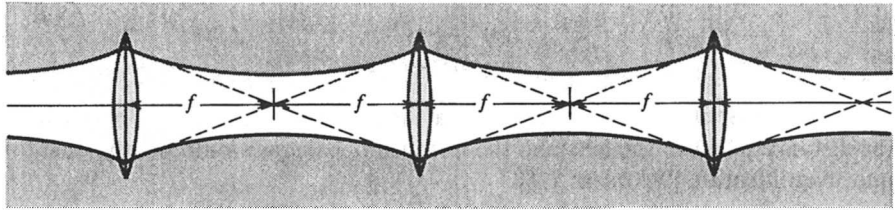
(a) $z_R \ll (z_2 - f)$ (b) $z_R = z_2 = f$

Figure 7.5. Gaussian beam relays. (a) Case 1. Highly divergent beams, similar to an image relay. (b) Case 2. Rayleigh range equal to the focal length.

are two possible definitions. One is that the divergence be as small as possible, the other is that the distance to the next beam waist be a maximum. In the first case, one can show [Problem 7.8] that z'_R is a maximum if $z_2 - f = 0$. (Thus, $\alpha^2 = f^2/z_R^2$ and $z'_R = f^2/z_R$.) In the second case z'_1 is a maximum when $z_2 = f + z_R$. That is, the beam is considered collimated when a positive focal length lens is positioned such that its focal point is one Rayleigh range beyond the beam waist. Inserting this condition into Eq. 7.15 yields

$$\alpha = \frac{|f|}{\sqrt{(z_2 - f)^2 + z_R^2}} = \frac{|f|}{\sqrt{2}z_R}, \quad (7.32)$$

and this result in Eq. 7.19 determines the location of the new beam waist:

$$z'_1 = f + \alpha^2(z_2 - f) = f + \frac{f^2}{2z_R^2}(z_R) = f + \frac{f^2}{2z_R}. \quad (7.33)$$

From Eqs. 7.32 and 7.33 we see that z'_1 and z'_R will be large if z_R is small. Thus, a long focal length lens accepting a highly diverged beam (small z_R) will give

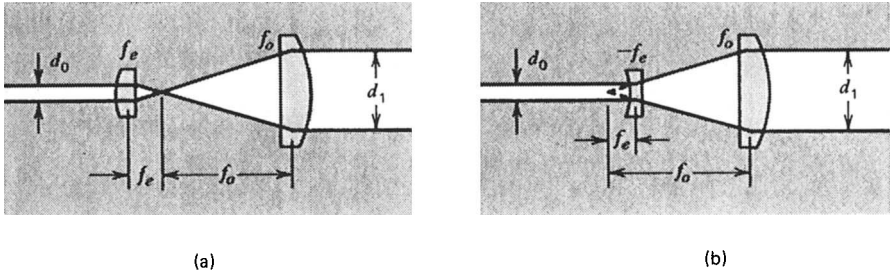


Figure 7.6. Gaussian beam collimation. (a) Keplerian telescope. (b) Galilean telescope. Eyepiece focal length, f_e ; objective focal length, f_o .

excellent collimation. If the initial beam is well collimated, however, then the beam must be focused with a short focal length lens before being recollimated with the long focal length collimation lens. This arrangement is nothing more than an inverted telescope (light goes in the eyepiece and out the objective (Fig. 7.6a)). It is also possible to reduce z_R with a diverging lens that forms a virtual beam waist in front of the negative lens. Together with the collimating lens this system is an inverted Galilean telescope (Fig. 7.6b).

We can use the information gained so far to calculate the increase in the collimation distance. The input beam is assumed to be well collimated. Therefore, the beam diameter at the eyepiece lens (focal length, f_e) is approximately equal to d_0 and the lens is well inside the Rayleigh range. This is the same as Case 1 of Section 7.2.2. so that

$$\alpha = \frac{f_e}{z_R},$$

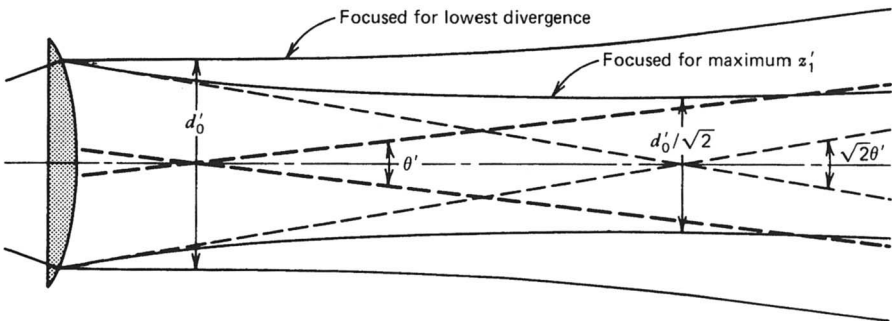


Figure 7.7. Comparison of two definitions of a collimated Gaussian beam. The bold lines show the asymptotes and $1/e^2$ points for a beam focused for smallest divergence, θ' . The fine lines indicate the asymptotes and $1/e^2$ points for a beam focused for maximum lens to beam waist distance. For a distance of about $2z_R$ from the lens this second definition maintains a smaller beam.

and

$$d'_0 = \alpha d_0 = f_e \theta,$$

$$\theta' = \frac{d'_0}{f_e},$$

$$z'_R = \frac{f_e^2}{z_R}, \quad \text{so that} \quad z'_R \ll z_R$$

$$z'_1 = f_e.$$

For maximum collimation we want $z'_2 = f_o + z'_R$, where f_o is the focal length of the telescope objective. It is not to be confused with f_0 , the diffraction term in Eq. 7.10. Using Eq. (7.32), $\alpha' = f_o/\sqrt{2}z'_R$. The output beam waist is then

$$d''_0 = \alpha' d'_0 = \alpha' \alpha d_0 = \alpha'' d_0$$

and the overall magnification is

$$\begin{aligned} \alpha'' &= \alpha' \alpha = \frac{f_e}{z_R} \times \frac{f_o}{\sqrt{2}z'_R} = \frac{f_e}{z_R} \cdot \frac{f_o}{\sqrt{2}(f_e^2/z_R)}, \\ \alpha'' &= \frac{f_o}{\sqrt{2}f_e} = \frac{M}{\sqrt{2}} \end{aligned} \quad (7.34)$$

where $M = f_o/f_e$ is the magnification of the telescope. The collimation has increased by a factor of $M/\sqrt{2}$, since the divergence has decreased to $\theta'' = \theta/\alpha''$ and the Rayleigh range has increased by

$$z''_R = \alpha''^2 z_R = \frac{M^2}{2} z_R. \quad (7.35)$$

Finally, it can be shown [Problem 7.9], that the beam waist is located at

$$z'_1 = f_o + \frac{M^2}{2} z_R = f_o + z''_R \quad (7.36)$$

In Example 7.C, the collimation of a helium–neon laser by a 40 power telescope is illustrated [Problem 7.10].

If instead of a maximum lens to beam-waist distance (z'_1), a minimum beam divergence is required, the above results are changed by replacing $\alpha'' = M/\sqrt{2}$ by $\alpha'' = M$ and $z'_1 = f_o + (M^2/2)z_R$ by $z'_1 = f_o$. Thus the beam waist is at the back

EXAMPLE 7.C. Collimate a helium–neon laser beam ($d_0 = 1$ mm, $\theta = 0.8$ mrad, $z_R = 1.25$ m) using a 40-power telescope ($f_o = 160$ mm, $f_e = 4$ mm (40 \times microscope objective)).

Solution: First,

$$\alpha'' = \frac{f_o}{\sqrt{2}f_e} = 28.28 \quad \text{and} \quad (\alpha'')^2 = \frac{40^2}{2} = 800.$$

The new beam characteristics are:

$$d_0'' = \alpha'' d_0 = 28.28 \text{ mm},$$

$$\theta'' = \theta / \alpha'' = 28.28 \text{ } \mu\text{rad},$$

$$z_R'' = (\alpha'')^2 z_R = 1 \text{ km (!)}.$$

Unless extreme care is taken with collimation optics, this diffraction limited performance cannot be achieved. Even with simple lenses, however, the collimation can be quite good.

If minimum divergence is required, then $\alpha'' = 40$ and $(\alpha'')^2 = 1600$. In this case $d_0'' = 40$ mm, $\theta'' = 20 \text{ } \mu\text{rad}$, and $z_R'' = 2$ km.

focal point of the telescope objective lens and the divergence is a factor of $\sqrt{2}$ smaller. In Fig. 7.7, the two cases are compared. As one can see from the figure, if collimation is needed close to the optical system, the beam waist can be smaller by a factor of $1/\sqrt{2}$ and collimated over a distance some $2z_R'$ from the lens. If collimation is needed in the far field (i.e., smallest beam divergence) then the focal point of the lens should be located at the original beam waist.

Of the two telescopes pictured in Fig. 7.6, the Keplerian type should be used if there is a need to spatially filter the beam (Ref. 7.10), a method for removing any spatial irregularities introduced by the optics before the collimator by putting a pinhole slightly larger than d_0' at the focus of f_e . For high power lasers focusing the beam to a small diameter could result in breakdown of the air by the high electrical field strength at the focus. In such cases, a Galilean type telescope should be employed (Fig. 7.6b).

7.2.5. Combinations of Lenses

It is possible using the above equations to calculate beam waist positions for a combination of lenses in a manner similar to the special case of beam-collimating telescopes. Instead of the ray-trace technique of the preceding

chapters, the approach is similar to the thin lens calculations used in introductory physics. The equations are applied once, an intermediate waist is found, then the equations are applied again (Example 7.D). In the example note that the final beam characteristics are the same as those in Example 7.B. In that example the lens was nearly 500 mm from the laser and the focal length of the lens had to be greater than 230 mm. By using two lenses, a similar result was achieved with shorter focal length lenses placed closer to the laser. In addition, the ability to move one lens relative to the other allows one to adjust the output beam parameters easily.

EXAMPLE 7.D. Using two lenses, obtain a “long” focus from a helium–neon laser. Assume beam waist at output mirror.

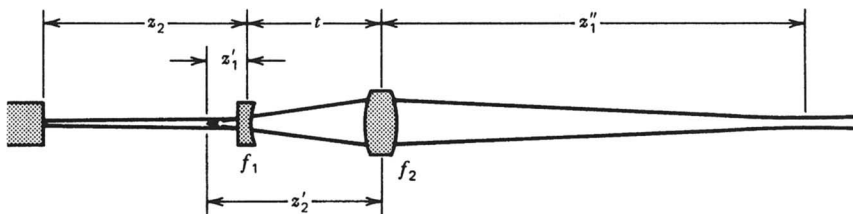
Helium–neon laser output:

$$d_0 = 1 \text{ mm}, \theta = 0.8 \text{ mrad}, z_R = 1.25 \text{ m}.$$

System dimensions:

$$z_2 = 135 \text{ mm}, \quad t = 58.42 \text{ mm},$$

$$f_1 = -50 \text{ mm}, \quad f_2 = 88.9 \text{ mm}.$$



Solution: For the first lens:

$$z_2 - f_1 = 185 \text{ mm}, \quad \alpha = \frac{|-50|}{\sqrt{(185)^2 + (1250)^2}} = 4 \times 10^{-2}$$

$$d'_0 = 4 \times 10^{-2} \times 1 \text{ mm} = 40 \text{ } \mu\text{m}; \quad \theta' = 20 \text{ mrad}; \quad z'_R = 2 \text{ mm}.$$

$$z'_1 = f_1 + \alpha^2(z_2 - f_1) = -50 + 16 \times 10^{-4}(185) = -50 + 0.3 = -49.7 \text{ mm}$$

z'_1 is negative (i.e., virtual), since the beam is diverging.

$$z'_2 = t - z'_1 = 58.42 - (-49.7) = 108.1 \text{ mm}.$$

For the second lens:

$$f_2 = 88.9 \quad \text{and} \quad z'_2 - f_2 = 108.1 - 88.9 = 19.2 \text{ mm.}$$

$$\alpha' = \frac{88.9}{\sqrt{(19.2)^2 + 2^2}} = 4.605.$$

$$d''_0 = \alpha' d'_0 = 4.605 \times 40 \text{ } \mu\text{m} = 184.2 \text{ } \mu\text{m}.$$

$$\theta'' = 20 \text{ mrad} / 4.605 = 4.34 \text{ mrad}.$$

$$z''_R = (\alpha')^2 z'_R = 42.4 \text{ mm}.$$

$$z''_1 = f_2 + (\alpha')^2 (z'_2 - f) = 88.9 + (4.605)^2 19.2 = 496 \text{ mm}.$$

Total distance:

$$\text{laser to beam waist} = 135 + 58.4 + 496 = 689 \text{ mm}.$$

The method given above is only one of several methods for calculating beam waist parameters. There are matrix based approaches (see Refs. 7.1 and 7.2) and chart based approaches (Ref. 7.4). In the case of periodic lens structures, such as a series of relay lenses, the matrix based approach may be more compact, whereas the chart-based approach may yield some insight as to poor choices of lenses and placements. The method given here is attractive because it is straightforward with certain useful approximations.

7.2.6. Paraxial Equations for Gaussian Beams

Following much the same mathematics used for the thin lens equations, it is possible to derive similar equations for the propagation of Gaussian beams at the interface of two media having optical power. Just as the paraxial equations for ray tracing bear a resemblance to the thin lens equations, so also are these paraxial equations similar to Eqs. 7.10 and 7.13. They are

$$(n' - z'_1 \phi)(n - z_2 \phi) = nn' - z_R z'_R \phi^2 \quad (7.37)$$

and

$$\frac{n - z_2 \phi}{n' - z'_1 \phi} = \frac{z_R}{z'_R}, \quad (7.38)$$

where $\phi = (n' - n)/r$ is the optical power of the interface as it is in paraxial optics (Section 4.1) and n' and n are the refractive indices of the refracting and

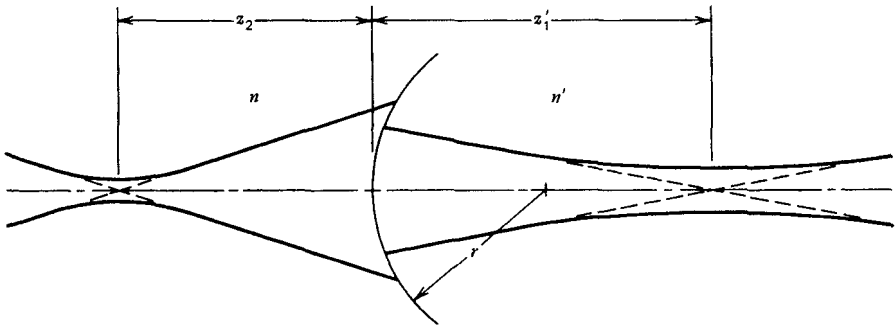


Figure 7.8. Paraxial ray trace of a Gaussian beam. The radius of curvature of the interface is r ; the refractive indices of the incident and refracting media are n and n' , respectively. The optical power of the surface is $\phi = (n' - n)/r$.

incident media, respectively as shown in Fig. 7.8. The distances z_2 and z'_1 are defined as before. Multiplying these equations by ϕ/nn' , one sees that $1/\phi$ replaces f and z/n in the respective media. Using these two equations, we may solve for $\alpha^2 = z'_R/z_R$ and find that

$$\alpha = \frac{\sqrt{nn'}}{\sqrt{(n - z_2\phi)^2 + z_R^2\phi^2}} \quad (7.39a)$$

and that

$$z'_1 = \frac{z_R z'_R \phi - z_2 n'}{n - z_2 \phi}. \quad (7.39b)$$

These are the paraxial counterparts to Eqs. 7.15 and 7.19. In Example 7.E, a Gaussian beam focused by a biconvex lens is analyzed two ways, a paraxial treatment and the thin lens method. The differences are small but significant [Problems 7.11–7.13].

For the special case of focusing a beam through a flat surface ($\phi = 0$), $\alpha = \sqrt{n'/n}$, where n is the incident medium refractive index and n' is that of the transmitting medium. If the incident medium is air ($n = 1$) the values of the beam parameters in the second medium are $d'_0 = \sqrt{n'}d_0$; $\theta' = \theta/\sqrt{n'}$; $z'_R = n'z_R$. The quantities preserved in all media are d_0/\sqrt{n} , $\sqrt{n}\theta$, and z_R/n , where d_0 , θ and z_R are the values for the beam in a vacuum.

If a Gaussian beam is focused inside of a material through a plane surface, where is the beam waist located? Setting $\phi = 0$ in Eq. 7.39b, $z'_1 = -n'z_2$. In terms of ray tracing the refraction of a ray of light along the asymptote at an angle of β with respect to the surface normal, as shown in Fig. 7.9, is refracted to

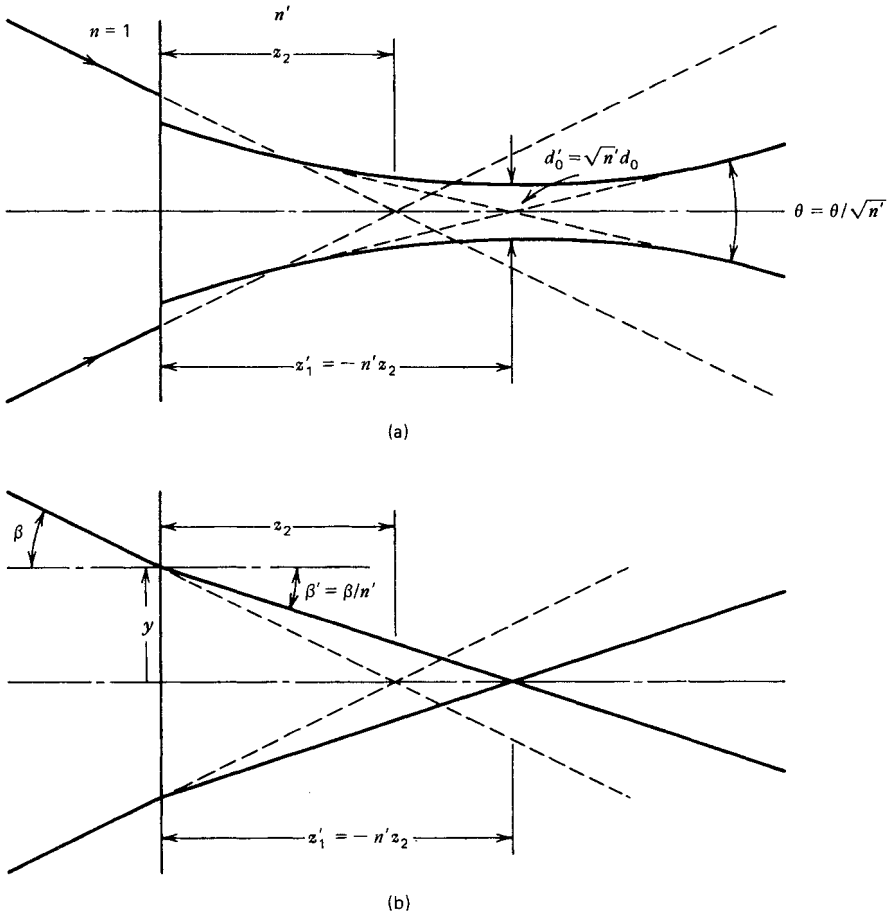


Figure 7.9. Focusing of a Gaussian beam into a plane interface. (a) Gaussian beam asymptotes, (b) ray trace of incident asymptotes. Both locate the beam waist at the same point.

an angle β' , where $\sin \beta' = \sin \beta/n'$ ($n = 1$). If d is the diameter of the beam at the surface then y is the radius of the beam at the surface, then

$$\tan \beta = \frac{d}{2z_2} \quad \text{and} \quad \tan \beta' = \frac{d}{-2z'_1}.$$

For small angles, setting the sines and tangents equal to their arguments, $\beta = n'\beta'$, we can write

$$\begin{aligned} \frac{d}{2z_2} &= \frac{n'd}{-2z'_1}, \\ z'_1 &= -n'z_2, \end{aligned} \quad (7.40)$$

EXAMPLE 7.E. A biconvex lens ($n = 1.5$, $R = -R' = 100$ mm, $t = 20$ mm (thick!)) is placed 100 mm from the output mirror of a laser ($d_0 = 1$ mm, $\theta = 0.8$ mrad, $z_R = 1.25$ m). Find the position and size of the beam waist after it passes through the lens. Compare it to the thin lens values.

Solution:

$$z_2 \phi = 100 \text{ mm}, \quad \phi_1 = (1.5 - 1)/100 = 1/200 \text{ mm}.$$

$$z_2 \phi = \frac{1}{2}, \quad z_R \phi_1 = 1250/200 = 6.25.$$

$$\alpha = \frac{\sqrt{1.5}}{\sqrt{(1 - \frac{1}{2})^2 + (6.25)^2}} = \frac{\sqrt{1.5}}{\sqrt{\frac{1}{4} + (6.25)^2}} = \frac{\sqrt{1.5}}{6.27} = 0.195,$$

$$\alpha^2 = 0.0382.$$

$$d'_0 = 59 \text{ } \mu\text{m}, \quad \theta' = 3.35 \text{ mrad}, \quad z'_R = 47.7 \text{ mm}, \quad z'_1 = 296 \text{ mm}.$$

The distance from the beam waist at the next surface is:

$$z'_2 = t - z'_1 = 20 - 296 \text{ mm} = -276 \text{ mm}.$$

$$\phi_2 = (1 - 1.5)/-100 = 1/200 \text{ mm} = \phi_1.$$

$$z'_2 \phi_2 = -1.381, \quad n - z'_2 \phi_2 = 1.5 - (-1.381) = 2.881.$$

$$z'_R \phi_2 = 47.7/200 = 0.2385.$$

$$\alpha' = \frac{\sqrt{1.5}}{\sqrt{(2.88)^2 + (0.239)^2}} = \frac{\sqrt{1.5}}{\sqrt{8.36}} = 0.424, \quad \alpha'^2 = 0.179.$$

$$d''_0 = 82.7 \text{ } \mu\text{m}, \quad \theta'' = 9.67 \text{ mrad}, \quad z''_R = 8.56 \text{ mm}, \quad z''_1 = 96.56 \text{ mm}.$$

Also

$$\alpha_{\text{total}} = \alpha\alpha' = 0.195 \times 0.424 = 0.0827.$$

In the thin lens treatment:

$$\frac{1}{f} = (1.5 - 1)\left(\frac{1}{100} - \frac{1}{-100}\right) = \frac{1}{2} \times \frac{2}{100} = \frac{1}{100}, \quad f = 100 \text{ mm}.$$

$$d''_0 = f \cdot \theta = 100 \text{ mm} \times 0.8 \text{ mrad} = 80 \text{ } \mu\text{m} \rightarrow \alpha = 0.08.$$

$$\theta'' = \theta/\alpha = 0.8/0.08 = 10 \text{ mrad}.$$

$$z''_R = 1250 \text{ mm} \times (0.08)^2 = 8 \text{ mm}.$$

$$z''_1 = f + \alpha^2(z_2 - f) = 100 + \alpha^2(0) = 100 \text{ mm}.$$

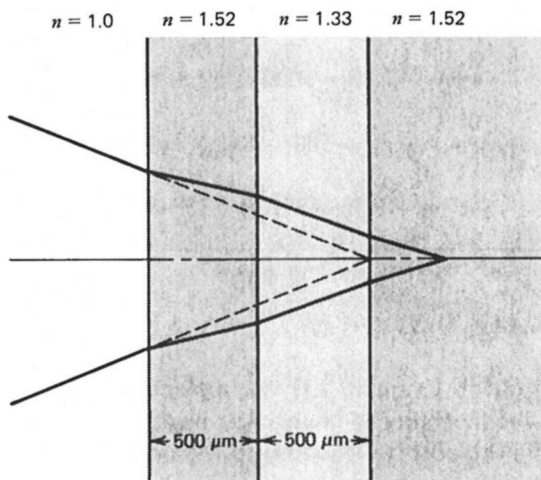
Thus, the thin lens values in this case give a good approximation to the paraxial figures.

the same relation just derived using Eq. 7.39b with $\phi = 0$ and $n = 1$. Thus, the position of a Gaussian beam waist after passage through a plane interface can be computed as if the asymptotes were refracted as ordinary light rays with their intersection locating the waist. As was shown earlier in the case of displacement of an image (Section 5.1.4), the beam waist is displaced by a parallel plate of thickness t and refractive index n by an amount

$$\Delta t = \frac{n - 1}{n} t.$$

An example of a calculation for focusing through parallel slabs of material is given in Example 7.F.

EXAMPLE 7.F. A helium–neon laser beam ($z_R = 1.25$ m) is focused by a short-focal-length lens ($f = 8$ mm) located well inside the Rayleigh range. The beam waist is located at the surface of a microscope slide. If a drop of water ($n = 1.33$) is put on the surface and covered with a $500\text{ }\mu\text{m}$ thick cover glass ($n = 1.52$) and the water entrapped between the glass and slide also has a $500\text{ }\mu\text{m}$ thickness, where is the new beam waist located?



Solution: In this case $\alpha = \sqrt{n'/n}$ for each interface. Since the power of the surfaces multiply the Rayleigh range in the expression for α , Gaussian beams can be treated very simply. All we need to do is follow the asymptotes of the beam in each region to their crossing at the beam waist.

1. Since $z_R, z_2 > f$, we may use Case 1 in Section 7.2.2. The original laser beam has (as we have seen earlier) the parameters $d_0 = 1$ mm, $\theta = 0.8$ mrad, $z_R = 1.25$ m. If $f = 8$ mm, then $d'_0 = \theta \cdot f = 0.8 \times 10^{-3} \times 8 \times 10^{-3} = 6.4\text{ }\mu\text{m}$. $\theta' = 0.125$ rad; $z'_R = 51.2\text{ }\mu\text{m}$, and $z'_1 = f = 8$ mm.

2. In the paraxial approximation, a parallel plate displaces an image a distance $[(n - 1)/n] \times t$, where t is the thickness of the plate (Example 5.1.4). Thus, to find the position of the beam waist we need only add the additional contributions to each layer.

- (a) The first glass surface is 1 mm from the original waist. The cover slip adds a distance of $[(1.52 - 1)/1.52] \times 500 \mu\text{m} = 171 \mu\text{m}$.
- (b) The water layer adds an additional $[(1.33 - 1)/1.33] \times 500 \mu\text{m} = 124 \mu\text{m}$.
- (c) Therefore, if there were air beyond the water, the beam waist would be $295 \mu\text{m}$ beyond the water–air surface. However, since the glass substrate is beyond this, the increase in distance into the surface will be nd , where d is distance in air, $295 \mu\text{m}$. Therefore the final beam position is $1.52 \times 295 \mu\text{m}$ or $448 \mu\text{m}$ beyond the original substrate surface.

3. An alternate calculation:

$$z'_1 = -1.52(-1000 \mu\text{m}) = 1520 \mu\text{m}.$$

$$z'_2 = t - z'_1 = 500 - 1520 = -1020 \mu\text{m}.$$

$$z''_1 = -1.33(-1020)/1.52 = 892.5 \mu\text{m}.$$

$$z''_2 = t - z''_1 = 500 - 892.5 = -392.5 \mu\text{m}.$$

$$z'''_1 = -1.52(-392.5)/1.33 = 448 \mu\text{m}.$$

7.3. CONJUGATE DISTANCES

As was pointed out in Example 7.D, the replacement of a single lens by two lenses provides the flexibility to change the position and size of the new beam waist. By moving the positive lens toward the negative lens, the beam waist is moved away from the two lenses. For example, by changing z'_2 in Example 7.D by 19.2 mm, from 108.1 to 92.9 mm ($z'_2 - f = 3.97$ mm there), the beam waist moves out to 1.69 m (see Example 7.G(a)). Since the distance between lenses was chosen to give $\alpha' = 20$ the beam waist is 0.800 mm. ($d''_0 = 20 \times d'_0 = 20 \times 40 \mu\text{m} = 0.8$ mm). If we erect a target at the beam waist ($z''_1 = 1.69$ m), the spot will be as small as possible with this lens configuration. But it is *not* the smallest achievable spot at this target distance! If the positive lens is now moved away from the first lens by 0.8 mm so that $(z'_2 - f)$ is changed to 4.77 mm, α' is reduced to 16.8 and the beam waist is reduced to $d''_0 = 0.672$ mm. ($\theta = 1.12$ mrad) and is now located at 1.44 m from the positive lens, between the lens and the target. What is most interesting is the size of the spot on the target. The target is 1.69 m – 1.44 m, or 0.25 m beyond the new beam waist, so that

using Eq. 7.4, $d_2'' = 0.734$ mm (see Example 7.G(b)). The spot beyond the beam waist is *smaller* than the beam waist when it was located at the target! It might seem that if focusing inside the target produced a smaller spot on the target, then an even shorter focus would produce even better results. But this is not so. Moving the second lens to a point where $\alpha' = 15$ ($z_2' - f = 5.58$ mm) results in a smaller $d_0'' = 0.600$ mm, but $d_2'' = 0.759$ mm. Thus, there is something special about the second configuration that was used. This effect is explained by the concept of conjugate distances.

EXAMPLE 7.G. (a) If the second lens of the combination in Example 7.D is moved from a point where $z_2' - f = 19.2$ mm to a point where $z_2' - f = 3.97$ mm, what is the beam diameter at the waist (d_0'') and where is it located (z_1'')?

Solution:

$$\alpha' = \frac{88.9}{\sqrt{(3.97)^2 + (2)^2}} = 20.$$

$$d_0'' = 20 \times 40 \times 10^{-6} \text{ m} = 0.800 \text{ mm}.$$

$$z_1'' = f + \alpha^2(z_2' - f) = 88.9 + 400(4) = 1.69 \text{ m}.$$

Thus the spot is 0.8 mm at ~ 1.7 m from lens.

(b) How does the spot size change on a target located at the beam waist location in part (a) when the lens is moved to a point where $z_2' - f = 4.77$ mm?

Solution:

$$\alpha' = \frac{88.9}{\sqrt{(4.77)^2 + 2^2}} = 16.8.$$

$$d_0'' = 16.8 \times 40 \times 10^{-6} \text{ m} = 0.672 \text{ mm}, \quad \theta'' = 1.19 \text{ mrad}.$$

$$z_1'' = 88.9 + 283(4.77) = 1.44 \text{ m}.$$

At the target

$$z_2'' = 1.69 - 1.44 \text{ m} = 250 \text{ mm}$$

and the spot size is

$$d_2'' = \sqrt{(0.672)^2 + (1.19 \times 10^{-3} \times 250)^2} = 0.734 \text{ mm},$$

smaller than when the beam waist was located at the target.

7.3.1. Conjugate Distance Equations

To illustrate this, some simple differentiation is necessary. However, if you are not familiar with calculus, the end result of this short derivation is still of use. In Fig. 7.10 we have set up the geometry for this derivation. Let d_2 be the diameter of the beam at a fixed distance r from the exit pupil of an optical system. The angle subtended by the spot at that distance is $\beta = d_2/r$. While r is fixed, z_1 and z_2 are not (i.e., the beam waist can be anywhere between the exit pupil and the target). The basic relations are (all of the primes on the variables have been dropped):

$$\begin{aligned} r &= z_1 + z_2, \\ d_2 &= \sqrt{d_0^2 + \theta^2 z_2^2}, \\ \beta &= \frac{d_2}{r} = \frac{\sqrt{d_0^2 + \theta^2 z_2^2}}{z_1 + z_2}. \end{aligned} \quad (7.41)$$

Since we wish to minimize β , let us take the derivative of β with respect to z_1 or z_2 and set it equal to zero:

$$\frac{d\beta}{dz_2} = \frac{1}{2} \cdot \frac{2\theta^2 z_2}{\sqrt{d_0^2 + \theta^2 z_2^2}} \cdot \left(\frac{1}{z_1 + z_2} \right) - \frac{\sqrt{d_0^2 + \theta^2 z_2^2}}{(z_1 + z_2)^2} = 0.$$

Rationalizing, we obtain

$$\begin{aligned} \theta^2 z_2(z_1 + z_2) &= d_0^2 + \theta^2 z_2^2 \\ \theta^2 z_1 z_2 &= d_0^2 \end{aligned}$$

and

$$z_1 z_2 = \frac{d_0^2}{\theta^2} = z_R^2. \quad (7.42)$$

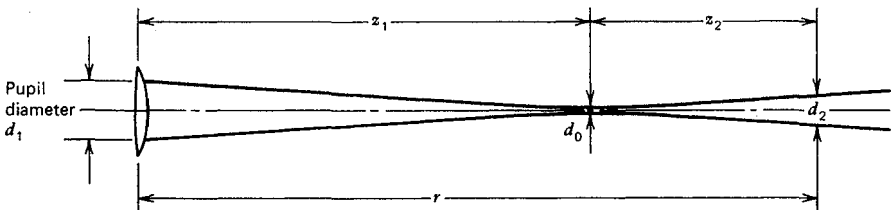


Figure 7.10. Geometry for the derivation of conjugate distances.

Thus, the minimum spot occurs when the product of the two distances (pupil to waist (z_1) and waist to target (z_2)) is equal to z_R^2 . When related in this manner, these two distances are called *conjugate distances*. The other distinguishing feature of these conjugate points is that the radius of curvature of the wavefront R is the same value for both points except for a change in sign and it is equal to their sums. That is,

$$R = |z_1| + |z_2|. \quad (7.43)$$

The benefit of analyzing a Gaussian beam system by computing and using the conjugate points is that one can calculate optimum values on the basis of input parameters. One can show that beam diameters at the conjugate points are related by

$$d_1 d_2 = \frac{4\lambda R}{\pi} \quad (7.44)$$

and that [Problem 7.14]

$$\frac{d_1^2}{z_1} = \frac{d_2^2}{z_2} = \theta^2 R. \quad (7.45)$$

Although the optimum focal length for a lens may not be equal to the lens–target distance, in the case of tightly focused beams for cutting or welding ($z_R \ll z_1$), z_2 is so small that one might as well consider the conjugate point at the beam waist [Problem 7.15]. But in cases of scanning systems and collimating problems, the conjugate point can be removed a considerable distance from the waist and this must be taken into account in a design problem. The most extreme case is the Gaussian relay case, where $z_1 = z_2 = z_R$.

7.3.2. Calculation of Conjugate Distance Parameters

There are two general cases that should be considered. The first of these cases assumes that the size of the new beam waist *or* the focal length of the lens and its position are known. With either condition the value of α can be calculated. Since z_R is already known and $z'_R = \alpha^2 z_R$, the product of the conjugate distances is found from

$$z'_1 z'_2 = z'^2_R = \alpha^4 z_R^2. \quad (7.46)$$

Also, since

$$z'_1 = f + \alpha^2(z_2 - f), \quad (7.20)$$

z'_2 , d'_2 , and R' can be easily determined [Problem 7.16]. Beyond a certain point it does little good to carry out the calculations if α is small and $z'_1 \gg z'_R$. In such a case $z'_2 = z'_R{}^2/z'_1 \ll z'_1$ and we may as well set $z'_2 = 0$ and $z'_1 = R$. But, even if z'_2 is substantial, the calculations are simple.

The more difficult case to compute is one in which only the target distance, which is equal to R' , and the beam diameter at the exit pupil, $d'_1 (= d_2)$, are given. Using Eq. 7.44 one can immediately check to see if the value of the spot at the conjugate point d'_2 is small enough for the application at hand:

$$d'_2 = \frac{4\lambda R'}{\pi d'_1}.$$

This can be used to find the ratio of the conjugate distances z'_2/z'_1 . Using Eq. 7.45 and setting the ratio equal to γ , we obtain

$$\gamma \equiv \frac{z'_2}{z'_1} = \left(\frac{d'_2}{d'_1} \right)^2 = \left(\frac{4\lambda R'}{\pi d'_1{}^2} \right)^2. \quad (7.47)$$

The quantity γ is useful in determining how important the effect of conjugate distances is. We can express z'_1 in terms of γ :

$$z'_1 = R' - z'_2 = R' - \gamma z'_1, \quad (7.48)$$

$$z'_1 = \frac{R'}{(1 + \gamma)},$$

If $\gamma \ll 1$, we can ignore the effect of conjugate distance and set $z'_1 = R'$ and $z'_2 = 0$. Note that $\gamma = (4\lambda R/\pi d'_1{}^2)^2$ is a product of the initial quantities and thus can be determined first. If, however, γ is determined to be close to or greater than unity, then z'_R can be determined from $(z'_R)^2 = z'_1 z'_2 = z'_1 \cdot \gamma z'_1 = \gamma R'^2/(1 + \gamma)^2$ and from this one can calculate that $\alpha = \sqrt{z'_R/z_R}$ [Problem 7.17].

7.4. LASER BEAM ILLUMINATION

Although most of this chapter has been concerned with finding the beam parameter relative to the waist region, not all problems require that the beam be focused on the surface in question. In some applications such as annealing and surface hardening, it is not necessary or useful to place the target at the beam waist. Such considerations as uniformity of illumination and tolerance of placement will play a part in the final design. It may be necessary to determine irradiance distribution in three dimensions to determine the best set up for a particular application.

7.4.1. Best Case Illumination

In contrast to those problems in which conjugate distances cannot be ignored, there are those in which the γ is small, and the new beam distance is approximated by $z'_1 = R' = f$ and $d'_2 = d_0$. These approximations in Eq. 7.44 gives

$$d'_0 \doteq d'_2 = \frac{4\lambda R'}{\pi d'_1} \doteq \frac{4\lambda f}{\pi d'_1}. \quad (7.49)$$

We see that, up to a point, as the lens is filled (d'_1 is made larger), the spot becomes smaller and thus the peak irradiance increases. Beyond this point the lens is overfilled and the focus spot begins to take on an Airy diffraction pattern (Section 1.7.3). In this region the spot size is the diameter of the central spot of the Airy diffraction pattern:

$$d'_{0\text{Airy}} = 2.44 \frac{\lambda f}{D}, \quad (7.50)$$

where D is the diameter of the focusing lens. The problem is this: If the lens is underfilled ($d'_1 < D$) then the spot size will not be a minimum, yet the lens is overfilled ($d'_1 > D$) then there will be a loss of power through the lens. Obviously there is some maximum point where the beam irradiance at the focus is a maximum.

Following Rhyins (Ref. 7.5) we define a truncation ratio as $k = d'_1/D$. That is, when $k = 1$, the $1/e^2$ points on the Gaussian beam are at the edge of the lens. For $k < 1$ the lens is underfilled; for $k > 1$ the lens is overfilled. Comparing the Gaussian and Airy cases:

1. *Gaussian case.* $d'_1 = kD$, where $k < 1$, $d'_0 = 4/\pi k \cdot \lambda f/D = (1.27/k)\lambda F$, where $F = f/D$, the f -number of the lens. Because the lens is underfilled, the power through the lens would be, except for reflection losses, the total power in the beam P_0 . Thus the irradiance of the spot is (using Eq. 7.49):

$$I = \frac{4P_0}{\pi d_0'^2} = \frac{4P_0}{\pi \times \frac{16}{\pi^2 k^2} \lambda^2 F^2} = \frac{\pi k^2 P_0}{4\lambda^2 F^2} \approx k^2 \frac{P_0}{\lambda^2 F^2} \quad (7.51)$$

as plotted in Fig. 7.11 in the region $k < 1$.

2. *Airy case.* $d'_0 = 2.44\lambda f/D = 2.44\lambda F$. The power through the lens can be calculated by integrating the Gaussian distribution over the lens area. As an approximation, we assume

$$P_t = P_0 \times \frac{D^2}{d_1'^2} = \frac{P_0 D^2}{k^2 D^2} = \frac{P_0}{k^2}. \quad (7.52)$$

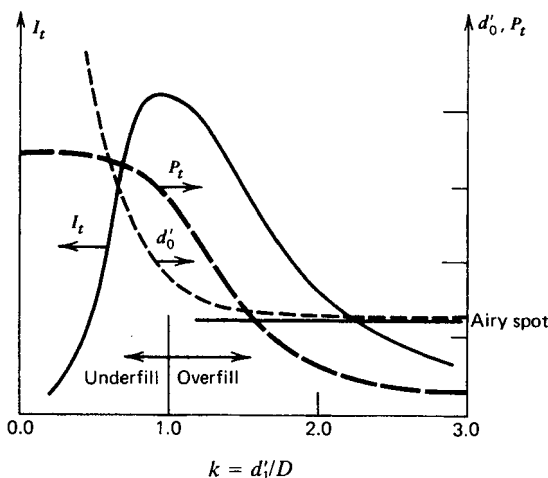


Figure 7.11. Plot of power (P_0), spot size (d'_0) and irradiance (I_t) as a function of lens filling factor, $k = d'_0/D$ (d'_0 = Gaussian beam diameter at the lens; D = lens diameter). $k < 1$ corresponds to an underfilled lens; $k > 1$ corresponds to an overfilled lens.

Since the spot size is constant, the irradiance is approximately

$$I = \frac{4P_t}{\pi d_0'^2} = \frac{4}{\pi d_0'^2} \frac{P_0}{k^2} = \frac{4}{\pi k^2} \cdot \frac{P_0}{[2.44\lambda F]^2} \approx \frac{1}{k^2} \frac{P_0}{\lambda^2 F^2} \quad (7.53)$$

as plotted in Fig. 7.11 for the region $k > 1$.

Thus in both regions the irradiance falls off as the truncation ratio departs from unity. More precise calculations show that maximum irradiance occurs at $k = 0.93$, a slight underfill. However, with this illumination, the spot has Airy rings and these may not be acceptable in certain applications. So the decision as to the size of a lens and the filling of its exit pupil is one that rests with the details of the application (including cost). In those cases where the Gaussian profile must be preserved, the lens must be underfilled to prevent any substantial diffraction due to the edge of the lens. As a rule of thumb, the beam diameter at the lens (d'_0) should be less than $\frac{5}{8}$ of the lens diameter (Ref. 7.6) corresponding to a truncation ratio of $k = 0.625$ [Problems 7.18 and 7.19].

7.4.2. The Irradiance Distribution of a Gaussian Beam

Equation 7.4 describes a boundary of the Gaussian beam, the hyperboloid of revolution for the focus of irradiance point of $1/e^2$ of the axis irradiance. This does not give a complete picture of the irradiance variation in the vicinity of the

beam waist. The irradiance distribution there can be described by an expression in terms of d_0 and z_R :

$$I(d, z) = \frac{I_0 \exp(-(2d^2/d_0^2)/(1 + (z^2/z_R^2)))}{1 + (z^2/z_R^2)} \quad (7.54)$$

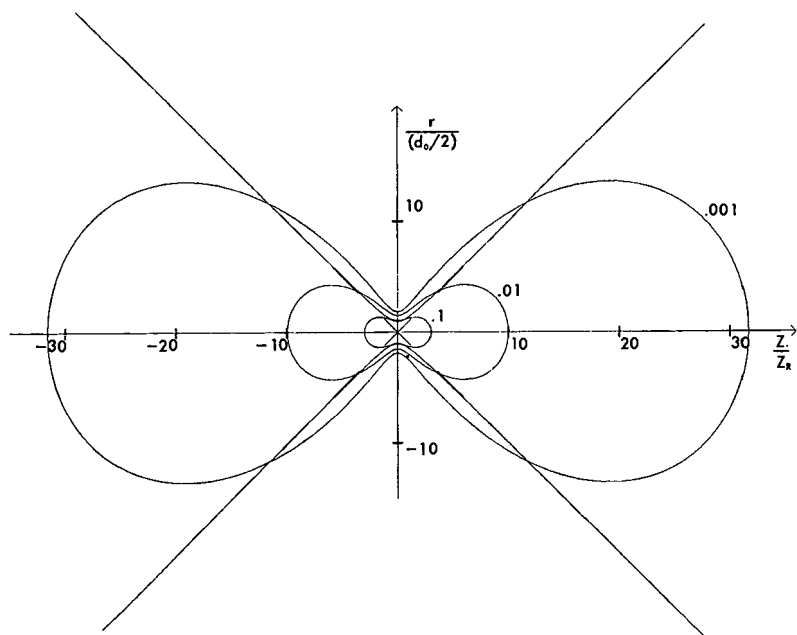
where $I_0 = I(0, 0)$, the irradiance at the center of the beam waist. A plot of the lines of constant irradiance (isophotes) is shown in Fig. 7.12. Note that there is a change in the character of contours from an elongated ellipsoid to a “dogbone.” This inflection point occurs when $I/I_0 = 1/e = 0.36788$. In terms of optical design, any effect that has an irradiance dependent threshold, such as in a photographic medium, will be more tolerant of a shift of the target out of the focus area if the beam power is adjusted to maintain the threshold value at $\sim 37\%$ of the central irradiance. The distribution shown here and the equations given here are for loosely (as opposite to tightly) focused beams; beams for which $\theta \leq \frac{1}{4}$ rad (or 14°).

7.4.3. Beam Shaping

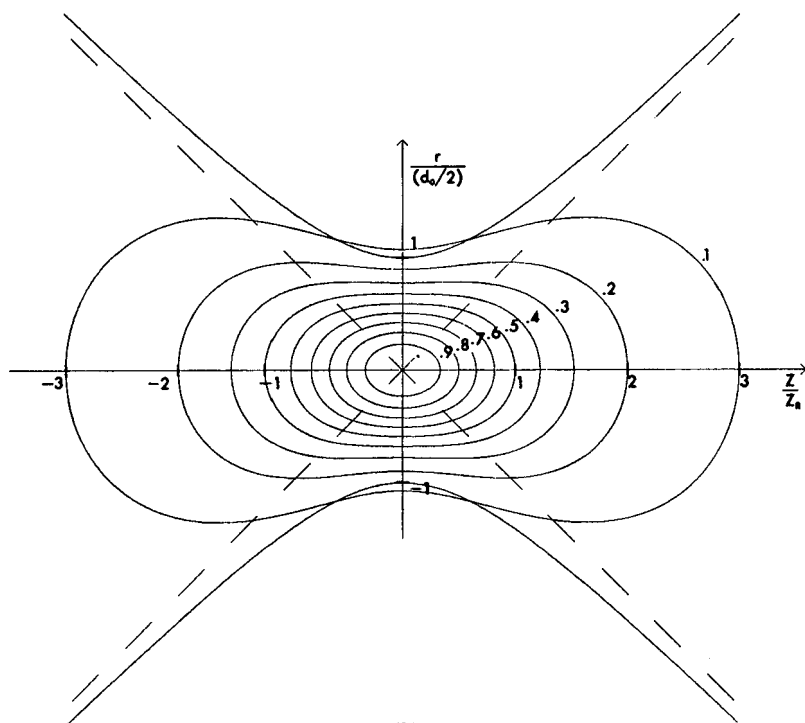
For some applications a circular Gaussian beam shape is not the most effective irradiance distribution. There are a number of methods of converting beam shapes from one form to another. The first of these, the cylindrical telescope, is used in some of the beam deflection schemes described in Chapter 8. In acousto-optic deflectors, the beam must be expanded and collimated, but only in one direction. If negative and positive cylindrical lenses are arranged to form a one-dimensional Galilean telescope, as shown in Fig. 7.13(a), the beam can be expanded and recollimated in one direction, while the original beam divergence and small beam cross sections are preserved in the other direction. It might seem easier to just use standard spherical lenses and expand the beam in two dimensions, but the economics of the beam deflectors require that the beam remain small so that it can pass through a thin slab of deflector material. Obviously, any telescope, when reversed serves as a beam compressor. Remember, however, that even though a telescope can compress and recollimate a beam, the divergence increases with decreased beam waist diameter. The restriction on the $d_0\theta$ product holds.

Another device for changing the beam size in one direction is the *prism beam compressor* (Ref. 7.7). Two prisms, arranged as shown in Fig. 7.13(b) reduce the beam width by refraction at the second slanted interface. If the refractive indices of the prisms are n_1 and n_2 , the compression due to the first prism oriented at the Brewster angle (Section 1.6.1) $c_1 = d_2/d_1 = 1/n_1$ [Problem 7.20] and that due to the second prism, $c_2 = 1/n_2$. The overall compression is the product of the two compressions and thus the product of the reciprocals of the two refractive indices,

$$c = c_1 \cdot c_2 = \frac{1}{n_1 n_2}. \quad (7.55)$$



(a)



(b)

Figure 7.12. Irradiance distributions of Gaussian Beams near a beam waist. (a) Contours of equal irradiance for three orders of magnitude; (b) between $0.1 I_0$ and I_0 .

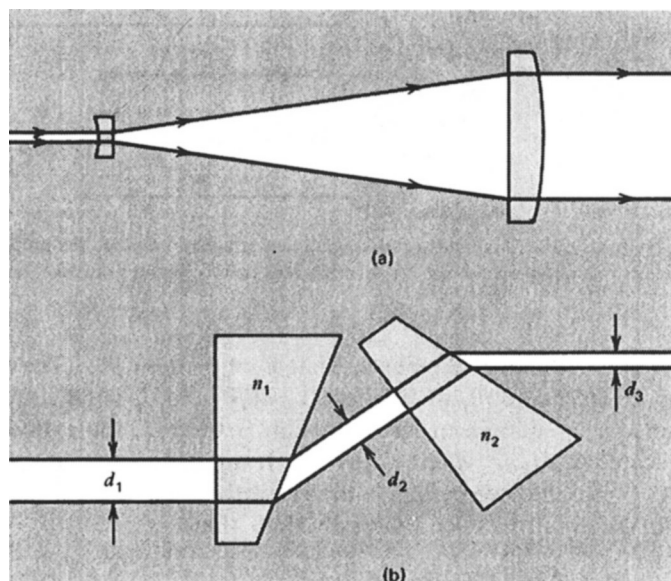


Figure 7.13. Beam expansion and compression. (a) One-dimensional telescope of cylindrical lenses used as a beam expander. (b) A two-prism beam compressor. Reversing the direction of light propagation changes the device from an expander into a compressor.

It is possible to arrange the prisms so that the output beam is parallel to the input beam. Since the prisms are cut to work at the Brewster angle, reflection losses are reduced. Some advantages to this assembly are that it is less costly to construct and more compact than a telescope, especially for modest compression ratios. Although the system can be used to vary the compression ratio by changing the angles of the prisms relative to the beams, this is done at the expense of reflection losses, especially if the surfaces where the beam is normally incident are anti-reflection coated. As was noted with the cylindrical telescope, the function of the system can be reversed by reversing the direction of the beam. The beam compressor is turned into a beam expander.

The two examples given above produce a fan-shaped beam, a beam with different divergences in two orthogonal directions. The beam spreads more rapidly in one dimension than in the other dimension, that is, θ_x is β times larger than θ_y ($\theta_x = \beta\theta_y$). This same situation is present in most semi-conductor diode (injection) lasers. In many designs it is necessary to convert the fan-shaped beam into a circular beam, with the same divergence for both dimensions. One method would be to use a cylindrical telescope or prism beam compressor to increase or decrease the divergence in one beam direction to match that of the other direction. Another method, shown in Fig. 7.14, uses a hybrid system, composed of both cylindrical and spherical lenses. Here a cylindrical lens increases the divergence of the beam in the more collimated direction to that of

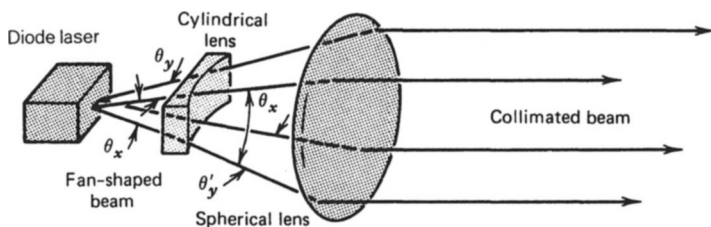


Figure 7.14. Hybrid system. A cylindrical diverging lens is used to diverge the more collimated of two divergences in a fan-shaped beam. A spherical lens then collimates the divergent beam.

beam in the less collimated direction. That is: $\theta'_y = \theta_x = \beta\theta_y$. The lens must increase the divergence of θ_y by β . Since $\alpha = \theta/\theta'$ and $\theta'_y = \beta\theta_y$, $\alpha_y = 1/\beta$. In addition, since a circular beam cross section is desired, the cylindrical lens should be located at a point where $d_{2x} = d_{2y}$. It can be shown that this location is $z_{2x} = \beta z_{Rx}$. Since the divergences in the two directions are different $\theta_x = \beta\theta_y$, so are the Rayleigh ranges. One can easily show that

$$z_{Ry} = \beta^2 z_{Rx}. \quad (7.56)$$

The Rayleigh range of the more collimated beam direction is large compared to the other direction, so that the cylindrical lens used to increase the divergence by β times would be inside of z_{Ry} . Thus the simple approximation of a short focal length lens inside the Rayleigh range can be used. For that case, $\alpha = f_y/z_{Ry}$ and $\alpha_y = 1/\beta$ determine the focal length of the lens: $f_y = z_{Ry}/\beta = \beta^2 z_{Rx}/\beta = \beta z_{Rx}$, where we have used Eq. 7.56 for the second equality. Once the divergences of the beam directions and the emerging beam diameters are the same, a spherical lens can be positioned, as in the collimation case, at $z_2 = f + z_{Rx}$ to produce a collimated spherical beam.

In some cases, (e.g., the diode laser), not only are the divergences different, but the location of the beam waists d_{0x} , d_{0y} are different. In this instance additional cylindrical lenses must be used to produce a common beam waist location and divergence (Ref. 7.8).

7.4.4. Flat-Top Distributions

Another irradiance distribution that is useful in some applications is the “flat-top” distribution, a variation across the beam that, at one edge, rises rapidly, remains constant across the center of the beam and falls just as rapidly, as shown in Fig. 7.15(a). If this distribution is rotationally symmetric about the beam direction, it is sometimes referred to as a “top hat” distribution, since a three-dimensional plot of irradiance in the x, y plane resembles a top hat (Fig. 7.15(b)). There are a number of different methods of producing this distribution. The simplest method is to use a mask to absorb or reflect the center of the beam and the edges of the beam. This method is inefficient since it reduces the total

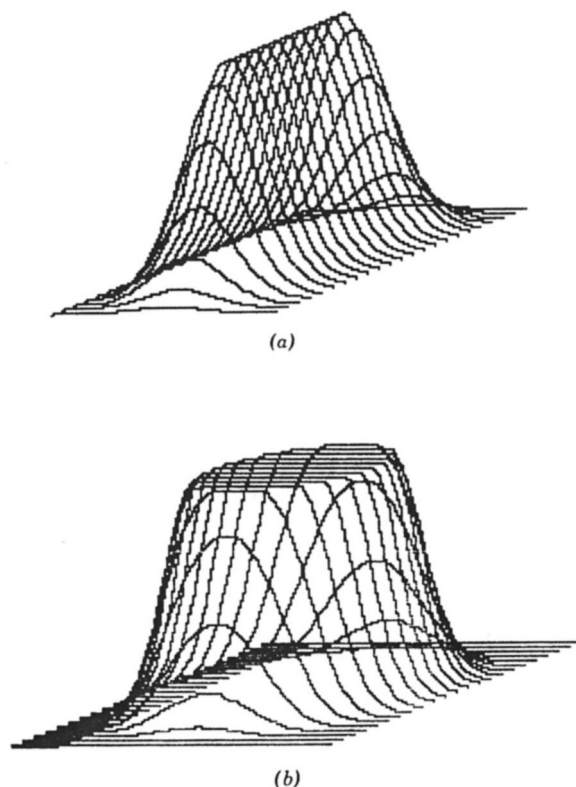


Figure 7.15. Non-Gaussian distributions. (a) Flat-top distribution; (b) Top-hat distribution.

energy in the beam. It also is difficult to use with high power lasers that could damage the mask. The work region must be located close to the mask, otherwise the distribution caused by diffraction will change in the far field.

One ingenious device that suffers only one of the drawbacks of the other methods (difficulty to produce) is the phase grating. The idea behind the phase grating is that a flat-top distribution in the far field is produced by an electric field distribution that has the form $\sin \phi/\phi$, as shown in Fig. 7.16(a). If a Gaussian beam illuminates a phase grating, a distribution of transparent material, which retards some parts of the beam by 180° with respect to the other parts, the transmitted electric-field amplitude will resemble the $\sin \phi/\phi$ function. In the far field, the distribution will close to a flat-top distribution, Fig. 7.16(b) (Ref. 7.9). By changing the periodicity of the grating and reducing the retardation to 144° , the flat-top distribution is enhanced. By combining the phase grating with a prism beam compressor a narrow flat-top distribution can be produced in the far field. Finally, it should be noted, a phase grating can be used to approximate other distribution. Since the relation between the phase grating output and the far-field pattern is that one is the Fourier transform

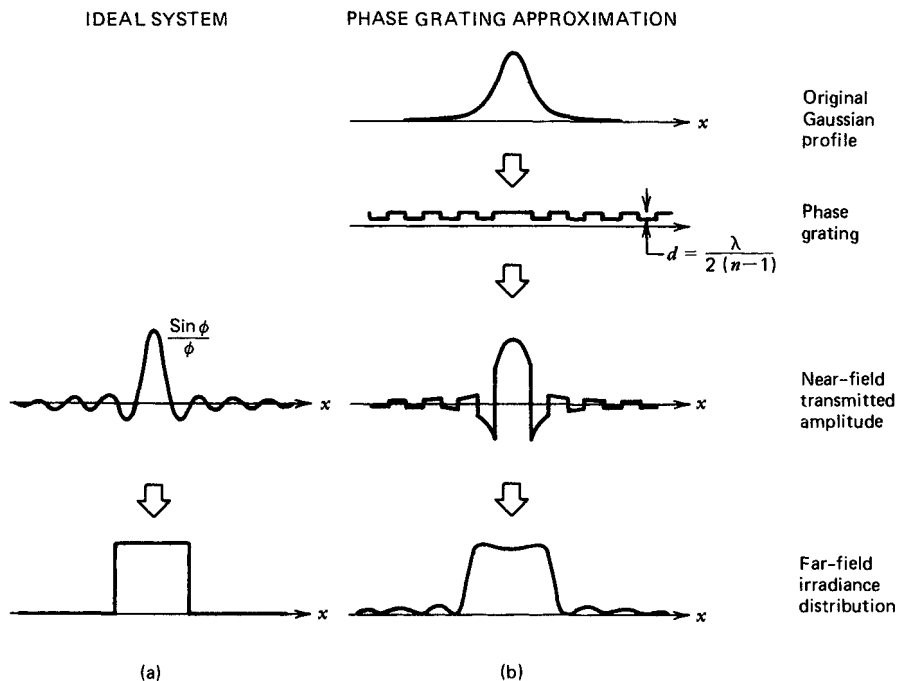


Figure 7.16. Binary phase grating. In an ideal system (a), a beam with an amplitude variation of $\sin \phi / \phi$ near a diffracting aperture will, in the far field, produce a flat-top irradiance distribution. An approximation (b) may be produced by illuminating a binary phase grating with a Gaussian beam.

of another, one must determine the closest phase grating approximation to the necessary output and construct a grating to produce it.

7.5. HIGHER ORDER MODES

The Gaussian beam is not the only irradiance distribution of radiation that can be produced by a laser, but it is the most fundamental. The resonant cavity referred to earlier may be able to support a number of other stable distributions, or modes, but none are as compact as the Gaussian beam and all contain nulls (regions of no radiation) within the irradiance distribution. The Gaussian beam is also called the TEM_{00} mode, a notation taken from the descriptions of electromagnetic field distributions in microwave cables and cavities. The other, higher order modes are labeled TEM_{mn} , where either m or n or both are greater than zero. In Fig. 7.17 there are shown a number of these distributions along with the TEM_{00} mode. Note the null lines in the distributions of the higher order modes. Calculations based on the electromagnetic field equations show that their intensity variations are described by polynomial functions multiplied by a Gaussian. A number of these have been plotted in Fig. 7.18. The

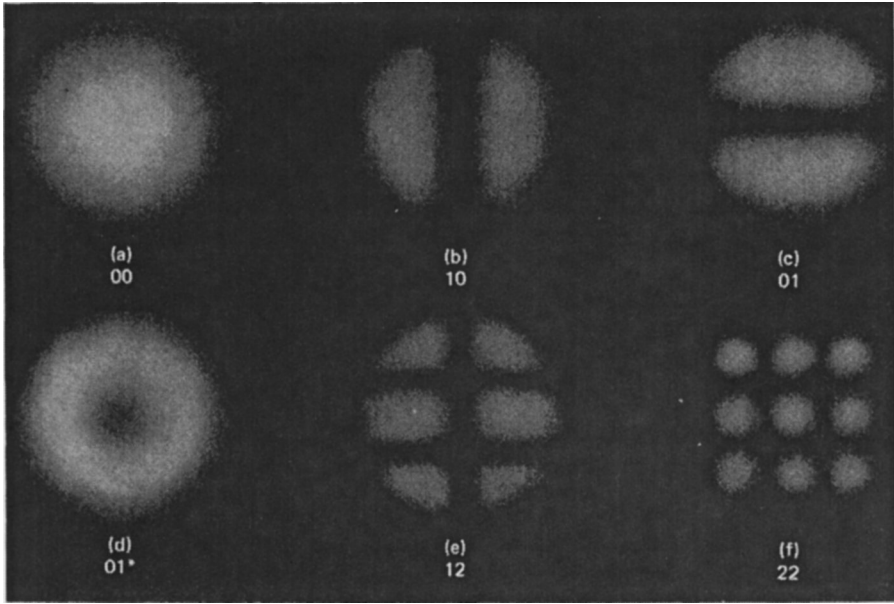


Figure 7.17. Transverse mode distributions.

descriptions of the propagation of higher order modes parallel those of the TEM_{00} modes, and many of the results can be derived by using the TEM_{00} mode values (Ref. 7.2).

For a higher order mode TEM_{mn} the beam spot size is a best fit to a rectangle of dimensions $d_x(0)$ by $d_y(0)$. There is a relation between the spot size and the divergence for higher order modes similar to that for the fundamental mode (Ref. 7.10).

$$d_{xm}(0)\theta_{xm} = \frac{4\lambda}{\pi}(2m + 1), \quad (7.57a)$$

$$d_{yn}(0)\theta_{yn} = \frac{4\lambda}{\pi}(2n + 1), \quad (7.57b)$$

where $d_{xm}(0)$ and $d_{yn}(0)$ are the beam diameters at $z = 0$ for the m and n orders respectively and

$$d_{xm}(0) = \sqrt{2m + 1}d_0, \quad (7.58a)$$

$$d_{yn}(0) = \sqrt{2n + 1}d_0, \quad (7.58b)$$

$$\theta_{xm} = \sqrt{2m + 1}\theta, \quad (7.58c)$$

$$\theta_{yn} = \sqrt{2n + 1}\theta, \quad (7.58d)$$

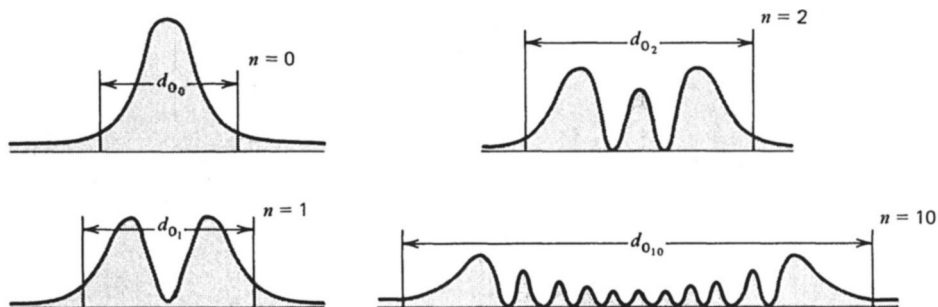


Figure 7.18. The profiles for higher-order modes. These are products of a Gaussian distribution and a polynomial (Hermite) function.

and d_0, θ are the TEM_{00} mode values [Problem 7.21]. Thus higher order modes have a larger spot size and divergence, but, by definition, the Rayleigh range

$$z_{Rm} = \frac{d_{xm}(0)}{\theta_{xm}} = \frac{d_0}{\theta} = z_R$$

remains the same. One difficulty with this measure of higher order mode spot size is that most of the energy is outside of the rectangle of dimensions $d_{xm}(0)$ by $d_{ym}(0)$. A practical alternative definition for higher order mode beam dimensions is the distance between the outermost radiance peaks (e.g., $n = 10$ in Fig. 7.18).

PROBLEMS

- 7.1. If the beam waist diameter of a TEM_{00} output of an argon laser operating at the 514.5 nm line is 1.5 mm, what are the divergence and Rayleigh range of the beam?
- 7.2. Derive the stability condition for laser resonators, $0 < g_1 g_2 < 1$, based on the definitions of g_1 and g_2 given in the text and the requirement that $z_R^2 > 0$.
- 7.3. A laser has mirrors whose radii of curvature are 3.33 m and -1.5 m. If they are 1 m apart, where is the beam waist located and what is the Rayleigh range?
- 7.4. A dye laser operating in the TEM_{00} mode at 620 nm has a full angle beam divergence of 1.2 mrad. The beam is to be focused to a diameter of 32 μm . Assume the initial beam waist is at the output mirror of the laser.
 - (a) If a 32-mm focal length lens is chosen, where should it be located?
 - (b) If a lens *must* be positioned 160 mm from the beam waist of the laser, what focal length is needed to produce the 32 μm spot size and where would the beam waist be located?

- 7.5. A 100- μm beam waist must be positioned 7.5 m from a CO_2 laser ($\lambda = 10.6 \mu\text{m}$) which has a divergence of 10 mrad. Find the focal length of the lens needed to accomplish this. (Assume beam waist is at the output mirror.)
- 7.6. A helium-cadmium laser ($\lambda = 442 \text{ nm}$) has a full angle divergence of 0.562 mrad. A positive lens is to be used to produce a 10 μm spot at a distance of 50 mm from the beam waist which is located at the output mirror of the laser. What is the focal length of the lens and where should it be placed?
- 7.7. Determine the relay geometry for two cases other than those given at the end of Section 7.2.3. (a) $z_2 = 3f/2$ and (b) $z_2 = f/2$. Express the distance between beam waists in terms of z_R .
- 7.8. Show that (a) the Rayleigh range is a maximum when $z_2 = f$ and that (b) a refocused Gaussian beam has a maximum lens-waist distance, z'_1 , when $z_2 = f + z_R$. Note that z_R is always positive.
- 7.9. Plot the beam waist position as a function of lens position along the beam axis in the vicinity of $z'_2 = f_o + z'_R$ (collimation point). Show that the collimation point is located at $z''_1 = f_o + z''_R$.
- 7.10. What is the maximum collimation distance for an argon ion laser operating in the deep blue at 457.9 nm with a 2 mm beam diameter given a 10 \times telescope made up of a +5 mm lens and a +50 mm lens?
- 7.11. A Gaussian beam whose wavelength is 442 nm and divergence is 3 mrad has its divergence increased by a thick plano-convex lens whose flat surface 15 cm from the beam waist faces the input beam, whose second radius of curvature is -30 mm , whose refractive index is 1.61 and whose thickness is 7.5 mm. What are the new beam characteristics and where is the new beam waist located? Use a paraxial calculation. How do these compare to a thin lens ray trace for an equivalent lens?
- 7.12. A 10 \times microscope objective ($f = 16 \text{ mm}$) is used to focus a helium-neon laser beam into the center of a 100- μm thick specimen ($n = 1.359$) mounted on a microscope slide of crown glass ($n = 1.517$). If a 0.5-mm thick cover glass (crown, $n = 1.517$) is laid on the specimen, how will this affect the focus? Use the paraxial form to determine the effect.
- 7.13. Compare the results (d_o, θ, z_R) of focusing an argon laser ($\lambda = 488 \text{ nm}$) beam ($\theta = 1.2 \text{ mrad}$) with the lens in Example 5.A as a thick lens using the paraxial treatment with that of a thin lens. The lens is 75 mm from the laser beam waist.
- 7.14. Using the expressions for d_1, d_2, R , and the conjugate relation $z_1 z_2 = z_R^2$, derive Eqs. 7.44 and 7.45.

- 7.15. Show that under the conditions of a strong focus: $z_R \ll z_1$, that the other conjugate distance z_2 is effectively located at the beam waist.
- 7.16. Using the equations presented in this chapter and the conjugate condition, $z'_2 z'_1 = z_R^2$, show that for a beam refocused by a thin lens of focal length f , that the conjugate variables in terms of the original beam variables d_0, z_2, z_R are

$$z'_2 = \frac{f^3 z_R^2}{[(z_2 - f)^2 + z_R^2][z_2^2 + z_R^2 - fz_2]},$$

$$d'^2_2 = \frac{d_0^2 f^2 (z_2^2 + z_R^2)}{[z_2^2 + z_R^2 - fz_2]^2},$$

$$R' = \frac{f(z_2^2 + z_R^2)}{z_2^2 + z_R^2 - fz_2}.$$

- 7.17. You are contracted to produce a laser show at an amusement park. Unfortunately, the best area for exhibit is a large sheet of cloth on the side of the park's star attraction, a roller coaster called "The Great National Screech Machine." The management would like to know how small of a spot you are going to project onto the sheet. To produce the size of display you want, you have to be 250 feet from the surface. The largest scanning mirrors you can use without slowing down the display noticeably are 10-mm-diameter mirrors. If you have a 5-watt argon ion laser whose output beam diameter for $\lambda = 488$ nm is 1.2 mm, what is the smallest spot you can project and what is the irradiance at the center of the spot? Is it useful to use conjugate distances here?
- 7.18. A helium–neon laser ($\lambda = 633$ nm, $\theta = 0.8$ mrad) is collimated by a 10 power telescope to illuminate a circular scanning mirror. How large must the mirror be so as not to change significantly the Gaussian profile?
- 7.19. A 10 power microscope objective ($f = 16$ mm) is used to converge a laser beam to a spot and then diverge. Variations and deviations in the irradiance of the beam profile causes light to be directed away from the central spot and into an area some distance from it. If a pinhole of the appropriate diameter is inserted at the beam waist, the Gaussian beam will be transmitted while light due to the non-Gaussian irradiance variation will be blocked leaving a smoothed Gaussian beam. This technique is called spatial filtering (Ref. 7.11). Assuming that a helium–neon laser ($\lambda = 633$ nm) has a 1 mrad divergence, what is the diameter of the smallest pinhole that can be used to filter the beam without introducing Airy rings into the pattern? Show, in general, that for a laser with a 1 mrad beam divergence, the product of the power of the lens

($\phi = 160 \text{ mm}/f$) and the diameter of the pinhole in microns is approximately 250 micrometers. (The power of a microscope objective, as defined here, is a dimensionless quantity.)

- 7.20. Show that the compression in dimension of a beam diameter by a prism oriented at the Brewster angle θ_B is equal to the reciprocal of the refractive index of the prism.
- 7.21. How much larger is the beam area at a distance of 10 m from a laser that is tuned to TEM_{59} than one tuned to TEM_{00} ?

REFERENCES

- 7.1. H. Kogelnik and T. Li, "Laser Beams and Resonators," *Appl. Opt.* **5**, 1550–67 (1966).
- 7.2. A. Yariv, *Quantum Electronics*, 2nd ed. (Wiley, New York, 1975), Chapters 6 and 7.
- 7.3. D. C. O'Shea, W. R. Callen, and W. T. Rhodes, *Introduction to Lasers and Their Applications* (Addison-Wesley, Reading, Mass., 1977), pp. 73–76. The g -parameter definitions have one sign change because the radius of curvature of concave mirrors is taken to be positive.
- 7.4. E. Stijns, "Some Nomographs for Gaussian Beams," *Appl. Opt.* **18**, 2827–29 (1979).
- 7.5. R. Rhyins, "Lenses for Low Order Modes," *Laser Focus*, April, 1974, 55.
- 7.6. P. Belland and J. P. Crenn, "Changes in the Characteristics of a Gaussian Beam Weakly Diffracted by a Circular Aperture," *Appl. Opt.* **21**, 522–7 (1982).
- 7.7. W. B. Veldkamp and E. Van Allen, "Compact, Collinear, and Variable Anamorphic Beam Compressor Design," *Appl. Opt.* **21**, 7–9 (1982).
- 7.8. H. Hanada, "Beam Shaping Optical System," U.S. Patent 4,318,594, 9 March 1982, assigned to Canon K. K.
- 7.9. W. B. Veldkamp, "Laser Beam I, Profile Shaping with Interlaced Binary Diffraction Gratings," *Appl. Opt.* **21**, 3209–3212 (1982).
- 7.10. W. H. Carter, "Spot Size and Divergence for Hermite Gaussian Beams of Any Order," *Appl. Opt.* **19**, 1027–29 (1980).
- 7.11. E. Hecht and A. Zajac, *Optics* (Addison-Wesley, Reading, Mass., 1974), 467–474.

8

Modulation and Scanning

(PUSHING A LASER-BEAM AROUND)

Until this chapter we have been concerned with the behavior of light in optical systems consisting of passive components, lenses, mirrors, and beamsplitters, that redirected or reformed the wavefronts. In all cases, the wavefront at a point in the system was constant in space and time. Now we will consider active components, components that can change the direction or amplitude of light as a function of time.

The intent of this chapter is to give a description of each technique and the limitations in the hardware that is used. The discussion should enable you to see the major considerations that must be addressed and provide guidelines for choosing between alternative approaches to a design. One of the difficulties presented by this subject is that the discussion is related to hardware, which is constantly being improved. Not all devices can be covered in a short chapter. Although most of this chapter will explore the modulation and scanning of laser beams, many, if not all, of the principles can be applied to imaging systems. The modulation of light, which will be explored first, will lead directly to scanning applications and finally to an example that includes both.

8.1. MODULATION OF LIGHT

The modulation, or control, of the irradiance of a beam or image can be represented schematically as in Fig. 8.1. A continuous source of radiation emits light that passes through a modulator; a signal sent to the modulator is impressed upon the light by causing some property of the light to vary in time. A simple signal, such as the square wave shown in the figure, helps to illustrate two basic characteristics of a modulator: *contrast*, or *extinction ratio*, and *frequency response*. Both of these characteristics measure the ability of the modulator to faithfully reproduce the input signal on the transmitted light wave.

The contrast refers to the depth of modulation on the beam compared to the signal modulation. Contrast may be expressed as it was in Section 6.5:

$$C \equiv \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \times 100\%, \quad (8.1)$$

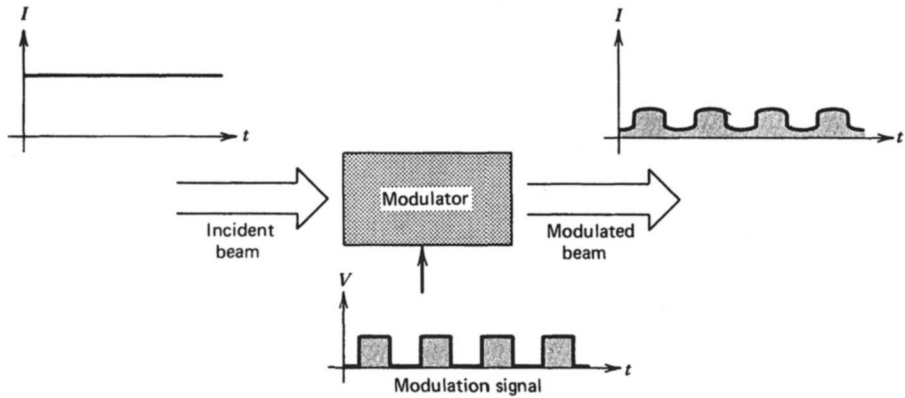


Figure 8.1. Modulation of a light beam.

where I_{\max} and I_{\min} are the maximum and minimum values of the modulated signal as shown in Fig. 8.2. Obviously if the contrast falls below a certain value, which usually depends on the noise in the system, the signal imposed on the light beam will be difficult to detect. Another way of expressing this same quantity is to give the ratio of the irradiance at maximum transmission in arbitrary units to that at minimum transmission, or extinction. This is usually expressed as a ratio such as 100:1, meaning the background signal is 1% of the maximum transmitted value.

In the modulation of light the frequency response refers to the ability of the system to reproduce a time-varying signal just as it does for other signal bearing systems. As indicated in Fig. 8.2, this property can be measured by the rise time of the system, the time T_r it takes for a signal to go from some low value (such as 10% of maximum signal) to some high value (such as 90% of maximum signal). It may also be expressed by a frequency, which is the reciprocal of the rise time ($\nu = 1/T_r$). While this single number may be useful, it does not fully characterize the response of a modulator. Perhaps the best description of the operation of a

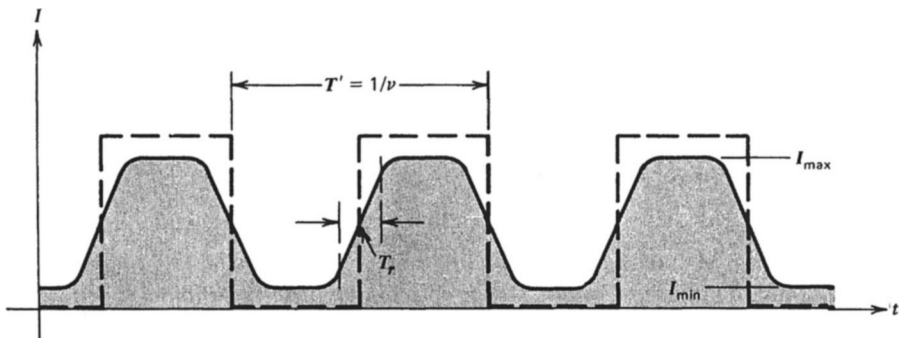


Figure 8.2. Parameters of a modulated signal.

modulator is a *frequency response curve*, a plot of contrast versus the signal frequency, as shown in Fig. 8.3. Thus the signal shown in Fig. 8.2 could be represented by point *A* on the curve at frequency ν' with contrast C' . Using this type of plot, the maximum frequency response of the system is determined by the minimum required contrast in the design. For example, if a contrast of $C'' = 15\%$ were needed then the maximum frequency response of the system, based on Fig. 8.3, would be ν'' . In addition to the above characteristics, the throughput (see Eq. 3.14), the power capacity, and the wavefront distortion of a modulator must be considered.

There are a number of properties of the electromagnetic wave that can be used in controlling the irradiance of a light beam: direction, polarization, phase, and amplitude. All of these will be used in some form in the modulators described here.

8.1.1. Source Modulation

In some systems the source itself is modulated. Usually these are not incandescent sources, since the maximum frequency response of an incandescent source is very low. This is because changes in light output are controlled by the thermal conductivity and heat capacity of the heated metals. Gas discharge sources including gas lasers can be modulated more rapidly. Limited by atomic collision processes, a gas laser can be modulated to about 500 kHz; this value varies with the type of laser and the excitation-versus-output curve. In comparison to the gas discharge lasers, semiconductor diode lasers can be modulated quite rapidly. Most commercial devices have rise times in the neighborhood of 10 nsec (100 MHz), but special devices can be obtained with frequency responses beyond the 30-GHz range.

In addition to direct modulation of source excitation, there are a number of methods involving intervention in the lasing process. Techniques known as modelocking and cavity dumping can provide nanosecond and picosecond long pulses at rates beyond 40 MHz. The actual repetition rate is governed by the length of the laser and is usually some submultiple of the frequency of oscillation of a light pulse traveling in the laser cavity, $\nu_0 = c/2l$ (l , the length of the laser). Normally this pulsed output is further modulated by techniques to be discussed

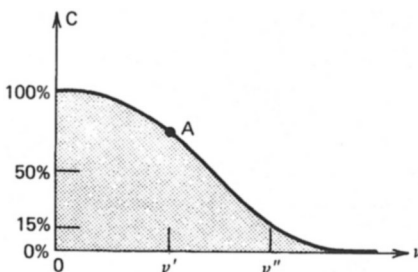


Figure 8.3. Frequency-response curve.

later. The advantage of such a combination is that it provides narrow pulses of high peak power, yet it only requires modulation at the much slower repetition rate of the pulse train. For further details, the reader should consult texts on laser operation for general principles (Ref. 8.1) and the specifications of individual laser systems for operating values.

8.1.2. Mechanical Modulators

For many years ship-to-ship signaling was done using a hand-operated shutter in front of a bright light. While the contrast was excellent, the frequency response was usually limited by the abilities of an operator to send Morse Code. Current mechanical modulators are mainly used to put a periodic signal on a beam. For example, many signal-sensitive, background-rejecting amplifiers (lock-in or phase-sensitive amplifiers) use some form of “chopper” to impress a frequency signature on a signal in the presence of extraneous radiation. The simplest of these is a slotted rotating wheel in front of a slit through which the light passes. Its advantage is its simplicity. By placing an LED and detector combination on opposite sides of the wheel, a reference signal can be easily generated. Obviously the frequency response of a motor-driven chopper is somewhat better than the early ship-to-ship signal device, but it leaves a lot to be desired. It shares a disadvantage common to all higher-frequency mechanical modulators; the modulation signal is confined to a single modulation frequency. In the case of the chopper, it equals the number of slots in the wheel times the rotation frequency of the motor. Another disadvantage of a chopper modulator is that it is bulky because of the motor and slotted wheel.

An alternate device that can be used for mechanical modulation is a mechanical scanner operating in a resonant frequency mode that directs light into the vicinity of a narrow slit. The exact form the modulation takes depends on the scanner, the shape of the beam, the slit, and any beam focusing optics. In Fig. 8.4 two modulation schemes are illustrated. In the first, a collimated beam is scanned across the slit to produce a quasi-sinusoidal signal. The exact

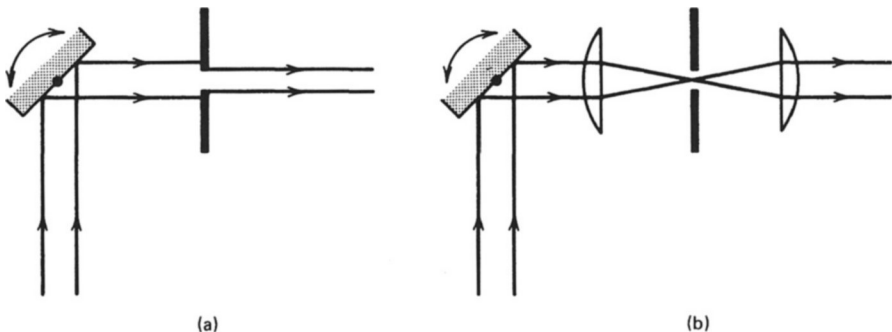


Figure 8.4. Mechanical modulators. (a) Quasi-sinusoidal modulation, (b) Square wave modulation.

modulation of the beam will depend upon the shape of the beam on the slit (uniform, Gaussian), the motion of the scanner (sinusoidal, sawtooth, ramp), and the shape of the aperture (pinhole, slit). The second scheme produces a square wave signal since the beam is focused onto the slit plane. If the slit is larger than the focus spot, the output will be a square wave whose rise time is approximately equal to the spot size at the slit divided by the beam scanning velocity at the slit.

These scanning mechanical modulators are smaller than most choppers and have a higher oscillation frequency. Although the scanning frequencies can be changed, they are usually fixed and are in the neighborhood of a few kilohertz. One must take into account the beam motion in the fixed slit device and the divergence of the light beyond the slit for the square-wave-type devices. These devices are somewhat more expensive than simple choppers. Thus mechanical modulators, whether rotating wheel choppers or scanner-slit combinations, can be used in optical design to produce square wave or near-sinusoidally modulated light beams in a simple and relatively inexpensive manner.

8.1.3. Electro-Optic Modulators

The concept of birefringence has been described in Section 1.6. In that section we assumed the amount of birefringence was set by the index of birefringence, $|n_{\text{ext}} - n_{\text{ord}}|$, and by the axes along which the crystal was cut. For example, by choosing to cut the crystal with optical axis in the plane of the plate and with a thickness d whose product with the index of birefringence equals a quarter wavelength of light, a *quarter wave plate* can be constructed. With the appropriate input polarization, linearly polarized light is converted to circularly polarized light with this device. By doubling this thickness a half wave plate, a polarization rotator for linearly polarized light can be made.

Since the retardation is the product, $d|n_{\text{ext}} - n_{\text{ord}}|$, an easy method of varying the transmission is to vary the index of birefringence. There are a number of liquids and crystals called electro-optical materials which, when a voltage is applied, will become birefringent. The degree of birefringence depends not only on the particular material but also on the strength of the electric field in the material. Thus, applying a voltage to an electro-optical material creates a variable wave plate that modifies the polarization of the incident light. If incident radiation, linearly polarized, either because of its origin (as in the case of many lasers) or passage through a polarizing filter, traverses a variable wave plate, the amount of light transmitted by a polarizer located after the wave plate and crossed to the original polarization direction will depend on the phase difference introduced by the waveplate. Thus the amount of light passing through the system is modulated by the applied voltage. Figure 8.5 shows a sequence of states of an electro-optic modulator. The curve of transmission *versus* applied voltage, shown in Fig. 8.6, is given by

$$I = I_0 \sin^2 \left[\frac{\pi}{2} \cdot \frac{V}{V_{\lambda/2}} \right], \quad (8.2)$$

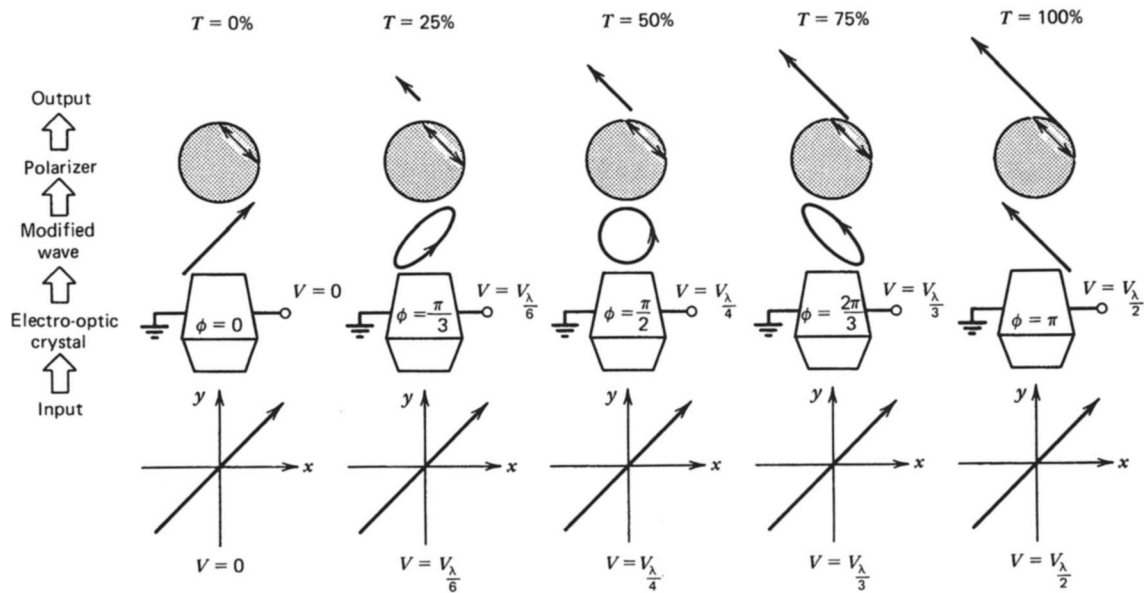


Figure 8.5. Electro-optic modulator output for different voltages impressed on the crystal.

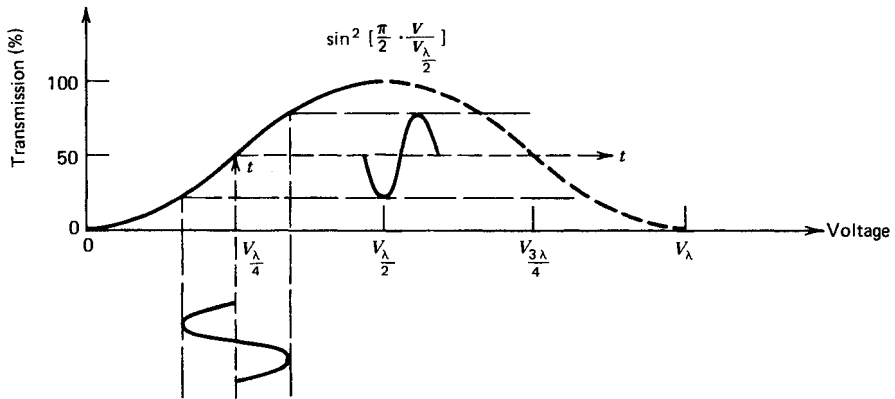


Figure 8.6. Transmission of an electro-optic modulator as a function of voltage impressed on the crystal.

where $V_{\lambda/2}$, the half wave voltage, is the voltage that produces a 180° phase difference. Obviously the curve is nonlinear. To obtain an approximately linear response, the system is biased at $V_{\lambda/4}$, the quarter wave voltage. This can be done electrically or it can be done optically by making the incident polarization circularly polarized through the combination of a linear polarizer and a quarter wave plate [Problem 8.1].

The frequency response of electro-optical modulators is considerably greater than that of mechanical modulators. It is limited by the RC time constant of the electrical circuit, the capacitance is due largely to the capacitor formed by the electro optical material sandwiched by the electrodes. The contrast of these modulators is limited by the extinction coefficient of the crossed polarizers. Together with the high voltages needed to provide the degree of birefringence, the relatively small attainable contrast ratio constitute the major disadvantages of an electro-optical modulator. When incorporating these devices into an optical design, care must be taken to see that the transmission properties and optical quality of the materials and the limiting apertures of the medium do not conflict with other requirements for the system.

8.1.4. Acousto-Optical Modulators

To understand acousto-optical modulation, it might be useful to return to Chapter 1 to review the material on diffraction gratings. In the case of normal incidence the light is diffracted according to the grating equation

$$m\lambda = d \sin \theta_d \quad (m = \text{an integer}) \quad (1.34)$$

where θ_d is the angle of diffraction and m is an integer. If the periodic variation in the grating profile is sinusoidal, only the zero and first orders ($m = 0, \pm 1$) will

occur. For non-normal incidence, Eq. 1.34 is modified to include the angle of incidence (Fig. 8.7):

$$m\lambda = d(\sin \theta_d - \sin \theta_i) \quad m = 0, \pm 1 \quad (8.3)$$

If the grating is a three-dimensional structure, a series of equally spaced planes, the constructive interference condition becomes more selective. The interference of radiation due to three dimensional grating is known as *Bragg diffraction*. A strong maximum occurs only if the angle of diffraction is equal to the angle of incidence ($\theta_i = -\theta_d$), giving for $m = 1$,

$$\lambda = d(\sin \theta_d - \sin(-\theta_d)).$$

$$\lambda = 2d \sin \theta_d. \quad (8.4)$$

This is the same Bragg equation used to analyze the results of x-ray diffraction from atomic planes in crystals.

Most diffraction gratings consist of rows of grooves in a material or a series of alternating transmitting or nontransmitting lines. However, any periodic spatial variation of the refractive index of a medium can also act as a grating. For example, if pressure is applied to a medium, the refractive index of the medium will change by some small amount. If this pressure is applied sinusoidally by sending a single-frequency sound wave through the medium, the moving sinusoidal refractive index variation that is produced in the material will act as a grating (Fig. 8.8). When the transverse extent of the sound wave is much greater than the wavelength of the sound, the sound waves in the medium serve as a Bragg grating, and light passing through it is subject to Bragg diffraction.

As an example, an ultrasonic sound wave at a frequency of 40 MHz is propagated through a block of fused quartz, whose sound velocity is 6 km/sec. The wavelength of the sound is then

$$\Lambda = \frac{v_s}{f} = \frac{6 \times 10^3 \text{ m/sec}}{4 \times 10^7 \text{ sec}^{-1}} = 150 \text{ } \mu\text{m}.$$

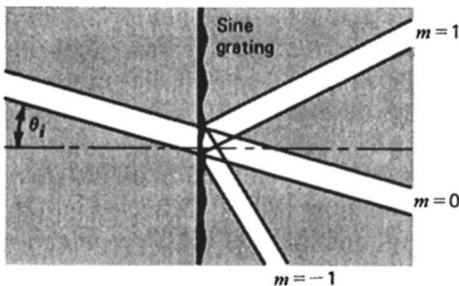


Figure 8.7. Diffraction of light by a sine grating.

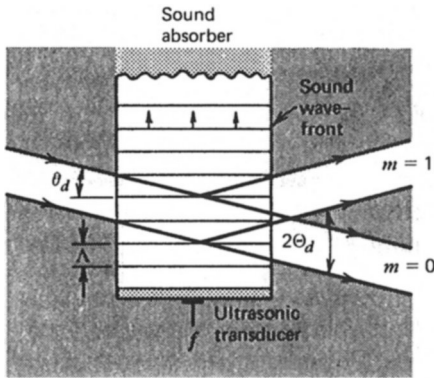


Figure 8.8. Diffraction of light in an acousto-optic modulator.

This spacing between compressions or rarefactions of the medium represent the grating constant of the thick grating created by the sound waves, that is, $d = \Lambda$ in the grating equation. The deflection angle θ_d can be found from

$$\sin \theta_d = \frac{\lambda}{2\Lambda}. \quad (8.5)$$

If a helium-neon laser ($\lambda = 0.633 \mu\text{m}$) beam is incident upon this grating, some of the beam will be deflected through an angle $2\theta_d$, where

$$\sin \theta_d = \frac{\lambda}{2\Lambda} = \frac{0.633 \times 10^{-6} \text{ m}}{300 \times 10^{-6} \text{ m}} = 0.00211.$$

Since $\sin \theta_d$ is small, $\sin \theta_d \cong \theta_d$, and

$$\begin{aligned} 2\theta_d &= 0.00422 \text{ radians} \\ &= 4.22 \text{ mrad} \\ &= 0.24 \text{ degrees.} \end{aligned}$$

As one can see, the diffraction angle is quite small in this case, but it is sufficient to create the modulation we require. (Note that the vacuum wavelength was used in the above calculation. It can be shown [Problem 8.2] that this is still correct after the beam has excited the acoustic medium, by using the law of refraction.)

If no ultrasonic wave is present in the medium, no light will be diffracted. As the sound-wave amplitude is increased, the amount of light diffracted into the first order increases. Thus, by modulating the amplitude of the ultrasonic wave the light beams (both zero and first order) are modulated. Beyond a certain

sound amplitude the grating becomes nonsinusoidal, and light is introduced into higher ($m = \pm 2, \pm 3, \dots$) orders.

The frequency response of an acousto-optic modulator is usually limited by the velocity of sound in the medium and the size of the incident light beam. Until the sound reaches the light beam there is no diffraction of light into the first order, since there is no disturbance present. The rise time of a signal will be equal to the time it takes the sound wave to cross the beam, that is, the beam dimension divided by the velocity of sound in the modulator:

$$T_r = \frac{d}{v_s} \quad (8.6)$$

For example, the rise time for a 1-mm diameter beam in a modulator whose sound velocity is 6 km/sec is 167 nsec. If the beam is focused to a small spot inside the modulator, the rise time can be reduced further, as shown in Example 8.A.

EXAMPLE 8.A. An acousto-optic modulator is to be used to put a 50-MHz signal on a light beam. Assuming a velocity of sound of 6 km/sec, what is the maximum beam diameter in the interaction region?

Solution: To transmit a 50-MHz signal requires a rise time

$$T_r < 1/(50 \times 10^6) = 2 \times 10^{-8} \text{ sec}$$

and

$$\begin{aligned} d &= T_r \times v_s = 2 \times 10^{-8} \text{ sec} \times 6 \times 10^3 \text{ m/sec} \\ &= 12 \times 10^{-5} \text{ m} = 120 \mu\text{m}. \end{aligned}$$

Note that if the beam is a Gaussian beam from a helium-neon laser, we would want to use the beam waist so that the divergence of such a beam is

$$\theta = \frac{4\lambda}{\pi d_0} = \frac{4 \times 633 \times 10^{-9} \text{ m}}{\pi \times 1.2 \times 10^{-4} \text{ m}} = 6.7 \text{ mrad}.$$

and the Rayleigh range is $d_0/\theta = 18 \text{ mm}$.

Since the beam is diverging rather strongly in the modulator, the Bragg condition will not be fulfilled for certain rays near the edge of the beam and therefore the amount of light diffracted into the first order will be reduced. (This phenomenon is explored in Example 8.B.)

EXAMPLE 8.B. An acousto-optic modulator ($n = 1.68$) has a transducer length of 55 mm and an acoustic wavelength of 0.1 mm. The input laser beam has wavelength $\lambda = 0.633 \mu\text{m}$ (helium–neon laser).

- What is the ultrasonic beam spread?
- What would be the diffracted laser beam profile if the initial Gaussian beam divergence is 10 mrad?
- How much misalignment would reduce the power of a 1 mrad beam to 50% of the power obtained when the beam is aligned?

Solution:

- A minimum in $I(\alpha)$ occurs when

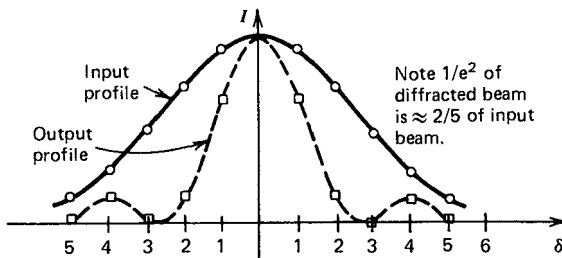
$$2\pi L \sin \alpha/2\Lambda = \pi.$$

There is considerable reduction when the quantity on the left is equal to $(\pi/2)$ ($(\pi L \sin \alpha)/\Lambda = \pi/2$). Thus, the beam spread is given by $\sin \alpha = \Lambda/2L = 0.1 \text{ mm}/2 \times 55 \text{ mm} = 1 \text{ mrad}$.

- We assume the input beam has a Gaussian profile with $1/e^2$ points at $\pm 5 \text{ mrad}$. Let this equal δ , so that $\alpha = \delta/n = 5 \times 10^{-3}/1.68$. This cannot be done in a closed form, but we can find the beam irradiance at various points and plot the two profiles (see below).

Angle, θ	Input Beam Irradiance, I_0	Reduction Factor, $I(\alpha)/I_0$	Output Beam Irradiance, $I(\alpha)$
1 mrad	$0.923I_0$	0.6935	$0.640I_0$
2 mrad	$0.726I_0$	0.1843	$0.134I_0$
3 mrad	$0.487I_0$	0.0003	$0.000I_0$
4 mrad	$0.278I_0$	0.0403	$0.011I_0$
5 mrad	$0.13I_0$	0.0312	$0.004I_0$

Note that the $1/e^2$ point of the diffracted beam is $\frac{2}{5}$ of the input beam.



- (c) Here $(\sin^2 x)/x^2 = 0.5$ and $x = 1.39$ rad (from tables or by trial and error). Hence $x = \pi L \delta / n \Lambda = 1.39$ or $\delta = 1.39 n \Lambda / \pi L = 1.35$ mrad. Thus the tolerance on positioning of an acousto-optic modulator can be severe.

It has been assumed in the above example that the sound waves emanating from the transducer are parallel waves. Actually, sound suffers the same diffraction effects as light. In the case of sound of acoustic wavelength Λ produced by a transducer of length L , the irradiance of the diffracted beam as a function of angle falls off as

$$I(\alpha) = I_0 \frac{\sin^2(2\pi L \sin \alpha / 2\Lambda)}{(2\pi L \sin \alpha / 2\Lambda)^2}, \quad (8.7)$$

where $\sin \alpha$ represents the angular deviation of the acoustic beam from that required for Bragg reflection. (If $\alpha = 0$, then all rays intersect the sound waves at the Bragg angle and $I(\alpha) = I_0$, i.e., no reduction of deflected beam irradiance.) There are a number of reasons why the power of the diffracted beam may be reduced. They are:

1. A small transducer of length L , resulting in a large acoustic beam spread.
2. A large divergence of the light beam (Bragg angle only partially satisfied).
3. A misalignment of the modulator relative to the beam by an angle δ . This is equivalent to a beam misalignment in the modulator material of δ/n .

We can compute the reduction in irradiance in the first order due to a non-zero α , whether it is

1. Intrinsic—due to transducer length L ($\alpha = \Lambda/L$),
2. Intentional—due to large input beam divergence ($\alpha = \theta/2$).
3. Instrumental—due to a misalignment ($\alpha = \delta/n$).

These effects are explored in Example 8.B. Note that in part (1) of the example, the angular divergence of the sound waves is smaller than the divergence of the helium-neon laser beam divergence given in part (2). Because of this only the central portion of the input beam is diffracted out of the zero-order beam, as illustrated in part (2) of the example (researchers use the appearance of a black band at the center of the undeviated beam as a sign of good modulator alignment). Note that in part (3), the tolerancing on the positioning and the maintenance of alignment must be carefully considered in any optical design. One further note on item (1) above: For best use of modulator and beam for modulation, the divergence of the light beam, which can be easily adjusted, should match the divergence of the acoustic beam.

TABLE 8.1. Comparison of Modulators

Type	Transmittance (%)	Maximum Frequency Response (MHz)	Rise Time (nsec)	Contrast	Visible Power Capacity (W)
Mechanical	100	0.005	1000 <	∞	5 <
Electro-optic	70	100	10	160:1 ^a	5
Acousto-optic	80	50	25	1000:1	20 <

^a The alignment of the electro-optic modulators affects the extinction ratio. This alignment includes the polarizers and analyzers of the system.

The acousto-optic modulator has a maximum frequency response that is a large fraction of the ultrasonic frequency, although the contrast is considerably reduced from that at lower frequencies. Overall the contrast in the diffracted beam is quite high since the light is diverted into a direction away from the initial beam direction. A contrast of 1000:1 is not unusual, but care must be taken with the surfaces and components of the system to prevent scattering light from reducing this ratio. Table 8.1 lists some of the features of the various modulators that have been discussed. All of these figures may be improved with time, so the table should be taken only as an indication of relative modulator performance.

8.2. SCANNING

Beside changing the power in a light beam, we would also like to be able to change its direction with time. The ability to scan a light beam is useful in applications ranging from laser displays to materials processing. After some basic concepts are examined, mechanical scanning systems, followed by acousto-optical and holographic scanning systems, are described. (Although it is possible to scan a laser by internal methods, the method is rather specialized and is not included here.)

8.2.1. Resolution—The Scan Equation

Scanning can be achieved in a number of ways: mirror galvanometer, rotating mirrors, and acousto-optical deflectors are among them. For the present all scanners can be characterized by three numbers. The first is the width of the scanner mirror or aperture, D , in the direction of the scan. The second number is

the aperture shape factor a , which determines the diffraction angle $\Delta\phi$ in the scan direction. That is:

$$\Delta\phi = a \frac{\lambda}{D}. \tag{8.8}$$

The values of a for a number of different aperture or mirror shapes are listed in Fig. 8.9. The values for slot and rectangular apertures were discussed in Section 1.7.3 and illustrated as Fig. 1.29. The third number is the maximum scanning angle. For example, in a mirror scanner, the maximum scanning angle is twice the mirror rotation angle.

One method of characterizing the quality of the scanning system is to determine the number of separately resolvable spots that could be produced by the scanner. By stating it in this manner, the parameter is independent of the optical system that follows the scanner providing the system does not reduce the maximum scan angle or cause additional diffraction by reducing the beam size at some point. The criterion for a resolvable spot is that used by Lord Rayleigh and was given in Section 1.7.3: the central maximum of one spot falls on the first diffraction minimum of the neighboring spot. For our case, this can be shown to be

$$\Delta\phi \cong \sin \Delta\phi = a \frac{\lambda}{D}. \tag{8.9}$$

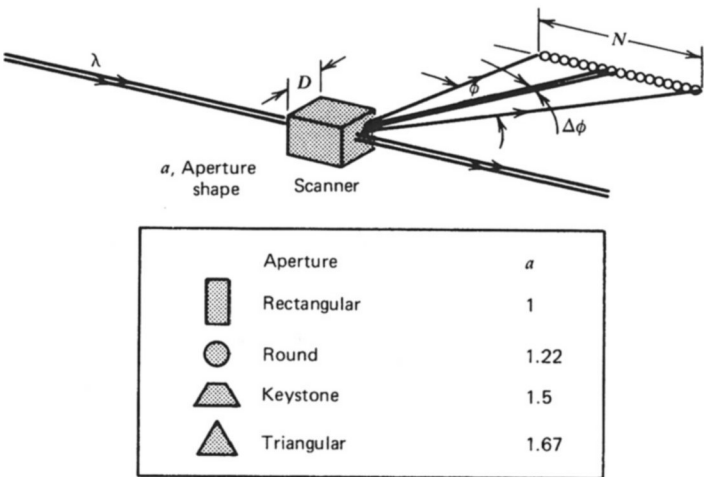


Figure 8.9. Parameters of scanners.

The number of spots N , and therefore the resolution, that the scanner can achieve is then (Fig. 8.9)

$$N = \frac{\text{total scan angle}}{\text{smallest resolvable angle}} = \frac{\phi}{\Delta\phi}. \quad (8.10)$$

Substituting for $\Delta\phi$,

$$N = \frac{\phi D}{a\lambda} \quad (8.11)$$

This simple relation is sometimes called the *scan equation*. It shows that the resolution is limited by both the source (λ) and the scanner (a , ϕ , and D) if the aperture is fully illuminated [Problem 8.3]. If the aperture is not fully illuminated (also referred to as underfilled), as in the case of laser scanning, then the beam divergence would be used. Except in some very exceptional cases, the aperture-shape factor will be a number comparable to those listed in Fig. 8.9, and the diameter will be the size of the beam within the aperture so the resolution will be less than in the case of a fully illuminated aperture. The considerations discussed in Chapter 7 for maximum focused spot irradiance apply here, although there may be a need to modify the results slightly to take into account noncircular scanner apertures.

8.2.2. Types of Scanning Geometries

Scanning geometries can be divided into three categories: objective scanning, post-objective scanning, and pre-objective scanning. The first of these, objective scanning, is quite simple: move the lens. The method has the obvious advantage of needing no separate scanner, but only some form of translator to produce the necessary scan. One example of *objective scanning* is shown in Fig. 8.10. The optical axis of a lens of focal length f is located a distance x from a collimated beam. Because the light is collimated and directed parallel to the optic axis it is focused a distance f beyond the lens and deflected a distance x onto the focal point. When the lens is now rotated about the *beam* axis, the beam focus will trace out a circle of radius x . Thus, by adjusting the distance between the lens axis and the beam axis, any size circle smaller than the diameter of the lens may be traced. More generally, any motion of the optic axis of the lens will produce exactly the same motion of the beam, as long as the beam remains within the lens aperture. When this system is used to cut holes in materials processing it is referred to as *trepanning*.

In Fig. 8.11 is shown a schematic of *post-objective scanning*. The beam enters the focus lens on axis. Beyond the lens, a scanner (here represented by a rotating mirror) redirects the beam to the target. There are several advantages to this system and one big disadvantage. The advantages are (1) the diameter of the focus lens need be only slightly larger than the size of the beam; (2) the scanning

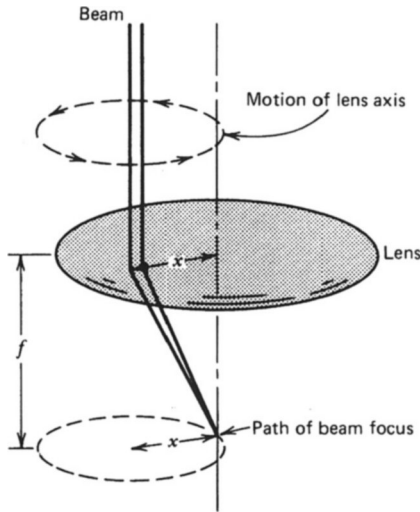


Figure 8.10. Objective scanning. This technique is called trepanning.

mirror can be somewhat smaller than the beam diameter, its size depending on the position of the scanner with respect to the mirror; and (3) the scan is linear with the scan angle ($L = R\phi$). The disadvantage is that the scan field is curved. For many applications, particularly in image scan systems and materials processing this characteristic will disqualify this type of scanning.

One means of achieving a flatter field is to use *pre-objective scanning*. A simple system is shown in Fig. 8.12. Although it is possible to obtain flat-field scanning using this system, the requirements on the components are considerably greater. The focus lens diameter must be considerably larger than the

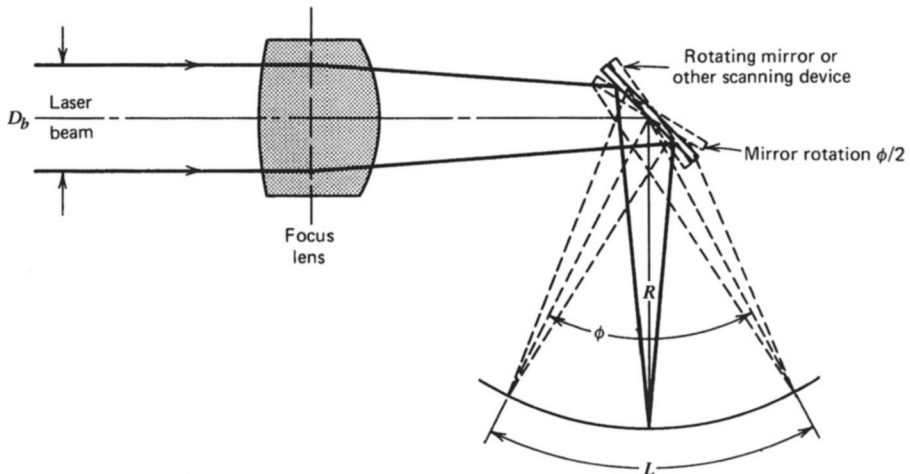


Figure 8.11. Post-objective scanning.

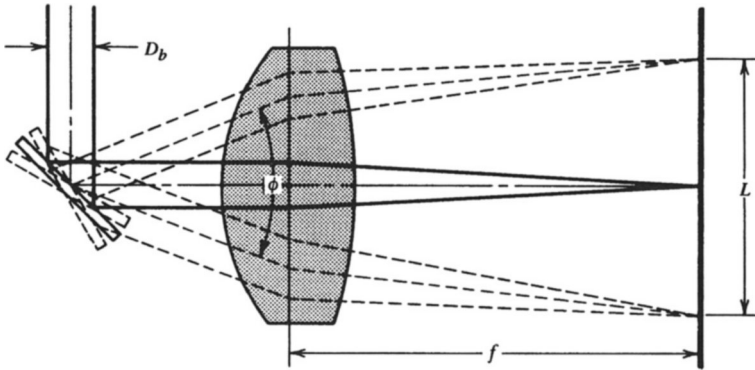


Figure 8.12. Pre-objective scanning.

beam diameter and well corrected for off axis images. Also, the scan distance is not a linear function of the scan angle unless care is taken to design a scan lens to provide for it.

One special pre-objective scanning geometry is the *telecentric geometry* shown in Fig. 8.13. In this geometry, the pivot of the scanning system is located at the front focal point of the lens. This means that the central ray from any beam will be refracted to a parallel ray by the lens, thus the beams all arrive normal to the target approximately one focal distance away. For Gaussian beams the target will remain inside the Rayleigh range reducing the tolerance on the target position. Notice that although the spot sizes are calculated using Gaussian beam relations, the beam directions and the scan lengths are calculated using standard geometrical optics with the central ray of the beam serving as the traced ray for each angle of the scanner [Problem 8.4 and 8.5].

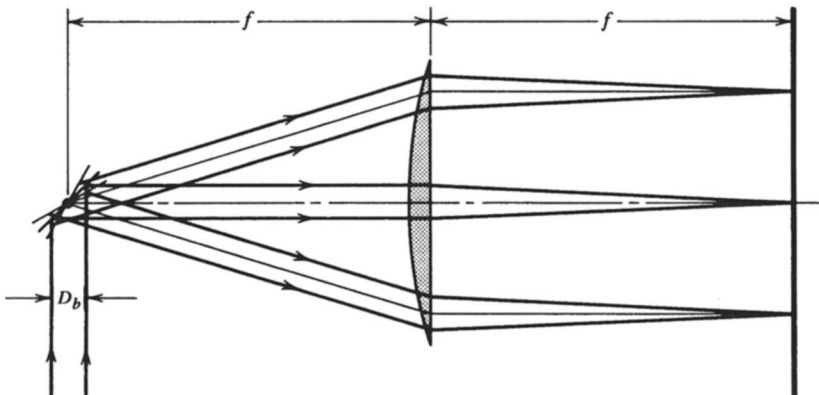


Figure 8.13. Telecentric scanning.

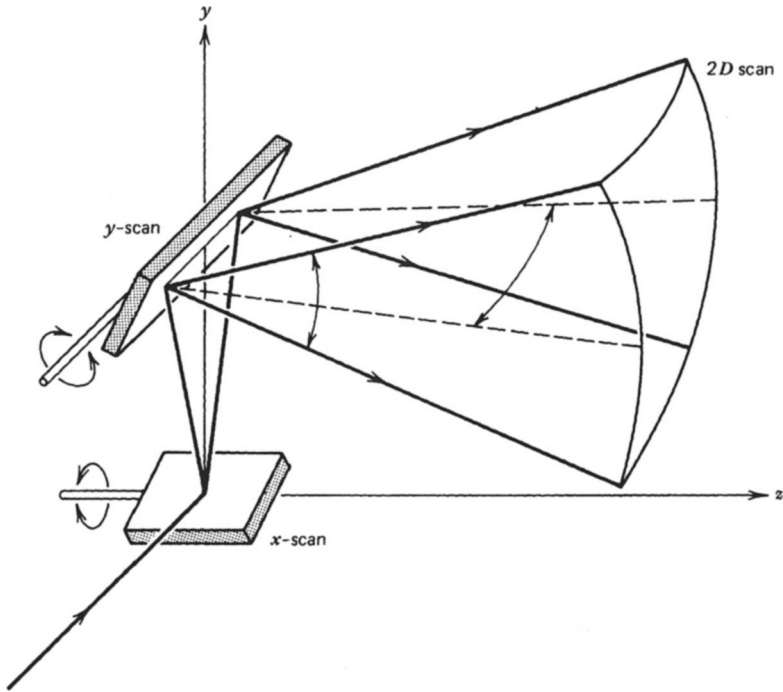


Figure 8.14. Two-dimensional scanning using two scanners with orthogonal rotation axes.

As noted earlier, moving the objective lens can provide a limited scan in two dimensions, covering an area slightly smaller than the area of the lens. For two-dimensional displays and pattern generation, the other two methods may be used. An x - y scanner can be constructed by combining two linear scanners at right angles as shown in Fig. 8.14. Another method of producing such a scan, especially for reading or writing two-dimensional information, is with a linear

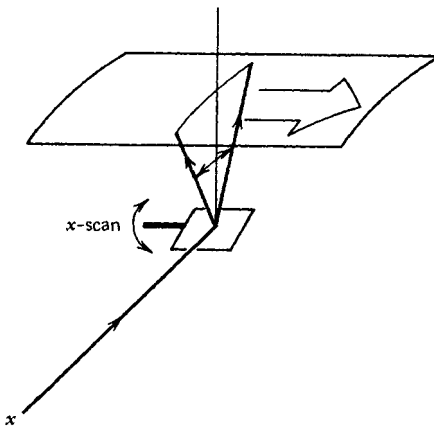


Figure 8.15. Two-dimensional scanning using a single scanner plus motion of the scanned medium. The moving medium may be curved to fit the path of the focused beam.

scan in one direction plus a target medium moving in an orthogonal direction. For post-objective scanners, the medium may be curved to fit the focal path, as shown in Fig. 8.15.

8.3. MECHANICAL SCANNERS

There are a number of different types of mechanical scanners. The choice of the specific type of scanner depends on the requirements of the optical system. For example, if an extremely high-speed, repetitive system is needed, a rotary scanner would be useful. On the other hand, a galvanometer scanner might be used in a system that required random access to specific points. In the following section we will look at two types of mechanical scanners, rotary scanners and nodding scanners.

8.3.1. Rotary Scanners

The simple mechanical scanner is a mirror rotating on a motor shaft. However, a single mirror, even if silvered on both sides will be unusable for a considerable portion of any rotation period. In place of a single mirror, one can use a series of mirrors attached to the faces of a symmetrical polyhedron to provide additional sequential reflecting surfaces. A variety of these mirror prisms is shown in Fig. 8.16.

Assuming the motor provides a constant angular velocity with a nearly wobble-free shaft rotation ($< 3''/\text{arc}$), the performance of a mechanical scanner is determined by the precision of the construction of the mirror prism. The specification of the mirror prism for a rotary scanner, or the evaluation of a manufacturer's stock scanner prism to see if it will meet the requirements of a particular optical system, is the major task to be performed when incorporating such a scanner into the design.

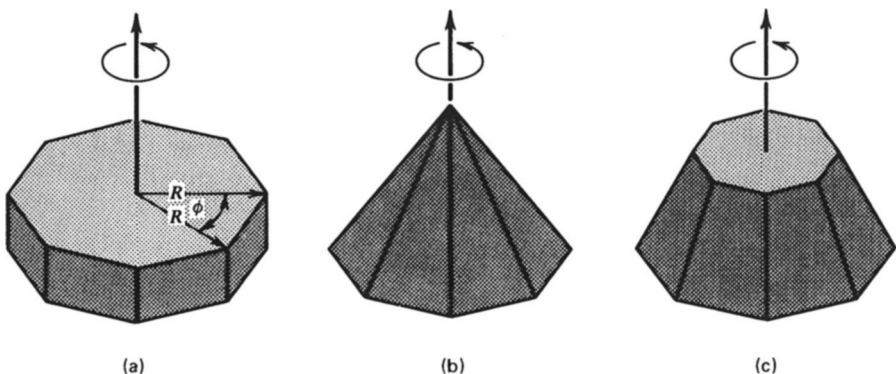


Figure 8.16. Three rotary scanner prisms.

Some of the considerations are listed in Table 8.2. In Example 8.C, the maximum number of facets for a given beam diameter and resolution is calculated. Usually the number chosen is smaller than the maximum, but too large a scan angle puts extreme requirements on the aperture sizes of the imaging systems and may reduce the amount of time the system is actually scanning (lower the duty cycle). The number of facets is usually some factor of 360 (6, 8, 10, 12, 18, etc.), simply because the fabrication of the corresponding angles is easier. Although it is possible to get a 17-facet mirror, there had better be a good reason why 18 facets would not work!

The perpendicular distance between the center of the prism and each facet surface must be held to a very tight tolerance (± 0.002 in.) if it is used in a post-objective scan, since the focus point will move in and out of the target plane if the facets are not equidistant from the rotation axis. It is better to specify this dimension with loose tolerance on the scanner radius (5.25 ± 0.025 in.), but require that all facets on a single prism be within ± 0.002 in. In a pre-objective scan system, the facets are usually illuminated by collimated or near-collimated light and the relative placement of the facets is not as critical (± 0.005 in.). The facet height is usually dependent on the facet width and must accommodate the beam size. The prism should be thick enough to prevent warping of the piece after finishing and leave enough space to account for "roll-off"—the unusable portion of the facet around the edges caused in the fabrication of the surface. The angular tolerance on any facet is specified as $360^\circ/n \pm m$, where n is the number of facets and m is the angular tolerance. Sometimes the value of m can be a good fraction of the facet angle ($360^\circ/n$), since angle variations can be compensated for by start-and-stop scan sensors. One of the most critical specifications is the tolerance on the angle between the facet and the mounting

TABLE 8.2. Rotary Scanner Specifications

1. Number of facets (n)—usually a common factor of 360° .
2. Facet to center distance.
3. Facet width.
4. Facet angular tolerance— $360^\circ/n \pm m$.
5. Mounting hole diameter.
6. Facet to mounting surface angle—determines tracking accuracy.
7. Mirror flatness—usually $\lambda/4$.
8. Reflectivity.
9. Surface—scratch and dig.
10. Material—usually aluminum—more exotic materials (e.g., beryllium) at higher scan rates.

EXAMPLE 8.C. A rotary scanner is to reflect a collimated beam from a helium-neon laser ($\lambda = 0.633 \mu\text{m}$). The beam diameter at the mirror prism is 5 mm. If we assume that a single facet is illuminated by the beam over a rotation angle equal to one half the prism facet angle (therefore the *scan angle equals* the prism facet angle), what is the number of facets needed to achieve a minimum resolution of 1000 (Rayleigh criterion)?

Solution: The scan equation requires that

$$N = 1000 = \frac{\phi D}{a\lambda} = \frac{\phi \times 5 \times 10^{-3} \text{ m}}{1 \times 0.633 \times 10^{-6} \text{ m}},$$

$$\phi = 0.127 \text{ radians} = 7.25^\circ.$$

If we choose $\phi = 9^\circ$ then $n = 360^\circ/9^\circ = 40$ facets. Note that fewer facets will give higher resolution since ϕ will be larger, but the scan angle will also become large.

surface. This specification determines if the facet normals all intersect the rotation axis at the same angle. The variation in this angle (called pitch angle error) introduces a beam wander from scan to scan, so that successive scans do not overlap one another. If there are no optical corrections (one will be shown later), the angle must be held to less than 10 arc seconds. In some applications, however, the design may *require* successive non-overlapping scans. In those cases, the specification must call out the sequence of angles and their tolerances [Problem 8.6].

The radius of the prism scanner depends upon the number of facets and the length of the facet. For example, in the case explored in Example 8.C, the number of facets was found to be 40. If the 5-mm beam were to stay on a single facet during the scanner rotation of $\phi/2 = 4.5^\circ$, the facet would have to be at least two beam diameters long, or 10 mm. Since the length of the facet can be seen from Fig. 8.16 to be equal to $2R \times \sin \phi/2$, then $R = 10 \text{ mm}/\phi$, using $\sin \phi/2 \cong \phi/2$. Thus, the radius of the prism is about 80 mm. This is a good-sized prism and if facet lengths much larger than 10 mm were needed or if extremely high speeds were required, the scanner design might have to be recalculated. The reason is that at very large radii and/or very high speeds, the centrifugal forces can be large enough to distort the optical surfaces, if not tear the prism apart. The dilemma comes from the fact that a large facet size gives good resolution, but it also increases the size of the prism, which increases strains. In some instances, these design values are accepted and a stronger material such as beryllium is specified as the scanner material.

The advantages of mechanical scanners is a constant scan velocity, which can also be quite high. This is combined with a low light loss and low distortion. The

disadvantages are cost and inaccessibility to triggering—in the sense that the scans cannot be started at a particular time, as is necessary when scanning a television raster. The high cost arises from the extremely tight tolerances required for many applications and from the stability required of the motor drive to maintain constant angular velocity.

In addition to the disadvantages cited above, there are a number of problems that must be considered if rotating prism scanners are to be used. Because of the high speeds, the greatest problem is the bearing wear, which limits the lifetime of most scanners. Even if the bearings do not fail, after some time a wobble may appear that can impair the usefulness of the optical system. All rotary scanners have a certain amount of scan dead time and most have long starting times because the moment of inertia and the required rotational velocities are large. Scanners sometime introduce vibration into the system and, in some cases, produce siren-like sounds. At very high velocities, air bearings and air turbine drivers must be used, which means that a compressed air source as well as electricity must be supplied to the optical system.

Some applications of mechanical scanners are listed in Table 8.3, and a number of optical design configurations is shown in Fig. 8.17(a), (b), and (c). Figure 8.17(a) shows a simple scanner system. It suffers from a large dead time. By using a truncated pyramid as shown in part (b), this drawback is removed, however the incident light is spread into six beams. By focusing the light onto one facet, the power delivered to the target could be increased. Part (c) of the figure shows a simple double-pass design that illuminates the scanner with parallel light yet allows the beam to enter the scanning region as a focused spot. This would be useful in systems where space is at a minimum. We will look at

TABLE 8.3. Applications of Mechanical Scanners

High-speed laser printers.
Web inspection.
Universal product code scanners.
Optical character recognition.
Facsimile.
Computer output microfilm.
Color print processing.
Phototypesetting and platemaking.
Map making.
Automated production, inspection, and testing.
Display.
High-resolution, large-screen projection.

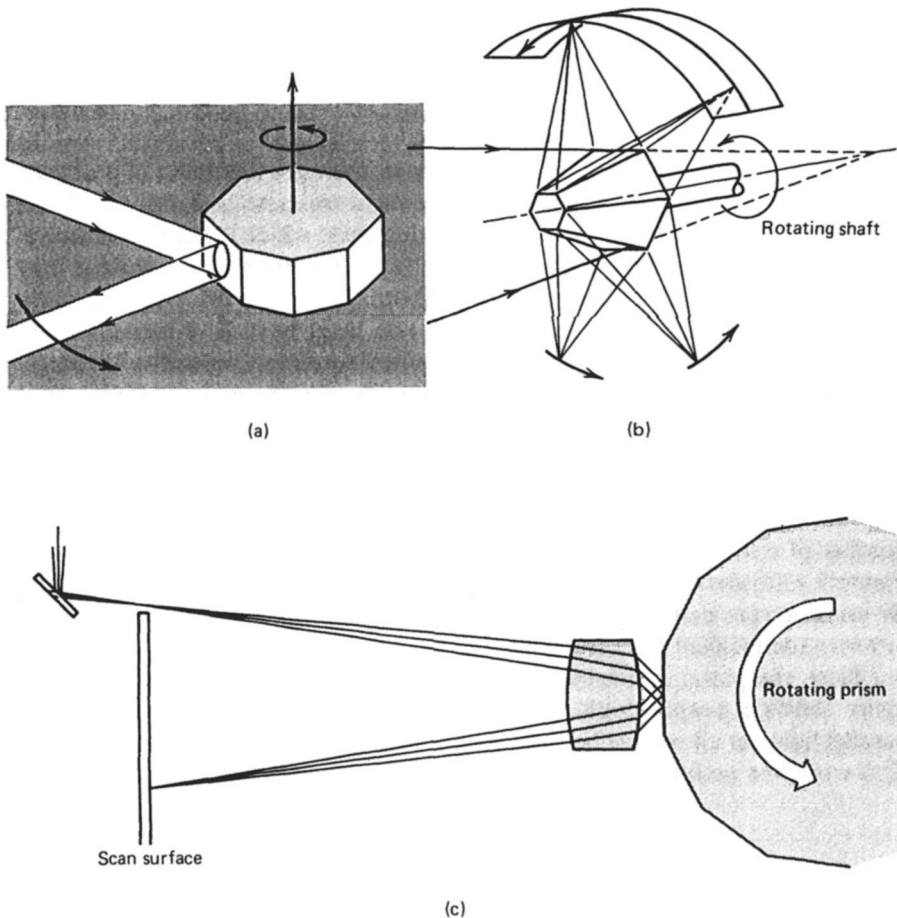


Figure 8.17. Three scanning geometries using rotary scanners. (a) Simple rotating polygon, (b) fully illuminated pyramidal scanner in a converging beam, (c) double pass scanner.

further considerations for a rotary scanner and certain optical compensations in an optical-design case study.

One problem with polygonal scanners not addressed so far is one brought about because the point of reflection moves as the prism moves. Figures 8.12 and 8.13 showed scans in which the point of rotation of the beam, also called the *center of deflection*, was in the plane of the mirror. This is usually true for the nodding scanners that will be discussed next, but it is not true for prism scanners. For example, in Figure 8.17(a) the beam is shown illuminating the center of the facet. At a later time, the edge of the facet will intercept the beam closer to the source as the edge of the facet swings into the beam so that the center of beam deflection moves closer to the source. Thus, for very small

changes in angles this center of deflection is located at the point where the beam intersects the facet. It can be shown that this point moves along the incident beam path between two extremes, the point where beam illuminates the center of the facet and that where the beam illuminates the edge. The maximum distance of motion of the beam rotation point can be shown [Problem 8.7] to be

$$\Delta = R \left(1 - \cos \frac{\phi}{2} \right) = R \left(1 - \cos \frac{\pi}{n} \right), \quad (8.12)$$

where R is the radius of the prism and ϕ is the facet angle, as illustrated in Figure 8.16(a), and n is the number of facets. From this one can see that the more facets on the prism, the smaller the amount of wandering of the center of deflection. But an increase in facet number also reduces the maximum scan angle and, therefore, limits the resolution that can be achieved. Another problem that arises is that the motion of the center of deflection in any scanning geometry such as a telecentric scan will complicate the simple analysis, which assumes the scan pattern is generated from a stationary deflection center.

8.3.2. Nodding Scanners

While the rotary scanner is one of the fastest scanning methods providing reliably repetitive scans at a fixed frequency, it cannot be used for every scanning application. There are less-expensive and more-flexible methods of scanning, which can be lumped under the general heading of nodding scanners—units that rotate a single mirror through a limited angle.

Four different types of these scanners are shown in Fig. 8.18. The *torsion rod scanner* (Fig. 8.18(a)) consists of a mirror mounted on a slender rod that can be sent into torsional oscillation by introducing a periodic signal into coils around a magnetic core. The changing magnetic field causes the armature that holds the torsion bar to rotate. The *taut band scanner* (Fig. 8.18(b)) is a more rugged scanner than the torsion-bar type. The mirror is mounted on a taut band or spring with a magnet attached to it. The magnet is positioned inside an electromagnetic coil. Just as in the case of a magnetic relay, the presence of a current in the coil will cause a movement of the magnet. The third type of scanner (Fig. 8.18(c)), the *galvanometer scanner*, also uses an armature-coil mechanism. The mirror is mounted on a rod attached to a moving element in the mechanism. One version, the moving coil type, uses the D'Arsonval movement found in ammeters and other rotating-pointer electrical meters. The other version, the moving iron type, is akin to a motor in that the coils are wrapped around the stationary piece, the stator, and the mirror is mounted on the motor shaft, the rotor. The second of these types is the more rugged and is used on most galvanometer scanners. The fourth scanner type (Fig. 8.18(d)), the *piezoelectric scanner*, has no coils or rotors and is therefore the most rugged of those illustrated. It consists of a mirror mounted on a two ceramic wafers that are bound together. A high voltage applied to the ceramics causes one ceramic

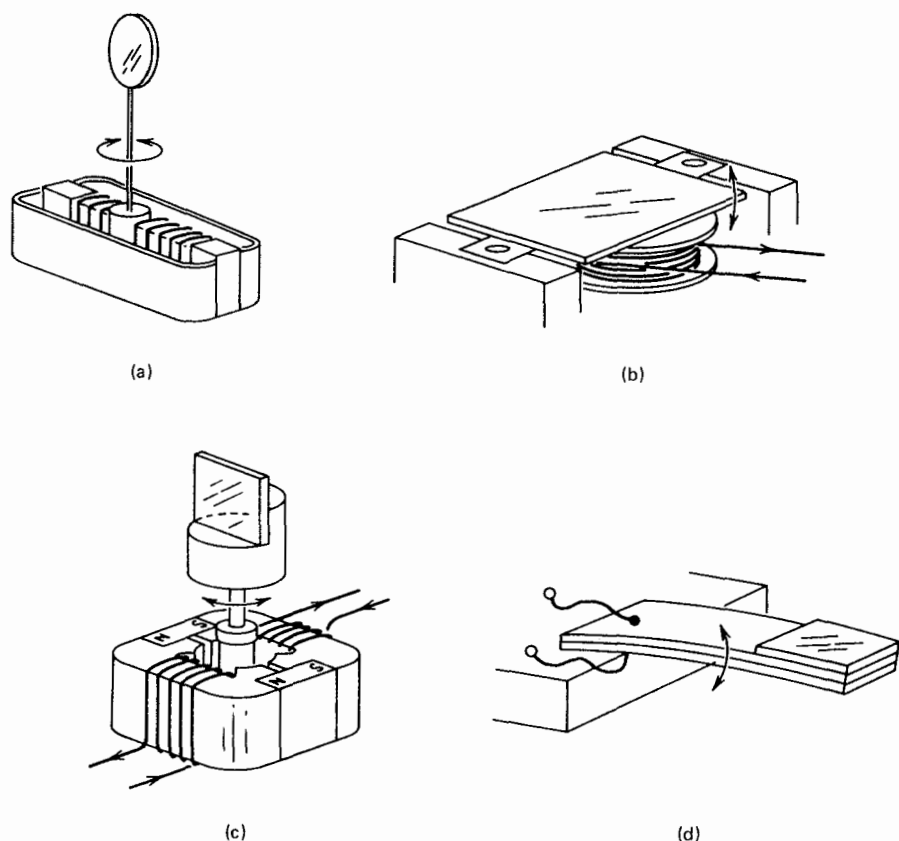


Figure 8.18. Four nodding scanners. (a) torsion-rod scanner, (b) torsion-bar scanner, (c) galvanometer scanner, (d) piezoelectric scanner.

to expand, while the other contracts, and since they are bonded together, this results in a bending of the structure. The angular motion of the mirror mounted on the wafer causes a scan with a center of deflection some distance from the wafer.

All of the nodding scanners can be described by certain common parameters that enable one to determine which type would best fit a particular application. They are:

1. *Maximum Rotation Angle.* This is usually the maximum angle through which the mirror can be rotated (peak-to-peak), rather than a center (no-signal position) to peak angle. The value can be only a few degrees for piezoelectric scanners or as much as 60° for some galvanometer scanners.
2. *Load-Free Resonant Frequencies.* These are the natural resonant frequencies of the scanner. Starting with an unloaded scanner (one without a mirror

- attached), the effect of the added inertia of a particular mirror is calculated and a loaded frequency is found. It is this frequency that determines the limits of the response of the scanning system.
3. *Sensitivity.* The amount of voltage or current needed to produce a unit deflection will determine the size and cost of the electronic interface to the scanner system.
 4. *Linearity and Repeatability.* These parameters specify how linear the mirror rotation will be. If the scanner is used without some feedback to check the position of the mirror, the repeatability specifies the tolerances on the location of the mirror if the same signal is applied at a later time.

Given the choices of scanners multiplied by numerous models for each type, it is useful to describe the calculations needed to evaluate a set of scanners. Rather than do this for all types of scanners, we will choose one, the galvanometer type, and determine the performance for a loaded scanner. After that, it will be easier to make some broad statements about the different scanner types.

With any scanner system, we require a minimum frequency response and resolution appropriate to the application. Stating the minimum frequency response of the system does not immediately give the required loaded resonant frequency for the scanner since a scanner can be operated under three different conditions, as suggested by Fig. 8.19. If the scanner motions correspond to frequencies below its resonant frequency, it is said to be operating *broadband*, since the operation is not greatly affected over a wide range of frequencies. In some cases the ability of a scanner to follow a complicated wave pattern can be increased by providing feedback circuits that modify the input signal on the basis of the position and velocity of the mirror. As a rule of thumb, a mechanical system can be considered to be operating broadband up to a frequency that is 85% of the resonance frequency. If it is operated in the region around the resonance frequency it is said to be operating *narrowband*, since signal deviations outside the resonance frequency region are difficult for the scanner to follow. At resonance, the required driving power is a minimum and the scan amplitude is greatest. Above the resonance region, the scanner will respond to fast rises, but more power is needed to achieve the necessary scanner motion, and there is a significant lag between signal and response. In this region, it is said to be

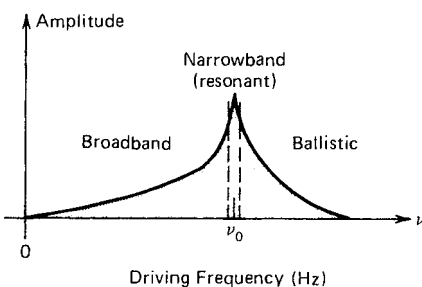


Figure 8.19. Frequency response of a periodic scanner.

operating *ballistically*. All of the above assumes the scanner has a single resonance frequency. If there is more than one resonance, the situation is complicated. For our purposes, we will assume a single loaded resonance at a frequency, ν_0 .

As an example, let us assume that we have decided that we need a scanner that has a broadband response up to 900 Hz and that our application also requires a resolution ($\phi/\Delta\phi$) of 1500 or better. Table 8.4 lists specifications for a number of commercial galvanometer scanners made by General Scanning and the moments of inertia for three round mirrors. The effect of the mirrors on the natural scanner frequency can be determined from Fig. 8.20. First, the ratio of the inertia of the mirror plus mount to the inertia of the armature must be determined. (Note that the effect of the mount inertia can be sizeable. If the moment of inertia is not given for a present or projected mirror mount, it can be determined by using the methods and formulae found in any college physics text.) The reduced (loaded) resonant frequency can be determined from the inertia ratio just calculated, the graph in Fig. 8.20, and the no-load frequency given in Table 8.4. For example, a 10-mm diameter mirror mounted on a G0606 scanner causes a reduction of the resonant frequency to 860 Hz (see Example 8.D).

The choice of a mirror diameter partly determines the resolution, as given by the scan equation (Eq. 8.11). The choice of scanner determines the total scan

Table 8.4. Specifications for General Scanning Galvanometer Scanners and Mirrors

Model	Natural Frequency, ν_0 (Hz)	Mechanical Rotation, θ	Armature Inertia, I_{arm} (g cm ²)
G0606	3200	6° (0.1 rad)	0.0016
G0612	2100	12° (0.21 rad)	0.0019
G102	2400	2° (0.03 rad)	0.0011
G108	1500	8° (0.14 rad)	0.0011
G112	1200	12° (0.21 rad)	0.0011
Mirror Diameter (mm)	Mirror Thickness (mm)	Mirror Inertia, I_m	Mirror + Mount Inertia ^a , I_m^+
5	0.5	0.00034	0.01234
7	1.0	0.0030	0.0150
10	1.0	0.0109	0.0229

^a Small mirror mount: $I_m = 0.012$ g cm².

Note: Specifications reproduced with permission of General Scanning Inc. Models 0606 and 0612 are no longer current models.

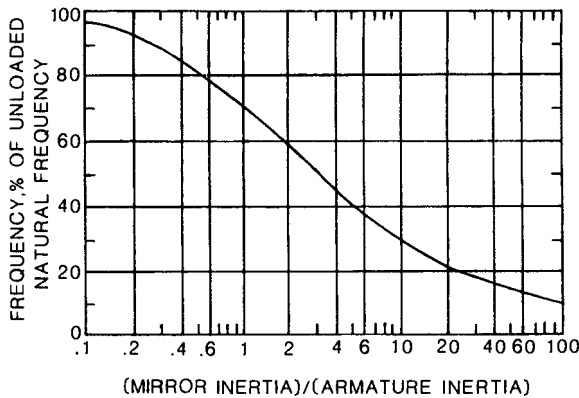


Figure 8.20. Curve for determining the loaded frequency of a General Scanning galvanometer as a function of the loading (mirror inertia/armature inertia). (Reproduced with the permission of General Scanning.)

angle ϕ , another factor in the equation. Thus another array of values can be calculated for the different mirror-scanner combinations. A 10-mm diameter mirror on the G0606 scanner has a resolution of 2590 (see Example 8.D).

Although we could tabulate all of the combinations, the most useful way to display all the data is to plot the values for the combinations, as given in Fig. 8.21. From the plot we can see that there are a number of combinations that will be useful. They range from the Model G108, 10-mm mirror with a resolution of 3600 and with a resonance at 900 Hz to the same scanner with a 5-mm mirror that provides 1800 spot resolution and a 1075 Hz resonance to the G0606

EXAMPLE 8.D. Calculate the loaded resonant frequency and resolution for a 10-mm-diameter mirror mounted on a General Scanning G0606 scanner from the specifications given in Table 8.4 and the frequency response plotted in Fig. 8.20. Assume $\lambda = 633 \text{ nm}$.

Solution: The inertia of the 10-mm diameter mirror plus mount is given as 0.0229 g cm^2 . The armature inertia for the G0606 is 0.0016 g cm^2 . Their ratio, 14.3, indicates from Fig. 8.20 that the G0606 resonant frequency of 3200 Hz will be reduced to 27% of this, or 860 Hz.

The resolution for this system is determined from $N = \phi D / a \lambda$. Here, $\phi = 2 \times 0.1 \text{ rad} = 0.2 \text{ rad}$. (Note that the scan angle is *twice* the mirror rotation angle!) $D = 10^{-2} \text{ m}$, $a = 1.22$ (see Fig. 8.9), and $\lambda = 0.633 \times 10^{-6} \text{ m}$. This gives

$$N = \frac{0.2 \times 10^{-2}}{1.22 \times 0.633 \times 10^{-6}} = 2590.$$

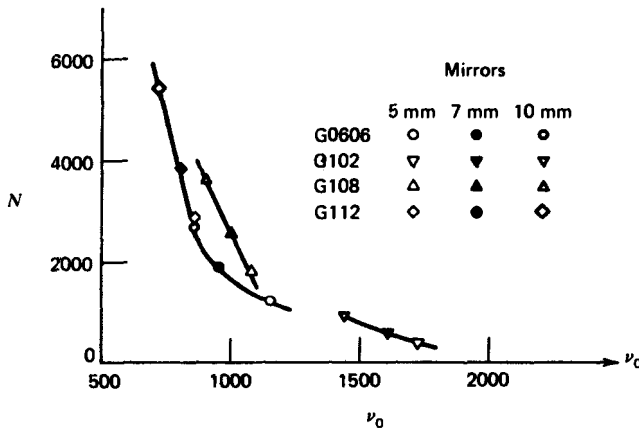


Figure 8.21. Resolution versus frequency for a number of galvanometer scanners with different mirror sizes.

model with a 960 Hz resonance and a resolution of 1800. Since we require broadband response at 900 Hz, 85 % of the resonance frequency must be greater than our required response. The only scanner capable of this is the Model G108 fitted with a 5-mm mirror. The other two scanner-mirror combinations fall below the frequency requirements for broadband operation. The calculation for the G0612 scanner has been left as an exercise [Problem 8.9].

Similar calculations need not be made for taut-band and torsion-bar scanners, since these are operated in the narrow band (resonant) modes. They are restricted to applications requiring sinusoidal scanning. Piezoelectric scanners can be used in broad band operation, but they have two drawbacks: the high voltages that are required to provide significant bends, and even "significant" bends do not amount to more than a degree or so. However, for high-frequency operation, they can approach the frequencies of the rotating polygon.

8.4. ACOUSTO-OPTICAL SCANNERS

Much of theory needed to describe an acousto-optical scanner has already been given in the section on acousto-optical modulators earlier in this chapter (Section 8.1.4). As was pointed out there, a beam sent through an acousto-optical modulator will be deflected through an angle of $2\theta_d$, where

$$2\theta_d \cong 2 \sin \theta_d = \frac{\lambda'}{\Lambda}, \quad (8.5)$$

λ is the wavelength of the light and Λ the acoustic wavelength. If we write the above equation in terms of the temporal frequency of the acoustic wave, $v_s = v_s/\Lambda$,

$$2\theta_d = \frac{\lambda}{v_s} \times v_s, \quad (8.13)$$

then it is obvious that the amount of beam deflection is directly proportional to the frequency of the ultrasonic wave, v_s . Also the lower the speed of sound in the medium, the greater the possible scan angle. Once the material is decided upon, a beam may be scanned by scanning the frequency of the ultrasonic wave. Since one can also change the amplitude of the ultrasonic wave, it would seem that an acousto-optic scanner could be used simultaneously as a scanner and a modulator. However, the conflicting requirements of response time and resolution usually do not allow this.

While acousto-optic scanning appears to be a simple and elegant method of scanning a beam of light, there are problems. Because the scanner is being used to change the beam direction, it would appear that it could only be used over a limited range since the Bragg condition could not be fulfilled for all angles. Example 8.B showed how critical beam misalignment could be. The problem is solved by manufacturing the scanner with a row of acoustic transducers attached to the side of the material rather than a single long transducer. It is then possible to introduce a frequency dependent phase difference between adjacent transducers. This results in an acoustic wave front which can be "steered" to the correct angle to deflect the input beam with maximum efficiency, as shown in Fig. 8.22.

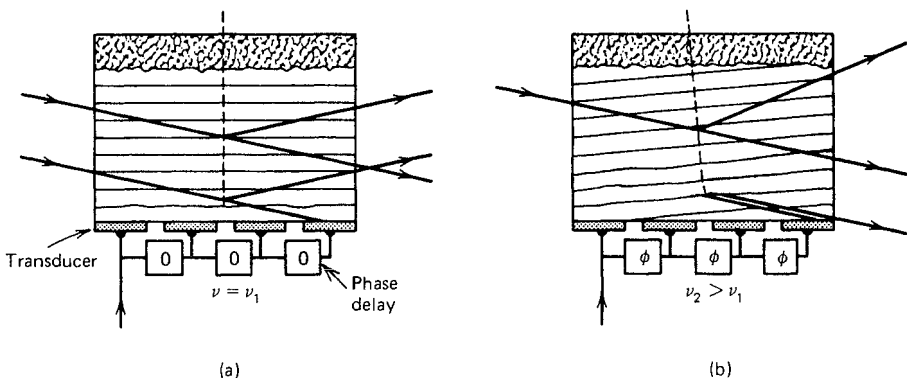


Figure 8.22. Steering of a light beam in an acousto-optic scanner. (a) incident beam and acoustic wavefronts oriented at the Bragg angle at frequency ν_1 (no phase delay imposed), (b) deflection of the same beam through a layer angle with a higher acoustic frequency ν_2 . (Phase delay between transducers to steer the acoustic wave so the Bragg condition is satisfied for the same input beam geometry as in (a)).

The magnitude of the scan is not very large. If v'_s and v_s are the maximum and minimum ultrasonic frequencies during the sweep, then the scan angle is given by

$$\phi = 2\theta'_d - 2\theta_d = \frac{\lambda}{v_s} (v'_s - v_s) = \frac{\lambda \Delta v_s}{v_s}. \quad (8.14)$$

For example, if a helium-neon beam is sent through a glass whose ultrasonic sound velocity is 6 km/sec and the acoustic wave frequency can be generated in a range of frequencies from 40 MHz to 70 MHz, the scan angle is 3.17 mrad. If a more exotic (and costly) material like tellurium dioxide is used, which has a sound velocity one tenth that of the figure given for the glass ($v_s = 0.62$ km/sec), the scan angle would be ten times larger, 32 mrad or 1.8° . There are a number of materials that are used for acousto-optic scanners and modulators. Depending on the time response, resolution, or cost requirements of the system, different materials are chosen.

Just as in the case of the other scanners, a large scan angle does not guarantee good resolution. As before we must compare the divergence of the beam with the scan angle. Using Eqs. 8.14 and 8.8,

$$N = \frac{\phi}{\Delta\phi} = \frac{\lambda}{v_s} \Delta v_s \times \frac{D}{a\lambda} = \frac{D}{v_s} \times \frac{\Delta v_s}{a}. \quad (8.15)$$

Note that the resolution is independent of wavelength, but dependent on D/v_s , which is the time τ it takes an ultrasonic wave to cross the beam. Thus the resolution can be increased by increasing the beam width in the direction of

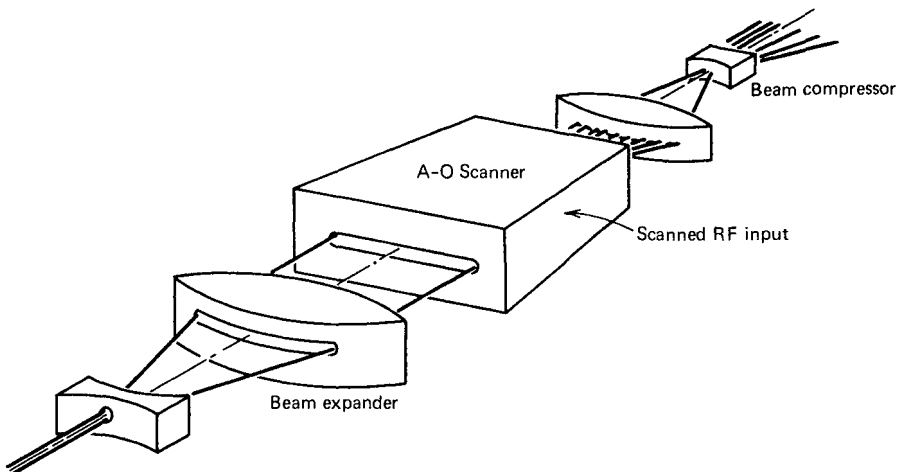


Figure 8.23. Beam expanders and compressors in an acoustic-optic deflection system.

sound propagation. Since scanners tend to be narrow perpendicular to the directions of sound and light propagation, to conserve material, space, and power, the beam is usually not expanded in both directions. Thus, a one-dimensional telescope made of cylindrical lenses (Section 7.4.3) may be used to recollimate the beam, increasing the beam diameter and thereby the transit time as shown in Fig. 8.23. Once the angular scan is imposed on the beam, it may be "down-collimated" using a similar cylindrical telescope in reverse. Now it is possible to see why scanning and modulation is not done using a single acousto-optic cell. A high frequency response is in direct conflict with the highest resolution, since the first requires a short transit time while the second requires a long one. A calculation of the resolution of an acousto-optic scanner is given in Example 8.E.

EXAMPLE 8.E. Given an acousto-optic scanner with a frequency scan from 40 MHz to 70 MHz and a sound velocity of 6 km/sec, what is the resolution:

- (a) If the beam output is directly from a helium-neon laser ($\lambda = 633 \text{ nm}$, $d_0 = 1 \text{ mm}$, $a = 1.27$)?
- (b) If a $10\times$ collimator is used?
- (c) If a $10\times$ collimator and tellurium dioxide (TeO) is used?

Solution:

- (a) $\tau = 1 \text{ mm}/6 \text{ km/sec} = 0.167 \mu\text{sec}$,
 $N = \tau \Delta v_s / a = 0.167 \mu\text{sec} \times 30 \times 10^6 \text{ sec}^{-1} / 1.27 = 3.9$.
- (b) $D = 1 \text{ mm} \times 10$. Therefore $N = 39$. Note that to get a 1000-spot resolution the material would have to be $\frac{1}{4} \text{ m}$ wide.
- (c) Since $v_s = 0.62 \text{ km/sec}$ for TeO, $\tau = 10 \text{ mm}/0.62 \text{ km/sec} = 16 \mu\text{sec}$,
 $N = 16 \times 10^{-6} \text{ sec} \times 30 \times 10^6 \text{ sec}^{-1} / 1.27 = 381$. For a resolution of 1000 the beam would be 26 mm wide.

A clever application of an acousto-optic scanner involves its use as part of a rotating prism system. One of the design restraints of a rotating prism is that the facet length must accommodate the scan length. If the facet is too small the beam will fall onto another facet before the scan is finished. But, as was noted earlier, larger facets increase the size and cost of the scanner. This can be remedied by introducing an acousto-optic scanner between the incident beam and the scanner. By synchronizing the acousto-optic scanner to the facet position it is possible to deflect the beam onto the center of the facet throughout the useable duration of the prism scan (Fig. 8.24). This is sometimes called *facet following* [Problem 8.10].

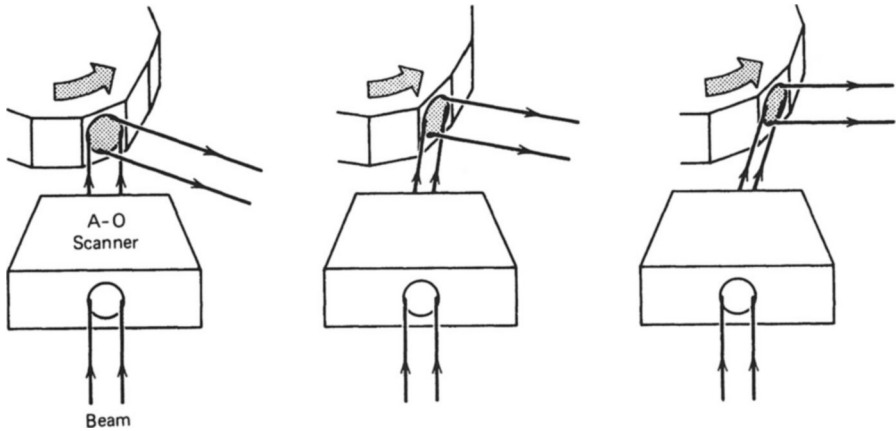


Figure 8.24. Facet following. Through the use of an acousto-optic deflector, a small beam can fully illuminate the facet of a rotating prism.

8.5. HOLOGRAPHIC SCANNERS

An alternative to a prism scanner is a scanner constructed of holograms. The holograms in this case are diffraction gratings made by interfering two laser beams in the surface of a photographic emulsion (Fig. 8.25(a)). The interference pattern that is formed in the film becomes a three-dimensional grating (Ref. 8.1), not unlike the grating formed inside an acousto-optic modulator. The efficient

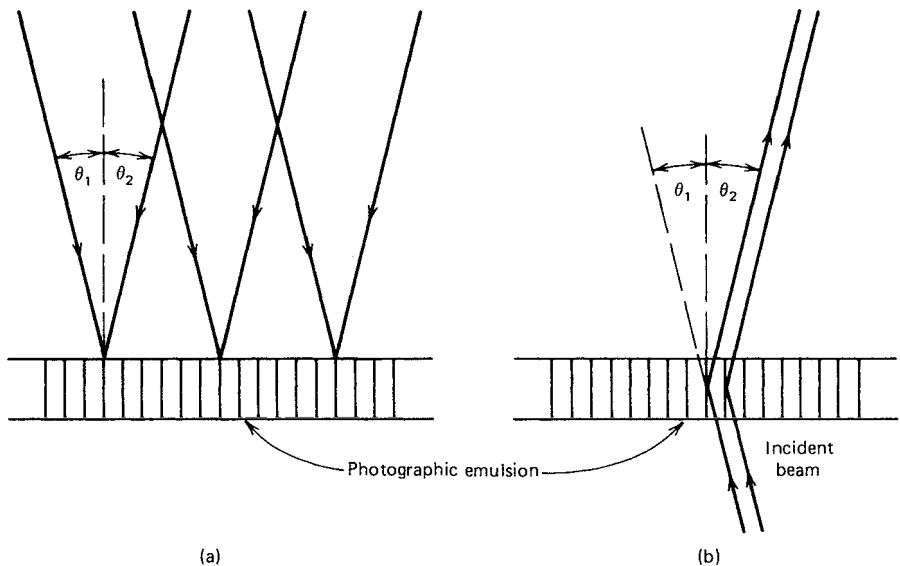


Figure 8.25. Holographic beam deflection. (a) construction of a holographic grating with two parallel beams, (b) reconstruction of one of the beams by illumination of the grating by a beam at the same angle as the other original beam. (Copyright 1982 by International Business Machines Corporation; reprinted with permission.)

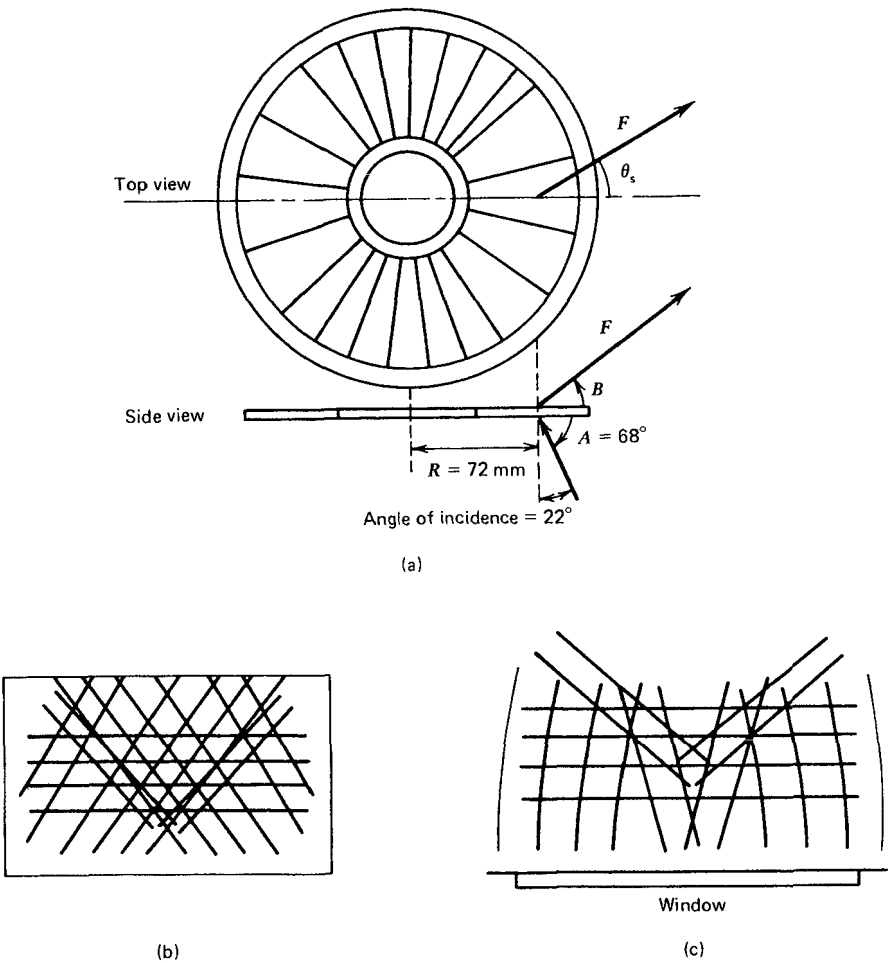


Figure 8.26. Mosaic holographic scanner. (a) Disk layout and dimensions, typical scan pattern (b) in a horizontal plane, (c) in a vertical plane. (Copyright 1982 by International Business Machines Corp.; reprinted with permission.)

diffraction of light in these gratings occurs only if the incident beam angle satisfies the Bragg condition (Fig. 8.25(b) and Eq. 8.4). If one of the beams used in construction of the hologram is a diverging beam, then the gratings not only redirect the beam but also focus the beam to a spot, so the hologram serves also as a lens. When the grating is moved, the beam is scanned through some distance.

A single hologram cannot replace a multifaceted scanner. However, a collection, or mosaic, of holograms with different orientations arranged on a circular disk, as shown in Fig. 8.26 produce a multiple scan pattern when the disk is rotated. This pattern can be considerably more varied than that of a prism scanner. It has the added feature that each of the scans can be focused to a

different point in space. The technique used to design these scanners is discussed in some detail in Ref. 8.2.

The use of holographic scanners in optical systems offers many advantages. One is the small volume of the scanner, another the ability to tailor the design to a specific application. Many tilt angles, focal lengths, and facet sizes can be easily placed on the same disk. In addition, the flexibility of the process allows changes to be incorporated in the design process. An example of one such design is a supermarket scanner, shown in Fig. 8.27, used to scan the Universal Product Code (UPC), the multiple bar pattern found on most consumer items in the United States. The beam from the laser is expanded, recollimated, and directed through a series of mirrors M_1 , M_2 , M_3 , and M_4 onto the disk at the appropriate angle of incidence. The diffracted and focused beam is redirected to the scan volume by mirrors M_5 and M_6 . The light scattered from the multiple-bar product code on the item is collected by the scanning facet and mirrors M_5 and M_6 and directed onto a detector by a focusing lens placed behind the input mirror M_3 . The light pattern detected by the scanning system is decoded to provide the cash register computer with the correct UPC number.

Because of the initial complexity of designing and fabricating a scanner for a particular application, a holographic scanner is not the answer to all scanner designs. Provided there is sufficient volume to justify the initial design and production costs, the holographic scanner may be the best choice. Once a master has been made and tested, the replication of additional copies is

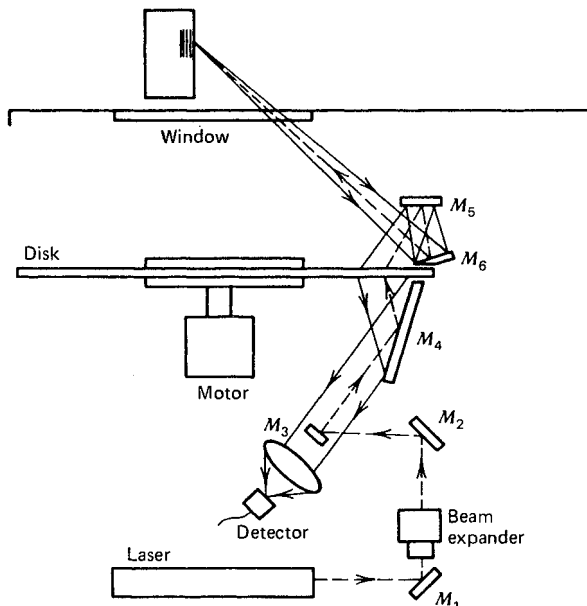


Figure 8.27. Supermarket scanner optical design. (Copyright 1982 by International Business Machines Corporation; reprinted with permission.)

relatively easy, since a replica is little more than a contact print of the original hologram. If there is not enough volume to justify its design, then a prism scanner is probably the best choice for repetitive, periodic scans.

8.6. ADDITIONAL TOPICS

Having listed the specific types of scanners and noted their characteristics, it is now possible to take a look at some topics that apply to most of these scanners. First, it is useful to compare and contrast the types of scanners we have discussed. In Table 8.5 we have listed the scanners we have discussed. Please note there is no attempted ranking of scanners according to resolution since this parameter is the result of a trade off between the design parameters, as we have seen in the case of the galvanometer scanners in Section 8.3.2. There is no magic choice of scanner for a particular application. The requirements of the individual design will limit (in some cases, force) one's choice of scanner.

8.6.1. Scan Magnification and Relay

For systems with relatively small deflections, a substantial scan length can be achieved by putting the target far away from the scanner (Fig. 8.28(a)). This gives rise to the dual problems of maintaining stability with long optical lever arms and constructing a workable system in a reasonable volume. While mirrors and heavy struts may accomplish the job, another approach is to increase the scan angle by adding additional scanners operating in synchronism with the initial scanner. For example, keeping in mind the Bragg condition for acousto-optic modulators, it is possible to magnify the deflection angle of the first modulator, by sending it into a second deflector that includes additional phase compensation to account for a changing input angle (Fig. 8.28(b)). Another method uses ganged mirror scanners (Fig. 8.28(c)). Note that each successive scanner must have larger mirrors, so that the speed of the entire system is governed by the last stage of magnification. There are many other

Table 8.5. Comparison of Scanners

	Galvanometer	Rotating Prism ^a	Taut-band	Piezoelectric	Acousto- optic
Scan angle	12°	90°	30°	2°	< 1°
Approximate maximum frequency (MHz)	0.010	0.05	0.001	0.50	20
Nature of scan	Random	Periodic	Harmonic	Harmonic	Random

^a Also holographic scanners.

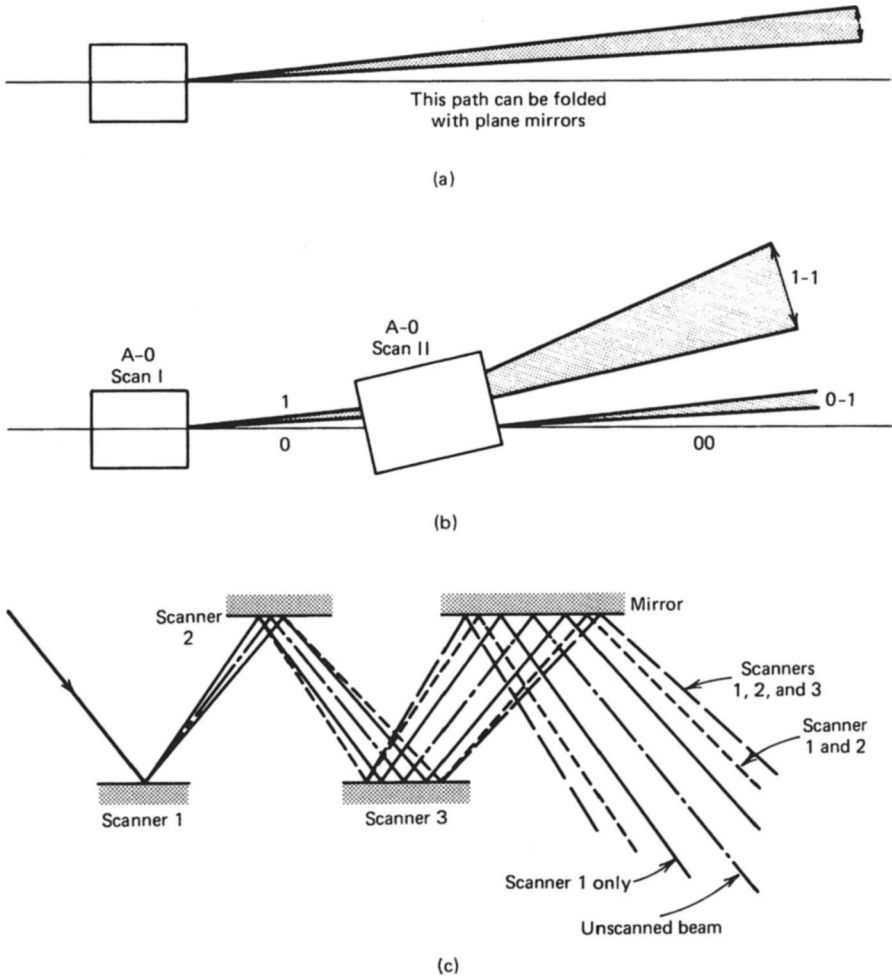


Figure 8.28. Scan magnification. (a) Increase scan length by putting target at a large distance from the scanner. (b) Ganged acousto-optic deflectors. (c) Ganged mirror scanners.

systems for scan magnification including the use of clever confocal mirror systems. The reader is referred to the references, especially Ref. 8.3, for specific examples.

In certain cases the problem is not really one of magnification, but translation of the center of deflection to a particular point in space (Fig. 8.29). This can be accomplished by a two-element relay system that translates the effective center of deflection a distance equal to four times the focal length of the imaging lens. The field lens is used to recollimate the beam after it had been focused by the first lens. By using different kinds of lenses in an afocal system a relay can be combined with scan magnification [Problem 8.11].

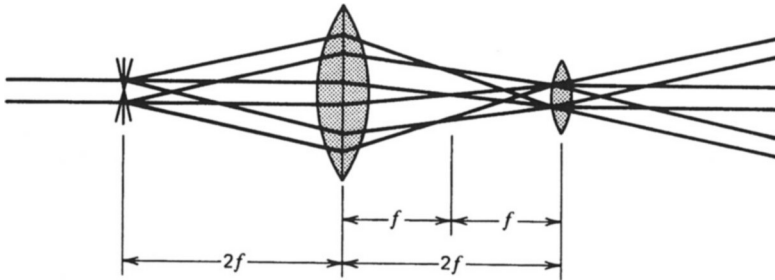


Figure 8.29. A scan relay.

8.6.2. Pattern Generation

It would be impossible to describe or list all of the various patterns that scanning systems can generate. Some are fairly simple to understand, others are highly elaborate and are generated by patented components. Beyond the single line scans as shown in the previous sections, we will look at two relatively simple patterns, a cross pattern and a double cross pattern.

In Fig. 8.30 we show the top and side view of a rotating prism scanner. The perspective and scan distances have been distorted for clarity. The mirrors (*A* and *B*) on either side of the lens are tilted upwards to direct the reflected rays to a common region above the plane of the prism so that the line scans cross in the focal region of both rays. Since the mirrors are at an angle with each other, the rays cross giving a single “X” as one facet illuminates one mirror and then the other.

If a tetrahedron is rotated about one of its threefold axes, in a square array of mirrors, as shown in Fig. 8.31(a), a double cross scan will result. The tetrahedron is irregular in that it is squashed to give 45°-angle mirrors with respect to the laser beam that illuminates all three mirrors from below. Note that any one facet and any two mirrors will form a single cross in a manner just described above. Thus each pair of mirrors properly oriented will form a cross. While it is possible to tilt the mirrors to give a common crossing point and a single “X,” usually they are tilted in such a way that two “X”s forming an overlapping double cross pattern in the focal region. By putting mirrors on the polygon at different angles (say 40°, 45°, and 50° mirrors), the complexity of the pattern can be increased to six crosses. Because the beam is illuminating all three facets, two or three of the mirrors are producing a scan line at any one time. Figs. 8.31(b) and (c) show a top view and illumination schedule for this system. A designer must consider whether simultaneous scans are the best method for his system. In some cases, the system could lead to ambiguous signals if more than one scan beam were picked up by a sensor. For one solution to this problem, see Section 10.7 on retrocollective geometry.

Multiple line patterns can be achieved by increasing the number of facets and putting a different tilt on each facet and then tilting each folding mirror by a

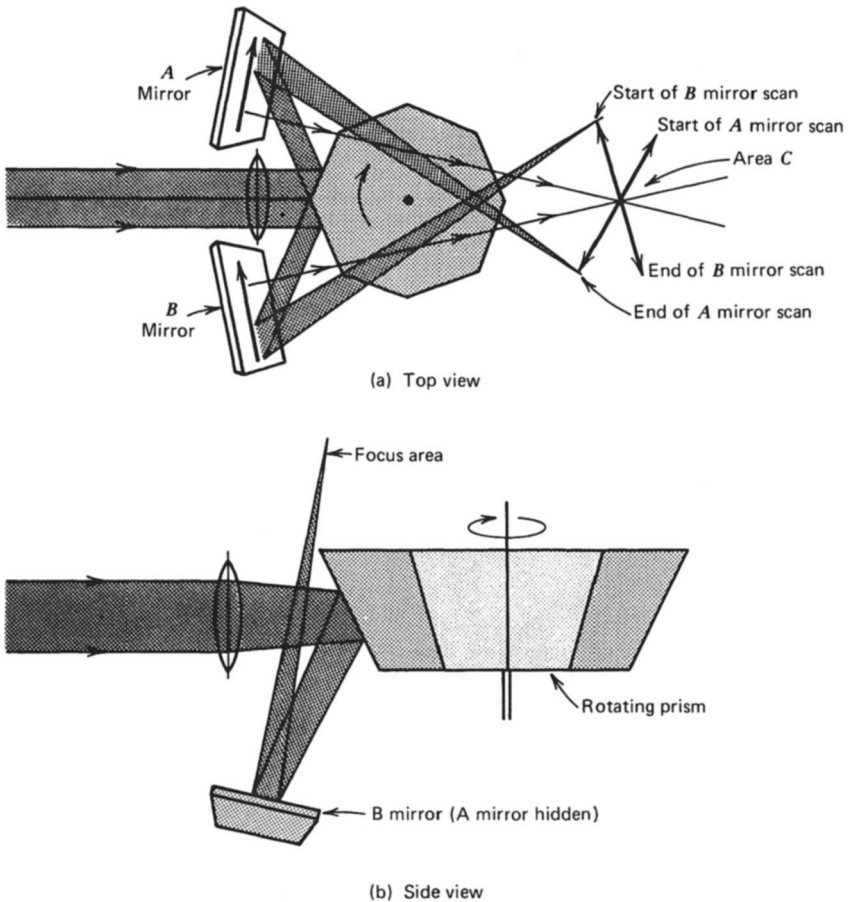


Figure 8.30. Scan pattern generated from a rotating polygonal prism.

different amount. The result, however, may resemble a spinning mirror ball in a dance hall more than an optical design element. And then, how does one interpret the resulting scan pattern?

8.6.3. Distortion

In many cases the scan velocity of a beam is constant with time. The information being imposed on the scanning beam is usually produced in a linear data stream (i.e., data rate is constant with time). Most lenses are corrected for distortion, so that the magnification is constant for all off-axis points. But most x - y scanners have a built in pincushion distortion if the scan image is projected on a plane rather than the interior of a sphere. Scanning lenses can be designed to include compensating barrel distortion, so that the resulting combination gives a scan

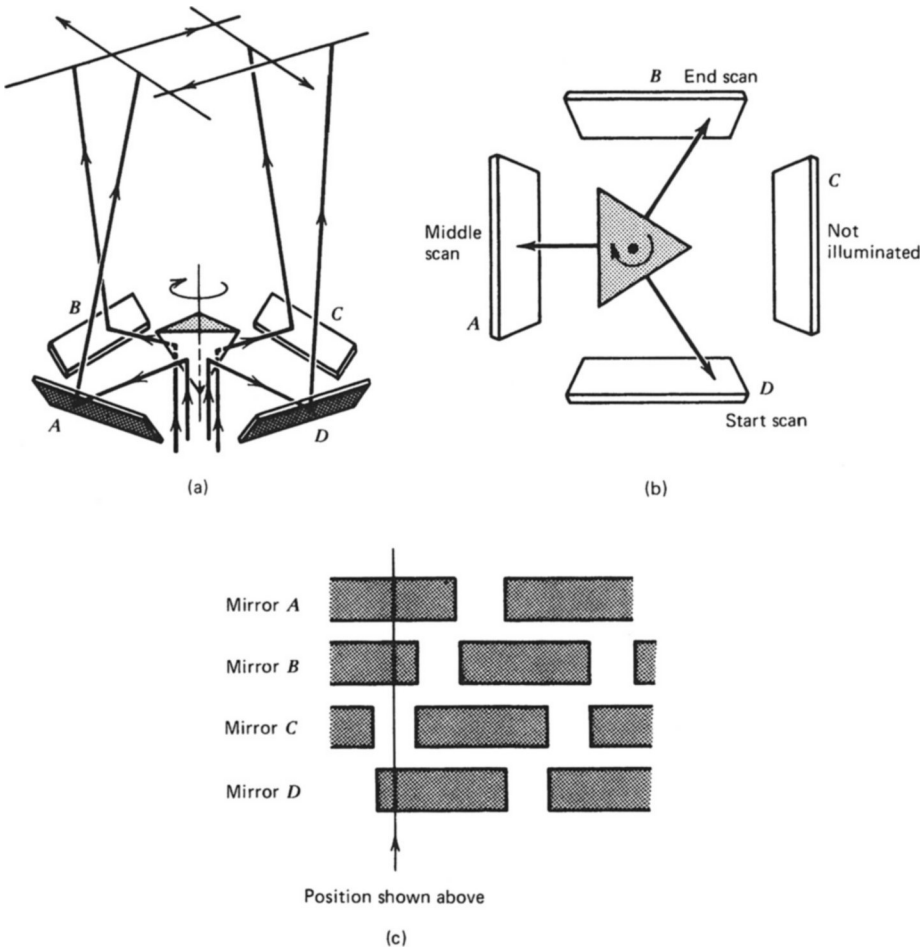


Figure 8.31. Scan pattern generated from a rotating tetragonal prism. (a) side view, (b) top view, (c) mirror illumination sequence.

line that is proportional to the scan angle. Thus scanning lenses have unusual characteristics making them undesirable for any application except the one for which they were designed.

8.6.4. A Scanning System for TV Display

One application that combines both acousto-optical and mechanical scanning is a *laser television display*. In Fig. 8.32 is shown one possible optical train for such a display. The beam from the laser is focused into (lens L_1) an acousto-optic modulator (A-OM) to increase the rise time of the modulator and provide better contrast. Rather than completely collimate the beam before the next stage,

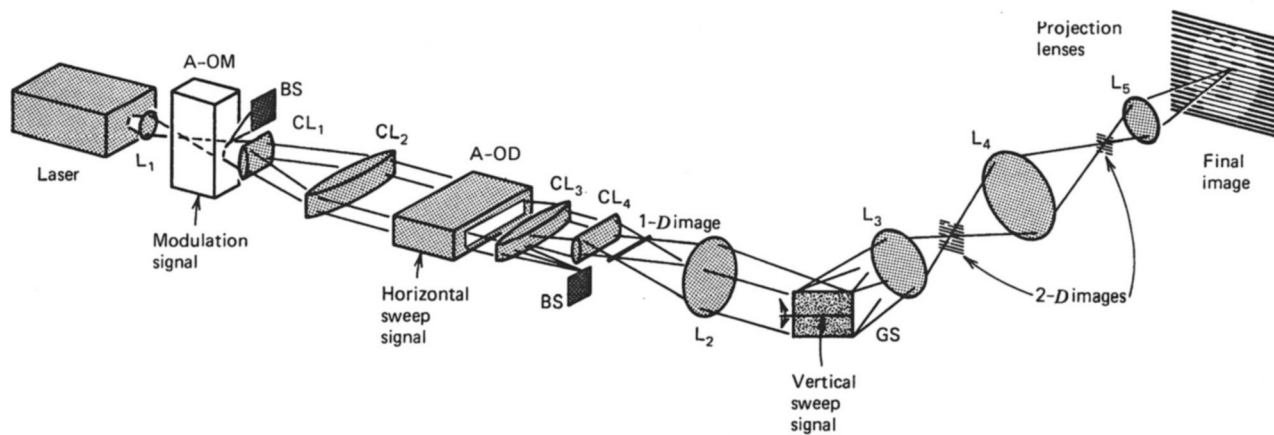


Figure 8.32. Laser television display. L—spherical lenses, CL—cylindrical lenses, A-OM—acousto-optic modulator, A-OD—acousto-optic deflector, GS—galvanometer scanner, BS—beam stop.

horizontal deflection, the cylindrical lens (CL_1) is used to return the collimation back to the original beam divergence in the vertical plane. Since CL_1 has no power in the horizontal direction the beam continues to expand to the width needed to achieve the required resolution (as indicated in Section 8.4, Eq. 8.14) in the horizontal direction. The large cylindrical lens (CL_2) provides the collimation before the beam enters the acousto-optical deflector (A-OD). Because of the different divergences in the two planes, two cylindrical lenses are needed to focus in the vertical beam dimension (CL_4) and to image the horizontal scan pattern (CL_3) for the zero and first orders of the deflector. A beam stop removes the zero order while the line that contains the horizontal scanning information is collimated by a spherical lens (L_2) before illuminating the mirror of the galvanometer scanner (GS). The vertically deflected light is then focused by L_3 to give a small two-dimensional television display. Two additional lenses (L_4 and L_5) act as a projection lens to produce a large two dimensional display of the raster pattern. By clever design some of these components can be combined to produce a more compact system. Additional refinements can be made by adding more components.

PROBLEMS

- 8.1. Describe how an electro-optic modulator can be biased to the linear operating point of the modulator by passing the radiation through a polarizer and quarter-wave plate combination.
- 8.2. Show that the expression for the diffraction angle θ_d (Eq. 8.5) is correct both inside and outside the acousto-optic modulator provided the wavelength of the light in the medium in which θ_d is measured is used in the calculation.
- 8.3. A helium-cadmium laser ($\lambda = 442$ nm) fully illuminates a rectangular scanning mirror whose width in the scan direction is 5 mm. If the maximum rotation angle of the mirror is 5° , what is the maximum resolution one could hope to achieve with this system?
- 8.4. Sometimes it is impossible to place a scanning mirror in a particular point but a beam must be scanned at that point. Using ray sketching, show that the center of rotation of a mirror can be transferred to another point in a system. What happens to the angular rotation of the mirror as the object and image distances are changed?
- 8.5. A TEM_{00} mode argon ion laser beam ($\lambda = 488$ nm) is focused to a 175- μ m beam waist spot on a scanning mirror. The scanned beam is to be relayed to an inaccessible point with no change in the beam waist or the original rotation angle of 5.7° . What must be the focal length and location of the lenses? Also sketch the system for maximum deflections.

- 8.6. A 12-sided polygonal prism is to be designed to provide a 1-in.-square, 12-line raster pattern with a helium-neon laser beam ($\theta = 1 \text{ mrad}$). The spot size at the screen is to be $100 \mu\text{m}$. If the scanner is operated in a telecentric mode, what should be the call out for pitch angles of the faces to produce the raster? What fraction of the total time that the beam is illuminating one face is needed to produce one line of scan?
- 8.7. Show that the maximum movement of the center of deflection along incident beam for a prism of radius R is given by Eq. 8.12.
- 8.8. Draw the path of the center of deflection by carefully drawing an octagon whose center to face distance is 50 mm.
- (a) The incident beam direction intersects the rotation axis of the octagon.
 - (b) The incident beam direction misses the center of rotation by 20 mm.
- 8.9. Calculate the resolution and loaded resonance frequency for the Series G0612 scan for each mirror and plot on a copy of Fig. 8.21.
- 8.10. Design a facet follower for a rotating octagon whose center-to-face distance is 20 mm. The octagon is illuminated by a helium–neon laser. Include assessments of the maximum scan angle, minimum scan angle. Calculate the resolution of your system. The lowest-sound-velocity material available to you is tellurium dioxide and the highest acoustic frequency change Δv_s is 40 MHz. You may not include in your resolution figure any part of the beam that would reflect back into the deflector. The location of the deflector relative to the prism axis and the beam direction must also be specified. Attempt to optimize this for maximum resolution. You will have to make reasonable assumptions on any requirements not specified above.
- 8.11. Design an afocal system of lenses to translate the center of deflection and magnify the scan by a factor of four. In terms of the initial input lens, give the total translation of the center of the deflection and the placement and focal lengths of all the lenses required for the task.

REFERENCES

- 8.1. D. C. O'Shea, W. R. Callen, and W. T. Rhodes, *An Introduction to Lasers and Their Applications* (Addison-Wesley, Reading, Mass., 1977).
- 8.2. L. D. Dickson, G. T. Sincerbox, and A. D. Wolfheimer, "Holography in the IBM 3687 Supermarket Scanner," *IBM Journal of Research and Development* **26**, 228–234 (1982).
- 8.3. L. Beiser, "Laser Scanning Systems," *Laser Applications*, Monte Ross, Ed. (Academic Press, New York, 1974).

9

Spectrometers

(HOW TO FIND OUT WHAT YOU'VE GOT)

In many optical systems the light must not only be collected and imaged as perfectly as possible, but also analyzed into its colors. Spectroscopy, in most of its forms is the physical separation of light according to its wavelengths. Some examples of this are:

1. The analysis of the origin or content of a light source as in forensic analysis and astronomy.
2. The analysis of a material by absorption spectroscopy.
3. The detection of a spectral signature in the presence of other radiation, as in the infrared spectral analysis of croplands.

The purpose of this chapter is to show some of the methods that can be used to analyze spectrally the radiation in an optical system. This is useful not only for the techniques themselves, but also because spectrometers represent somewhat more complicated optical systems than we have considered thus far. Many optical systems incorporate some spectral analysis components and some are built specifically for spectral analysis. Most of these systems can be classified on the basis of the component used to resolve the light into its components: filter, prism, or diffraction grating. We will not cover spectroscopic techniques that employ multiple beam interference, such as Michelson and Fabry–Pérot interferometers. However, these techniques can be used to construct extremely high resolution spectrometers (Ref. 9.1).

9.1. FILTER SPECTROMETERS

One of the simplest and least expensive methods of isolating a spectral band is a filter. While it can be as simple as a piece of colored glass or as elaborate as a multilayer dielectric filter, it is always compact and easy to mount. For this reason, a spectrometer built around a filter or series of filters can be a very attractive solution for certain spectral analysis problems.

9.1.1. Bandwidths and Optical Density

The width of the band of color passed by a filter or other spectrum resolver, can be expressed either in terms of the band of wavelengths or frequencies passed by this element. We start with the relation between frequency and wavelength,

$$\lambda\nu = v, \quad (1.3)$$

where v is the velocity of light in the medium. The relations between wavelength and frequency bandwidths, assuming that $\Delta\lambda$ and $\Delta\nu$ are sufficiently small, are

$$\Delta\nu = -\frac{v}{\lambda^2} \Delta\lambda \quad \text{and} \quad \Delta\lambda = -\frac{v}{\nu^2} \Delta\nu. \quad (9.1a,b)$$

On substituting Eq. 1.3 into Eq. 9.1a, one obtains the relationship

$$\frac{\Delta\nu}{\nu} = -\frac{\Delta\lambda}{\lambda}, \quad (9.2)$$

showing that the fractional bandwidth is the same, whether expressed in terms of wavelength or frequency. The negative sign only indicates that $\Delta\nu = \nu_1 - \nu_2$ is negative when $\Delta\lambda = \lambda_1 - \lambda_2$ is positive. All of the above assumes that $\Delta\nu$ or $\Delta\lambda$ is small compared to ν or λ [Problem 9.1].

Just as in electronics, one can think of a light filter in terms of the band of frequencies that transmits. A high-pass filter will pass higher frequencies (shorter wavelengths); for the visible region this would be a blue or blue-green filter, whose transmission curve is shown in Fig. 9.1(a). A low-pass filter passes lower frequencies (longer wavelengths), and in the visible region this would be a red or yellow filter (Fig. 9.1(b)). A band-pass filter passes a narrow region of the spectrum around some central wavelength (Fig. 9.1(c)).

Although many transmission curves of filters and spectrometers are plotted as the percent transmission versus wavelength, some applications require that the transmission be plotted over several orders of magnitude. A logarithmic scale in units of *optical density* is used in such cases. The optical density of a material is defined as

$$D = -\log_{10} \tau. \quad (9.3)$$

where τ is the transmission of the material. When the transmission is unity, then its log is zero and the optical density is zero. If the transmission is reduced to 10% ($\tau = 0.1$), the log will be -1 and the optical density will be 1. For each power of 10 reduction in the transmission, the optical density increases by one unit. Thus if only one thousandth of the incident light is transmitted by a photographic emulsion, for example, its optical density would be 3 [Problem 9.2].

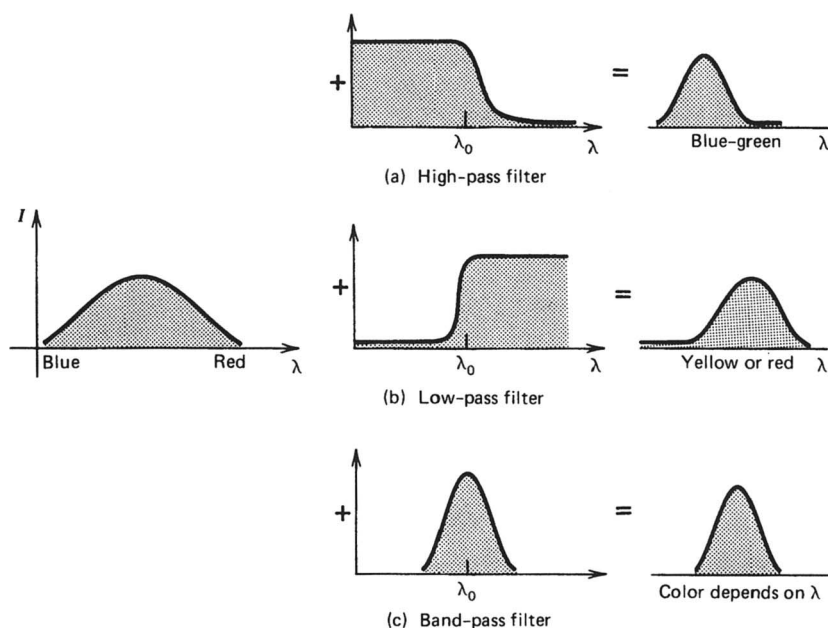


Figure 9.1. Transmission curve for filters. (a) High-pass filter, (b) Low-pass filter, (c) Band-pass filter. The characteristic wavelength, λ_0 , locates the cut-off, cut-on, or center wavelength of the filter, respectively.

9.1.2. Absorption Filters

Any absorbing material dissolved in a medium can be used as an absorption filter. The gelatin sheet used in theatre lighting is an example. However, such filters are not stable over a long time and can deteriorate under high illumination levels (sometimes rather spectacularly!).

In most optical designs, the filters are made of glass since glass filters are stable and well characterized with standard absorption curves. A set of these filters mounted on a rotating wheel can be used to measure the transmission characteristics of samples. Other designs incorporate an array of filters in front of several detectors for simultaneous measurement.

In some cases the filter can be made part of another optical element. For example, the plastic focusing lens in a laser scanner was made with a plastic that had been dyed red. Thus, the lens served as a low-pass filter, absorbing most of the ambient light but allowing the lens to transmit and focus the red light of a helium–neon laser ($\lambda = 633 \text{ nm}$) onto a detector.

9.1.3. Dielectric Filters

As was noted earlier, the deposition of a thin layer of a dielectric can sharply change the reflective properties of a substrate. The effect is even more dramatic

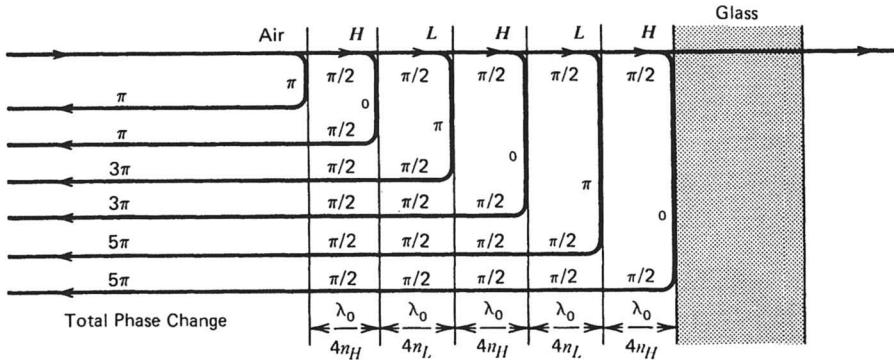


Figure 9.2. Structure of a dielectric mirror. All layers have an optical thickness of a quarter of a wavelength. The alternating structure of high and low refractive indices provides the additional phase changes to give constructive interference between all reflected rays.

when a series of layers of alternating high- and low-refractive-index materials are deposited on a surface.

9.1.3.1. Dielectric Multilayer Reflector. In Fig. 9.2, the structure of a multi-layer dielectric coating and the phase changes due to reflection and transmission at each interface are illustrated. This figure is similar to Fig. 4.9 in Section 4.2.2. The reader is referred to that section for a more detailed discussion. An examination of the structure shows that at the design wavelength, λ_0 , the phase changes from alternating interfaces are 180° out of phase with each other. Since all of the layers have an optical thickness of a quarter of the design wavelength, two passages through a layer adds an additional 180° phase change. The combination of interface phase changes and optical path-length phase changes produces constructive interference between all reflecting rays and creates a mirror of extremely high reflectivity and very low absorption at λ_0 .

What happens away from the design wavelength? At half the design wavelength, for example, all of the layers will cause a 180° phase shift, and since a reflected wave passes through a layer twice, all reflected waves are shifted by a multiple of 360° and therefore there is no net phase change except that occurring upon reflection from the interface. Since the refractive indices of the layers alternate, reflections from successive interfaces are out of phase with one another causing destructive interference. So at $\lambda_0/2$ the reflectivity is small compared to that at the design wavelength. A plot of the reflectivity as a function of wavelength is shown in Fig. 9.3. Note that there are regions of the spectrum above and below λ_0 where the multilayer dielectric coating acts as a high pass or a low pass filter, depending on the band of incident wavelengths. Filters can be designed to provide just such properties throughout the visible spectrum and beyond. Only the absorption of the materials and the substrate limit the design.

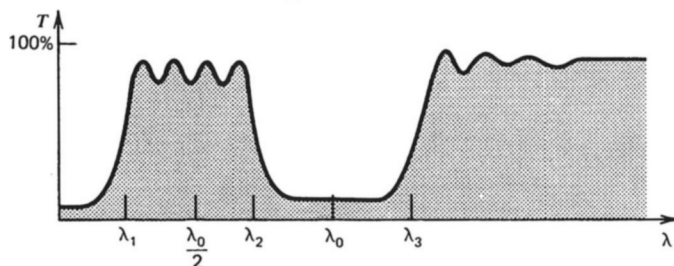


Figure 9.3. Transmission as a function of wavelength for a dielectric mirror. The mirror was designed for high reflectivity at λ_0 . At wavelengths other than λ_0 , this multilayer can act as a filter that passes high frequencies (for wavelengths in the vicinity of λ_2) or passes low frequencies (for wavelengths in the vicinity of λ_1 or λ_3).

9.1.3.2. Fabry–Pérot and Multiple-Cavity Filters. A *Fabry–Pérot spectrometer* is an optical instrument that can be used as a variable band-pass filter or a spectrometer. It consists of two highly reflective mirrors parallel to each other and separated by a distance, d (Fig. 9.4(a)). This arrangement forms a resonant optical cavity that transmits specific frequencies and rejects all the rest (Ref. 9.1). The condition for resonance is that the mirror separation be an integral number of half wavelengths. The transmission of such a filter is shown in Fig. 9.4(b). The frequency difference between adjacent transmission bands can be shown to be [Problem 9.3]

$$\Delta\sigma = \frac{c}{2d}, \quad (9.4)$$

where c = speed of light.

A dielectric filter similar to a Fabry–Pérot interferometer can be constructed of two highly reflective multilayer mirrors, like those described above, separated by a thick dielectric layer, as shown in Fig. 9.5(a). A somewhat different structure is made up of a set of two or three smaller cavities as shown in Fig. 9.5(b). The performance of these filters is dependent on the number of layers and

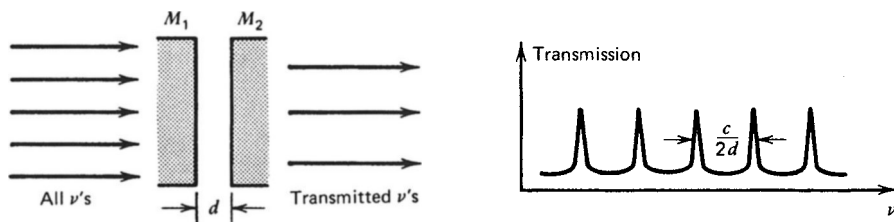


Figure 9.4. Fabry–Pérot filter. (a) Consists of two highly reflective mirrors separated by a distance d and aligned parallel to each other. (b) Transmission of light through a Fabry–Pérot filter. The transmission occurs at frequencies spaced $c/2d$ apart.

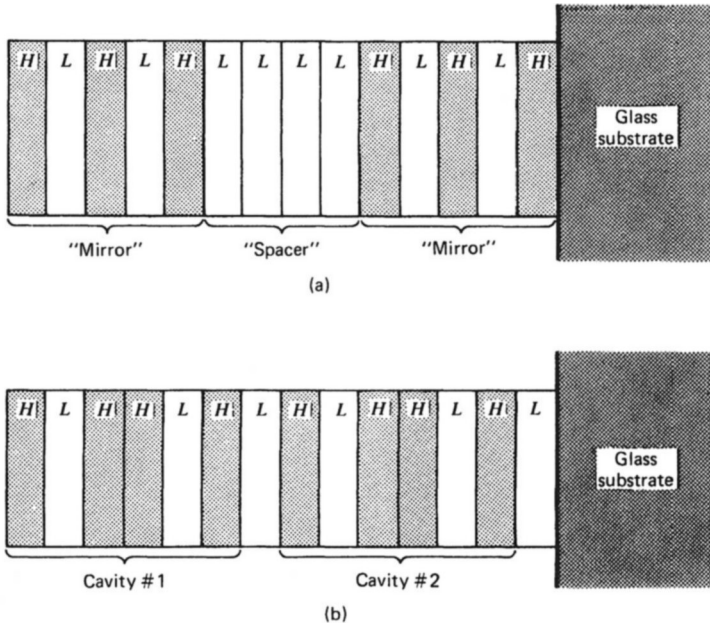


Figure 9.5. Structure of band-pass dielectric filters: (a) single cavity Fabry-Perot type; (b) multiple cavity type. Each layer is one quarter wavelength optical thickness ($nd = \lambda_0/4$). H —high refractive index; L —low refractive index.

their arrangement. The pass bands can be as small as 0.1 nm with peak transmittances above 50%. In addition to these values, a designer should be aware of the amount of rejection of unwanted radiation away from the pass band. Some multilayer dielectric filters are mounted on glass filters to block radiation from neighboring transmission bands.

Even though these filters are designed to transmit about a particular wavelength, it is possible to “tune” a filter by tilting it with respect to the direction of incident light. The center of the pass band can be shifted to shorter wavelengths: from λ_0 to $\lambda' = \lambda_0 \cos \theta$, where θ is the angle between the filter normal and the direction of light propagation [Problem 9.4]. It should be

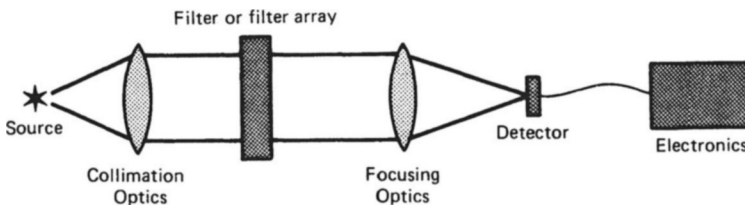


Figure 9.6. A simple filter spectrometer.

pointed out that although this procedure is convenient for small tilts, there is a considerable reflection loss at higher angles of incidence.

Pictured in Fig. 9.6 is a spectrometer built around a set of filters. The spectrometer could be used to investigate either a source or the transmission of a sample. Although this design would have limited resolution, it could be an inexpensive solution to a spectral analysis problem that did not require much data. Dielectric filters could also be used to scan a narrow band of wavelengths by repeatedly tilting the filter through a range of angles corresponding to the range of wavelengths to be scanned.

9.2. PRISM SPECTROMETERS

The basic design for a prism spectrometer can be illustrated by Figure 2 from Newton's *Opticks* published in 1704. It is reproduced in Fig. 9.7. The spectrometer contains a source, the sun; a dispersing element, the prism; and a detector, the eye. All prism spectrometers are variations on this arrangement.

9.2.1. The Dispersion of a Prism

The passage of light through a prism deviates the beam by an angle δ . If the light is monochromatic, then δ will only have one value, but if the beam contains many wavelengths, the beam will be dispersed into its colors. For, as was shown in Chapter 1, the dispersion of glass, the variation of refractive index with wavelength, causes light falling at an angle on an air-glass interface to be separated into its colors. A calculation of this separation for a flint glass (670472) was given in Example 1.A.

In that calculation and in succeeding work on achromatization of lenses, the dispersion was given in terms of the V -number, defined in Eq. 1.5. The dispersion can also be written as the change in refractive index as a function of wavelength, $dn/d\lambda$. To be able to find this quantity would require a detailed knowledge of the refractive index function, $n(\lambda)$. However, $dn/d\lambda$ can often be approximated by

$$\frac{\Delta n}{\Delta \lambda} = \frac{n_F - n_C}{\lambda_F - \lambda_C}, \quad (9.5)$$

which can be calculated from the V -number as shown in Example 9.A.

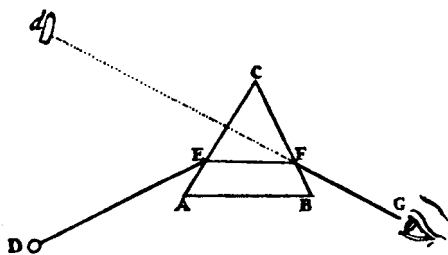


Figure 9.7. Figure 2 from Newton's *Opticks*, published in 1704, contains all of the basic components of a spectrometer: source, dispersing element, and detector. The dispersion is indicated by the elongated image of the sun representing its spectrum.

EXAMPLE 9.A. A flint glass (670472) (Table 1.1) prism with an apex angle of 30° (Example 1.A and Problem 1.5) is used in a spectrometer at minimum deviation ($\delta = 56^\circ 50'$). Calculate the angular dispersion and resolution at the center of the visible spectrum if the base length of the prism is 60 mm.

Solution: Because data (Table 1.1) is available we can do somewhat better than the Eq. 9.5. It is useful to compare the approximate value with a more accurate one.

From Eqs. 9.5 and 1.5

$$\frac{\Delta n}{\Delta \lambda} = \frac{n_F - n_C}{\lambda_F - \lambda_C}, \quad \frac{1}{V} = \frac{n_F - n_C}{n_d - 1}$$

$$\frac{\Delta n}{\Delta \lambda} = \frac{n_d - 1}{V(\lambda_F - \lambda_C)}$$

$$= \frac{0.67}{47.2(0.4861 - 0.6563)}$$

$$= -0.083 \mu\text{m}^{-1}.$$

Using Table 1.1 in the vicinity of 550 nm (*e* line at 546.1 nm),

$$\frac{\Delta n}{\Delta \lambda} = \frac{n_e - n_d}{\lambda_e - \lambda_d} = \frac{1.67341 - 1.67003}{0.5461 - 0.5876} = \frac{0.00428}{-0.0415}$$

$$= -0.103 \mu\text{m}^{-1},$$

$$\frac{\Delta n}{\Delta \lambda} = \frac{n_F - n_e}{\lambda_F - \lambda_e} = \frac{1.68000 - 1.67431}{0.4861 - 0.5461} = \frac{0.00569}{-0.0600}$$

$$= -0.095 \mu\text{m}^{-1}.$$

As can be seen from these calculations, the dispersion changes considerably across the visible for some glasses. Eq. 9.5 is only an approximation. The average of the dispersion values is 0.099, suggesting $dn/d\lambda$ be approximated by $0.1 \mu\text{m}^{-1}$.

The angular dispersion is

$$\frac{d\delta}{d\lambda} = \frac{\sin 30^\circ \cdot 0.1 \mu\text{m}^{-1}}{\cos 56.84^\circ \cos 56.84^\circ} = 0.167 \text{ rad}/\mu\text{m}^{-1}.$$

The resolution for the prism is

$$R = \frac{dn}{d\lambda} \cdot B = 0.1 \mu\text{m}^{-1} \times 60 \text{ mm} = 6000.$$

The ease of separation of light into its colors by a prism is measured by the angular dispersion of the prism—the rate at which the final refraction angle changes with wavelength, $d\delta/d\lambda$. This quantity is made up of two contributions, one due to geometry, the other to the material dispersion just described, that is,

$$\frac{d\delta}{d\lambda} = \frac{d\delta}{dn} \frac{dn}{d\lambda}. \quad (9.6)$$

It can be shown that when the deviation δ is a minimum that $\gamma = \varepsilon$ in Fig. 9.8 and using the geometry illustrated in Fig. 9.8 and Snell's Law [Problem 9.5] that the geometrical contribution has the form

$$\frac{d\delta}{dn} = \frac{\sin \alpha}{\cos \gamma \cos \delta}, \quad (9.7)$$

where α is the apex angle of the prism and γ is the angle of refraction. Therefore, the angular dispersion for a prism is

$$\frac{d\delta}{d\lambda} = \frac{\sin \alpha}{\cos \gamma \cos \delta} \cdot \frac{dn}{d\lambda}. \quad (9.8)$$

9.2.2. A Prism Spectroscope

In order to sort light into its various wavelengths, all rays from a source must be incident upon the prism face at the same angle, that is, the incident light must be collimated. If the incident light were to have a range of angles of incidence, there would be no unique angle of refraction for each wavelength and the purpose of the spectroscope would be defeated.

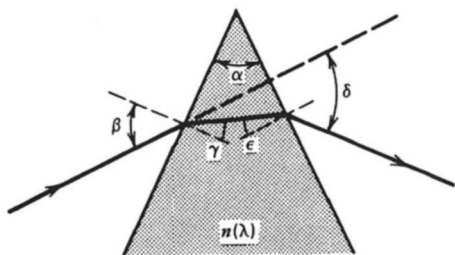


Figure 9.8. The geometry of refraction of light by a prism. The angle α between the two faces is called the apex angle.

In most spectrometers these problems are solved by imaging the source onto a narrow slit and placing a lens one focal length away from the slit to collimate the emerging light before it passes through the prism. A sketch of a simple prism spectrograph is shown in Fig. 9.9. The collimating lens then provides a beam of light whose angle of incidence is constant across the prism so that the refraction angle is the same for all rays of a given color. On the other side of the prism another lens focuses the refracted beams. Because each color is incident at a different angle to this lens, each is focused to a unique location in the focal plane of the lens, forming monochromatic images of the slit. The slit is chosen because its image for one color will not overlap with images due to other colors. Note that if the focusing lens is not achromatic, the focal length will vary with wavelength and the image plane must be tilted to provide a focused image at all wavelengths [Problem 9.6].

The ultimate limitation on the resolution of any spectrometer (assuming a sufficiently narrow slit) is the diffraction of the light by one of the elements. If we assume all of the lenses in the system are larger than the prism face, then the resolution is determined by the size of the prism. In spectroscopy the resolving power is defined as the ratio of the wavelength being observed to the smallest wavelength difference that can be resolved:

$$R = \frac{\lambda}{\Delta\lambda}. \quad (9.9)$$

As before we use the Rayleigh criterion as a convenient means of determining the resolution of a prism instrument. It is argued that if two wavelengths are deflected by angles that differ by less than the diffraction angle due to light passing through the prism, they cannot be resolved. Thus we need to equate the difference in deflection angle $\Delta\delta$ to the diffraction angle θ to derive an expression for the resolution.

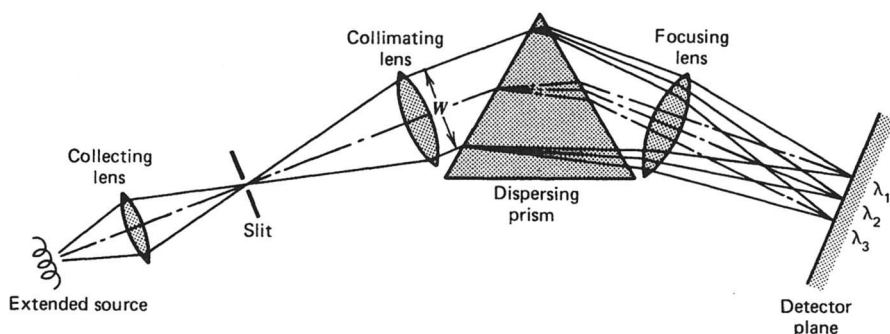


Figure 9.9. A simple prism spectroscope. The source is imaged onto a narrow slit and the emerging light is collimated by a lens. The refracted beams, emerging at different angles are focused by a second lens to give produce separate monochromatic images of the slit.

From diffraction theory we know that for large apertures the diffraction angle that represents the Rayleigh criterion is well approximated by:

$$\theta \doteq \sin \theta = \frac{\lambda}{W}, \quad (9.10)$$

where W is the beam width as labeled in Fig. 9.9. The angular separation for two wavelengths λ and $\lambda + \Delta\lambda$ according to Eq. 9.8 is

$$\Delta\delta = \frac{\sin \alpha}{\cos \gamma \cos \delta} \cdot \frac{dn}{d\lambda} \Delta\lambda.$$

Equating the above expression with the diffraction angle given in Eq. 9.10, we obtain

$$\frac{\lambda}{W} = \frac{\sin \alpha}{\cos \gamma \cos \delta} \cdot \frac{dn}{d\lambda} \Delta\lambda.$$

Solving for the resolving power R yields

$$R = \frac{\lambda}{\Delta\lambda} = W \frac{\sin \alpha}{\cos \gamma \cos \delta} \cdot \frac{dn}{d\lambda}. \quad (9.11)$$

It can be shown (Ref. 9.1, for example) that at the angle of minimum deviation—an illumination geometry that is symmetrical with respect to the incident and refracted angles—the geometrical factor, $W \cdot \sin \alpha / \cos \gamma \cdot \cos \delta$ in Eq. 9.11 is equal to the length of the base of the prism B in the direction of propagation and that Eq. 9.11 simplifies to

$$R = \frac{dn}{d\lambda} B. \quad (9.12)$$

Thus the resolving power can be increased by using a high-dispersion prism with a large apex angle and a large base. Note that the focal lengths of the lenses have nothing to do with the resolving power provided they are diffraction limited, an assumption that may be unwarranted at high resolving power. In this minimum deviation geometry, γ , the angle of refraction, is equal to half the apex angle α . In Example 9.A the resolution and angular dispersion of a flint glass prism are determined.

In Fig. 9.10(a) two prism spectrometer geometries are shown. The first of the two arrangements, known as a *Littrow geometry*, provides additional dispersion by reflecting the light back through the prism. In the second case, the prism bulk is reduced by coating a mirror on the back surface of the prism. The unfolded system corresponds to the geometry shown in Fig. 9.9, but Fig. 9.10(b)

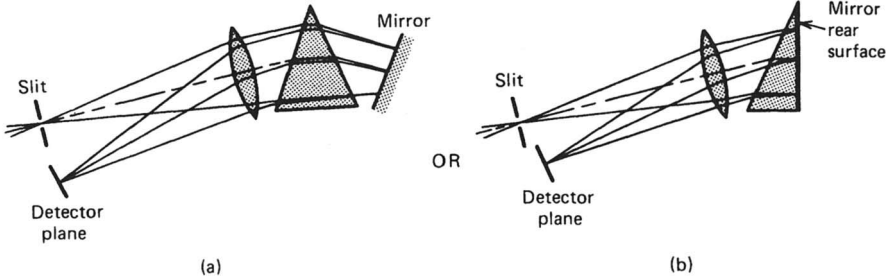


Figure 9.10. Littrow prism spectrometer geometries. (a) Provides two passes for increased angular dispersion; (b) has a more compact geometry and uses a smaller prism.

obviously is a more compact arrangement. As devices for analyzing light, prism spectrometers are reasonably simple with none of the problems of overlapping orders that we will encounter in the grating spectrometers. However, prisms can be somewhat bulky, and at higher resolving powers the required prism is expensive.

9.3. GRATING SPECTROMETERS

Grating spectrometers come in many shapes and sizes; all offer higher dispersion with a compact dispersing element. Any regular periodic structure can be used to resolve light into its constituent colors provided the period is small enough. A grating for a spectrometer is usually a series of finely spaced grooves on a glass substrate. In many instances, the surface is metallized to produce a reflection grating. Diffraction gratings can also be created using holographic techniques. In this section the basic equations for the diffraction grating are given, its resolution is discussed, and some grating spectrometer geometries are described.

9.3.1. The Grating Equation and Angular Dispersion

Monochromatic light of wavelength λ incident on a grating of periodicity d at an angle β with respect to the normal to the grating surface will be diffracted into several angles δ_m , according to the grating equation (Eq. 1.35 with differing angle notation)

$$m\lambda = d(\sin \delta_m - \sin \beta), \quad m = 0, \pm 1, \pm 2, \dots, \quad (9.13)$$

where m , an integer, is the order of the diffracted beam. In Fig. 9.11(a) the incident beam and five diffracted orders are shown. The zero order corresponds to a direct transmission through the grating surface. The same geometry holds for a reflection grating, but all beams would be reflected about the grating surface as if it were a mirror, as illustrated in Fig. 9.11(b). The beams are shown

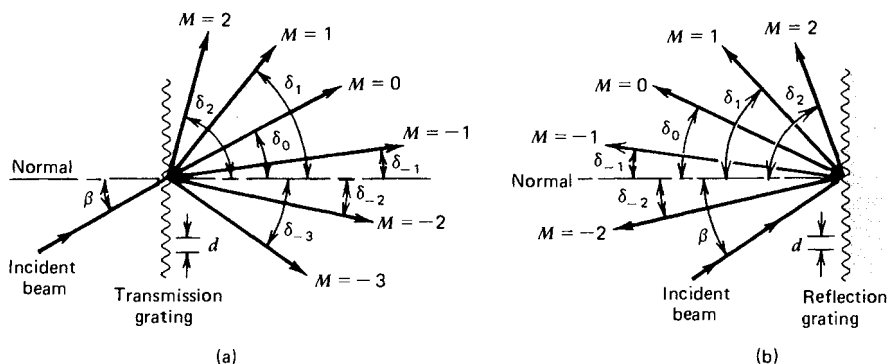


Figure 9.11. (a) A transmission diffraction grating. All beam angles are measured with respect to the grating normal. There are five orders shown here, the zero order being a direct transmission of the light through the grating. (b) A reflection grating. The arrangement of the beams is the same as the transmission grating except they are reflected about the plane of the grating.

for only one wavelength. However, in a spectrum there is a range of wavelengths. In some cases there are problems of overlapping orders. For example, note that light at 300 nm in the second order will have the same diffraction angle as light of 600 nm in the first order. In fact, any wavelength with product $m\lambda$ equal to the same value will be diffracted at the same angle if all the radiation is incident at the same β . In spectroscopy it is sometimes necessary to either prefilter or predisperse the radiation to eliminate overlapping orders. For example, a high-pass filter at the entrance slit to a spectrometer will remove the 600-nm light in the first order and transmit the 300-nm second-order radiation [Problem 9.7].

Angular dispersion of a grating is defined in the same way that it was defined for the prism, $d\delta/d\lambda$. Taking the derivative of both sides of Eq. 9.13 with respect to λ yields

$$m = d \cos \delta_m \frac{d\delta_m}{d\lambda}.$$

The dispersion is thus given by

$$\frac{d\delta_m}{d\lambda} = \frac{m}{d \cos \delta_m}. \quad (9.14)$$

9.3.2. Resolving Power of a Grating

Just as a prism, the limits of resolution are determined by diffraction, which can be related to the width of the diffracted beam. In the m th order, a beam diffracted at an angle δ_m by a fully illuminated grating has a beam width of

$W = Nd \cos \delta_m$. The width of the grating is given by Nd , N being the number of periods and d the periodicity (Fig. 9.12). Using the same analysis as in Section 9.2.2, the diffraction angle can be expressed in terms of the beam width as

$$\sin \delta \doteq \Delta \delta = \frac{\lambda}{Nd \cos \delta_m}. \quad (9.15)$$

From Eq. 9.14, the angular separation between two different wavelengths is

$$\Delta \delta_m = \frac{m}{d \cos \delta_m} \cdot \Delta \lambda. \quad (9.16)$$

Since the diffraction angle represents the smallest angular separation of resolved wavelengths, we equate Eqs. 9.15 and 9.16.

$$\frac{\lambda}{Nd \cos \delta_m} = \frac{m \cdot \Delta \lambda}{d \cos \delta_m}. \quad (9.17)$$

Solving for the resolving power, we find that the diffraction angle cancels and a simple expression remains

$$R = \frac{\lambda}{\Delta \lambda} = mN. \quad (9.18)$$

Thus only the total number of grooves in the grating and the order of the diffraction determines the resolving power of the grating. However, to achieve useful, compact gratings, an extremely fine ruling must be produced.

If the expression for the order of the grating given in the grating equation (Eq. 9.13) is substituted into the expression for the resolving power just derived, it

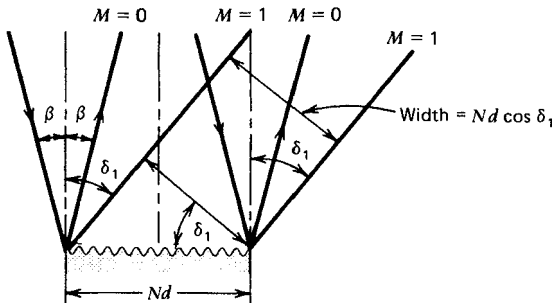


Figure 9.12. Geometry for determining the resolving power of a grating. The effective width of the grating is $Nd \cos \delta_1$, where the width of the grating is Nd . (N , the number of grating lines, and d , their separation.)

can be shown [Problem 9.8] that the resolving power of a grating of width W will be less than $2W/\lambda$.

Please note that unless the entire grating is illuminated, the maximum resolution cannot be achieved. Underfilling a grating in effect reduces the total number of grooves in the grating (N is smaller); an overfilled grating just throws light away. Thus, consideration must be given to the radiometry of the system when designing an optical instrument containing a grating. In many cases the improper illumination of a grating can be traced to poor design of the initial light-collection optics.

9.3.3. Grating Blaze

A diffraction grating can be produced by photoreducing a regular pattern of lines or by scribing grooves on a glass plate or metal surface. Usually a grating made without consideration for the type or shape of the periodicity will direct most of the incident radiation into the zero order (normal reflection or transmission) with progressively smaller amounts going into higher orders.

However, by shaping the “grooves” or other periodic structure in a manner that provides constructive interference between neighboring steps of the grating, the power diffracted into a nonzero order can be considerably enhanced, making the grating a more efficient disperser of radiation. The grating is called a *blazed* grating. In Fig. 9.13, the grating grooves are shown with broad faces on one side whose normals all are at an angle of α with respect to the grating normal. For light entering along the face normals—that is, at an angle α with respect to the grating normal—the additional path traveled by one ray reflecting off one facet compared to a ray reflecting off the neighboring facet is $2d \sin \alpha$. If the wavelength of the incident light is equal to this path-length difference, constructive interference will take place and the amount of light diffracted will be very high. At this wavelength λ_B , known as the *blaze wavelength* where

$$\lambda_B = 2d \sin \alpha, \quad (9.19)$$

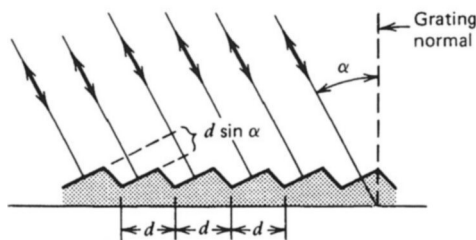


Figure 9.13. Blazed grating geometry. The facets on the grooves are oriented to provide constructive interference at a wavelength equal to twice the facet distance $d \sin \alpha$. This wavelength is called the blaze wavelength of the grating.

upwards of 75 % of the light can be diffracted into the first order. Away from the blaze wavelength, the region of efficient diffraction extends from about $\frac{2}{3}$ to $\frac{3}{2}$ of λ_B [Problem 9.9].

In addition to mechanical methods, gratings are now manufactured by holographic techniques, whereby periodic interference patterns are recorded on photographic films. If these holographic gratings are recorded on emulsions thicker than the wavelength of the exposing radiation, then they can also behave like blazed gratings. Just as with the mechanically ruled gratings, the geometry of the grating illumination is critical to achieving high efficiency. Some holographic gratings when mounted on a rotating substrate can be used as scanning elements (Section 8.5).

9.3.4. Grating Errors

Besides intentional modification of the basic diffraction gratings by blazing, the manufacturing methods produce errors that show up as additional or “ghost” slit images accompanying the desired images. The most easily identified of these errors is the *Rowland ghost*, which is due to a periodic error in the screw of the ruling engine. This error adds another periodicity besides the intended grating period and results in faint sidebands to a strong line. In the breadboarding stages of a design, a grating may be examined for ghosts by illuminating the entire surface [Problem 9.10] with a low-power helium-neon laser. If the periodic errors that occur have a period close to the grating period, the resulting ghosts, called *Lyman ghosts*, are more difficult to find since they may be quite far from the diffraction order related to the ghost.

Because no periodic device is needed to create them, holographic gratings are free of such ghosts. However, a mosaic of such gratings (to make a very large grating) can have an accidental interference pattern that might produce ghosts.

9.3.5. Grating Spectrograph Geometries

Although there may be an instance when a transmission grating geometry (Fig. 9.11(a)) is required in a design, most designs incorporate some type of reflection geometry because of the efficiency of reflection gratings and the compactness of the design. Modifications of the following reflection geometries to transmission versions can be made by reflecting the diffraction paths and all subsequent optical elements about the grating plane.

The simplest spectrograph geometry, shown in Fig. 9.14(a), consists of a collimator followed by the reflection grating and a focusing lens placed in the path of one of the diffracted orders. In the figure, the ray paths for three different wavelengths are shown (for all other geometries in this chapter only a single wavelength ray path will be shown). Figure 9.14(b) illustrates a Littrow geometry similar to that used in the prism instruments. Obviously to work well a blazed grating should be used. Because the slit and detectors would have to occupy the same region if they were placed on the optic axis of the lens, the slit is

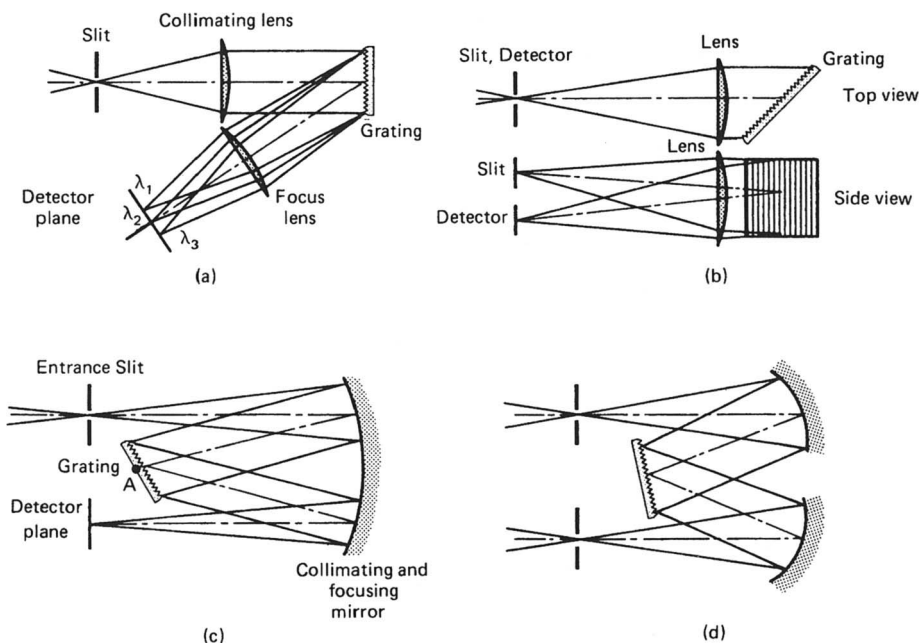


Figure 9.14. Grating spectrograph geometries. (a) Simple geometry. Focus lens is placed in the path of the diffracted order with detector at its focus. (b) Littrow geometry. (c) Ebert–Fastie geometry. (d) Czerny–Turner geometry.

placed above the optic axis and the detector below it. Another compact geometry is the *Ebert–Fastie*, shown in Fig. 9.14(c). The single spherical mirror serves as both the collimating and focusing element. The placement of the grating next to the entrance and exit points gives a tight geometry that can fit inside a cylinder slightly wider than the grating and whose length is equal to the focal length of the mirror.

A standard configuration for many monochromators is the *Czerny* (pronounced churn'-ee)–*Turner* geometry, shown in Fig. 9.14(d). In this geometry, which is similar to the Ebert–Fastie, the single mirror is replaced by separate focusing and collimating mirrors. This reduces the required diameter of the mirrors to approximately the same as the grating diagonal. For some laser spectroscopy applications the scattered light at the laser wavelength is so strong it can be detected as an additional background at other wavelengths. To reduce this unwanted background, two monochromators are placed in series (the exit slit of the first is the entrance slit of the second) with both gratings set to pass the same wavelength. The second monochromator is used to reject the laser wavelength background and allows the detection of weaker radiation associated with a laser-induced process.

A concave grating can provide both dispersion and focusing all in one element. Using a concave grating in a design allows a designer to reduce the

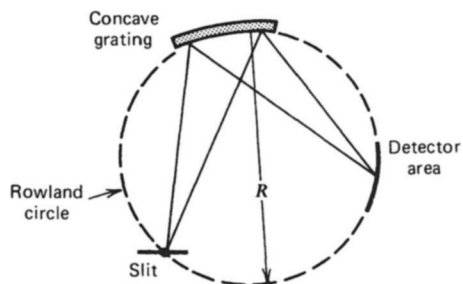


Figure 9.15. The Rowland circle. The entrance slit, the concave grating, and the detector plane are located on a circle whose diameter is equal to the radius of curvature R of the grating. All images of the slit will be found in a spectrum located on the circle.

weight of the system and eliminate lenses. This is particularly useful in a system to be used in the vacuum ultraviolet region of the spectrum, where the glass would absorb the radiation. The optimum geometry, shown in Fig. 9.15, consists of the source, the grating, and the detector all located on a circle, called the *Rowland circle*. The Rowland circle is constructed with a diameter equal to the radius of curvature of the grating. A number of different mountings have been found that take advantage of certain features from linear dispersion (Abney mounting), to wide spectral coverage (Paschen mounting), to compact construction (Eagle mounting). All are described in Refs. 9.2 and 9.3.

Finally, the geometries for holographic gratings are determined by the geometry used to produce the interference pattern that created the grating. Since there are a multitude of geometries that can produce high-contrast interference patterns, there is a great deal of flexibility available when designing a holographic grating geometry. By the same token, however, all subsequent use is constrained to that geometry. Any deviation from the recording geometry will drastically reduce the efficiency and usually increase the aberration.

9.4. SPECTROMETER SYSTEMS

The combination of the source, dispersion element, and detector may represent the most basic type of spectrometer, but for many applications it is not sufficient to provide the needed information.

9.4.1. Sources

We can divide spectrometric sources into two classes, broadband sources and narrowband sources. Broadband sources may be incandescent lights in the visible region, "global" heated ceramic sources in the infrared, or the ultraviolet output of radiation from a synchrotron particle accelerator. Normally, these broadband spectral sources are used for the study of absorption spectra since they provide a continuous background against which absorption lines can be observed.

In comparison with the broadband sources that are part of the spectrometer, narrowband sources tend to be sources under study. Examples of narrowband sources include gaseous discharges at low pressure, laser emissions, and the fluorescence and phosphorescence of materials. Stars provide examples of both types of sources. They can be studied for the broad continuum, which gives some indication as to the size and age of the star, or for the discrete lines in the spectrum, the positions of which give a measure of the velocity of the star relative to the earth. In addition, absorption bands in the continuum indicates the presence and species of interstellar matter between earth and the star.

9.4.2. Spectrum Format

Depending on the information sought, the dispersing element of a spectrometer can remain fixed during a measurement or can be moved to supply additional information. For example, when a photographic plate is used, the grating or prism presents the entire spectrum to the detector and there is no need to move the dispersing element. Also, if a single spectral line is being monitored by a detector, the dispersing element can remain stationary.

If, however, the detector consists of a light-sensing element with a slit to define the band to be measured, a spectrum can be recorded only if the grating or prism is moved, scanning the spectrum past the slit. Although it is possible to move the detector and slit past the spectrum, this method is cumbersome.

An example of one scanning arrangement is the *sine bar drive* shown in Figs. 9.16(a) and (b). In part (a) the input and output beam geometry is shown. Because there is a fixed entrance slit and fixed exit slit, the angle γ between these beams is constant. If the angle of incidence is β and the diffracted beam angle is δ_m , the sum of these two angles must be

$$\gamma = \beta + \delta_m. \quad (9.20)$$

Thus the grating equation (Eq. 9.13) can be written as

$$m\lambda = d(\sin \delta_m - \sin(\gamma - \delta_m)). \quad (9.21)$$

When the grating is set to reflect the zero order through the exit slit, $\beta = \delta_0 = \gamma/2$, since this is a case of reflection. As the grating is rotated from the zero order position through an angle θ , β changes from $\gamma/2$ to $\gamma/2 + \theta$, that is,

$$\beta = \frac{\gamma}{2} + \theta.$$

Using Eq. 9.20.

$$\delta_m = \gamma - \beta = \gamma - \frac{\gamma}{2} - \theta = \frac{\gamma}{2} - \theta.$$

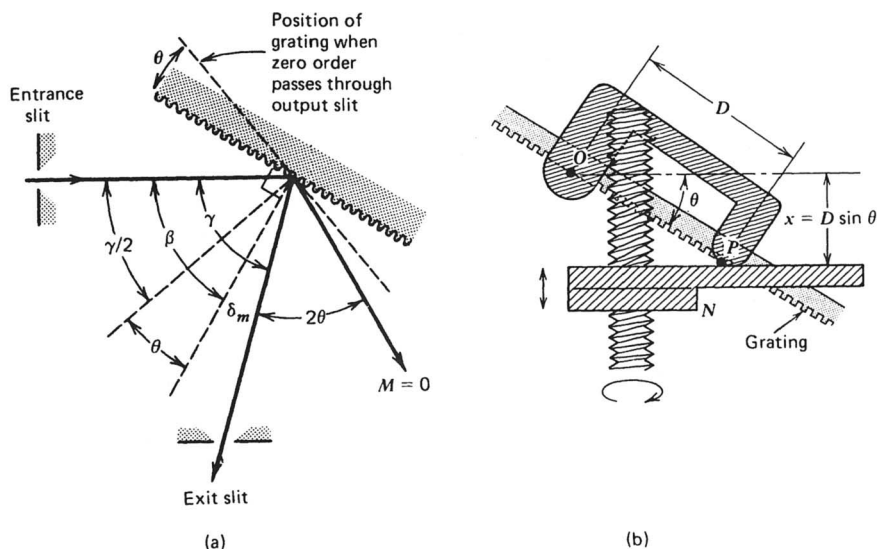


Figure 9.16. The sine bar drive. (a) Input and output beam geometry, (b) drive mechanism. This linkage changes the linear motion of the nut N on the screw into a circular motion of the grating so that the nut position is equal to the sine of the grating angle. O is the rotation point of the grating, P is the point where the linkage touches the nut on the screw and D is the distance between the two points.

Substituting for δ_m in Eq. 9.21.

$$m\lambda = d \sin\left(\frac{\gamma}{2} - \theta\right) - d \sin\left(\frac{\gamma}{2} + \theta\right),$$

$$m\lambda = -2d \sin(\theta) \cos\left(\frac{\gamma}{2}\right). \quad (9.22)$$

Thus the argument of the sine function is the angle of grating rotation from the zero-order (i.e., reflecting) position. In Fig. 9.16(b) a bar fixed to the grating at the point of rotation of the grating O has a pivot point P a fixed distance D from O . Point P pushes against another bar that is attached to a nut on a long, finely threaded screw. As the screw is turned, the point P moves a distance equal to $D \sin \theta$. Therefore, if the screw is turned at a constant rate, the sine of the grating rotation angle θ changes at a constant rate and the wavelength that is diffracted at angle γ to the input beam changes linearly with time [Problem 9.11]. The sine bar drive is an example of one of the many scanning mechanisms available to the optical designer. In some cases an extremely rapid scanning of the spectrum is needed and a sine bar drive would be too slow. Then a rotating optical element in the diffracted beam such as the polygonal prism or mirror might be used.

9.4.3. Spectrometer Control

If only the frequency or the wavelength of some spectral detail is needed, a well-calibrated dial or photographic plate will suffice. However, if the relative responses at various wavelengths are to be found, then the system must be more complex. The greatest complexity occurs if an absolute measurement of a physical quantity is required.

The simplest type of spectrometer system is an *open-loop system* in which there is no attempt to control the system. In Fig. 9.17(a) the source output is analyzed, detected, and plotted. If, for example, the source power drops during the scan, there is no method to compensate for this. However, if a portion of the source output is sensed and sent to the electronics (dotted lines), variations in the source power can be adjusted for. In Fig. 9.17(b) a *closed-loop system* is shown that provides a comparison between two examples, an unknown and a reference. Here, the loop is closed by the chopper, which tells the electronics which sample is being analyzed by the spectrometer. In this case the loop provides a means of rapidly comparing two samples (one a standard) with a single optical system.

In these systems the loop can be closed at any number of points—at the source, at the spectrometer, or any other element. In earlier days the operator of the system was an important, if sometimes uncertain, part of the loop. Whether he or she was reading a photographic plate or adjusting the gain on an amplifier to produce a readable strip chart with a pen recorder, the volume of information and the number of variables to be controlled were manageable. Today this is no longer true. Integrated circuit chips or microprocessor-based computers can provide much of the control necessary to increase the volume and reliability of

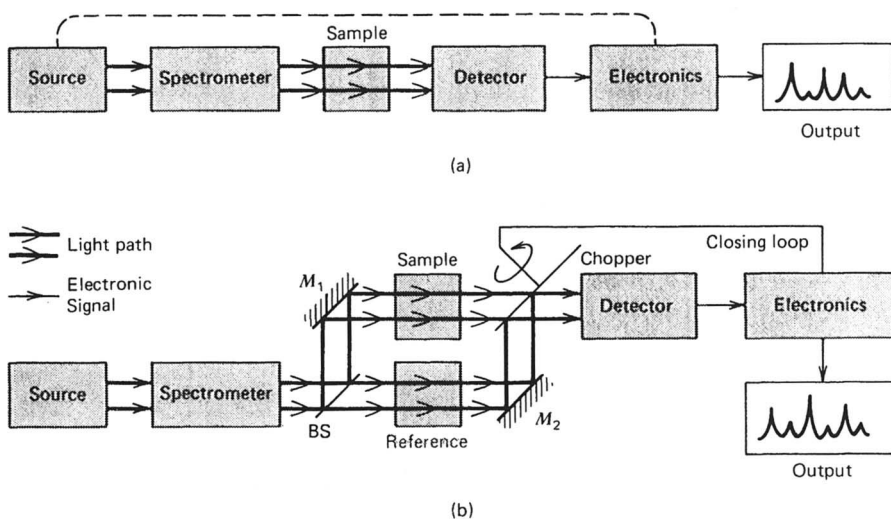


Fig. 9.17. Spectrometer systems. (a) Open-loop system. (b) Closed-loop system using a reference to compensate for system variations.

the data from a spectrometer system. However, as we discuss in the final chapter, a blind faith that some control system can compensate for an inefficient or sloppy optical design is a way to assure yourself a measure of future grief.

PROBLEMS

- 9.1. Given two wavelengths λ_1 and λ_2 whose separation $\Delta\lambda$ is not a small fraction of these wavelengths, what is the correct expression for the frequency difference $\Delta\nu$ for such a band? Using $\lambda_1 = 488 \text{ nm}$ and $\lambda_2 = 514 \text{ nm}$, calculate the approximate (from Eq. 9.1a) and exact frequency difference between these two lines.
- 9.2. What is the optical density of a filter that transmits $\frac{1}{3}$ of the incident energy? $\frac{1}{2}$? If a neutral density filter (a filter whose attenuation is nearly constant across a large range of wavelengths) is marked with an optical density of 1.5, what is its transmission?
- 9.3. Using Eq. 9.1 and the fact that the Fabry–Pérot mirror separation is an integral number of half wavelengths at a transmission band, derive Eq. 9.4. (Hint: each transmission band corresponds to a different integer.)
- 9.4. What is the size of the tuning range that can be achieved by tilting a dielectric filter designed to pass the 514.5-nm argon-ion laser line through an angle of 30° ? What happens if a dielectric filter is placed in convergent light?
- 9.5. Derive Eq. 9.7 using the geometry in Fig. 9.8 and Snell's Law. First show that

$$\sin \beta' = \sin \alpha \cdot \sqrt{n^2 - \sin^2 \beta} - \cos \alpha \sin \beta$$

and then find $d\delta/dn$.

- 9.6. If the lens in Example 5.E were used to focus the light from a prism spectroscope, what angle with respect to the plane normal to the optic axis of the lens would be necessary to provide the best focus for the F , d , and C lines? (Small angle approximations are in order here.)
- 9.7. What band of visible wavelengths would have to be suppressed in a spectrograph to record signals only from the second-order spectrum in the range of 500–700 nm?
- 9.8. Show that the resolving power of a diffraction grating is always less than $2W/\lambda$, where W is the width of the grating Nd .
- 9.9. What is the blaze angle of a grating if the spacing between steps is 480 nm and the grating is designed to be efficient between 320 and 720 nm.

- 9.10. Why is it necessary to illuminate an entire grating to see Rowland ghosts? What if only a small portion of the grating were illuminated with a low-power laser?
- 9.11. Using the arguments given in relation to Fig. 9.16, show that if the screw on the sine bar drive is turned at a constant rate, the output wavelength at an angle of γ with respect to the input beam changes linearly with time.
- 9.12. How could you use a diffraction grating as a beam combiner to superimpose two monochromatic images at two different wavelengths? Assume the wavelengths are 500 and 600 nm and the grating constant is 1000 lines/mm.
- 9.13. A simple test is to be made as a biological assay. A certain substance is released by a chemical reaction. The concentration of this substance is an indication of the level of activity of the biological system. The substance absorbs strongly at 650 nm and is nearly transparent at 500 nm. Devise a simple system to determine the concentration of the substance and therefore the activity. Assume that at maximum activity the substance will attenuate 650-nm radiation by 3 O.D. for a 1-cm path length and attenuate 500 nm by only 0.2 O.D. for the same concentration. (While it is not necessary for the simple design requested here, the irradiance through a sample is given by $I(\lambda) = I_0(\lambda)e^{-\alpha(\lambda)cx}$, where $I(\lambda)$ and $I_0(\lambda)$ are the transmitted and incident irradiances of wavelength λ , $\alpha(\lambda)$ is the absorption coefficient curve characteristic of the absorbing substance, c is the concentration, and x is the path length, here, 1 cm.)

REFERENCES

- 9.1. E. Hecht and A. Zajac, *Optics* (Addison-Wesley, Reading, Mass., 1974).
- 9.2. R. A. Sawyer, *Experimental Spectroscopy* (Dover, New York, 1963).
- 9.3. R. J. Meltzer, "Spectrographs and Monochromators," in *Applied Optics and Optical Engineering*, R. Kingslake, Ed., Vol. V, Chapter 3, pp. 47–84 (Academic Press, New York, 1969).

10

Detectors

(HOW TO RECORD WHAT YOU'VE FOUND)

In all the optical systems that we have discussed, the optical train terminates at some detector or detector plane. In the preceding chapter on spectrometers the simple example illustrated in Figure 2 of Newton's *Opticks* ends with the eye as the detector. For some systems this is an adequate detector. It is a convenient, widely available device that records images in full color and real time under a wide range of light levels. It suffers from certain disadvantages such as a connection to an imperfect and highly selective memory bank whose retention time is dependent on the type of object being observed. We will discuss the eye first and then a host of other detectors. The number of detectors that are available today is considerable, and it would be impossible to cover all of them in a book, let alone a single chapter in a textbook. We have, therefore, selected a number of common and useful detectors and left some of the more exotic types for you to explore later. Because this chapter consists of a discussion of the many detectors available for the optical designer, a narrative has been added to the material to provide a thread of continuity through this series of descriptions.

You have been hired by an aerospace firm to do some optical design work of a yet unspecified nature. During your first day on the job the supervisor drops by the office to give you some idea of what you will be doing. "We're in a slack period right now and the contract we were anticipating hasn't been signed, so we don't have a definite task for you yet. But you will probably do the detectors for this project. I'll let you know when I get some more information." The supervisor's voice trails off down the hall while you ask yourself some of the questions you would like to have asked him then. What kind of source will the system be looking at? Is the radiation in the visible? The infrared? The ultraviolet? How long do I have to detect the signal? How strong will it be after it gets through the optical system? Do I have to record an image or just an irradiance level? The questions rise to the surface like bubbles, as though the supervisor had uncapped a pop bottle. Maybe the system will require a TV camera on an elaborate scanning platform similar to the ones on the Voyager series spacecraft. Wouldn't that be a fantastic project to begin a job with? But then maybe it will be a very simple job that could be done by eye.

10.1. THE EYE—THE ULTIMATE DETECTOR

In most of the optical systems the information is ultimately put in a form that can be read by a human. (There are, of course, exceptions to this statement—an optical system integrated into a robot, for example.) We already discussed the eye as a simple optical instrument in Section 1.5 and as a color-sensitive detector in Section 3.5 on photometry and spectrometry. In Table 10.1 we list some of the most common characteristics used to specify detectors. Here, we look at the eye as a device and compare these characteristics of a detector to those of the human eye.

The spectral response of the human eye, which extends approximately from 400 to 700 nm, defines the visible spectrum, the range of wavelengths used in the field of photometry. The eye's sensitivity is somewhat harder to specify because it is so adaptable; it responds over a large variation of illumination levels—from

Table 10.1. Representative Photodetectors

Detector Types	Peak Sensitivity (A/W)	Response Curves	Frequency Response (Hz)	Resolution (line pairs per mm)
Eye	NA	Fig. 3.16	20	5
Thermopile	55 V/W	Flat	1	NA
Pyroelectric	2400 V/W	Flat	5×10^5	NA
Photographic Plate	NA	Fig. 10.4	Note 1	200
Electrophotographic	NA	Note 2	10	Note 3
Photomultiplier	10^5	Note 4	10^9	NA
Vidicon Tube	NA	Note 4	10^6	Note 5
Image Converter	NA	Note 4	10^3	20
Solar Cell	0.3	Fig. 10.17	5×10^3	NA
Photodiode Photovoltaic	0.3	Fig. 10.17	10^6	NA
Photoconductive	0.3	Fig. 10.17	3×10^7	NA
Avalanche Photodiode	100	Fig. 10.17	10^9	NA
Charge Coupled Diode Array	0.3	Fig. 10.17	$10^7/N$ $N = \text{of diodes}$	30

NA: Not applicable in most cases.

Note 1: Must calculate irradiance I and determine time to achieve necessary exposure.

Note 2: Extends from far UV to green, peaks about 400 to 450 nm.

Note 3: Depends on optics and toner.

Note 4: See curves in Fig. 10.8. (For problems, choose the most effective surface.)

Note 5: Approximately 500 lines (horizontal) \times 350 lines (vertical) per frame.

bright sunlight to starlight. The eye is an extremely sensitive detector, for the sensation of light can be caused by only a few photons—as good as the best photomultiplier tubes. Since the advent of the laser, the maximum illumination that is safe in various applications has come under considerable scrutiny. Since it depends on the situation, the wavelength of the radiation, and the coherence of the source, the maximum power input for the human eye is not easily determined and is still a subject of study. The minimum signal that the eye can detect is affected by its past history, as anyone who has walked into a dark room on a sunny day knows. Even when the eye is dark adapted it will produce a signal spontaneously in the absence of any radiation. This is called “eigengrau” (German for “self gray”).

If there is one quality that the eye lacks when compared to other detectors, it is good temporal frequency response. The eye will perceive a succession of images as continuous if their frequency is above about 25 Hz—the actual value varies with illumination levels. This has been useful for the creators of motion pictures and television programs, but it means that for the observation of events happening in less than a tenth of a second we must depend on faster detectors than our eye.

The angular acceptance of the eye is not easy to specify with a single number. Because our peripheral vision, which has very little resolution, is combined with high-resolution central vision, the overall angular acceptance of the eye is about 120° , while that of detailed vision is only 5° . The resolution is governed by the regions of sensitivity on the *retina*, shown in Fig. 10.1. The high-resolution cone receptors occupy the central 0.25-mm diameter circle, the *fovea*, while the low-resolution rod receptors are distributed over the rest of the sensitive region of the retina. The cones are responsible for the high illumination response given by the photopic curve (Fig. 3.16), whereas the rods provide the low-light-level sensitivity for dark-adapted vision.

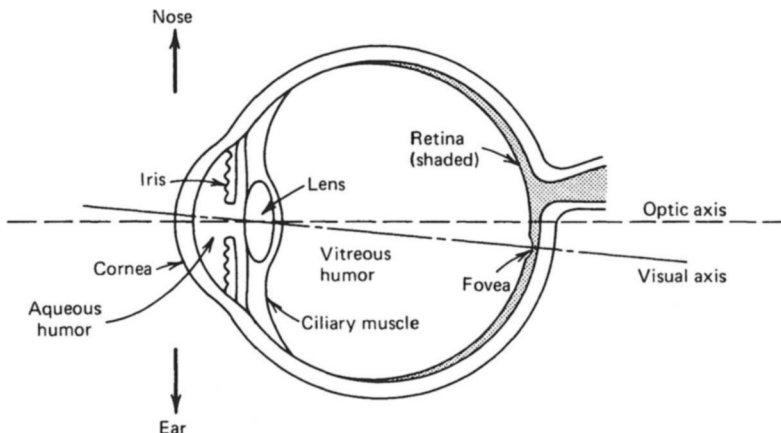


Figure 10.1. Schematic horizontal cross section of the human eye. Note the visual axis through the center of the lens and the fovea is not coincident with the optic axis of the eye.

All in all, the eye is a formidable optical system providing zoom focusing, adaptation to a wide range of light levels, central and peripheral imaging, and dual responses to provide both high-level and low-level detection—all in a one-inch-diameter sphere! Any optical designer worth his or her salt would be proud to design a system like that. Of course, they wouldn't have the luxury of millions of years of field trials.

The supervisor sticks his head in the door of the office. "I've got a new project for you. We have to measure the total radiation coming from a target for the new Space Shuttle package. See what you can throw together on commercial detectors that we could use." He gives a little wave and tears off down the hall as the questions begin to rise once more. What sort of target? Does it want to be seen or will it try to hide from the detectors? Is it moving? What does he mean by total radiation? The last time you asked those questions the reply was: "Sorry, friend, but that's all they've told me so far". Do we want detectors to measure total radiation from the target? That means we'll need a broadband detector.

10.2. THERMAL DETECTORS

Radiation, whatever its wavelength, consists of packets or quanta of energy, called photons. The amount of energy per quantum is given by $E = h\nu$, where h is Planck's constant, an extremely small constant equal to 6.625×10^{-34} J-sec and ν is the frequency of radiation. Radiation detectors that convert this energy to an electrical or chemical effect on a quantum-by-quantum basis are called *quantum detectors*. The eye is an example of a quantum detector operating by photochemical conversion. Detectors can also convert radiation into heat. These are called *thermal detectors*. In comparison to most quantum detectors, thermal detectors are simple devices. They consist of an absorbing surface and a means of measuring the amount of heat absorbed. It is possible, for example, to cause the radiation to heat a gas in a cell and then measure the pressure change due to the heating (Golay cell). Because heat will change the resistance of metal wire or a semiconductor (thermistor), the presence and amount of radiation can be detected when the element is part of an electric circuit (bolometer). In this section we describe two thermal devices of particular interest, thermoelectric detectors and pyroelectric detectors. In later sections we describe the quantum detectors.

10.2.1. Thermoelectric Detectors

All thermal detectors depend on a surface that absorbs radiation over a large range of wavelengths and converts the radiation to heat. To the extent that the absorbing surface does not reflect or scatter any of the incident radiation, the device will be a good radiation detector. Beyond this the usefulness of the

detector depends on the ability of the device to measure the temperature increase in the absorbing surface.

One method of measuring the amount of heating in a material is to attach a thermocouple to it (Fig. 10.2(a)). The temperature difference between the thermocouple junction attached to the absorbing surface and a reference junction at a constant temperature produces a voltage that can be recorded with a microvoltmeter. If the meter has been calibrated with a known power input, the incident power from any source can be found. By combining a series of thermocouples into alternating measuring and reference junctions, the amount of thermoelectric power generated is increased and the sensitivity of the device is also increased (Fig. 10.2(b)). This multiple junction detector is called a *thermopile*. A sensitive thermopile, which can be quite expensive, will measure incident power levels in the microwatt range. To provide accurate readings at low power levels it is necessary to ensure that any radiation due to the surroundings and background has been subtracted from the total signal. At very low levels, the absorbing surface is cooled to reduce the thermal noise that would interfere with the signal.

The thermopile offers no realistic possibilities for image detection; however, its linearity over a wide range of power levels and its uniform spectral sensitivity over a range from the near ultraviolet to the infrared make it extremely useful for

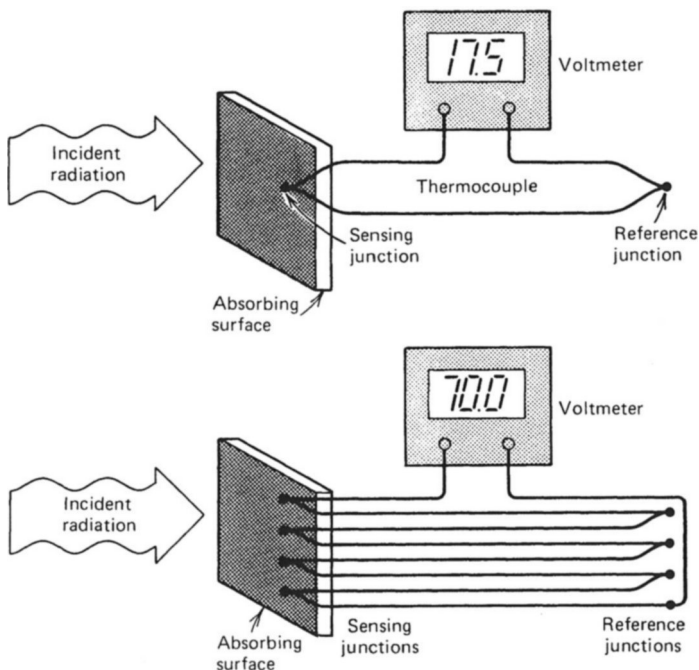


Figure 10.2. Thermocouple-based radiation detectors. (a) Simple thermocouple detector. (b) Thermopile.

power measurements on a variety of sources. Perhaps its greatest drawback is its slow response time, which depends on the size of the absorbing surface but is usually about one second for most thermopiles.

10.2.2. Pyroelectric Detector

If a ferroelectric crystal, a crystal that has a permanent dipole moment, is heated, its dipole moments will change (*pyroelectric effect*). This causes a change in the voltage across the crystal, and if the opposite surfaces of the crystal are connected in an external circuit, a current will flow until the two sides of the crystal reach the same potential (Fig. 10.3). The amount of the current is a measure of the heat absorbed by the crystal and, therefore, of the amount of incident radiation. However, the current exists only as long as the crystal is heating up (or cooling down). Once a new stable temperature is reached and the voltage drop across the crystal is zero, the current is zero.

The current is proportional to the time rate of change of temperature,

$$I = p(T)A \frac{\Delta T}{\Delta t}, \quad (10.1)$$

where $p(T)$ is the pyroelectric coefficient of the crystal and A is the electrode area. The mass of the crystal can be written as: $m = \rho V = \rho A d$ where ρ is the density of the crystal, V is its volume and d is its thickness. Recalling the definition of specific heat,

$$c_p = \frac{\Delta E}{\Delta T} \cdot \frac{1}{m} = \frac{\Delta E}{\Delta T} \cdot \frac{1}{\rho A d}. \quad (10.2)$$

Solving for ΔT in this equation, and substituting into Eq. (10.1), we obtain

$$\begin{aligned} I &= p(T)A \frac{\Delta E}{\Delta t} \cdot \frac{1}{c_p \rho A d} \\ &= W \frac{p(T)}{\rho c_p d}, \end{aligned} \quad (10.3)$$

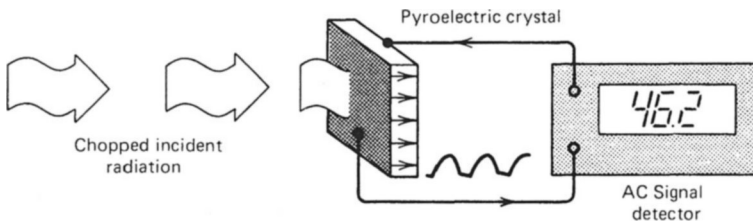


Figure 10.3. Pyroelectric detector. Chopped (modulated) radiation produces a change in the spontaneous polarization of the crystal. The generated AC signal is then measured.

where the $W(= \Delta E/\Delta t)$ is the instantaneous power causing the current I and the second term is a constant of the detector. Note that since a current occurs because of a change in the signal, the input radiation must be chopped, pulsed, or modulated to be measured.

Because the detecting crystal can be very small yet still give a substantial signal, the response time of the detector can be extremely short. Thermal detectors with frequency responses exceeding a megahertz are possible. Because the spectral response is broadband like any other thermal detector, it is useful in high-speed infrared systems. Although it would seem that pyroelectrics are ideal detectors, pyroelectric crystals are fragile and, in some cases, hygroscopic. Therefore, unless special care is taken in the placement and packaging of these devices, their reliability may be a problem in the overall design of the system.

At a weekly project review meeting you find out that the project has been expanded. Along with the broadband detectors, they want some simple imaging system.

This time there is an opportunity to get more information. Do you have some idea about the wavelength coverage? How good do the pictures have to be? Is this supposed to be real-time imaging?

The supervisor and the project director are able to provide the following guidelines. Most of the imaging is to be done in the near infrared and the pictures would be recovered at a later time. Therefore, there is no need for immediate transmission of the pictures. As for the resolution, no information has been forthcoming. It appears that the imaging is to be used to make sure that the device was looking at the correct target.

"Sounds like a simple infrared camera," you mumble under your breath.

10.3. PHOTOGRAPHIC MATERIALS

After the human eye, the most familiar light detection device is photographic film. Today we tend to dismiss the complex photographic process because we can drop the roll of film off at the drugstore and get a finished set of prints a day later or wait a few minutes until the camera spits out a nearly instant picture. The simplicity of our everyday experience with photography gives way to the true complexity of the subject once one sees the kinds and types of emulsions available and the variety of processing conditions. It is impossible to go beyond the rudiments of the photographic process here. Considerable information is available from technical manuals, books, and from the manufacturers of film.

The advantages of photographic systems include an ability to record images and to integrate the light flux over a period of time to capture a weak signal. Photographic materials need no electrical power to work (except for motion pictures, where power for the film transport is required). These materials are compact and lightweight compared to other roughly comparable imaging systems such as television cameras.

10.3.1. Silver-Based Emulsions

Although there is now a variety of other photographic methods available to the optical designer, detection of radiation by silver halide grains embedded in an emulsion is still one of the best methods available. The prime reason is that only a few photons are needed to sensitize a silver halide crystal. The energy needed to finish the conversion of silver bromide or silver iodide to elemental silver is provided by the chemicals in the development process. Prior to development, the information exists as a latent image in the form of the sensitized grains on the plate or film.

Because the silver halide grains by themselves are sensitive only to blue and near-ultraviolet radiation, dyes must be added to absorb the radiation in other parts of the spectrum and to provide a pathway to sensitize the silver halides. If the spectral response is extended into the green region of the spectrum, the film is called *orthochromatic*; if the region of sensitivity extends across the entire visible spectrum the film is called *panchromatic*. A representative set of spectral sensitivity curves is shown in Fig. 10.4. Infrared films are available, but they must be handled and stored carefully because of the risk of fogging the plates due to heat.

Once a film has been exposed, the amount of darkening in a particular region is dependent on a number of factors: (1) the irradiance on the exposed region, (2) the wavelength of the radiation, (3) the length of the exposure time, and (4) the conditions for developing the film. Assuming the film is sensitive at the radiation wavelength, the relationship of the other three factors to the amount of darkening can be expressed as a curve called the *Hurter–Driffield* or *H–D curve* such as the one shown in Fig. 10.5. The vertical axis is the optical density, D (see Section 9.1.1); the horizontal axis is the logarithm of the exposure, H , the product of the irradiance and the exposure time, $E_e t$.

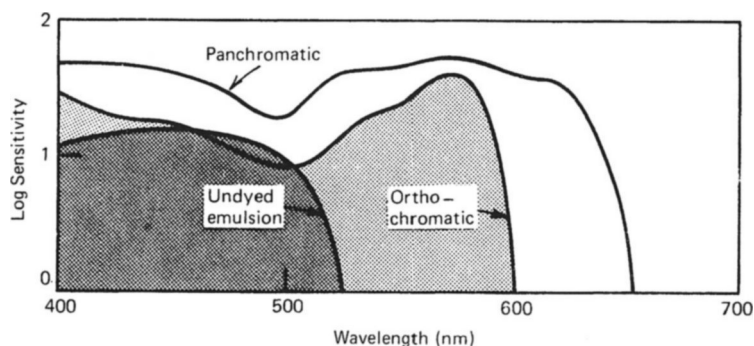


Figure 10.4. Photographic sensitivity curves. Each photographic emulsion has its own peculiar sensitivity curve. The curves illustrated here are intended to be representative of three classes of film sensitization.

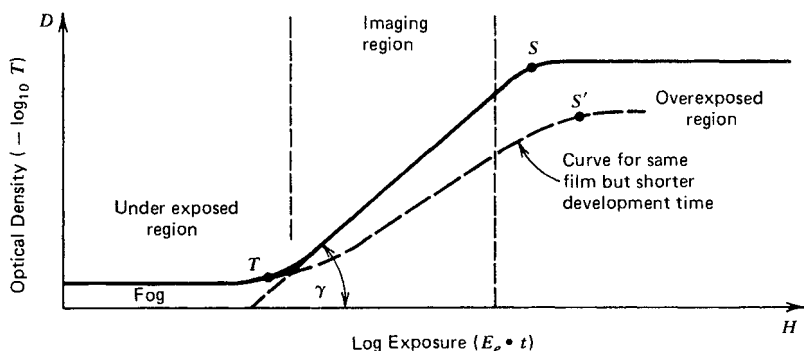


Figure 10.5. Characteristic Hurter-Driffield curve of a photographic emulsion. Note that developing for a shorter period of time reduces the contrast and increases the exposure range.

To understand this curve, examine its three regions. On the left side of the curve the exposure is small and the density is constant. This is the underexposed region. The nonzero density in this region is due to a small reduction in transmission when the plate is developed even if it has not been exposed. This level is called the *fog level* of the plate. The point T where the curve begins to bend upward is the *toe* of the curve. At the other end of the exposure axis is the overexposed region. Here, any additional radiation has little effect on the film since almost all crystals have been converted to silver. (There is an effect called *solarization* that causes the curve to fall for extremely large exposure values so that additional irradiation can actually reduce the density. Although solarization has artistic uses, it is rarely used in optical design.) The point where the overexposed region bends over is the *shoulder* of the curve (point S in Fig. 10.5). The region between the shoulder and the toe is the *imaging region*, the linear portion of the H-D curve, where variations in irradiance are translated into optical density variations. If the slope of the linear portion of this region is very steep, then only a small variation in exposure will cause a large change in optical density, yielding a high contrast image. If the slope is less steep, the range of exposure will be increased with lower contrast resulting. The amount of contrast is specified by the slope of the linear portion of the H-D curve, designated by the Greek letter gamma:

$$\gamma = \frac{\Delta D}{\Delta H}. \quad (10.4)$$

The gamma of a particular piece of film depends not only on the type of film, but also on the development time and the development temperature. If the temperature is taken to be the standard 20°C (68°F), a plot of γ versus development time provides the designer with an idea of the range of γ that can be achieved with reasonable development times.

There are a number of effects that can influence the quality of the image. For example, it has been assumed that the exposure, H , is the product of the irradiance times the exposure time. Thus any point on the H - D curve can be achieved in a number of ways. One could reduce the irradiance and expose for a longer time or increase the irradiance and shorten the exposure time. However, for very short and very long exposure times this reciprocal relation does not hold; the total exposure H must be somewhat more than is needed at intermediate times. This is known as *reciprocity law failure*. Other effects, called adjacency effects, occur when exposed regions are close to one another. The proximity of a heavily exposed region can influence the rate of development of an adjacent region of lower exposure.

Assuming that the optical system you have designed is absolutely superb, what can you expect of photographic materials when recording an optical image? It depends on the resolution of the film. This quantity can be specified by a number of methods. One is to determine the maximum number of dark-light line pairs per millimeter that the film can record. This evaluation can be done visually, although it becomes very subjective, or by scanning the photographic image with a narrow slit and detector. The resolution of standard photographic film rarely exceeds 100 line pairs/mm. For work that requires the recording of extremely fine interference patterns (e.g., holography), resolutions upwards of 1500 line pairs/mm are required. Perhaps most useful is a modulation transfer function (MTF) of the film. This plot of contrast versus spatial frequency (see Section 6.5.3) provides an evaluation of performance over the useful spatial frequency range.

10.3.2. Other Photographic Materials

In contrast to the silver-halide-based materials, all other photographic methods require an abundance of light. Thus these methods are useful only when the source is a large bank of fluorescent lights, a focused laser beam, a bright ultraviolet light, or a similar source. In some cases the light must be rich in ultraviolet wavelengths to produce a photochemical change. For example, compounds with nitrogens attached to a benzene ring (diaz materials) break down in the presence of UV radiation. Because there is no “graininess” to these materials, they generally have a higher resolution than silver halide emulsions.

Polymerization of compounds by irradiation is another high-illumination means of producing images. If a gelatin containing a dichromate is illuminated, the dichromate causes cross-linking in the gelatin. This reduces the solubility of the material. Development then consists of rinsing the materials in water. The result is a layer whose thickness is proportional to the amount of illumination. This amounts to a “phase photograph,” that is, the image is recorded as an optical path in the emulsion rather than a pattern of absorbing materials. This type of process is particularly useful in holography and in photolithography. Other photopolymerizers contain an absorbing material called an initiator. Light absorbed by the initiator is channeled to the monomer in which it is

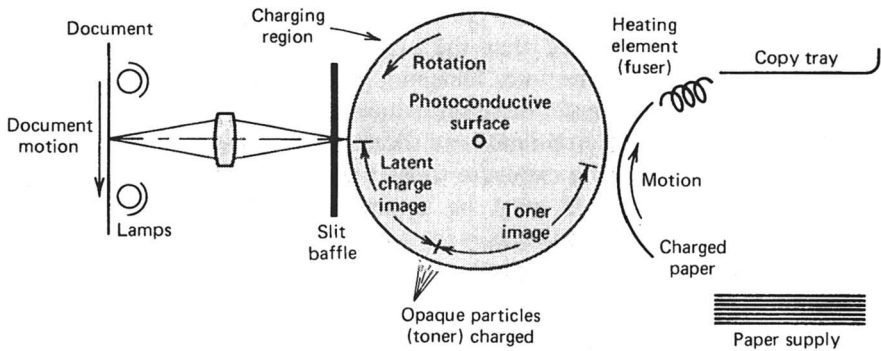


Figure 10.6. Electrophotographic imaging. Either a portion of a brightly illuminated object or the entire object is imaged onto a photoconductive surface. The latent image built up as a surface charge distribution is changed to a distribution of opaque particles (toner) and then transferred to a charged sheet of paper. The toner is melted on the paper by local heating to create the permanent copy. For document scanning, drum rotation and paper motion must be synchronized to reduce distortion.

embedded, causing polymers to form. The spectral sensitivity is controlled by the absorption spectrum of the initiator. Thermal effects can also be used to produce local color changes. In all cases the amount of light required is significantly greater than with conventional photographic materials.

Another method that requires high irradiance is *electrophotography* (Fig. 10.6). One type is used in Xerox (TM) copying machines. A photoconductive surface is sprayed with charge and exposed to a bright image. In those areas where the illumination is high, the conductivity increases, enabling charges to leak away. Thus, charges remain only in the darker region. If the surface is then sprayed by a dust of colored particles (toner) charged opposite to that of the surface, the particles will be attracted to the unilluminated and weakly illuminated regions. The colored particle distribution is finally transferred to paper by static electricity techniques and permanently fused to it by heating. Because of charge migration and fill-in, the electrophotographic process cannot approach the contrast ratios and resolution of the other processes discussed in this section. However, its convenience and rapidity make it particularly useful for photocopying.

There are certain disadvantages to any of these photographic methods. They all need some sort of processing to bring out the latent image. There is no electrical signal that can be used for real-time processing. And while the process time can be reduced, there is no real-time information. Still, if some form of archival image is needed in an inexpensive format, then a photographic method is the method of choice.

One of your office mates sits down and puts his feet up on your desk.

"Have you heard the latest wrinkle in the project?"

"No. You mean they are not going to use infrared cameras?"

"I understand they want real-time information and images. It looks like we are going to need a low-light-level television. What do you know about them?"

"Not a whole heck of a lot. Are you sure? You're not pulling my leg, are you?"

"No. That's what the supervisor told me last night. Go ahead and ask him."

"I think I will," I say, reaching for the telephone.

"On second thought, don't bother. Project management is doing project reviews for the next two days."

You begin to review the basics of photodetectors and imagers.

10.4. PHOTOEMISSIVE DETECTORS

Photoemissive detectors are extremely sensitive light detectors that use the photoelectric effect to provide an electrical signal that is proportional to the incident light. After describing the effect and its use in photomultiplier tubes, we look at two devices for imaging radiation: image tubes and converter tubes.

10.4.1. Photomultipliers

Some of the easiest elements to ionize are the alkali metals, those elements that are one atomic number greater than the rare gases. A certain amount of energy, called the *work function*, is needed to pry an electron away from the atom and surface on which it is held. As Einstein explained in 1917, to produce a free electron the energy of a single light quantum must be sufficient to overcome the binding energy of the electron to the surface. Once the electron is free it must be removed from the vicinity of the surface if it is not to reattach itself. This removal can be accomplished by the presence of a voltage difference between the photoemissive surface and a collection point for the electrons. If the collection point is more positive than the surface, emitted electrons will be swept to the collection point by the electric field.

The construction of a *photodiode* is shown in Fig. 10.7. The photoemissive surface, or *cathode*, is usually deposited on the inside of a glass tube or on a metal surface that is connected to the outside through a metal feedthrough. The positive collecting surface, or *anode*, is also connected to the outside by a feedthrough. By connecting a voltage source between the anode and cathode and measuring the current that is generated in the external circuit, the amount of radiation on the photocathode can be determined [Problem 10.2].

Since the amount of energy in a photon determines whether an electron may be emitted from the surface, and since the amount of energy in a photon is proportional to the frequency of the light, therefore, the sensitivity of the photocathode must vary with the color of the incident light. By combinations of

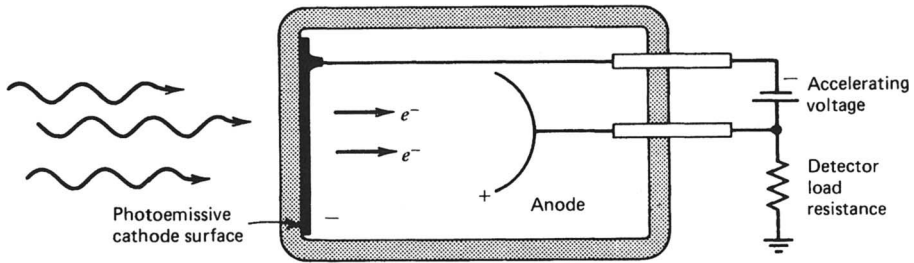


Figure 10.7. Photoemissive diode. Photons incident on the cathode cause electrons to be emitted. An accelerating voltage between the cathode and anode directs the electrons to the anode where they are collected. The photocurrent through the load resistor is a measure of the amount of light falling on the device.

various alkalis and other metals, a number of photocathode surfaces can be produced that provide high sensitivity over a range of visible and near infrared wavelengths, as shown in Fig. 10.8. In the ultraviolet region this sensitivity can be extended only if the glass substrate on which the surface is deposited is replaced by quartz, since glass absorbs below 350 nm. Even in the best of circumstances a photosurface has only a 25% quantum efficiency, that is, only one out of every four photons produces a photoelectron.

The output of a photodiode might be sufficient at high light levels, where a large amount of photoelectrons provide a substantial current. However, there are more rugged and less-expensive semiconductor devices (Section 10.5) that can do the same job. It is only by multiplying the original photoelectrons that we can obtain the extremely high sensitivities needed for operation at very low light levels. The device that accomplishes this is illustrated by a schematic diagram of a *photomultiplier tube* in Fig. 10.9. A series of secondary electron-emitting surfaces called *dynodes*, are placed between the anode and the cathode, and an accelerating electric field is imposed between the dynodes. An electron emitted from the photocathode is accelerated to the first dynode, where its kinetic energy is sufficient to knock off a number (usually 2 to 4) of secondary electrons. These electrons are in turn accelerated toward the next dynode to produce additional secondary electrons. A cascade of 8 to 10 dynodes can produce a gain of a million or more, depending on the voltage across each dynode. This results in a sizeable amount of current in the form of a pulse of electrons at the anode for every photoelectron emitted by the cathode. To provide the proper distribution of voltage between dynodes, a chain of resistors serves as a voltage divider for the high voltage imposed on the tube, as indicated in Fig. 10.9. Since there is always a small current flowing in this circuit it is sometimes referred to as a *bleeder chain* and the current as bleeder current.

The output of a photomultiplier tube can be measured in two ways. One method is to integrate the current in all the pulses over some period of time; the average current becomes the resultant electrical signal. This is the simplest method of using the output of a photomultiplier tube. This method, however,

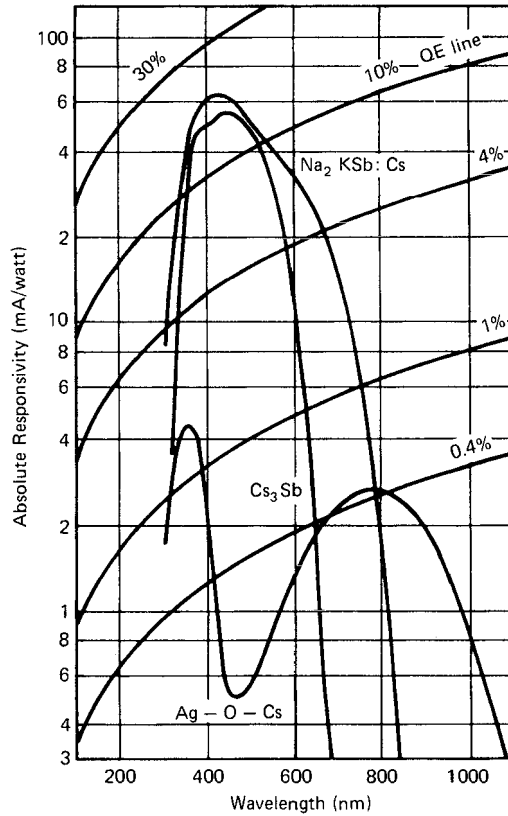


Figure 10.8. Photocathode sensitivity curves for selected combinations of photoemissive materials. (Reproduced with permission, RCA Corporation.)

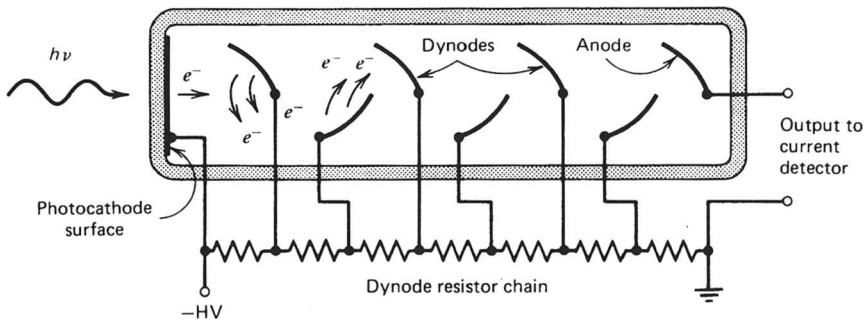


Figure 10.9. Photomultiplier tube. The current from a photocathode is amplified by accelerating the photoelectrons into a series of dynodes that emit several electrons for each electron hitting the surface, creating a cascade of electrons that can be integrated at the output as a current or detected as a current pulse.

includes noise due to the dynode chain. The second method, used in low light level applications, employs standard nuclear instrumentation counting electronics to count the individual pulses. Discrimination against some noise sources is done simply by setting the counters to reject pulses whose amplitudes are smaller or larger than the range of pulse heights expected for photoelectron pulses.

The advantages of a photomultiplier tube are its high sensitivity (on the order of 100 kA/W), its linearity over 10 orders of magnitude, and its fast response. The response time is usually less than 100 nsec, with some tubes operating in the 10-nsec range. Its disadvantages are that it is fragile and that it requires a high-voltage power supply for operation. For low-light-level operation, the cathode must be cooled to about -40°C to reduce the number of thermally excited photoelectrons that could be counted as signal.

One technique that reduces the overall size and weight of the photomultiplier tube is to fashion a continuous dynode chain out of a length of hollow glass tubing coated on the inside with dynode materials. When 2 kV are applied to the tube, current amplification can take place just as in the discrete dynode case. These devices are called continuous-channel electron multipliers. (The commercial detector that uses this method is called a Channeltron.) In all these cases, the output of a photomultiplier tube is a measure of the total amount of radiation falling on the photocathode. These tubes cannot produce or amplify an image as image tubes can.

10.4.2. Image Tubes

The first surface of an image tube is the same as that of a photomultiplier tube. The difference is that the spatial distribution of the light pattern delivered by the optical system is preserved for further processing. There are two types of image tubes, the *storage tube*, which is used to produce an electronic signal like those used in television cameras, and the *converter* or *amplifier tube*, which produces an image of a scene that could not be seen by the unaided eye. Infrared viewers, sometimes called snooperscopes, and low-light-level viewers are devices that use the second type of tube.

10.4.2.1. Storage Tubes. An image storage tube, shown schematically in Fig. 10.10, consists of two sections, a read section and a write section. The write section consists of a photoemissive surface, similar to those described for photomultipliers, and a storage surface. By applying magnetic or electric fields between these two surfaces, photoelectrons emitted from the cathode are directed to a corresponding point on the storage surface, just as a lens directs light rays from an object point to a corresponding image point. At the storage surface the electrons produce a spatially non-uniform charge distribution that mirrors the light distribution on the photocathode. This charge pattern is read by sweeping an electron beam across the storage surface in a set of closely

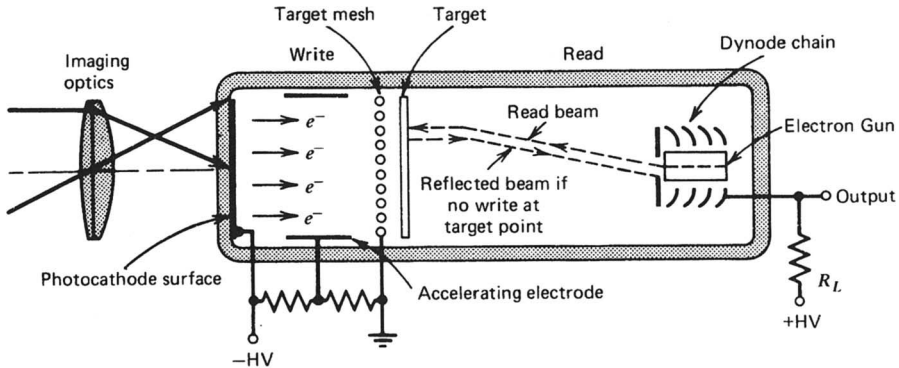


Figure 10.10. Storage tube (Image orthicon). The image on the faceplate is transferred to the target surface at unit magnification then read serially by a raster-scanned read beam.

spaced parallel lines (raster pattern). The beam deposits a charge density on the surface that is sufficient to bring it back to a uniform charge distribution. Since the amount of charge required to produce a uniform distribution depends on the irradiance that produced the non-uniformity, the variation in the amount of charge deposited provides a signal which can be amplified and redisplayed to produce a television image.

The raster pattern scan of the storage surface is an obvious limitation to the resolution of a system that incorporates a storage tube as a detector since density of lines determines the amount of detail that can be displayed. Many formats use the standard television scanning rates: 525 lines per frame, 30 frames per second (American standard), 625 lines per frame, 25 frames per second (European standard). Only about 92% of the lines are displayed, since at the end of the scan the spot must be returned to the starting point (vertical blanking). Further, the number of visually distinct image lines is only 70% of this or about 340 lines (American) and 370 lines (European). The horizontal resolution is on the order of 500 lines per frame. The exact resolution is dependent on the electrical bandwidth of the transmitting and receiving devices and on the level of illumination at the object. Higher resolution tubes are available for special applications.

There are other limitations imposed by these tubes. If a portion of the image is too bright, the surrounding region is affected, resulting in a flare in the image. If the object does not vary or the camera is not moved, the image can be temporarily imprinted as an after-image. This happens to some degree from frame to frame since the electron beam does not perform a complete erasure during each scan of the storage surface. In certain cases this incomplete erasure can affect the resolution of the system as much as the raster pattern. The spectral response of the tube is that of the photocathode unless a filter is placed in front of a tube as is done with dielectric filters to separate the red, green and blue images for color television.

10.4.2.2. Image Intensifiers and Converters. In contrast to the storage tubes, which convert an optical image to an electrical signal, image intensifiers and converters produce a visible image. An image intensifier takes a dimly lit scene, such as a landscape lit only by starlight, and produces an image that is easily viewed. An image converter does the same thing by converting scenes illuminated by infrared or ultraviolet radiation to visible pictures.

The front end of these tubes is similar to those of storage tubes, an emissive photocathode and an electrostatic lens to focus the electron distribution onto a special surface, as shown in Fig. 10.11. The difference in the two types of tubes is that in place of a storage surface, a phosphor screen receives the image. With a suitable acceleration voltage present, the photoelectron can excite a phosphor on the screen and produce an increase in the illuminance of the image. If the incident wavelengths are invisible, but the photocathode still responds to them, the excitation of the visible phosphors will convert the image on the photocathode to a visible image.

To obtain further gain, these tubes can be cascaded by placing the photocathode of a second stage next to the phosphor screen of the first. For good optical coupling the substrates for both surfaces are made of bundles of optical fibers. Another method of producing additional amplification is to use a *microchannel plate* inside the image tube. Consider this device as many short lengths of channel multiplier tubing stacked side by side to produce a disk with the channels parallel to the axis of the disk, as shown in Fig. 10.12. Accelerated photoelectrons entering a channel will be multiplied before impinging in the phosphor screen attached to the other side of the microchannel plate.

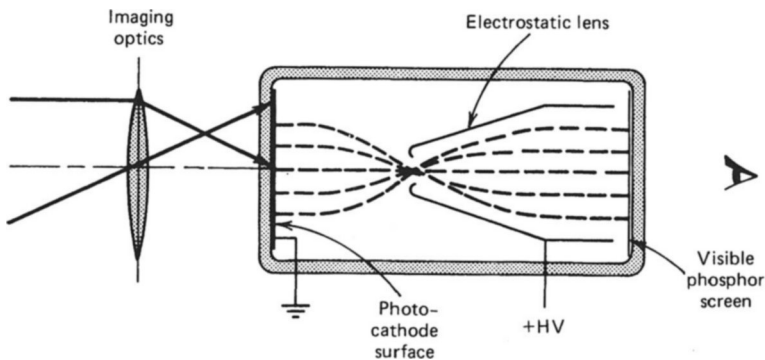


Figure 10.11. Image intensifier and converter. If the image on the faceplate is in the visible and the brightness is increased by reproducing it on a bright phosphor screen, this device is an intensifier. If the image contains only radiation in the infrared or ultraviolet, the device is a converter.

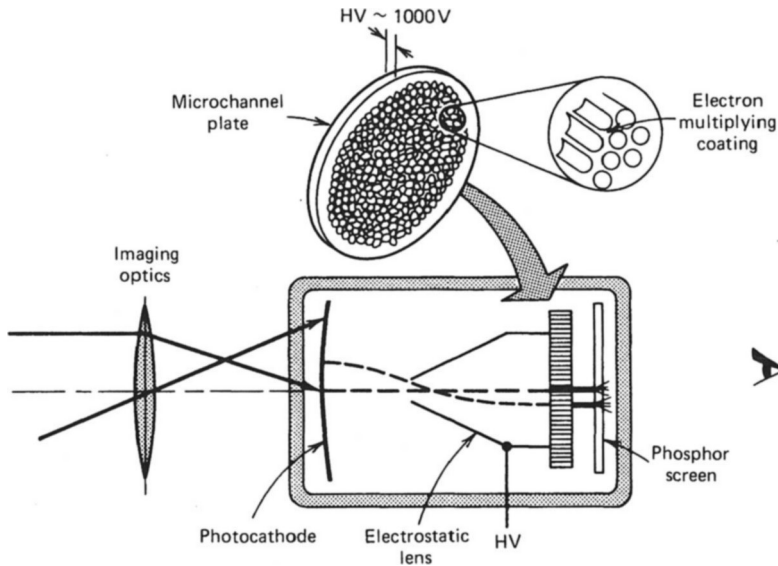


Figure 10.12. Microchannel plate. This element used inside image converters and intensifiers provides additional electron amplification similar to a channel multiplier (Section 10.4.1). The resolution of the image is governed by the size and distribution of channels in the plate.

There is a memo from your supervisor lying on your desk when you come in one morning.

SUBJECT: Changes in Detection System for Project # RF 7-225-4.

Because of budgetary difficulties and some unforeseen cost overruns, the imaging capability of this device will have to be reduced. All further work on the infrared sensitive high-resolution camera system is to be terminated. A survey of the best possible low-cost alternatives to this system is to be initiated. Management wants a detailed comparison of the specification scale down for alternate systems within 5 working days.

Penciled on the bottom of your copy is a short note: "Sorry about this. That design of yours looked awfully good."

So you reach out, pull a catalog of semiconductor detectors off the shelf and begin to thumb through it.

10.5. SEMICONDUCTOR DETECTORS

The full explanation of the physics of semiconductors and their relation to metals and insulators can be found in texts on solid state physics and physical electronics. For our purposes, semiconductors are materials whose conductivity is between that of metals and insulators. Of importance to optical design, many semiconductors change their conductivity when they absorb radiation.

The distribution of electrons in a solid can be pictured in terms of a series of energy bands, each band holding only a limited number of electrons. The lowest energy bands contain those electrons tightly bound to the nuclear cores of the atoms that make up the solid. At higher energies are those electrons that are loosely bound to the cores. At low temperatures, when there is no excitation, the solid is in its ground state and all electrons are in the lowest energy bands. These are the *valence bands* of the solids as illustrated in the schematic diagram in Fig. 10.13. Above the highest valence band is the first empty band. An electron in this band is no longer bound to a single atom and is free to move through the solid, contributing to its conductivity. This is the *conduction band*.

In a semiconductor the amount of energy needed to raise an electron across the energy gap between the valence and conduction bands is small. Even thermal excitations cause some population of the conduction band. The promotion of an electron to the conduction band has another effect on the solid. The loss of an electron to the conduction band results in a vacancy in the valence band. Since this vacancy is also mobile in electric fields, its effect must also be accounted for in understanding the properties of semiconductors. This vacancy, appropriately called a *hole*, acts under an electric field very much like an electron with a positive charge.

In a semiconductor like cadmium sulfide, an absorbed photon creates an electron-hole pair. If an electric field is imposed on the illuminated sample, the resistance of the material will drop because of the additional radiation-excited charge carriers. Cadmium sulfide can change its conductivity by as much as 10^{11} upon illumination. A photoconductor such as this in an electric circuit can be used to detect radiation.

The range of spectral sensitivity of CdS is in the visible region. It is an inexpensive light detector that is used for simple light-sensing applications. Most other solid semiconductors are sensitive in parts of the infrared region.

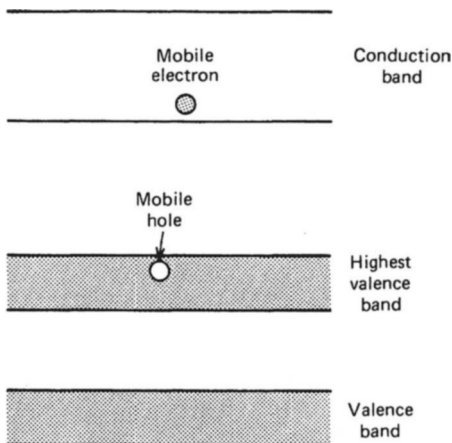


Figure 10.13. Energy band scheme for a semiconductor. When an electron acquires sufficient energy to be promoted to the conduction band, a mobile vacancy is left behind in the valence band.

While they are more sensitive and responsive than the thermal detectors discussed in Section 10.1, their sensitivity is not constant with wavelength and their time response depends on the level of illumination. Detectors, such as those made of InAs, PbSe, PbS, or InSb and numerous metal impurities in germanium, are sensitive in the infrared between 2 and 40 micrometers. However, in this region the heat of the detector itself produces noise, so they must be cooled to liquid nitrogen or liquid helium temperatures. In the visible and near infrared region of the spectrum, the frequency response and sensitivity can be improved by using a different type of semiconductor detector, one that is based on a junction geometry like that found in diodes and transistors.

10.5.1. Photovoltaic Detectors

Semiconductors (such as CdS) that contain no impurities are called intrinsic semiconductors; their properties are dependent only on the properties of the material itself. In contrast, most of the semiconductors used in electronics and optics have small amounts of other elements deliberately mixed with them. These impurities, or *dopants*, provide additional electrons or holes, thereby changing the conductivity of the material. If additional mobile electrons are added, then it is a negative or *n*-type material; additional holes produce a *p*-type material. By combining *n*-type and *p*-type materials in various geometries, semiconductor diodes, transistors, and related devices can be constructed. These same basic elements yield a whole host of low-cost, rugged optical detectors.

The basic structure for these devices, as it is for their electronic counterparts, is the diode. If *n*-type material is deposited next to *p*-type material, the excess electrons and holes in the vicinity of the junction between the two types of materials combine. That is the electrons in the *n*-type material diffuse across the junction and combine with the holes in the *p*-type material. This changes the relative energies of the conduction and valence bands, stopping the flow of holes and electrons across the junction, causing a permanent local electric field across the junction that maintains a charge carrier balance in the center of the junction (Fig. 10.14).

A quantum of radiation incident upon the junction region is absorbed and an electron-hole pair is created. If the junction field separates the carriers before they can recombine, then in an open circuit a nonzero voltage appears across the diode. Alternatively, in a closed circuit, current will flow, as shown in Fig. 10.15. If a voltage or current measuring device is added to the circuit, we have a light detector. A detector that supplies the voltages or current without any additional voltage source is said to be operating in a *photovoltaic mode*. The exact characteristics of the detector depend on the diode geometry. The rest of this chapter describes these geometries and their corresponding detectors.

A semiconductor detector can convert radiation into electrical power. The device that does this best is the *solar cell*. To get the greatest signal from a semiconductor the junction region is made as large as possible. This is achieved by making the junction the same size as the device, as shown in Fig. 10.16. By

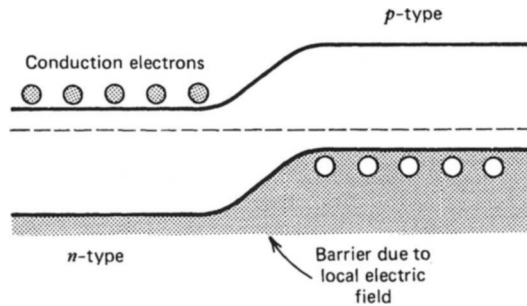


Figure 10.14. *p-n* Junction diode. The conjunction of two semiconductors with excess carriers of opposite sign results in electrons from the *n*-region combining with holes in the *p*-region. The reduction of charges in both regions produces a stabilizing field that shifts the energy of the bands and provides barriers to additional migration of electrons and holes.

depositing a thin (about $150\ \mu\text{m}$) *n*-type layer on a thick *p*-type substrate that is backed by a conducting electrode, light incident on this broad area creates electron-hole pairs that provide electrical power. The conducting electrode on the *n*-face is a network of thin conductors that connect to a common conductor. The efficiency of production depends on the spectral content of the radiation and on the thickness and composition of the thin *n*-type layer, since the depth of absorption of radiation depends on wavelength. The radiative-to-electrical energy conversion takes place only in the very narrow junction region. The spectral response of these and other semiconductor diode detectors covers most of the visible region, but varies with the size of the energy gap between the valence and conduction bands in these materials. A typical response curve is shown in Fig. 10.17.

For optical designers, solar cells can serve as inexpensive detectors of high light fluxes for such applications as laser beam scanning detectors, detectors to protect other light sensors against high illumination levels, and other simple

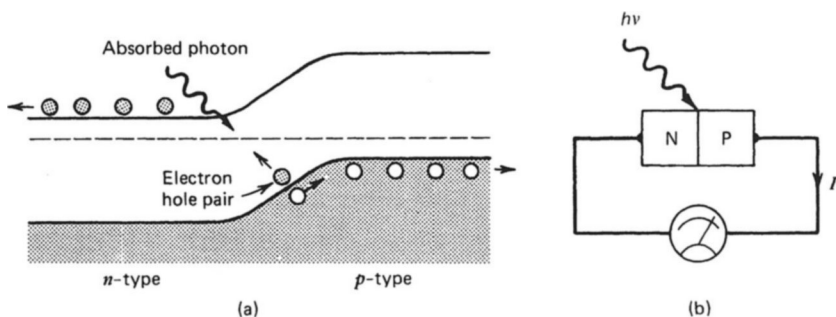


Figure 10.15. Photodiode. (a) A quantum of radiation absorbed in the junction region creates an electron-hole pair. The presence of these additional charges can cause a voltage to appear across the diode if it is an open circuit, or a current to flow if it is short-circuited as in part (b).

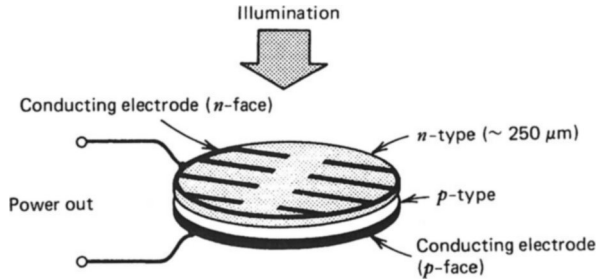


Figure 10.16. Solar cell. To convert the largest amount of radiation to electrical energy, the top layer of the diode is made as thin as necessary to cause the bulk of the radiation to be absorbed in the junction region. To reduce shadowing, the top electrodes are made as thin as possible and as close as needed to collect the photo-generated electrons.

tasks. There are several drawbacks to these devices. Since the field that is generated is dependent on the level of illumination, the response is nonlinear. Also, the frequency response of solar cells is usually less than 5 kHz.

If the area of the solar cell is reduced considerably, the overall signal will be reduced, but the frequency response will be enhanced. These photodiodes have increased linearity (about 10^4 times greater than a solar cell and much lower noise). The frequency response is inversely proportional to the size of the active area A of the photodiode,

$$\nu_c = \frac{K}{A}, \quad (10.5)$$

where $K = 0.1 \text{ MHz cm}^2$. If a number of these diodes is deposited on a common substrate and in a close geometry, as shown in Fig. 10.18, the detector can be

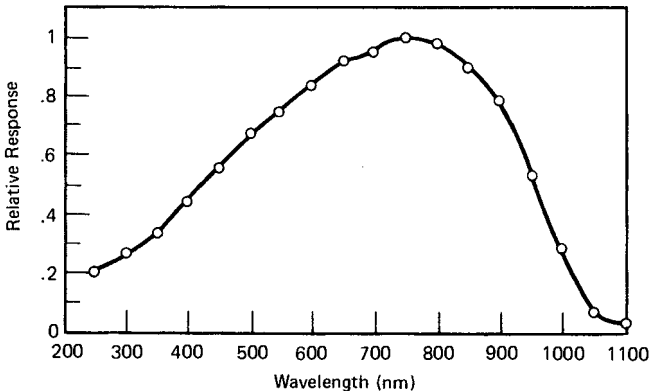


Figure 10.17. Spectral response of silicon photodiodes. A typical curve for a semiconductor photodiode. (Data courtesy of EG & G Reticon.)

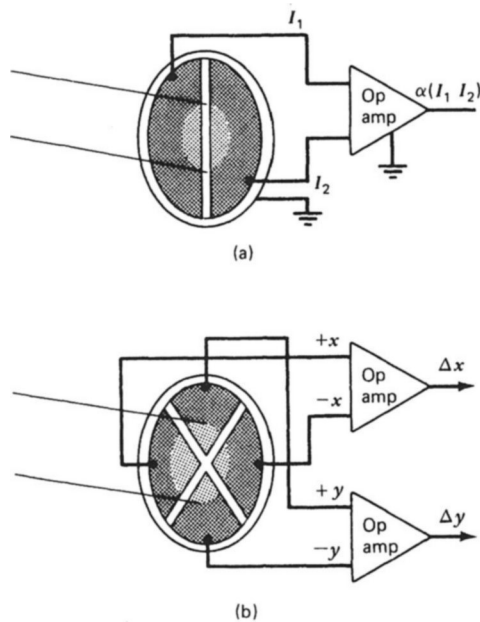


Figure 10.18. Sectioned semiconductor detectors. (a) A split-cell detector can be used to determine one-dimensional misalignment. Unequal illumination of the two cells results in an error signal at the output of the operational amplifier. (b) A quadrant detector can provide alignment information on two axes. Note that if the illuminating beam is too small, information on the center of the beam will be lost.

used to align and center an optical system. In Fig. 10.18(a) it can be seen that even illumination of the diode will result in equal outputs to the operational amplifier inputs. Any deviation from equal illumination of both detectors will produce unequal detector outputs and an error signal from the operational amplifier. Similarly, the quadrant detector shown in Fig. 10.18(b) will give an x -axis error but no y -axis error. These error signals can be used to close a control loop by providing signal to a control device such as a scanner to hold the beam at the centered position.

10.5.2. Photoconductors

One of the principal attractions for using semiconductors in the photovoltaic mode is that no external power source is needed. If, however, faster response times and more sensitivity are required, then a bias voltage is applied to the diode, and the device is operated in the *photoconductive mode*. In this mode the diode is reverse-biased by a voltage on the order of 20 to 45 volts. This increases the field in the junction region and creates a volume in which there are no mobile carriers, known as the *depletion layer*. In this region a newly created electron-hole pair will be separated and swept away before recombination can

occur. Because an electric field is present, the time response of such devices can be reduced to 10 nsec or so. The linearity of these devices remains good. The only disadvantage is that the reverse-biased diode is noisier than its unbiased cousins. Sensitivity, about 0.4 A/W, is approximately the same for devices operated in the photovoltaic mode.

To increase the size of the junction, and therefore the depletion layer, some diodes are sometimes manufactured with an additional layer which has the same conductivity as the intrinsic (pure) material. The layers are in a sequence *p-i-n*, where *i* stands for intrinsic. The device is called a *PIN photodiode* since it describes the diode construction. The quantum yield for such devices can be quite high, as much as 0.8 electrons per photon.

10.5.3. Avalanche Photodiodes

There is a large gap in sensitivity between inexpensive semiconductor photodiodes at approximately 1 A/W and costly photomultipliers at 1000 A/W. This gap is bridged by *avalanche photodiodes*, whose sensitivity is in the neighborhood of 10 to 1000 A/W. In a way, the avalanche photodiode is the semiconductor equivalent of a photomultiplier tube, since it also operates by current multiplication. If the reverse bias on a photodiode is steadily increased, a voltage is finally reached at which the diode no longer works as a one way conductor. At this breakdown, or Zener, voltage the current increases rapidly with small increases in reverse bias, as indicated at point V_Z in Fig. 10.19. If a diode is biased at a point V_B just above the shoulder of the curve, any additional carriers in the junction produce a large change in the amount of reverse current. On a microscopic basis, an absorbed photon creates an electron-hole pair. Each carrier, accelerated by the large bias field, collides with the semiconductor crystal lattice. The increased kinetic energy is converted to the creation of additional carrier pairs, just as electrons in the photomultiplier generate more electrons by colliding with a dynode. It is this method of carrier multiplication

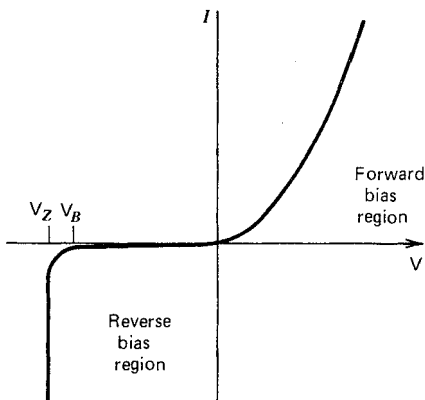


Figure 10.19. Characteristic curve for a semiconductor. Avalanche photodiodes are operated with a bias voltage V_B . Small changes in the illumination on the photodiode result in large changes in the current through the diode.

that gives the avalanche photodiode its name. The quantum efficiency of these devices approaches 90% at the peak response wavelength of 900 nm with response times in the vicinity of 0.5 to 2 nsec. Because the device operates on the shoulder of the current-versus-voltage curve, the response is linear over a small operating range, and the temperature of the device must be carefully controlled to maintain this linearity. Another disadvantage of this device is that the reverse current is never zero when the diode is in the dark. The dark current, usually on the order of 100 nA, serves as a lower bound for the sensitivity of avalanche photodiodes.

Thus, there are a number of semiconductor detectors that are useful in the visible and near infrared region of the spectrum. They are compact, relatively inexpensive and require little or no power with the exception of the avalanche types. About the only thing they cannot provide is images.

You have just left the airline ticket and other bills with the secretary after returning from the annual meeting of the Optical Society of America when one of your office partners hails you in the hall.

“Have you decided how many elements you will need in the array?” she asks with a grin.

“Run that by me again. I haven’t the faintest idea what you are talking about.”

“Well, there’s a rumor that they want to replace your new imaging package with a diode array. I heard that at the project review meeting they decided that the single diode sensors weren’t going to give us enough information to be able to track properly.”

You drop your armload on the desk, and thumb through the file drawer for the folder containing the diode array material. This will have to be a fast job. If the past is any indicator, specifications will be needed by Friday and a breadboard of the system in a week.

10.6. PHOTODIODE ARRAYS

With the exception of a few electronic devices to transmit electrical signals, all of the basic concepts needed to understand the semiconductor diode array have already been presented. In brief, a *diode array* is a linear or two-dimensional array of diodes hooked to an electronic circuit that can store and read out the photodiodes response in an orderly manner. It is, in a sense, a cross between an image tube, a photographic plate, and a photodiode.

The construction of these arrays is very similar to that of large-scale integrated circuits. When examined under a microscope they resemble the large random-access memories for computers. An example of a two-dimensional array and the finished chip packaged in a dual in-line pin arrangement is shown in Fig. 10.20. Basically a photodiode array consists of a series of photodiodes

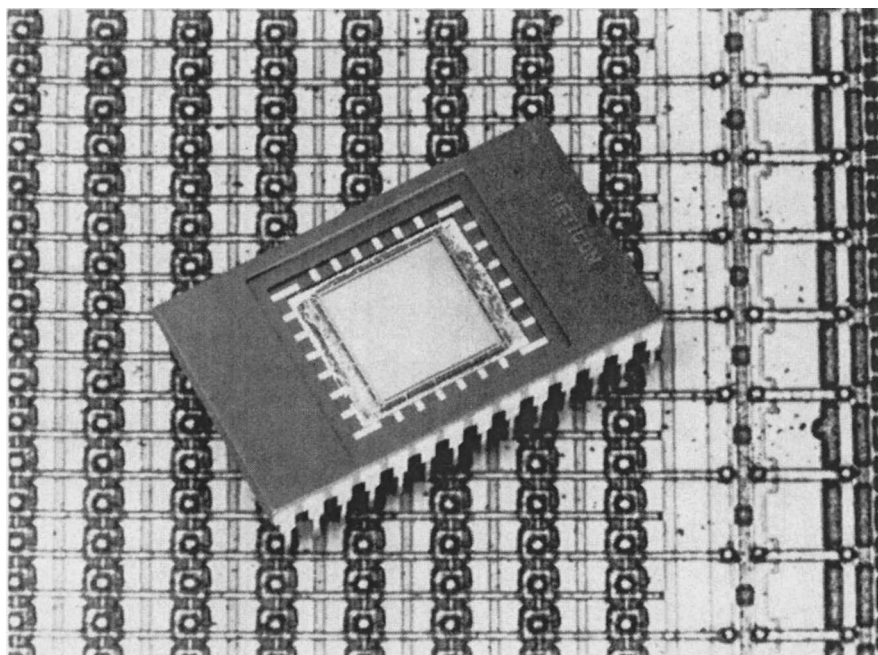


Figure 10.20. Silicon photodiode array. A 256×256 area array in a standard integrated circuit dual inline package atop a magnified view of the surface of the photoreceptor. The spacing between photodiodes is $40 \mu\text{m}$. (Courtesy of EG & G Reticon.)

deposited in a linear or square geometry. The diodes themselves either have a small capacitance or they are constructed with an accompanying capacitor as shown in Fig. 10.21. All diodes and capacitances are connected to a common line and to a common switch. In the simple schematic, a video line is connected to the diode when a pulse is sent from a shift register to close the switches. The shift register is nothing more than a clocked counter that provides control pulses at each of its output lines in sequence, enabling it to scan a large number of circuits, one after another. At the start of a scan of a photodiode array the shift register connects the diodes and capacitor to the video line, charging the capacitor and creating a depletion layer in the photodiode. This is repeated for all the diodes in the array. At this point the array is analogous to an unexposed photographic plate. During exposure incident photons are absorbed, creating electron-hole pairs just as in single photodiodes. Here, however, the carriers reduce the charge on the capacitor in the diode circuit, thereby providing a means of integrating the signal received by the diode. The pattern of charge reductions in the photodiode circuits is the equivalent of the latent image in the photographic plate before development. To “develop this plate,” the shift register is again scanned. As each of the diode circuits is recharged, the current on the video line not only provides the charge, but also, in the act of recharging,

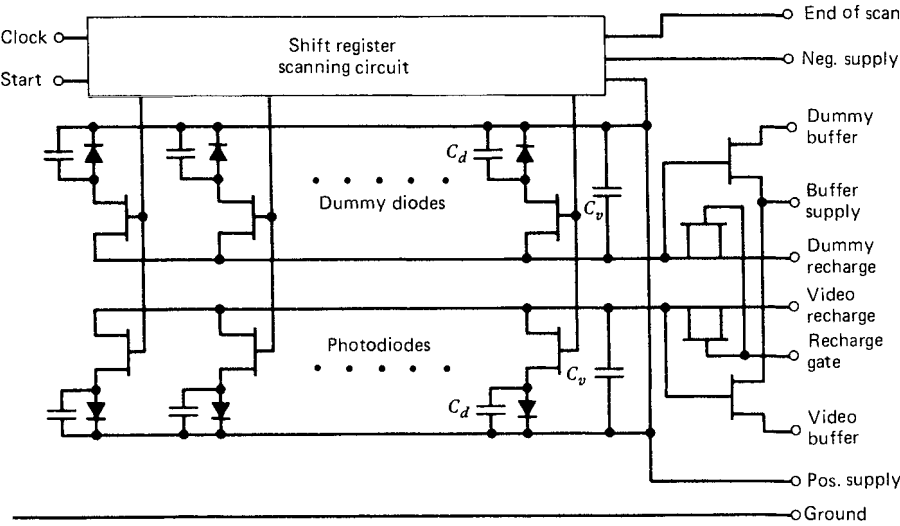


Figure 10.21. Schematic of a diode array readout circuit. In this version, the diodes are clocked out one at a time onto the video line by a digital shift register. (Courtesy of EG & G Reticon.)

measures the amount of illumination on that diode since the last scan. This is similar to the action in the storage tube in which the signal was obtained by recharging the storage surface (Section 10.4.2.1). The shift register scan sends a sequence of pulses on the video line to be analyzed. The pulse height providing information on the illumination level and the pulse position on the location of the diode.

Another method of reading out the pattern stored in the diode circuits is the *charge coupled diode array* or *CCD array*. The circuit for the CCD array, shown in Fig. 10.22, is the same in all respects as the photodiode array except that an analog shift register is used in place of the digital shift register. An analog shift

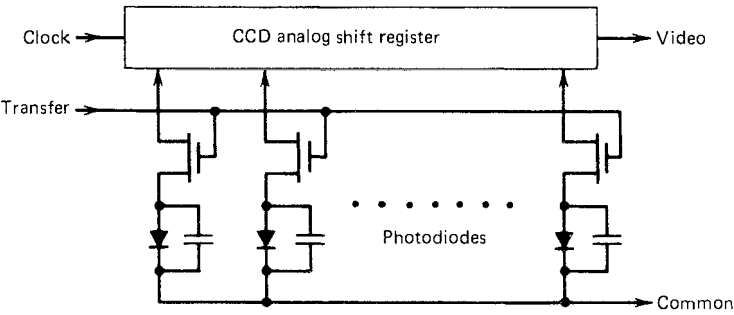


Figure 10.22. Charged coupled diode array. In this circuit all diodes are read simultaneously into an analog shift register that stores the position and amount of signal to be clocked out later to a video line. (Courtesy of EG & G Reticon.)

register preserves the size of the signal read from each diode. Instead of sequential readout of the diode circuits, all diodes are read into the analog shift register simultaneously. The pulses are then clocked out of the shift register in sequence. If the pulses are then sent into an analog-to-digital converter, the pulse heights are then available for storage in a computer for further analysis. Note that the readout of the array is simultaneous in this second case—a beneficial feature for high-speed imaging applications. The maximum sample rate of these devices is about 40 MHz. At the higher speeds the diodes are read by two shift registers, each register accessing alternate diodes. At the end of the scan the contents of the registers are then stored separately in the correct locations or interleaved to produce the proper sequence of data.

The response and spectral sensitivity of these arrays are similar to their discrete relatives (Fig. 10.17), but they suffer additional noise problems from thermally generated carriers during long integration times and from switching transients caused by readout. The first of these can be reduced by cooling the array to temperatures well below room temperature (-80°C in some cases). A 10^4 -fold reduction in noise can be realized by cooling to -40°C . The second problem, switching transients, can be eliminated by reading out a parallel set of unilluminated dummy diodes and subtracting their signals from the signals of the illuminated set.

The separation between the photodiodes determines the limiting resolution of these devices. Most moderate resolution arrays are made with center-to-center distances of $25\text{ }\mu\text{m}$ (0.001 in.); high-resolution arrays are made with $15\text{-}\mu\text{m}$ spacing. If one calculates a value of 70% of the interline frequency, as in the case of storage tubes, these devices have maximum spatial frequencies of approximately 28 lines/mm for the moderate resolution arrays and about 45 lines/mm for their high-resolution counterparts. The actual performance for a system, of course, depends on combined transfer function of all components.

When processing the signal from an array, it is sometimes necessary to take into account any variation in the diode-to-diode response. This is particularly true in spectroscopic and astronomical applications where relative irradiance measurements are important. Usually this calibration is performed by illuminating the entire array uniformly, then reading and storing the response of each diode in a computer memory. Later, the response of each diode is scaled to the response recorded under uniform illumination.

The photodiode array constitutes an excellent marriage between optics and semiconductor technology. It can be used in industrial applications as a noncontact sensor for sorting and measuring as well as for pattern recognition. In scientific spectroscopy it serves as an electronic photographic plate with the advantages of linearity and computer compatibility. In the areas of graphics and communication it can replace television and other display technologies. The photodiode array is not the answer to all optical design problems. The cost of the most useful devices is still relatively high. This is particularly true if the application requires computer storage and accompanying software. However, just as in the case of other semiconductor products, the prices of these devices

will almost surely fall. There are, for example, prototype 35-mm cameras with CCD arrays and tape recorders being produced with the prospect that they will replace the photographic silver halide technology. Also, as was noted earlier, CCD arrays resemble random access memories (RAMs) for microcomputers. Small primitive cameras using dynamic RAMs whose surface can be exposed to light are available for microcomputers.

Your office partner appears at the door to the office. "Guess what?" she asks, her face beaming.

"What?" you echo, trying not to encourage her.

"I'm the new project supervisor."

"Well, congratulations!" you offer. "But what's happened to the old supervisor?"

"He's been made Vice President of Research. Which brings me to why I am returning to this humble office." She pauses and waits for you to ask her her reasons. Receiving no question, she continues. "I'd like you to become project director while we shake the bugs out of this. What do you say?"

"Gee, I don't know. Why me?"

"Because we need someone who knows something about detectors and you're the expert around here. Oh, by the way, you're going to be pretty busy so you'd better put together a request for someone to help you. Then you can teach them all that you've learned since you've been here."

10.7. DETECTOR GEOMETRIES

Most of the examples in the text have shown a detector located at the image plane or focal plane of an optical system. If it is a nonimaging application, then it is enough to collect as much light as possible and direct it onto the detector. For example, if the integrated or total transmission of a transparency is being measured, the illuminating source and not the transparency should be imaged onto the detector. If, however, the content of the transparency is to be analyzed by a diode array, then the image should fall on the detector plane.

Except for a brief mention in Example 3.A, there has been no discussion of detector geometries that respond to light scattered from a surface. In many materials applications and scanning applications these geometries are used because the surfaces of the materials are opaque. Some of the scattered-light geometries are shown in Fig. 10.23.

The simplest geometry is *proximity placement*—locating the detector next to the scattering point. A strip photodiode detector, such as the one shown in Fig. 10.23(a), can be used to pick up image information by detecting the light from a scanned laser beam that is scattered by a document placed next to the detector surface. Unless the detector is extremely uniform over its entire length,

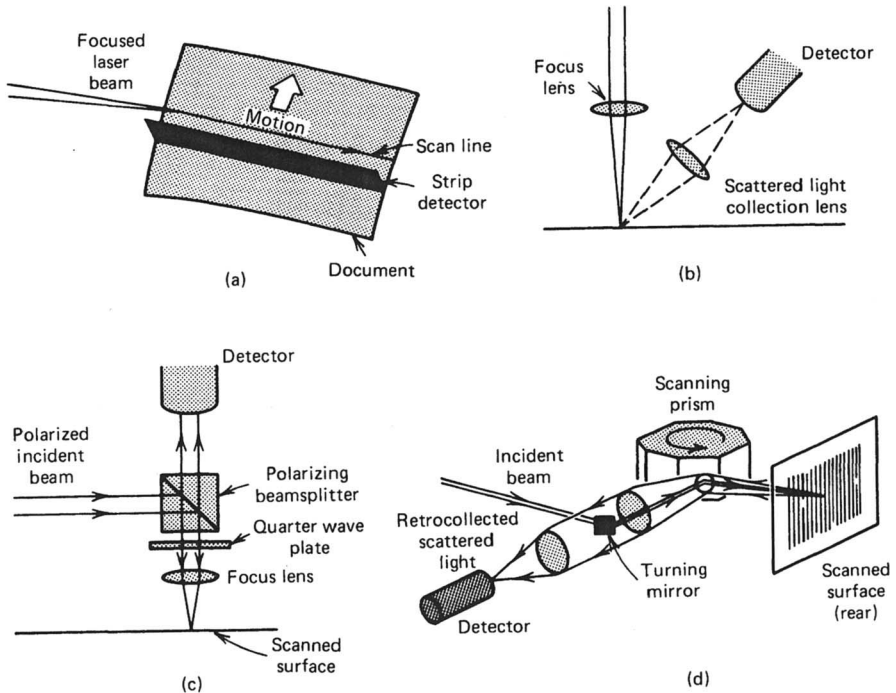


Figure 10.23. Detector geometries. (a) Proximity placement. Detector is positioned close to the scattering point. (b) Imaged detection. A second lens is used to image the scattered light onto the detector. (c) Beamsplitting geometry. The quarter wave plate is used to rotate the plane of polarization of the reflected light so that it will pass through the polarizing beamsplitter instead of being reflected. (d) Retrocollective geometry. The focus lens also serves as the collecting lens for back-scattered light, while the turning mirror blocks the direct reflections.

only high contrast objects, such as line drawings and documents, can be reliably scanned by this system. In many applications (Example 3.A is one) it is not possible to put the detector close to the scattering point. In such cases the scattered light is collected by a lens and the scattering point is imaged onto the detector (Fig. 10.23(b)). Because most light is either backscattered or forward-scattered along the beam, a detector is usually placed near the input beam or in the path of the reflected beam. In the case of reflection the directly reflected beam is blocked out because the strength of this part of the beam is indicative of the quality of the surface and not the information printed on it. (Certainly, information can be encoded by changing the texture of the surface and while this is considerably more durable than printing, it is usually more expensive.)

Another important method used in a *reflecting* or *scattering geometry* employs a polarizing beamsplitter and quarter-wave plate (Fig. 10.23(c)). The beamsplitter is constructed to reflect light of one polarization and transmit the other polarization, circumventing the usual problem with beamsplitters, that

only one quarter of the light is ever used. Light of one polarization is reflected by the polarizer and then converted to circular polarization by a quarter-wave plate (Section 1.6). Upon reflection the handedness of the circular polarization is reversed (say, right- to left-handed) and after its next traversal of the wave plate it has been rotated through 90° . Thus it is polarized in the correct sense to be transmitted by the polarizing beamsplitter to a collection lens and focused on the detector. Any absorption, scattering, or depolarization will reduce the signal to the detector and provide information on the scanned surface. Geometries similar to this one are used on laser-based videodisc readers.

To take advantage of forward and backward scattering, a beam can be directed onto a focusing lens by a small turning mirror. The scattered light is then collected by the same lens, while the direct reflection is blocked by the turning mirror (Fig. 10.23(d)). This arrangement is called *retrocollective detection* and is probably the most useful collection geometry, both in terms of efficiency and flexibility. Since the light is collected in a solid angle centered about the scanning beam, this geometry is ideal for post-objective scanning. It can be used to scan a document or bar code without determining which illuminating beam is being read. In most other geometries the scattered light would have to be collected by another lens.

There are descriptions of numerous detection systems in the technical journals, such as *Applied Optics* and *Optical Engineering*. A study of these systems can reward the student with insight as to how others have solved their problems. Also, attention to new detectors in product magazines and at technical exhibits can provide additional design options.

PROBLEMS

- 10.1. A Q-switched ruby laser ($\lambda = 694.3 \text{ nm}$) producing 1 Joule in 15 nsec is fired at a photoemissive surface. If the surface were 20% efficient in converting photons to photoelectrons (and could stand the jolt!), what would be the resulting current? The charge on an electron is 1.6×10^{-19} coulombs and an Ampere is 1 coulomb per second.

Problems based on detectors are somewhat more difficult to devise since they are either fairly easy calculations or almost impossible design problems that would depend on a survey of commercial devices. For the following problems, refer to Table 10.1. For each problem state (1) the type of detector you would use, (2) any special characteristics, (3) the reasons for using it, (4) other possibilities, and (5) additional questions that would have to be answered to be able to do further specification.

- 10.2. A detector is needed to determine when an edge of a cardboard sheet is in the proper position for folding on an assembly line. The light source is a helium–neon laser.

- 10.3. A detector is needed for a guidance system based on acquiring and holding the image of a star on the detector.
- 10.4. A detector is needed to measure the output power of a number of sources against a standard blackbody radiation source.
- 10.5. A detector is needed for rapid assessment of the quality of images based on the presence or absence of sharp edges in the image.
- 10.6. A detector is needed to turn on a streetlight when it gets dark.

11

The Design Process

(HOW TO PUT IT ALL TOGETHER)

One of the delights of optics is the ease with which most systems can be understood. For the most part, all one has to do is to trace a light ray through the system. This is probably why the convention of light initially entering a system from the left is used: You “read” an optical schematic as though you were reading a sentence. Of course, in devices with several subsystems this may be somewhat harder to do. But usually there are separate drawings or color codings to allow one to understand what is happening. For example, in Fig. 11.1 the block diagram for a laser system is shown along with some considerations that must be addressed in the process of designing a system. It is all nice and linear; the design follows the light.

Although the optical system may be linear, everything in the system is connected to everything else! As we have seen in previous chapters, a specification in one area has an effect on the range of choices in another area. (See, for example, the choice of acousto-optic transducer material in Section 8.4.) The same system depicted in the block diagram in Fig. 11.1 is represented in Fig. 11.2 in terms of the design process. Designing an optical system is not so much like writing a sentence, where the words and letters are strung out in a linear fashion, as it is like solving a crossword puzzle, where entering an answer at 5 down affects the answer for 17 across.

In this chapter a procedure for designing optical systems is presented. It is based upon fairly standard methods of optical design. However, the un-experienced designer should be cautioned that this procedure resembles the scenario of an actual optical design in much the same way that the statistics of

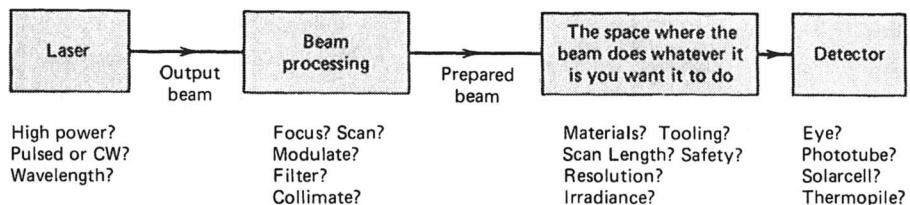


Figure 11.1. Block diagram of a laser scanning system. Listed are some of the considerations that should be addressed in the course of a design.

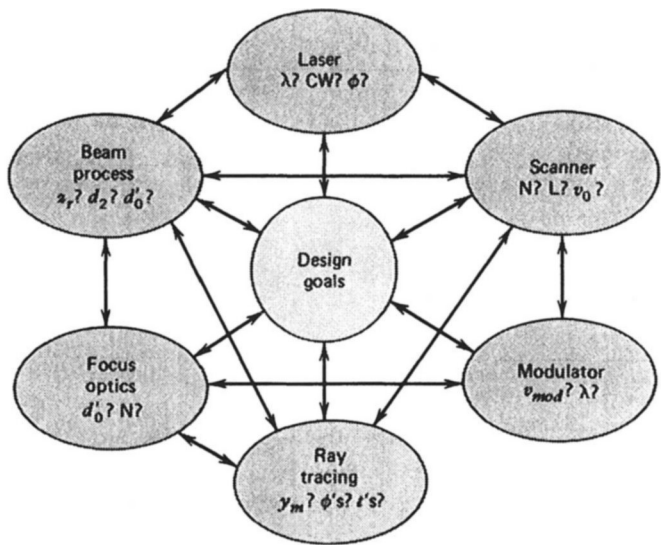


Figure 11.2. The design process of a laser scan system.

an average person resembles any individual—some things are similar, some things are very different, but no one matches the profile of an average person. Most interesting, the final product may bear little resemblance to the initial concept.

11.1. SETTING DOWN OBJECTIVES

Why does anyone start an optical design? It could be that someone (you, your boss, or someone else) recognizes a need that is not filled by the present instrumentation or methods. It may be that a device can be produced more cheaply or with less components through some new design. As the requirements for a system are better understood simplifications can take place that could not be permitted in a new, untested design. An example of this is the generation of new and simpler designs for supermarket scanners. In Fig. 11.3 two rotating mirror scanners are shown. The earlier design contains many facets and produced an extremely complicated scan pattern. The later design consists of only two mirrors yet provides comparable reliability.

Although there is a whole range of approaches to generating a design, a useful way to characterize various approaches is to label the extremes. There is the technical approach, which is to produce a design whatever the cost. One might call this the “brute force” method. This requires advanced components designed to the limits of current technology to be used to build a larger or more intricate version of earlier designs. The opposite extreme is the marketing

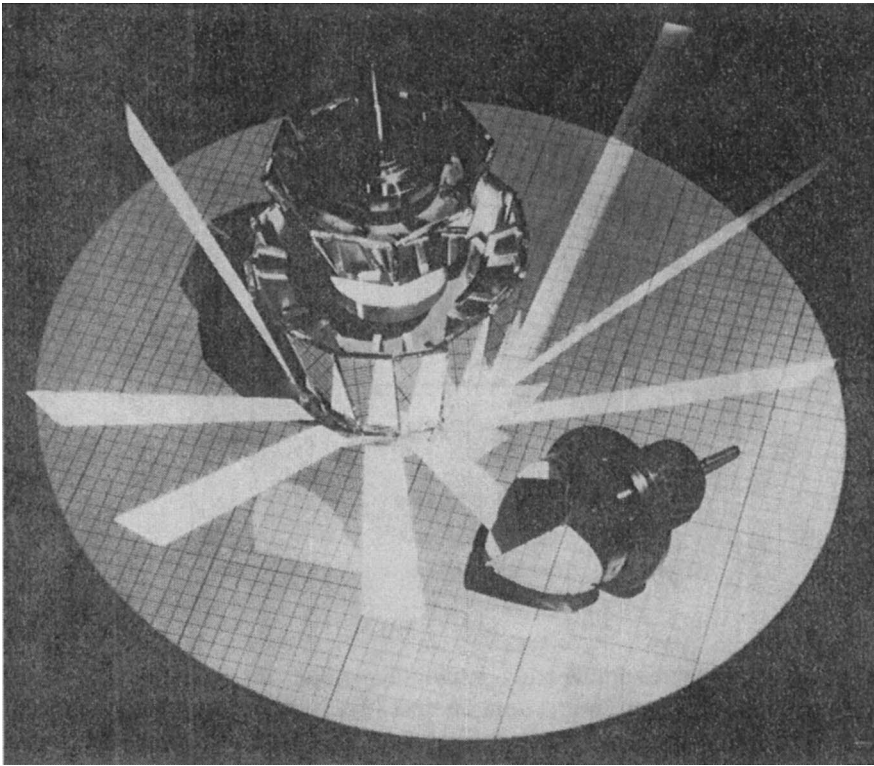


Figure 11.3. Two scanners for scanning the Universal Product Code (UPC), the code seen on many supermarket products. On the left is the older version. On the right is a newer, simpler version. (Scanners courtesy of Spectra-Physics, Inc.; photo by Vincent Mallette.)

approach, which expects a new design at no additional cost. This approach requires that stock components and assemblies from other instruments be lashed together in an attempt to produce a new and better instrument (or at least one with a few more saleable features). Between these extremes is a balanced approach. This approach appeals to simplicity through clean design in which a judicious choice of state-of-the-art components is used to produce a better optical system. The value and reason for each component is questioned. In addition to the case studies of the laser printer and slide projector optics presented at the end of the chapter, there are scattered throughout the chapter some examples of commercial systems that in one way or another demonstrate this balanced approach.

One caution: Beware of the excellent, unacceptable product. There are some areas where a good idea is not accepted because those who design the device do not understand the attitude and practices within the field. For example, in the

field of medicine, a device that requires too large of a deviation from accepted practice without results verging on the spectacular may not be adopted in routine clinical practice despite some obvious benefits. This is true in other fields as well. Several years ago an instant motion picture system was developed. While the system was well executed, competition from video tape devices overwhelmed the device in the market place.

Where do you start in an optical design? The fulcrum point for all designs is the determination of the design objectives. Additional time spent on this step can save considerable time and money later on. First, the overall objective should be determined. It should probably be no more than one sentence long. That is the easy part. Next the specifications for the system should be set down. This is considerably harder since not all of the performance characteristics can be known in advance. Still, the exercise should prove valuable since it will pinpoint some of the unknowns that must be investigated in the course of the design. Beyond the technical specifications that should be spelled out, non-technical aspects of the design should be considered. These can include human engineering. Questions regarding how the device will be used, who will use it, and who will maintain it must be raised and answered. Perhaps the hardest part of this process is assessing the limitations of the device. This is an attempt to find out where the problems will most likely occur and what parts of the design are not critical and therefore may be traded off against more demanding specifications.

The question then is: Where do I start? In some cases, an idea or new component triggers the design area, and the designer is off and running. (Sometimes this is done without stating the objectives, as described above, causing someone's later distress.) In other cases, a survey of the literature and present instrumentation is useful. (Refs. 11.1 and 11.2.) Information on the state of the art in basic research and development can be found in technical journals like the *Journal of Applied Physics*, the *Journal of Quantum Electronics*, and others. Applied research work is found in *Applied Optics*, government laboratory and agency reports, National Technical Information Center (NTIC) reports, and the specialized volumes such as those produced at topical meetings held by SPIE, the International Optical Engineering Society, and by the Optical Society of America. In addition, announcements and data on commercial devices and components are found in the semi-technical journals such as *Laser Focus*, *Lasers and Applications*, and *Photonics Spectra*.

The final part of this first step should be an evaluation of the proposed system in comparison with existing systems. Particular attention should be paid to differences in those parameters that give increased capability or require higher tolerances or a new design. These should then be balanced against those parameters that allow reduced tolerances or rely on an existing design. The trade-offs between these choices should be made as clear as possible. Finally, one should ask the question: Why hasn't this been done already? A good assessment of the objectives should give you the answer.

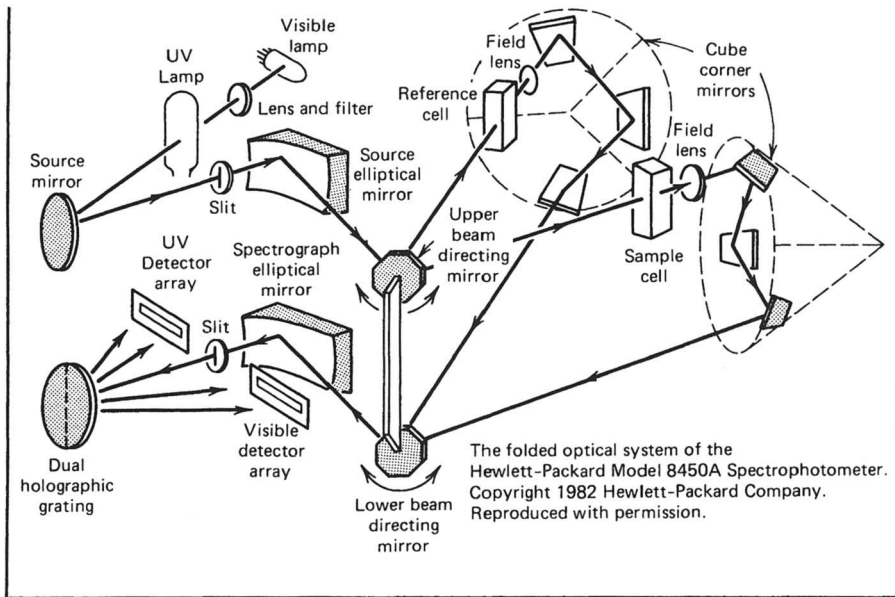
11.2. PRELIMINARY DESIGN

Generating a design is somewhat like riding a bike. It looks easy to those who watch and to those that know how, but it is very difficult for those who are learning. Sometimes a straightforward assembly of components is all that is required; other times a modification of an existing design is made to meet new specifications. The literature search used in determining objectives could be of use here. There are, of course, a whole set of noncritical evaluation procedures that range from brainstorming (a highly overrated group process in which ideas are tossed out with the hope that they will trigger additional ideas,

EXAMPLE 11.A. Measurement of the absorption spectrum of materials has been done one wavelength at a time by passing the output from an ultraviolet or visible source through a spectrometer (see Fig. 9.17). Design an instrument that will allow the simultaneous measurement of absorption at many wavelengths using new technology.

Solution: Instead of passing light through a spectrometer one wavelength at a time, two sources, a tungsten-halogen lamp for visible light and a deuterium lamp for ultraviolet were combined as a single source and focused onto a sample or reference cell by an elliptical mirror and a directing mirror (see figure). All of the light transmitted by the cell is returned via a set of these mirrors in the form of a cube corner reflector to another directing mirror and focused by a second elliptical mirror onto the slit of a spectrometer. The spectrometer consists of a dual holographic grating that disperses the light into visible and ultraviolet spectra and images these spectra onto two corresponding linear photodiode arrays. Transmission spectra of the sample and reference cells are taken by moving the directing mirrors from the reference to the sample position. The absorption spectrum is then generated by taking the ratio of the transmitted sample irradiance to that of the reference signal at each wavelength.

This design is almost impossible without recent advances in technology. The two diode arrays provide a simultaneous measurement of spectra from 200 to 400 nm and from 400 to 800 nm. The two holographic gratings side-by-side on the same substrate provide efficient use of light in each of the two spectral regions and are corrected for aberrations in their respective regions. The ellipsoidal mirrors were manufactured by a new technique called replication (see Section 6.3.6). Finally, microprocessor electronics enable the designer to handle all of the information produced by the spectrometer efficiently. The device was designed by engineers at Hewlett-Packard as the 8450 series UV/VIS spectrophotometer (Ref. 11.3). (Figure is copyright 1982, Hewlett-Packard Company and is reproduced with permission.)



one of which will do the job) to empathy (“If I were on a light beam traveling through the system, what would I see?”). Sometimes inverting a system, as is done in the spectrometer described in Example 11.A, can produce new designs. A study of some of the patents in the field can trigger new ways of looking at designs using a specific trick in a different application or by improving on an existing set of patents. While the infringement on patents with slight modifications to confuse the issue is unethical, the improvement of existing devices and methods is one of the benefits of a working patent system. When claiming a patent, a thorough understanding of the optical design is at least as important as making the design work. There have been cases where the inventor did not understand what he or she was doing and claimed the wrong features for patenting.

Once a design, or set of alternate designs, has been decided upon, a rough system layout of all components should be done. Ray sketching of an unfolded system provides information for component placement on the optic axis. Paraxial ray traces should be run on the system to identify all stops, pupils, and windows. Folded systems should then be drawn to assure that the components will fit and no new limiting apertures are introduced. In parallel with the ray traces any calculations on active components (e.g., aperture needed by an acousto-optic modulator to achieve a target resolution) should be carried out. Once these have been completed, an initial evaluation of the system should be done. If there are parallel designs, some may be dropped at this point.

The next step is to evaluate the hardware that is needed to produce the design. Some of the components are stock items that can be ordered from a manufacturer’s catalog; others must be custom made. When it comes to

nonstandard components, there is always the question whether they should be fabricated in-house or purchased from a supplier. If a part does not depart too far from like items being produced by others, it probably should be bought rather than made. Although the evaluation of such choices is one of the jobs of manufacturing engineers, the designer should not divorce himself or herself from such considerations. Sometimes the designer can suggest a small translation or redesign of a component that will reduce fabrication and assembly costs considerably.

With the introduction of hardware it is possible to do a second evaluation, again reducing the number of parallel designs. First, a cost analysis should be done to see if one can afford to build the system. Then a benefit analysis should be made: Even if you can afford to build it, is it worth it?

11.3. THE BREADBOARDING PROCESS

In some larger organizations the design and breadboarding processes are done by two separate groups. In many cases, however, the designer has the responsibility of producing a prototype device, usually on an optical table or optical bench. This process of producing a preliminary working system, taken over from electronics, is called *breadboarding*. At times, there is a tendency to wait until all the components needed are available before assembling the system. This is a poor practice since it wastes time and reduces the opportunity to discover unforeseen problems early. In many cases, substitute components such as simple lenses can be used in the breadboard to give some idea of the final arrangement. There are times when preliminary testing not only uncovers problems, but also generates new design possibilities. Another enjoyable and exciting aspect of optical design is the technical dialogue between designers as they tinker, modify, and invent new apparatus. (At times it is great fun.) Some modifications may require a return to new preliminary designs or perhaps a revision of the component list or specifications.

Once everything is assembled and the necessary tests have been completed, a third evaluation is made. If the results are poor, a careful step-by-step evaluation of the breadboard design should determine the problem and, one hopes, indicate the nature of the changes needed to improve performance. If the results are indifferent, then a return to some of the earlier design stages may be necessary. Sometimes it is at this stage that you learn the answer to the question: Why hasn't someone done it already? If the results are good, the system can be sent on to development for reduction to manufacturing procedures. Again, depending on the size and structure of the organization, the designer may be expected to contribute to a final design. Beyond the considerations given in this text the designer will have to worry about strain-free mounting of components, thermal compensation through choice of materials, and perhaps, microprocessor control of optical elements.

11.4. CASE STUDY—LASER PRINTER

Although the objective for the IBM 3800 printer is not stated specifically in an article detailing its design [Ref. 11.4], the description of the device indicates that the optical designers at IBM sought to produce a digital nonimpact printer to record high-speed computer output. In the quick survey of laser scanning systems the authors indicated that laser scanning systems used for producing high-accuracy photomasks for integrated circuits and data recording systems for computer output microfilm (COMs) were considered. The preliminary design, as shown in Fig. 11.4, consisted of a rotating polygonal mirror scanning a modulated laser beam across a rotating photoconductive drum using an electrophotographic process similar to the process described in Section 10.3.2 and illustrated in Fig. 10.6. Early analysis showed that the pitch angle errors (errors due to mirror facet normals not lying in a plane perpendicular to the axis of rotation of the rotating polygon) were the limiting factor in the design. The correction techniques being used at the time incorporated expensive electro-optic and acousto-optic deflectors, and in some cases, would not completely solve the problem. In order to be able to use the rotating polygonal scanner on a production basis some method had to be devised to overcome the pitch angle error problem.

The output from this system would be formed on the drum surface as a series of characters formed from a dot matrix scan. The characters were to be formed inside an 18×24 matrix similar to that shown in Fig. 11.5. The horizontal separation between dots is 0.14 mm (71 dots/cm) and the vertical separation was 0.176 mm (56.7 dots/cm). To create an overlapping pattern so that vertical lines

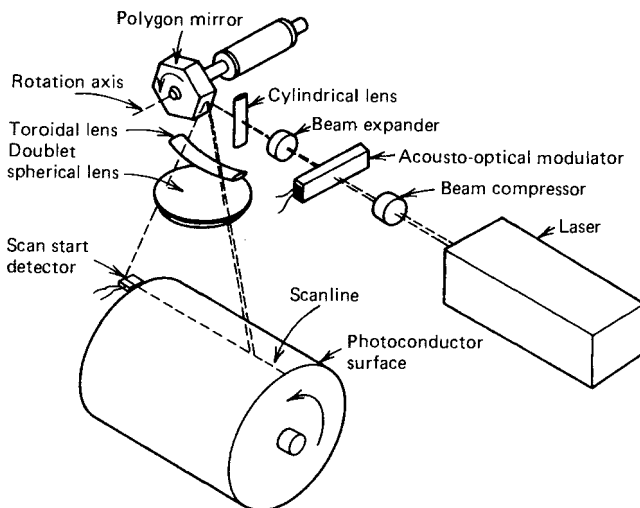


Figure 11.4. Preliminary design of a pre-objective scanning laser printer. (Copyright 1977 by International Business Machines Corporation; reprinted with permission.)

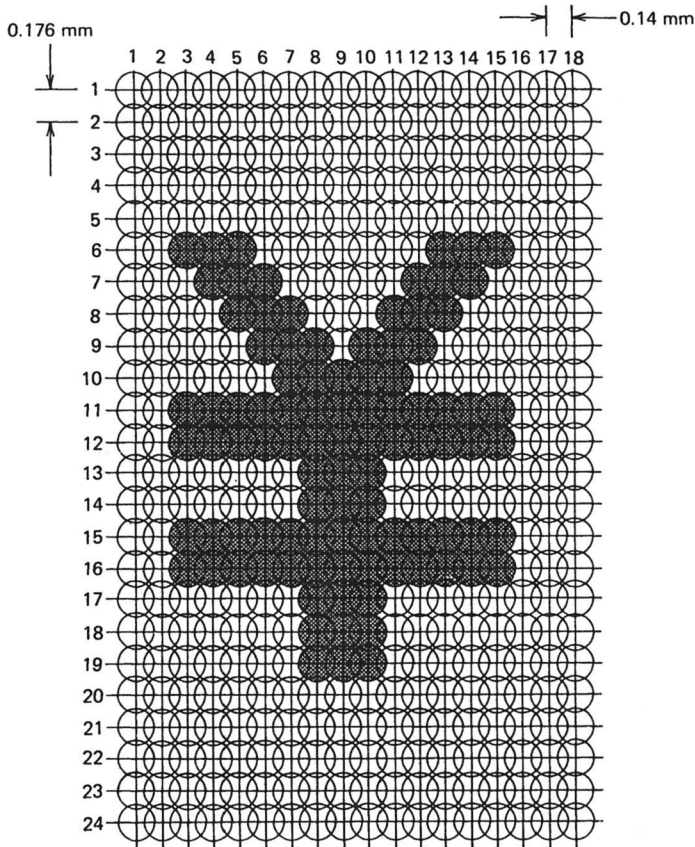


Figure 11.5. Structure of the 18×24 dot-matrix character cell formed by laser beam scans. In the horizontal rows there are 71 dots/cm and in the vertical columns 56.7 dots/cm. (Copyright 1977 by International Business Machines Corporation from the *IBM Journal of Research and Development*.)

appeared to be continuous, a spot diameter of about $250\text{ }\mu\text{m}$ is required. (It should be noted here that the $250\text{-}\mu\text{m}$ beam diameter will not necessarily yield a spot $250\text{ }\mu\text{m}$ on the paper. The nature of the electrophotographic process and the threshold of the effect (see Section 7.4.2) can change this value.) Standard character line widths were set at 0.4 mm (3 dots) for a vertical line and 0.35 mm (2 dots) for a horizontal line. While the requirements for a small spot size were high, it conferred a number of advantages, among them an increased flexibility of font and type size that a matrix with fewer and larger dots could not produce.

Another requirement growing out of the objectives was an extremely fast writing speed. If the printer must generate 31,000 characters per second, and each character cell is made up of 18×24 or 432 addressable dots, the writing

rate if 432×31000 or 13.4×10^6 dots/sec. The time for one dot to be written is then the reciprocal of this or 75 nsec.

If a commercial acousto-optical modulator could be used, then the rise time would have to be somewhat better than 75 nsec. Assuming that fused quartz is used in the modulator and that the transit time T_r must be at least 50 nsec, the beam size inside the modulator, according to Eq. 8.6, must be

$$\begin{aligned} D &= T_r v_s, \\ &= 50 \times 10^{-9} \text{ sec} \times 6 \times 10^5 \frac{\text{cm}}{\text{sec}} \\ &= 0.3 \text{ mm}. \end{aligned}$$

Thus, a standard helium–neon laser with a 1 mm output beam diameter must be compressed by a factor of 3 or 4 before being sent through the modulator.

The resolution of the system must be estimated. Although the spots are to be separated by $140 \mu\text{m}$ horizontally and $176 \mu\text{m}$ vertically, these distances are determined by the rise and fall time of the modulator for the horizontal direction and the accuracy of the scan line placement on the drum for the vertical direction. Indeed, the spot size was made intentionally larger than this to produce continuous lines. If we assume that the beam must scan nearly the entire width of a $14\frac{7}{8}$ in. by 11 in. computer printout page, use 14 in. (356 mm) as the scan length, and we take 0.14 mm as the horizontal separation between spots, then the resolution required of the system is

$$N = \frac{356 \text{ mm}}{0.14 \text{ mm}} = 2543 \text{ spots}.$$

Although a six-sided mirror is shown in the original article, the scan angle would be extremely large and the optics would also have to be large to accept such a large angle. In the printer, an 18-sided mirror is used to reduce the scan field angle to a practical 34° (0.6 radians). Using this value we can estimate the diameter of the beam illuminating a facet from the scan equation (Eq. 8.11). If $a = 4/\pi = 1.27$, $\theta = 0.6$ rad and $\lambda = 633 \text{ nm}$, then

$$\begin{aligned} D_0 &= Na\lambda/\theta \\ &= \frac{2543 \times 1.27 \times 633 \times 10^{-9} \text{ m}}{0.6} \\ &= 3.4 \text{ mm}. \end{aligned}$$

The actual facet length was determined by an iterative procedure of trying out certain dimensions and then computing the effects of the design. Ultimately it was found that the facet length was a function of the radius of the polygon, the

scan angle, the beam diameter at the facet, the beam wander on the facet, and the amount of “roll-off” on the mirrors (Section 8.3).

The beam, having been compressed to satisfy risetime requirements, must be expanded about a factor of 10 to satisfy resolution requirements. A schematic of this portion of the system is shown in Fig. 11.6. One can show [Problem 11.1] that if the acoustic wavelength is about 0.1 mm, the refractive index of the cell is 1.67 and the length of the transducer is 55 mm, the angular tolerance on satisfying the Bragg condition for the incident light beam is about 0.2 mrad. An analysis of the tolerances on the lenses of the compressor telescope initially used in the design showed that small tilts produced an unacceptable wavefront distortion thereby reducing the efficiency of the modulator. To counteract this the lenses were redesigned with different refractive indices that reduced the spherical aberration and coma to zero and the sensitivity to tilts to an acceptable level.

When the collimated beam illuminates the polygon facet, it is reflected into the lens and focused onto the photoconductor. The position of the dot along the scan line depends upon the position of the facet. If the mirrors do not form a perfect equal-sided, equal-angle polygon, the spot will not be positioned at the same point on the scan line after the scanner has rotated to the same relative position for the next facet. This defect is easy to correct, since a small detector located next to the drum can signal the electronics when to begin to send data to the modulator. Thus the data stream always begins at the same point for each scan.

The more difficult error to correct is that due to facets whose surfaces are not parallel to the axis of rotation. This pitch angle error causes the light that should be directed into the plane containing the optic axis of the focus lens and the scan line on the drum to be reflected into the lens at an angle 2ϕ to this plane, where ϕ is the tilt angle between the facet surface and a line parallel to the rotation axis. This is also the angle between the facet normal and the optic axis

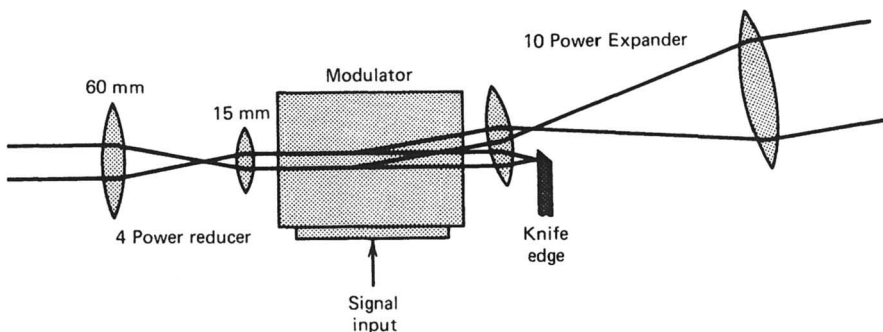


Figure 11.6. The beam compressor, modulator, and beam expander for the laser printer. (Copyright 1977 by International Business Machines Corporation from the *IBM Journal of Research and Development*.)

at the center of the scan [Problem 11.2]. If the focal length of the focusing lens is f_s , the deflection of the focused spot from the desired scan line is

$$\Delta x = f_s \tan 2\phi. \quad (11.1)$$

The exact value of the spherical lens is not known. However, if a 1 meter lens is used to focus the beam, the spot size for a collimated beam with a beam diameter of 3.4 mm (Eq. 7.44) is the required 250 μm . If we also require that the pitch angle error produce a spot displacement a factor of ten smaller than the 140 μm spot separation, say, no more than 15 μm , then the pitch angle error can be determined from

$$\begin{aligned} \tan 2\phi &< 15 \mu\text{m}/1 \text{ m} \\ &< 15 \mu\text{rad} = 0.3 \text{ arc seconds.} \end{aligned}$$

A specification that all facets of the polygon be held to a pitch angle error of about 0.3 arc seconds would result in a very expensive component and in the rejection of this design because of cost. The solution chosen by the designers of the device was to use an optical trick based on the fact that any ray through the front focal point of a lens becomes a ray parallel to the optic axis after passing through the lens (Section 1.3). This was accomplished not with spherical lenses, since the beam had to be scanned in the plane perpendicular to the rotation axis [Problem 11.3], but with two cylindrical lenses. In Fig. 11.7 side and top views of the optical path are shown. In the side view, in which the path has been unfolded, the collimated beam is focused into a line on the facet by a cylindrical lens L_1 and collected by another cylindrical lens L_2 one focal length distance from the facet. Thus any facet tilt will direct the focused bundle out of the scan plane, but the second cylindrical lens (L_2 in Fig. 11.7) recollimates the beam parallel to the optic axis of the spherical lens L_3 . All rays parallel to the optic axis are focused on the optic axis and therefore on the scan line. Thus the addition of the cylindrical lens pair reduces the overall sensitivity of the system to pitch angle error.

There is a limit to the amount of tilt that can be tolerated, otherwise any set of mirrors mounted in approximately the correct orientation would serve. The remaining problem is that the line focus does not stay on the facet for all angles of rotation of the polygon. If one draws a circle through the edges of the polygon, it can be shown that the distance from the facet surface to the cylindrical focus point is

$$S = r \left(1 - \cos \frac{\theta_1}{2} \right), \quad (11.2)$$

where r is the radius of the polygon and θ_1 is the polygon angle ($360^\circ/\text{number of facets}$). There are really two errors introduced by the geometry. One causes defocus of the spot on the drum since the line image is not located on the facet.

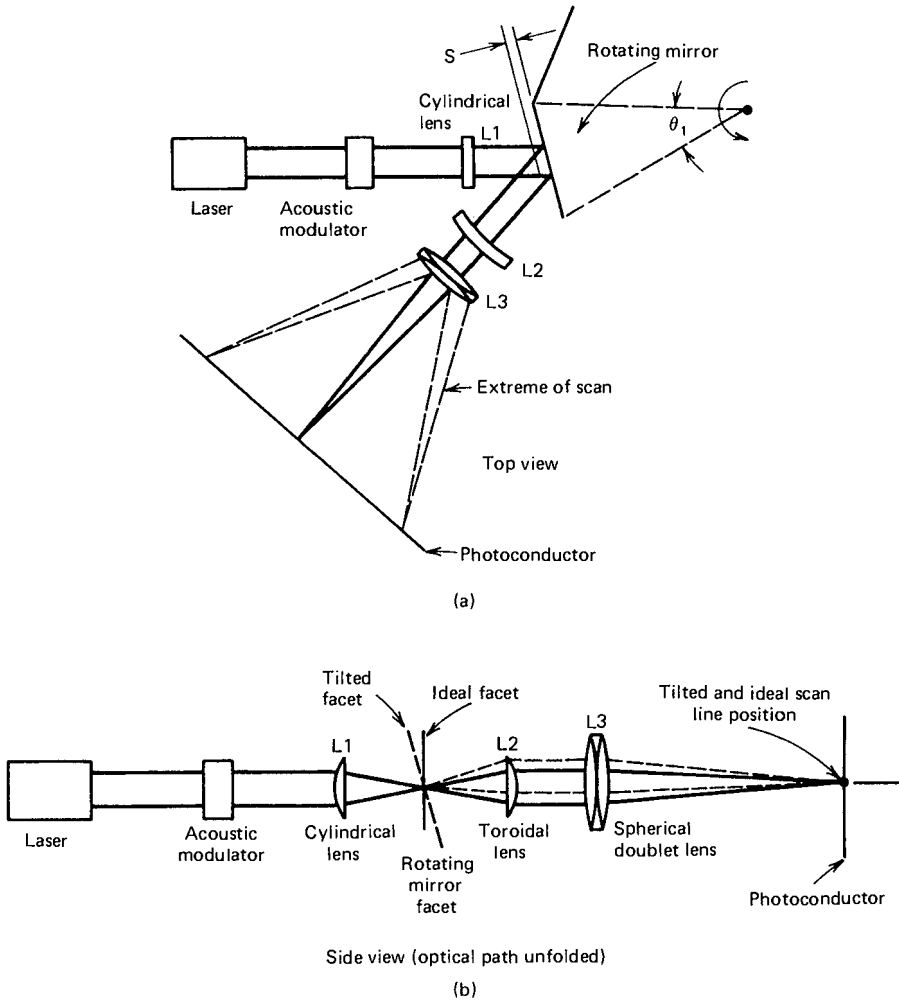


Figure 11.7. Two views of the polygonal scan geometry. (a) An overall view of the folded system. (b) An unfolded geometry that illustrates the use of the cylindrical/toroidal lens correction of pitch angle error (facet tilt). (The reducer-expander lenses shown in Fig. 11.6 are omitted in both views here.) (Copyright 1977 by International Business Machines Corporation from the *IBM Journal of Research and Development*.)

This is corrected by bending the second cylindrical lens into a toroid so that the optical path length from the laser to the drum is nearly constant for all scan angles. The other error causes the focus point to be located above the optic axis because near the facet edge the facet intercepts the beam before it comes to a focus. If the facet has a tilt of ϕ , then the focus is located a distance $\Delta x_1 = S \tan 2\phi$ above the optic axis of the toroidal lens. One can think of this distance as a small object Δx_1 in height (Fig. 11.8). When the light from an

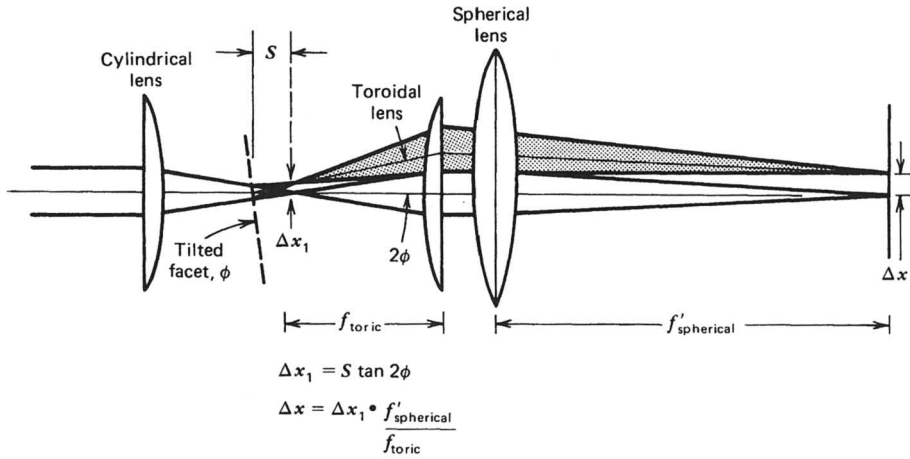


Figure 11.8. Scan line deflection due to sag S . (Copyright 1977 by International Business Machines Corporation from the *IBM Journal of Research and Development*.)

object located at the focal point of a lens is collimated and sent to another lens, the magnification of the system is the ratio of the focal length of the second lens to the focal length of the first [Problem 11.4]. Thus the error Δx_1 is magnified by f_s/f_t , where f_s is the focal length of the spherical lens and f_t is the focal length of the toroidal lens. The scan line error in this case is $\Delta x = \Delta x_1 f_s/f_t = S f_s \tan 2\phi/f_t$. The increase in pitch angle error tolerance G is the amount of deflection without correction divided by the error introduced by the defocus S ; that is

$$G = \frac{f_s \tan 2\phi}{S \cdot \frac{f_s}{f_t} \tan 2\phi} = \frac{f_t}{S} \quad (11.3)$$

Thus, the larger the toroidal lens focal length and the smaller the radius of the polygonal mirror the greater the tolerance in the pitch angle error for the polygonal scanner. For the system described in the article, a gain of 80 in pitch tolerance was achieved. If the required value of 0.3 arc seconds could be relaxed by a factor of 80, then specifications could be written for a scanner that had a pitch tolerance of 0.4 minutes of arc (1.2 mrad), a considerably looser specification.

The final element in the optical train is the spherical focusing lens. Besides focusing the beam to a spot on the detector, the lens must also linearize the spot velocity near the ends of the scan. Since the lenses are considerably smaller than the scan line, there is no possibility of a telecentric scan. As has been pointed out before (Sections 6.3.3. and 8.6.3), scanning can introduce a geometrical distortion into an image and this was compensated for by designing the opposite distortion into the lens. This lens could not be optimized with a single set up, but had to be evaluated with ray traces at different scan angles to

produce the best possible spot over the entire scan line. According to the authors, the final lens design produces “a scanning spot that maintains a positional accuracy of ± 0.013 cm over the 35.6-cm field.”

There is little that can be said about the breadboarding of this item or about any additional parts of the design process, except that the device has been introduced and has been on the market for a while. Included in the article was a note on the methods of alignment used with this device:

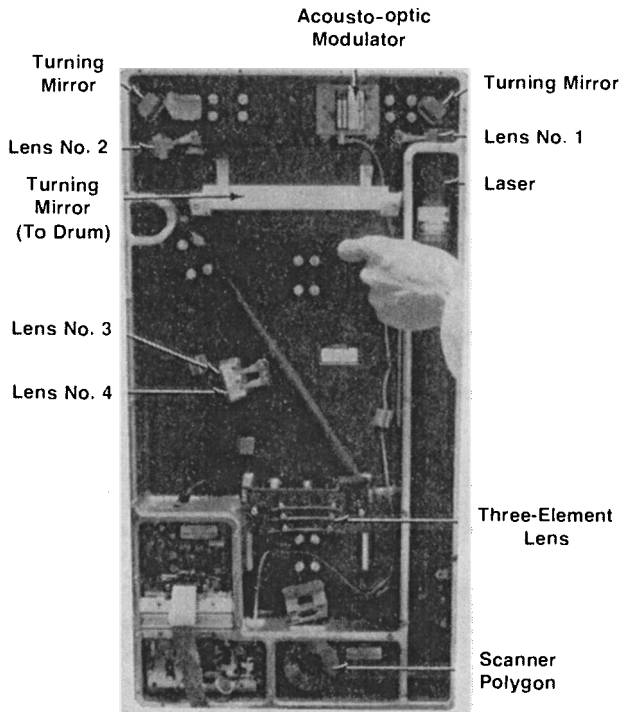
Because the laser subsystem contains a multiplicity of optical elements in a three-dimensional configuration, optical components and mounts with dimensional tolerances small enough to obtain the required alignment in assembly without adjustments are not practicable. Adjustments for many of the optical components are therefore supplied. The lateral centering of the components is accomplished by using the subsystem laser and an X - Y position sensitive light detector that can be mounted in a series of accurately located holes in the subsystem base casting. Focus adjustment is determined by the use of a small rotating slit scanner for measuring the laser beam diameter. Magnification, distortion and scan line displacement are measured by means of a fixture that holds a series of photodiode detectors in the scan plane.”

Another approach to a laser printer is described in Example 11.B.

EXAMPLE 11.B. Design a laser printer using post-objective scanning.

Solution: While pre-objective scanning can provide a near-flat focal surface, it requires a large, complex scan lens that must both produce a flat field and correct for angular distortions. The alternative is post-objective scanning. As was noted in Chapter 8, this method leads to a curved focal surface. However, if the spot size is not too small, the Rayleigh range may be sufficiently large that the spot does not change much over the scan region. For example, if the “depth of focus” is taken to be $-z_R/2 < z < z_R/2$, the spot size increases by only 12% of the waist diameter. If a waist diameter is $125\text{ }\mu\text{m}$, one can show [Problem 11.5] that a spot size of $140\text{ }\mu\text{m}$ can be maintained over a scan length of 298 mm (paper width) provided the center of deflection is more than 566 mm from the center of the scan plane and the scan angle is less than 28° .

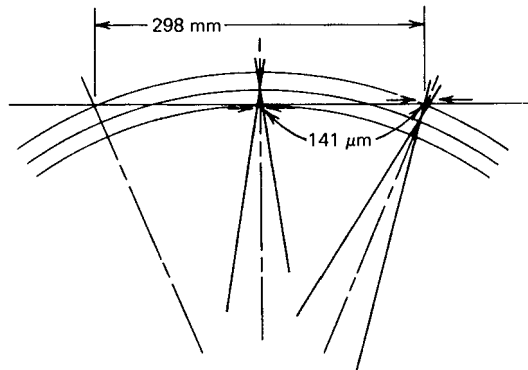
In the actual system shown in the figure, the center of deflection is located about 850 mm from the scan surface and the total scan angle is about 20° to compensate for other inaccuracies in the system. An acousto-optic cell is used for modulation and a polygonal scanner is used for scanning as in the case study in the text. Cylindrical lenses to focus the beam onto the scanner surface are used to correct for pitch angle error. However, instead of constructing a large toroidal lens to handle angular



(b)

correction, the designers have surrounded the second cylindrical lens with an afocal system, a short Galilean telescope whose positive spherical lens reduces the angle of incidence of the beam as it traverses the cylindrical lens and whose negative spherical lens restores the beam to its original deflection and convergence.

In the photograph of the system, note the use of the triangular leaf springs to hold the lenses in place. This type of a design requires close



tolerances on the layout and machining of the baseplate that holds all of the components. Alignment can be done using the turning mirrors in the corner of the case. The acousto-optic modulator is adjusted only in the Bragg angle. Because post-objective scanning is used the facet center to rotation axis distance has to be tightly controlled (see Section 8.3.1).

The triad of lenses consisting of the second cylindrical lens surrounded by the afocal angle reducer is hidden by a plate that holds the three lenses against alignment posts in the base plate. The system, the Model 2680 Laser Printer, was designed by engineers at Hewlett-Packard, Inc. (Ref. 11.5). (The figures are copyright 1982 Hewlett-Packard Company and are reproduced with permission.)

11.5. CASE STUDY—SLIDE PROJECTOR DESIGN

Not all of the optical systems to be designed these days are laser-based systems. The one illustrated above was chosen because there was sufficient information in the open literature and it illustrated some of the points that had been made in earlier chapters in a concrete manner. This section is intended to describe many of the considerations that go into the design of a slide projector with less mathematical detail than in the previous section.

In contrast to the somewhat demanding, but easily specified objectives of the laser printer, the specifications for a slide projector are more difficult to determine. Some are a matter of judgment. Starting with the light source for the projector, the wattage and filament structure of the bulb must be specified. This sounds easy enough. The question must be converted to a radiometric one: How bright does the screen have to be and how large a screen has to be filled? But this is then translated to a sociological question: Where do most people put the projector and the screen? At this point it may be marketing specialists and others who provide the data necessary to be able to translate radiometric calculations into a specification on a projector bulb.

To provide some basis for the design, the following specifications, subsequent calculations, and arguments are taken from a set of notes by Professor John McDonald of the University of Reading with his kind permission.

1. The screen is to be 2.65 m by 1.76 m at a distance of 7.76 m from the projector.
2. The screen illuminance (without a slide in position) is to be in the range of 100 to 150 lumens/m².
3. The slide size is 34 mm × 22 mm.

The luminance of a source has the same definition as its radiance (Eq. 3.7) except it is given in photometric units:

$$L_v = \frac{\phi_v}{A_s \Omega}$$

Thus the luminous flux in the projector is

$$\phi_v = L_v A_s \Omega. \quad (11.4)$$

The total power needed on the screen is the illuminance multiplied by the screen area. Using the higher of the two values, 150 lm/m², the total flux on a screen whose area is 4.66 m² is 700 lm. If one assumes that 40% of the radiation will be lost in the optical train from scattering and unwanted reflections, about 1160 lm is required.

A standard source of illumination for projection systems is the tungsten-halogen lamp. Similar to an ordinary incandescent lamp, it is filled with a halogen gas, usually iodine, that prolongs the life of the filament and prevents deposition of tungsten on the glass envelope. These lamps usually have a rectangular, flat dense filament with about a 2:1 aspect ratio. If the lamp is located near the center of curvature of a concave mirror, the filament and its real image form a square source. For such lamps luminance L_v is nearly the same for all sizes of the lamp; it is about 30 lm/sr-mm². The only difference between lamps of different wattage is the size of the filament. A number of these lamps and some parameters particular to this design are listed in Table 11.1.

If the values $\phi_v = 1160$ lm and $L_v = 30$ lm/sr-mm² are substituted into Eq. 11.4, the product of the source area times the collection solid angle is

$$A_s \Omega = \frac{\phi_v}{L_v} = 39 \text{ mm}^2\text{-sr}.$$

Table 11.1. Tungsten-Halogen Lamp Parameters

Watts	Filament Size	Area \times 2(mm ²)	Required Solid Angle (sr)	Condenser Collection Semiangle (deg)
50	3.1 \times 1.6	9.92	3.93	—
100	4.4 \times 2.2	19.36	2.01	53.1
150	5.4 \times 2.7	29.16	1.34	40.8
250	7.0 \times 3.5	49.0	0.80	30.2
500	10.0 \times 5.0	100.0	0.39	20.6

If a source area equal to twice the filament area is used, the solid angle required for lamps of various wattages can be found. The collection half angle α shown in Fig. 11.9 can be shown to be equal to $\sin^{-1}(\Omega/\pi)$ [Problem 11.6]. The values of α for each of the lamps has been calculated except for the 50-W lamp which cannot produce sufficient power to illuminate the screen even if the half angle were an impossible-to-achieve 90° . Based on the value of α , the design for the condenser for the 500-W lamp would be the easiest and the 100-W lamp would be the most difficult. In addition, the 500-W lamp would require a larger lamp holder and a fan for cooling. Even if a lamp can be cooled by convection, there are other reasons for including a fan in the design. A cooled lamp has a longer lifetime. Also safety regulations on the maximum temperature of the case must also be met. But it should be noted that adding a fan also contributes to noise in the room and increases the cost of the projector.

Köhler illumination (Section 3.4.3 and Fig. 3.15(c)) is to be used to provide uniform illumination of the screen as shown in Fig. 11.9. It is necessary to have some idea of the size of the projection lens so that the image of the filament will fill the lens. Since the magnification is 2.65 m/35 mm or 76 times, the image may be considered at infinity and the slide near the focal point of the lens. The chief ray angle \bar{u}_0 and the object length \bar{y}_0 are related as

$$\tan \bar{u}_0 = \frac{\bar{y}_0}{f} \quad (11.5)$$

If one assumes the screen will be filled and that the chief ray is drawn from the corner of the screen to the center of the projection lens, then $\tan \bar{u}_0 = 1.59/7.76 = 0.205$ and the slide semidiagonal is 20.5 mm giving the projection lens focal length as

$$f = \frac{\bar{y}_0}{\tan \bar{u}_0} = \frac{20.5 \text{ mm}}{0.205} = 100 \text{ mm.}$$

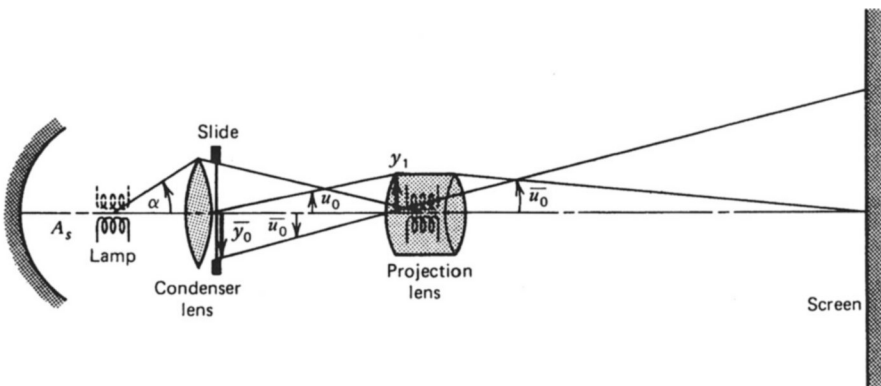


Figure 11.9. Schematic design of a slide projector optical system.

The size of the projection lens can be determined from the optical invariant. The semidiagonal of the 7-mm square source is 5 mm and the $\sin \alpha = \sin 30^\circ = 0.5$. Thus

$$I = 5 \text{ mm} \times 0.5 = 2.5 \text{ mm}$$

Since the slide semidiagonal \bar{y}_0 is 20.5 mm, the paraxial marginal ray angle is $u_0 = I/\bar{y}_0 = 2.5 \text{ mm}/20.5 \text{ mm} = 0.122$. This corresponds to an f -number $(1/2u) = 4.1$ and a projection lens radius of 12.2 mm. To fill the projection lens then the source with a 5 mm semidiagonal must be magnified -2.44 times by a condenser lens. It is impossible to do this with a single lens because the marginal slope angle for an axial ray from the source will be

$$u_1 = \frac{u_0}{M} = \frac{0.5}{-2.44} = -0.205$$

and the focal length of the condenser lens would come from a y - u trace

$$u_1 = u_0 - y_1/f_{\text{cond}}.$$

$$0.205 = 0.5 - \frac{20.5}{f_{\text{cond}}}.$$

The lens would have a focal length

$$f_{\text{cond}} = 29 \text{ mm}$$

and an f -number

$$F = \frac{f_{\text{cond}}}{2y_1} = \frac{29 \text{ mm}}{41 \text{ mm}} = 0.7,$$

an extremely high-speed lens. A lens such as this could produce a spherical aberration of 50 mm! Obviously a compound lens construction is required. Commercial slide projectors often use thick aspheric components as shown in Fig. 11.10. The cost of an aspheric mold can only be justified by high volume production and the quality of molded optics is inferior to polished elements. We will assume for this design that this is a special projector with a small production run and look for solutions using only spherical surfaces.

One approach is to use one or more lenses with centered and aplanatic surfaces, similar to those in the system illustrated in Fig. 6.22. An aplanatic meniscus lens (centered surface followed by an aplanatic surface) has no spherical aberration or coma, the dominant aberrations of a condenser. Their drawback is that while each aplanatic meniscus reduces the slope angle by the

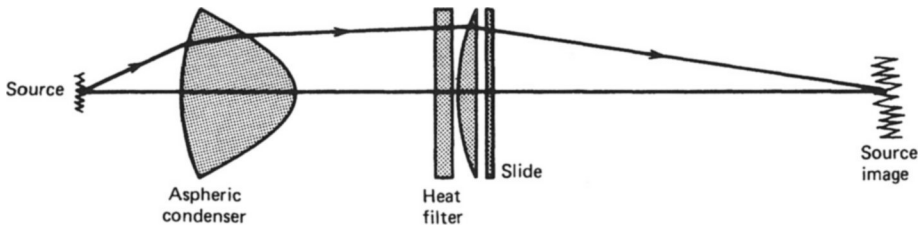


Figure 11.10. Slide projector condenser constructed with a massive aspheric collection lens.

reciprocal of its refractive index [Problem 11.7] it cannot cause the ray slope to change sign and form an image of the source. Thus one additional lens must be used. If the slope angle to the lens delivered by the two menisci is such that the additional lens is an equiconvex lens working a unit magnification $u_0 = -u_1$, the spherical aberration will be a minimum (a condition similar to the biconvex lens with an infinite object distance illustrated in Fig. 6.18). If the slide were located right after the single lens then the height of 20.5 mm for the marginal ray and a slope angle of -0.205 , the image distance will be 100 mm and at unit magnification this corresponds to $2f$. Thus the equiconvex lens should have a focal length of about 50 mm. If its refractive index is 1.517 (BK7), the spherical aberration is a considerable improvement over the former single lens. Using the aplanatic formula (Eq. 6.47), paraxial ray trace formulae (Eq. 5.13), some arbitrary choices of reasonable thicknesses and separations, a possible design layout, one used by Professor McDonald, is shown in Fig. 11.11 [Problem 11.8]. A heat filter, usually a dielectric mirror with short wavelength transmission and long wavelength reflection, is inserted somewhere between the lamp and the slide.

Rays at the edge of the slide suffer a substantial amount of spherical aberration, 8.2 mm. This is a large fraction of the source image diameter of 24.4 mm. The effect of the spherical aberration due to the condenser lenses on the marginal rays from the edge of the slide is that the central (chief) ray in the

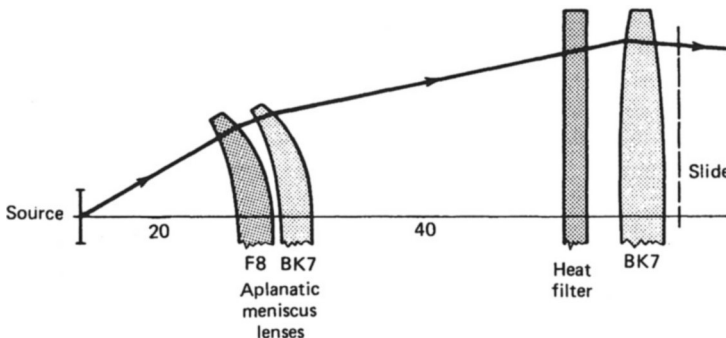


Figure 11.11. Possible design layout of a slide projector condenser.

cone of light directed through the entrance pupil crosses the axis before reaching the pupil. The effect of spherical aberration is to, in effect, shift the stop for marginal rays to EP_A from the entrance pupil for on-axis rays (EP_0) (Fig. 11.12). Two questions arise with respect to the projection lens:

1. Can light through the shifted entrance pupil get through the projection lens and onto the screen?
2. What effect on the aberrations of the projection lens will this pupil shift have? (In Chapter 6 it was shown that the aberrations were a function of the stop position.)

From question 2 we see it is not sufficient to simply get the light onto the screen if on the way the image quality is seriously affected. Question 1 can be addressed by opening the projection lens up, beyond the $f/4.1$ computed above and acknowledging a significant amount of vignetting at the edge of the field. The answer to question 2 is “perhaps not too much” if we can design the projector lens so it is insensitive to stop shift.

It is likely that the equiconvex lens also exhibits higher order aberrations (not true for the menisci). This could be determined by exact ray tracing. The design could be amended, perhaps by allowing the menisci to depart slightly from aplanatic. This will introduce some aberration, but obtaining greater initial bending and reducing the power of the equiconvex lens.

The design of the projection lens would usually be carried out by modifying an existing design. One source of designs is a collection of patent lenses compiled by Cox (Ref. 11.6). In doing so the spherical aberration and coma should be reduced as much as possible to make the lens insensitive to the pupil aberration introduced by the condenser system. Rapid evaluation with some simple ray-tracing programs can produce an acceptable design before subjecting it to one of the major design programs. Because the image distance is so far removed from the lens, it is useful to turn the design around and make the screen the object point and the slide the image point.

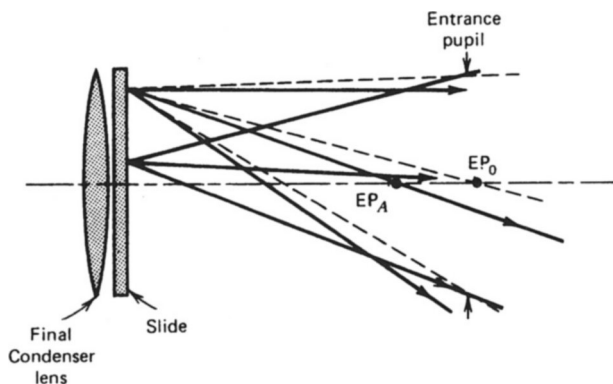


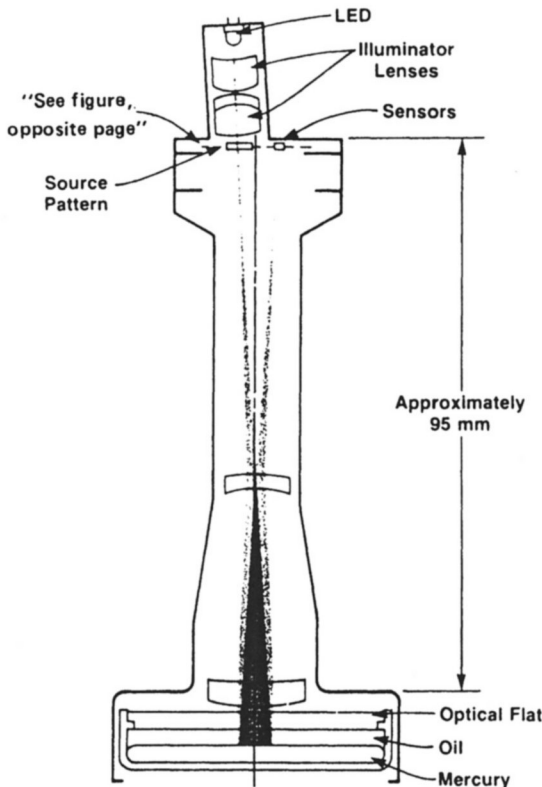
Figure 11.12. Condenser aberration can be translated as an aberration of the pupil function if the spherical aberration causes the marginal ray of a lens system to cross the axis at a point EP_A removed from the entrance pupil determined by the paraxial chief ray (EP_0).

removed from the lens, it is useful to turn the design around and make the screen the object point and the slide the image point.

In many cases a zoom lens is used to incorporate flexibility into the design. Despite one's initial tendency to design the largest possible lens aperture to get the highest possible illumination, the lens cannot be too large or else the depth of focus is too small. This would require that the projector always be positioned so the projection direction was normal to the screen. Although this is a common set up of a slide projector, there are sometimes audience arrangements that require a tilted screen. Also a small depth of focus would require continuous adjustment of the focus to compensate for small variations from slide to slide. In the design of the lens the fact that most slides are mounted in paper frames must

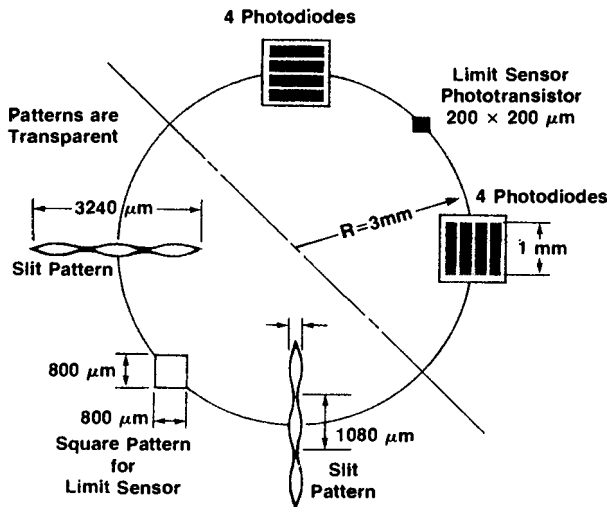
EXAMPLE 11.C. Design an optical tiltmeter.

Solution: The instrument shown in the figure is one solution to the problem of sensing a tilt on a platform. The innovative part of the design is not in the gravity-sensing mechanism, a captive pool of mercury covered



with oil to damp surface vibration, but in the source-detector pattern that sits above the reflective surface. Instead of a single x - y quadrant detector described in Section 9, the designers used two sets of sinusoidal masks and linear arrays of four diodes to measure tilt in two orthogonal directions.

The light emitting diodes (LED) illuminates the two sinusoidal patterns and a square pattern for a limit sensor. This pattern is formed into a collimated beam by the pair of lenses near the mercury mirror. These same lenses reimage the reflected light onto the diode arrays and limit sensor on the opposite side of a circle centered on the optic axis of the lens pair. If there is no tilt in the system, then the two interior photodiodes are equally and strongly illuminated and the two photodiodes on the outside of the array are equally but weakly illuminated. When the instrument is tilted, the sinusoidal pattern shifts on the diodes and the phase shift of the sinusoidal pattern can be detected by summing the differences in photocurrent between the first and third diodes with the difference between the second and fourth diodes. This results in a high degree of accuracy. The limit sensor is included to prevent a misreading of the sinusoidal pattern if the tilt were too large. This method of detection results in less than 1 arc



second resolution in a range of 240 arc seconds. This development by optical engineers at Hewlett-Packard, Inc., means that a distance or angle measuring device incorporating one of these tiltmeters need only be leveled to ± 120 arc seconds. The additional adjustments are compensated for electronically using the readouts from this tiltmeter. (The figures are copyright 1982 Hewlett-Packard Company and are reproduced with permission.)

be taken into account. Here the pillow contour of the film produces a curvature of the object field that must be translated into a flat image field.

Then there is the image that appears up on the screen. How good does it have to be to satisfy most people? The eye can resolve about 1 minute of arc. However, that cannot be translated directly into some resolution figure unless someone is willing to specify where the viewer is going to sit. Besides, there are other things that can cause loss of resolution: the distortion of the film due to the heat of the lamp, the tilt of the slide in its holder (sometimes called the film gate, or just gate), the flatness of the screen, and tilt of the slide projector. Usually people like their pictures to have somewhat more contrast than real life. This means a lens designed so that the contrast remains high at low spatial frequencies would produce a more pleasing picture to most viewers than a lens whose MTF dropped off more rapidly at low frequencies but imaged high spatial frequencies. In addition, the lens should be designed so that no aberration stands out above all others.

For such a device the basic attack is to set up a photometric solution first, assuring that the image will be bright enough to be useful, then design the lens to deliver the best possible compromise image. Other optical systems may demand a different ordering of priorities (Example 11.C).

11.6. CONCLUSIONS

No design is ever finished; it is stopped. Since most designs are a series of compromises with time, money, and nature, the scope of the design and the time devoted to it depend upon these limitations. In addition, the rapid innovations in the fields of optics render older designs obsolete and make possible hitherto unheard of performance. Only through continual awareness of the advances in the field through reading the literature, attending conferences and short courses, and daily practice of this delightfully complex art, can a designer ensure that he or she is ready to make light work!

PROBLEMS

- 11.1. Show that the angular tolerance for the acoustic modulator in the laser printer is as given in the text.
- 11.2. Show that the angle between the facet normal and the optic axis at the center of the scan is equal to the tilt angle between the facet surface and a line parallel to the rotation axis.
- 11.3. Explain why spherical lenses could not be used in the tilt error correction section of the case study presented in the text. Draw diagrams to illustrate your explanation.

- 11.4. Show that the magnification of the pitch angle error Δx is equal to the ratio of the focal length of the spherical lens to that of the second cylindrical (toroidal) lens.
- 11.5. Show that if the “depth of focus” of a laser beam is defined as $z_R(|z| < z_R/2)$, the beam diameter is only 12% greater than at the beam waist. Also show that a spot size of $140\text{ }\mu\text{m}$ can be maintained over a scan length of 298 mm for a helium–neon laser with a $125\text{-}\mu\text{m}$ beam waist, if the center of deflection is more than 566 mm from the scan plane and the scan angle is less than 28° .
- 11.6. If Ω is the collection solid angle, show that the collection half angle α is equal to $\sin^{-1}(\Omega/\pi)$.
- 11.7. Show that a positive slope angle may be reduced in magnitude by a positive aplanatic meniscus in the amount of the reciprocal of a refractive index.
- 11.8. Starting with a 30° collection half angle and producing a -0.205 final slope angle into the projection lens, calculate the radii of curvature of the lenses in the condensor system based on the glass choices indicated in Fig. 11.10.

Additional Problems. In most issues of the journal *Applied Optics* there is a patent review column. Almost every issue presents a few optical systems that can be appreciated from the short abstract and the accompanying patent illustration. Pick a patent and obtain a copy to analyze the innovation to the extent possible. Try to understand why the design is important and, perhaps, how this innovation could be applied in a different context. (Copies of patents are available at regional government repositories and libraries of major universities.)

REFERENCES

- 11.1. R. Kingslake, *Optical System Design* (Academic Press, New York, 1983). An expansion of the author's undergraduate and summer school courses. Contains many examples, most of them imaging systems.
- 11.2. W. J. Smith, *Modern Optical Engineering* (McGraw-Hill, New York, 1966). Contains more detail than this text with an emphasis on engineering.
- 11.3. G. W. Hopkins and A. Schwartz, “An Optical System for Full-Spectrum Measurements,” *Hewlett-Packard Journal* **31**, 17–20 (1980).
- 11.4. J. M. Fleischer, M. R. Latta, and M. E. Rabedeau, “Laser-Optical System of the IBM 3800 Printer,” *IBM Journal of Research and Development* **21**, 479–483 (1977).
- 11.5. L. M. Hubby, Jr., “Optical System Design for a Laser Printing System,” *Proceedings SPIE* **390**, 79–84 (1983).
- 11.6. A. Cox, *A System of Optical Design* (Focal Press, New York, 1964).

Index

- Abbe illumination, 112
- Abbe number, *see* *V*-number
- Abbe sine condition, 193, 211
- Aberration, 11, 48, 185–215
 - chromatic, 5, 162–169
 - reduction, 204–219
 - spherical, 20, 186–190, 197, 204–207, 215, 217, 223, 387–389
- Absorption filter, 315
- Achromat, 168, 213
- Achromatization, 165–169, 213
- Acousto-optic modulation, 51, 276–282, 309
- Acousto-optic modulator, 276–282, 309, 375, 377, 378, 380, 382–384
- Acousto-optical deflector, *see* Acousto-optical scanner
- Acousto-optical scanner, 298–302, 305, 306, 309
- Afocal system, 32
- Airy diffraction pattern, 46, 48, 257, 258
- Amici prism, 132, 133
- Amplifier tube, 350–353
- Analyzer, 37, 38
- Anamorphic system, 196
- Anastigmat, 198
- Angle of minimum deviation, 323
- Angle solve, 172
- Angular dispersion, 320, 321, 324, 325
- Angular field of view, 72
- Angular magnification, 28–32, 53
- Anode, photomultiplier, 347
- Anti-reflection coating, 46, 124–127, 130, 131
- Aperture shape factor, 283, 284
- Aperture stop, 68–74, 105, 148, 205–208, 214, 215
- Apex angle, 321, 323
- Aplanat, 198, 211
- Aplanatic point, 213
- Aplanatic surface, 211, 212, 215, 387
- Applied Optics*, 366, 371, 393
- Aspheric surface, 215–219, 387, 388
- Astigmatism, 194–197, 205–207, 213
- Avalanche photodiode, 337, 359, 360
- Axial color, 163
- Axial intercept plot, 187, 188
- Axial ray, 66, 74, 78, 179, 181
- Ballistic response, 295
- Band-pass filter, 314–319
- Bandwidth, 314
- Barrel distortion, 202, 203, 215
- Beam compressor, 259, 261, 378
- Beam displacer, 134, 135
- Beam divergence, 231–239, 241–251, 253–255, 261, 262, 265, 266
- Beam expander, 259, 261, 378
- Beam shaping, 259, 261–264
- Beamsplitter, 45, 46, 128–130
 - polarizing, 129
- Beam waist, 231–248, 250–255, 256–260, 262, 265, 266
- Bending, 204, 205
- Birefringence, index of, 274
- Birefringent crystal, 35–37, 39–41, 274
- Blackbody radiation, 3
- Blaze, 327, 328
- Blaze wavelength, 327, 328
- Bleeder chain, 348, 349
- Blur spot, 223, 224
- Bolometer, 339
- Bragg angle, 281, 299, 303, 378
- Bragg diffraction, 277, 281, 298, 303
- Bragg equation, 277, 281, 298
- Breadboarding, 374
- Brewster angle, 34, 38, 54, 259
- Brightness, 92, 98–100
- Brightness theorem, 100
- Broadband response, 295
- Cadmium sulfide, 354
- Calcite, 37, 40, 53
- Canada balsam, 37

- Candela, 115
- Candle, *see* Candela
- Cardinal points, 148–151
- Cathode, photomultiplier, 347–355
- Cavity dumping, 272, 273
- CCD array, *see* Charge coupled diode array
- Center of deflection, 292, 293, 306
- Channeltron, 350, 352
- Charge coupled diode array, 362, 363
- Chief ray, 14, 66, 78, 191, 192, 196–198
- Chopper, 273, 274
- Chromatic aberration, 5, 162–169
- Circle of least confusion, 194
- Circularly polarized light, 33, 40
- Closed-loop system, 333
- Closing, ray trace, 67, 147, 148, 163, 181, 182
- Coddington trace, 196
- Collimation, Gaussian beam, 241–245, 261, 262
- Color, 32
- Coma, 191–195, 197, 204–208, 211–213, 215, 387
- Compressor, laser beam, 259, 261, 300, 301, 378
- Computation, optical, 219–221
- Computer programs, ray tracing, 81, 82, 169–170, 219–221
- Condenser lens, 386–388
- Conduction band, 354–356
- Confocal beam parameter, 233
- Conjugate, 28
- Conjugate distance, 252–256
- Constant deviation mirror, 122
- Contrast, 226, 270, 272, 282
 - photographic plate, 344
- Converter tube, 350–353
- Cornea, 27
- Corner cube reflector, *see* Cube corner reflector
- Cosine fourth law, 95, 101, 102, 105, 106
- Critical angle, 9
- Crown glass, 5–7, 34, 167
- Cube corner reflector, 123, 140, 372, 373
- Curvature, 146, 165, 167, 205, 206
 - center of, 12, 14, 20, 211
 - radius of, 20, 46, 146, 152, 165, 167, 248
- Cylindrical lens, 81–86, 196, 259, 261, 262, 301, 310, 311, 379–384
- Czerny-Turner spectrometer, 329
- Decentration, 222
- Depletion layer, 358
- Design process, 368–392
- Detector, 38
 - diode array, 337, 372, 373
 - electrophotographic, 337, 346, 375, 376
 - frequency response, 337
 - image converter, 337
 - photodiode, 337
 - photoemissive, 347–353
 - photographic, 337, 342–346
 - photomultiplier, 337, 347–350
 - photovoltaic, 355–358
 - pyroelectric, 337, 341, 342
 - quadrant, 358
 - quantum, 339
 - resolution, 337
 - response, 337
 - semiconductor, 353–366
 - sensitivity, 337
 - solar cell, 337, 355–357
 - split-cell, 358
 - thermal, 339–342
 - thermoelectric, 339–341
 - thermopile, 337–341
 - vidicon tube, 337
- Detector geometry, 364–366
- Diazo material, 345
- Dichromate, 345
- Dielectric filter, 315–319, 351
- Dielectric mirror, 124, 127
- Diffraction, 7, 41, 43, 46
 - Bragg, 277, 281, 298, 303
 - circular aperture, 46, 48, 257, 258
 - pattern, 48, 257, 258
 - single slit, 46, 48
- Diffraction grating, 44, 49–51, 276, 277, 324–330
- Diode array, 337, 372, 373
- Diode laser, 259, 261
- Diopter, 63
- Dispersion, 5, 52, 162, 319–321, 323
 - angular, 320, 321, 324, 325
 - crown glass, 162
 - flint glass, 8, 320
- Distance of most distinct vision, 28
- Distances, sign conventions, 61
- Distortion, 202, 203, 205–207, 215, 381
- Divergence, laser beam, 231–239, 241–251, 253–255, 261, 262, 265, 266
- Dopants, 355
- Doped semiconductor, 355–359
- Double plot, 220
- Dove prism, 132, 135, 137
- Dynode, 348–351
- Ebert-Fastie spectrometer, 329
- Eigengrau, 338
- Electric field, 42, 43, 45

- Electromagnetic radiation, 1, 32
- Electron, 347–356
- Electro-optic modulator, 274–276, 282
- Electrophotography, 337, 346, 375, 376
- Elliptically polarized light, 33, 40
- Encircled energy plot, 184, 185, 224
- Entrance pupil, 68–74, 158–162, 207–210
- Entrance window, 71, 72, 76, 113, 159–161
- Equivalent air thickness, 154, 155
- Exact ray tracing, 177–227
- Exit pupil, 68, 70, 72, 76, 78, 102–106, 159–161
- Exit window, 71, 72, 102–106, 159–161
- Expander, laser beam, 259, 261, 300, 301, 378
- Extended source, 92–95, 98–105
- Extinction ratio, 270, 276
- Extraordinary ray, 37
- Eye, 27–28, 30, 114, 337–339
 - resolution, 338
 - sensitivity, 2
- Eyepiece:
 - microscope, 29
 - telescope, 30
- Fabry-Pérot filter, 317, 318
- Facet following, 301, 302
- Fast axis, 40
- Fiber optics, *see* Optical fibers
- Field curvature, 198–202, 213–215
- Field flattener, 214, 215
- Field stop, 71, 72, 76, 105, 113, 148
- Field of view, 72
- Filter:
 - absorption, 315
 - band-pass, 314–319
 - dielectric, 315–319, 351
 - Fabry-Pérot, 317, 318
 - high-pass, 314–317
 - low-pass, 314–317
 - multiple cavity, 317, 318
- Filter spectrometer, 318, 319
- First-surface mirror, 123, 124
- Flat-top distribution, laser beam, 262–264
- Flint glass, 8, 320
- f*-number, 76, 77, 162, 209, 210, 221, 222, 387
- Focal length:
 - lens, 15, 59, 61, 78–81, 148–151, 162, 163, 209, 210, 386, 387
 - mirror, 12, 20
- Focal point, 67, 149, 151
 - mirror, 12
- Fog level, Hurter-Driffield curve, 344
- Folded optical system, analysis, 135–138
- Folding, optical system, 18, 19, 135–138
- Fovea, 338
- Fraunhofer, Joseph von, 4
- Fraunhofer lines, 4
 - table, 3
- Frequency, light, 4, 32, 42
- Frequency response, 270–272, 279, 282
 - scanner, 295–298, 305
- Fresnel's equations, 124
- Galilean telescope, 243, 245, 259, 261, 383
- Galvanometer scanner, 293, 294, 296–298, 305, 310, 311
- Gamma, Hurter-Driffield curve, 344
- Gas discharge, 2
 - mercury, 3
- Gas laser, 272
- Gaussian beam, 230–266
 - focusing, 237–241, 245–260
 - higher order modes, 264, 265
- Gaussian constants, 156–162
- Gaussian profile, 230, 231, 258
- Gaussian relay, 241, 242
- Geometrical optics, laws of, 7
- Glass:
 - crown, 5–7, 34, 167
 - flint, 3, 6–9, 51, 167
 - select types, 6, 7, 54
- Glass line, 7
- Glass map, 5, 7
- Glass number, 5, 54
- Golay cell, 339
- Grating, diffraction, 44, 49–51, 276, 277, 324–330
- Grating blaze, 327, 328
- Grating equation, 49, 51, 54, 276, 277, 324, 331
- Grating error, 328
- Grating spectrometer, 324–330
 - geometry, 328–330
- Half-wave plate, 40, 274
- H-D curve, *see* Hurter-Driffield curve
- Height solve, 172
- Helium:
 - d-line, 5–7
 - lines, 3
- Helium-neon laser, 234, 239, 240, 245, 246
- High-pass filter, 314–317
- Higher order mode, Gaussian beam, 264, 265
- Hole, 354–356
- Holographic grating, 328, 330, 372, 373
- Holographic scanner, 302–305

- Holography, 345
 Hurter-Driffield curve, 343-345
 Hydrogen spectrum, 2
- Illuminance, 384, 385
 Illumination:
 Abbe, 112
 Köhler, 112, 386
 laser beam, 256-266
 Illuminator, 111-114
 Image, 11, 17
 formation, 12
 intermediate, 25, 30, 130, 155
 magnification, 24
 orientation, 22
 real, 12, 24, 30
 virtual, 2, 13, 17, 18
 Image converter, 337
 Image displacer, 134, 135
 Image evaluation, 223-227
 Image inversion, 121, 132, 133
 Image reversal, 121, 133
 Image space, 68
 Image tube, 350-353
 Incidence, plane of, 8, 30, 125
 Index of birefringence, 274
 Index of refraction, *see* Refractive index
 Intensity, 88, 90-92
 Interference, 41, 44-46, 49
 Interferometer, Michelson, 44-46
 Intrinsic semiconductor, 355, 359
 Inverse square law, 90
 Iris, 27, 28
 Irradiance, 38, 39, 44, 88, 90-101
 Irradiance distribution, Gaussian beam, 258-264
- Journal of Applied Physics*, 371
Journal of Quantum Electronics, 371
- Keplerian telescope, 243, 245
 Köhler illumination, 112, 386
- Lagrange invariant, *see* Optical invariant
 Lambertian source, 93
 Lambert's law, 93
 Laser, 4, 272, 273
 beam waist location, 239
 Rayleigh range in, 234
 Laser beam compressor, 259, 261, 300, 301, 378
 Laser beam expander, 259, 261, 300, 301, 378
 Laser beam illumination, 256-266
Laser Focus, 371
 Laser printer, 375-384
 Laser scanning system, 368-384
 Laser television display, 309-311
Lasers and Applications, 371
 Lateral color, 164, 215
 Length, optical system, 59, 60, 73
 Lens:
 cylindrical, 81-86, 196, 259, 261, 262, 301, 310, 311, 379-384
 human eye, 27
 negative, 16
 objective, 29
 positive, 16
 toroidal, 380-382
 Lenses:
 combination of, 24, 26
 types, 15
 Lensmaker's equation, 24, 151, 152, 165, 204
 Light, 1, 32
 speed of, 4
 wave nature, 42-44
 Light emitting diode, 109, 273, 391
 Linearly polarized light, 33-41
 Line pair, 225
 Line spread function, 224, 225, 227
 Liquid mirror, 128
 Littrow geometry, 323, 324, 328, 329
 Longitudinal axial chromatic aberration, 164
 Low-pass filter, 314-317
 Lumen, 88, 90, 114-116
 Lyman ghost, 328
- Magnesium fluoride, 124, 126, 127
 Magnification, 28, 58, 192, 193
 angular, 28-32, 53
 combination of lenses, 25
 lateral, 22, 24
 longitudinal, 22
 transverse, 22, 24, 158
 Magnifier, 28-30
 Malus, Etienne, 39
 Malus, law of, 39
 Marginal ray, 14, 67, 78, 195
 Matrix method of ray tracing, 169
 Mechanical modulator, 273, 274, 282
 Mechanical scanner, 273, 274, 288-298, 305, 309, 369, 370
 Mercury lines, 3
 Meridional ray, 177-184, 191, 195, 196, 213
 Merit function, 221
 Michelson interferometer, 44-46
 Microchannel plate, 352, 353
 Microscope, 5, 17, 29, 30
 Minimum deviation, angle of, 323

- Mirror, 15, 18
 - concave, 11, 20, 21
 - constant deviation, 122
 - convex, 12, 20
 - dielectric, 124, 127
 - first-surface, 123, 124
 - focal length, 21
 - liquid, 128
 - plane, 119–128
 - second-surface, 123, 124, 130
 - spherical, 11, 12, 20
- Mirror equation, 21–22
- Mirror images, 119–123
- Modelocking, 272, 273
- Modulation, 270–283
 - acousto-optic, 51, 276–282, 309
- Modulation transfer function, 226, 227, 392
- Modulator:
 - acousto-optic, 276–282, 309, 375, 377, 378, 380, 382–384
 - electro-optic, 274–276, 282
 - mechanical, 273–274, 282
- Monochromatic radiation, 32
- Multiple cavity filter, 317, 318

- Narrowband response, 295, 298
- Near point, 28, 30, 53
- Newton, Isaac, 36, 336
- Nicol prism, 37
- Nodal points, 150, 151
- Nodding scanner, 293–298, 305
- n -type material, 355–357

- Object, 11
- Object distance, 24
- Objective, telescope, 30
- Objective scanning, 284, 285
- Object space, 68
- Oblique ray, 182
- Offense against sine condition, 191–193
- Opening, ray trace, 67, 147, 179
- Open-loop system, 333
- Optical axis, crystal, 35, 37, 40–44
- Optical computation, 219–221
- Optical density, 314
- Optical elements, combinations of, 17
- Optical Engineering*, 366
- Optical fibers, 10
- Optical instruments, 26–32
- Optical invariant, 76, 156–158, 186, 201, 387
- Optical pathlength, 5
- Optic axis, 10
- Opticks*, 36, 336
- Orders, diffraction grating, 50, 51

- Ordinary ray, 35
- Orthicon, 350, 351
- Orthochromatic, 343

- Panchromatic, 343
- Parabola, 20
- Paraboloid, 215, 217, 218
- Parallel plates, 134, 135, 152–156, 251, 252
- Parallel ray, 67, 78, 80
- Para-principal trace, 196
- Paraxial approximation, 19
- Paraxial equations, Gaussian beams, 247–252
- Paraxial rays, 19, 20
- Paraxial ray tracing, 144–175, 373
- Patents, 373
- Pathlength, optical, 5
- Path length difference, 46
- Pattern generation, 307–309
- Pellicle, 130
- Pentaprism, 132
- Petzval curvature, 198–202, 213–215
- Petzval radius, 198
- Phase factor, 42
- Phase grating, 263, 264
- Photoconductor, 354, 358–360
- Photodiode, 337, 356, 363
 - spectral response, 357
- Photodiode arrays, 360–364
- Photoemissive detector, 347–353
- Photoemissive diode, 347, 348
- Photographic plate, 337, 342–346
- Photometry, 90, 114–116, 337
- Photomultiplier, 337, 347–350
- Photonics Spectra*, 371
- Photopic curve, 114, 115, 337
- Photopolymerizers, 345
- Photovoltaic detector, 355–358
- Physical optics, 32, 42
- Pickup, 173
- Piezoelectric scanner, 293, 294, 298, 305
- Pincushion distortion, 202, 203, 215
- PIN photodiode, 359
- Pitch angle error, 290, 375, 378–381
- Plane of incidence, 8, 30, 125
- Plane mirror, 119–128
- Plane wave, 43–45, 48
- Plot:
 - axial intercept, 187, 188
 - encircled energy, 184, 185, 224
 - ray intercept, 188, 189, 197
- Point source, 88–91, 95–98
- Point spread function, 223, 224
- Polarization, 32, 41, 125
 - methods of, 34–39

- Polarization (*Continued*)
 - modifiers, 39-41
 - rotator, 41, 274
- Polarization axis, 38
- Polarized light, 32, 33, 37, 40, 41
- Polarizer, 37
 - birefringent crystal, 37
 - pile of plates, 34, 35
 - sheet, 34-36
- Polarizing beamsplitter, 129
- Polygonal scanner, 288-293, 369, 370, 375, 377-383
- Porro prism, 132, 154-156
- Post-objective scanning, 284, 285, 382-384
- Power (optical):
 - lenses, 63, 65, 165-167
 - surface, 146, 170, 247
- Power (radiometric), 88-90
- Pre-objective scanning, 285, 286
- Principal plane, 79, 80, 150, 151
- Principal surface, 149-151
- Prism, 18, 130-134
 - Amici, 132, 133
 - Dove, 132, 135, 137
 - mirror, 9, 10
 - penta-, 132
 - Porro, 132, 154-156
 - right-angle, 131-133, 137
- Prism scanner, *see* Polygonal scanner
- Prism spectrometer, 319-324
- Proximity detector placement, 364, 365
- p*-type material, 355-357
- Pupil:
 - entrance, 68-72, 74, 76, 78, 158-162, 207-210
 - exit, 68-70, 74, 76, 102, 106, 159-162
- Pyroelectric detector, 337, 341, 342
- Pyroelectric effect, 341
- Quadrant detector, 358
- Quantum detectors, 339
- Quarter-wave plate, 40, 274, 276
- Q-U* trace, 178-183
- Radiance, 88, 92, 93
- Radiation, monochromatic, 32
- Radiometer, 107
- Radiometry, 87-107
- Radius of curvature:
 - sign conventions, 61
 - wavefront, 233, 234, 255, 256
- Range finder, 139, 140
- Raster pattern, 351
- Ray, 7
 - axial, 66, 74, 78, 179, 181
 - chief, 66, 68, 71, 78
 - marginal, 67-69
 - meridional, 177-184, 191, 195, 196, 213
 - oblique, 182
 - parallel, 67, 78, 80
 - paraxial, 19, 20
 - sagittal, 191, 192
 - skew, 177, 183, 185, 193, 195, 196
 - tangential, 191, 192
- Ray height, sign conventions, 61
- Ray intercept plot, 188, 189, 197
- Rayleigh criterion, 49, 283, 284, 322
- Rayleigh range, 232-234, 236-248, 250, 251, 254-256, 259, 262, 266
- Ray sketching, 1, 10-18, 135, 373
- Ray tracing:
 - backward, 74, 75
 - conventions for, 10-17, 19-21
 - exact, 177-227
 - paraxial, 144-175, 373
 - programs, 81, 82, 169-170, 219-221
 - reflective surfaces, 170-172
 - reverse, 74, 75
 - thin lens, 56-84
 - sign convention, 20, 61
- Ray-vertex perpendicular, 178-181
- Reciprocity law failure, 345
- Reduced thickness, 154, 155
- Reducing aberrations, 204-219
- Reflectance, 126-128
- Reflection, 34
 - law of, 8, 24, 53, 125, 134, 135, 144, 146
- Reflectivity, 125-128
- Reflector, cube corner, 123, 140, 372, 373
- Refraction, law of, 8, 24, 53, 125, 134, 135, 144, 146, 153, 180
- Refractive index, 4, 9, 124-127, 319
 - birefringent crystal, 35, 37
 - sign conventions, 20, 61
 - table, two glasses, 3
- Replication, 218, 219, 372
- Resolution:
 - acousto-optical scanner, 301
 - angular, 48
 - diode array, 337, 363
 - eye, 338
 - photographic film, 345
 - scanners, 282-284, 377
 - spectrometer, 322
 - television, 351
 - vidicon tube, 337, 351
- Resolution chart, 225

- Resolving power:
 - grating, 325–327
 - spectrometer, 322, 323, 325–327
- Resonant frequencies, scanner, 294, 296–298, 305
- Resonator, unstable, 234
- Retina, 27, 338
- Retrocollective detector geometry, 365, 366
- Retroreflector, 123, 140
- Right-angle prism, 131–133, 137
- Rise time, 271, 279, 283
- rms spot size, 223
- Roof, prism, 133, 134
- Rotary scanner, *see* Polygonal scanner
- Rowland circle, 330
- Rowland ghost, 328
- Sag, *see* Sagitta
- Sagitta, 199, 216, 217
- Sagittal fan, 194
- Sagittal focus, 194
- Sagittal plane, 191
- Sagittal ray, 191, 192
- Scaling, 158, 159, 162, 169, 209, 210
- Scan angle, 300, 305
- Scan equation, 282–284, 377
- Scan magnification, 305, 306
- Scanner, 25
 - acousto-optical, 298–302, 305, 306, 309
 - galvanometer, 293, 294, 296–298, 305, 310, 311
 - holographic, 302–305
 - mechanical, 273, 274, 288–298, 305, 309, 369, 370
 - piezoelectric, 293, 294, 298, 305
 - polygonal, 288–293, 369, 370, 375, 377–383
 - taut band, 293, 294, 298, 305
- Scanning, 282–305
 - objective, 284, 285
 - post-objective, 284, 285, 382–384
 - pre-objective, 285, 286
 - telecentric, 286
 - two-dimensional, 287, 288, 309, 375–382
- Scanning angle, 283
- Scanning geometry, 284–288, 292
- Scan relay, 305–307
- Scattered light detector geometry, 365, 366
- Schmidt corrector, 216, 217
- Schott glass catalog, 54
- Scotopic curve, 114
- Searchlight, 108
- Second-surface mirror, 123, 124, 130
- Seidel aberration, 186
- Semiconductor detector, 353–366
- Semiconductor laser, 259, 261, 272
- Shoulder, Hurter-Driffeld curve, 344
- Sign convention, ray tracing, 20, 24, 61
- Silicon monoxide, 124
- Sine bar drive, 331, 332
- Sine condition, offense against, 191–193
- Sine grating, 276, 277
- Skew ray, 177, 183, 185, 193, 195, 196
- Slide projector, 384–392
- Slit, 2
- Slope angle, sign convention, 61
- Slow axis, 40
- Snell's law, 8, 24, 53, 125, 134, 135, 144, 146, 153, 180, 213
- Sodium, *D* lines, 2–5
- Solar cell, 337, 355–357
- Solid angle, 89
- Solve, 147, 172, 173
- Source:
 - extended, 92–95, 98–105
 - incandescent, 2, 3, 272
 - laser, 272, 273
 - point, 88–91, 95–98
- Source modulation, 272, 273
- Spatial filtering, 245, 268
- Spatial frequency, 226
- Spectral luminous efficiency, 114
- Spectrometer, 313–334, 373
 - filter, 318, 319
 - grating, 324–330
 - prism, 319–324
- Spectrometer control, 333, 334
- Spectrometer sources, 330, 331
- Spectrometer system, 330–334, 372, 373
- Spectroradiometry, 116
- Spectrum, 50, 51, 337
 - electromagnetic, 2
- Speed of light, 4
- Spherical aberration, 20, 186–190, 197, 204–207, 215, 217, 223, 387–389
 - longitudinal, 187–190, 223
 - transverse, 187–190, 223
- Spherical wave, 43–45
- Split-cell detector, 358
- Spot diagram, 184, 185, 223
- Stability condition, 234
- Star test, 224
- Stop:
 - aperture, 68–74, 105, 148, 205–208, 214, 215
 - field, 71, 72, 76, 105, 113, 148
- Stop shift, 389

- Stop shift equations, 205–207
- Storage tube, 350, 351
- Tangential fan, 194
- Tangential focus, 194
- Tangential plane, 191
- Tangential ray, 191, 192
- Taut band scanner, 293, 294, 298, 305
- Telecentric geometry, 286
- Telecentric scanning, 286
- Telescope, 30–32, 132, 155, 156, 243–245, 261
 - Cassegrainian, 218
 - Galilean, 243, 245, 259, 261, 383
 - Keplerian, 243, 245
 - Richey-Chrétien, 218
 - Schmidt, 217
- Television scanning rates, 351
- Tellurium dioxide, 300, 301
- TEM₀₀ Mode, 230, 232, 234, 264–266
- Thermal detector, 339–342
- Thermistor, 339
- Thermocouple, 340
- Thermoelectric detector, 339–341
- Thermopile, 337–341
- Thin lens equation, 23, 31, 62
 - Gaussian beams, 235
 - Gaussian form, 52, 235
 - Newtonian form, 52, 235
- Third order aberration, 186
- Through-focus transfer function, 227
- Tiltmeter, 390–392
- Tilts, 222
- Toe, Hurter-Driffield curve, 344
- Top hat distribution, laser beam, 262, 263
- Toroidal lens, 380–382
- Torsion rod scanner, 293, 294, 298
- Total internal reflection, 9, 10, 37, 52, 131
- Transfer equation, 62
- Transmission, law of, 7
- Transmittance, 96, 98
- Transverse axial chromatic aberration, 164
- Trepanning, 284, 285
- Truncation ratio, 257, 258
- Tungsten filament, 2
- Tungsten-halogen lamp, 372, 385, 386
- Two-dimensional scanning, 287, 288, 309, 375–384
- Universal product code, 304, 370
- Unstable resonator, 234
- UPC, *see* Universal product code, 304
- Valence band, 354–356
- Vertex, 10
 - lens, 149
 - mirror, 12, 14
- Vidicon tube, 337
- Vignetting, 101–111, 113
- Vitreous humor, 27
- V-number, 5–7, 51, 163, 165–167, 319, 320
- Wave:
 - monochromatic, 42
 - plane, 43
 - spherical, 43–45
 - transverse, 32
- Wavefront, curvature, 233, 234, 255, 256
- Wavelength, 2, 32, 42, 43
 - light, 4
- Wave plate:
 - half-, 40, 274
 - quarter-, 40, 274, 276
- Window:
 - entrance, 71, 72, 76, 113, 159–161
 - exit, 71, 72, 102–106, 159–161
- Work function, 347
- Xerographic detector, *see* Electrophotographic detector
- X-ray, 2
- y - nu trace, 144–148
- Young's experiment, 44, 47
- y - u trace, 66
- Zener voltage, 359
- Zinc sulfide, 128, 129