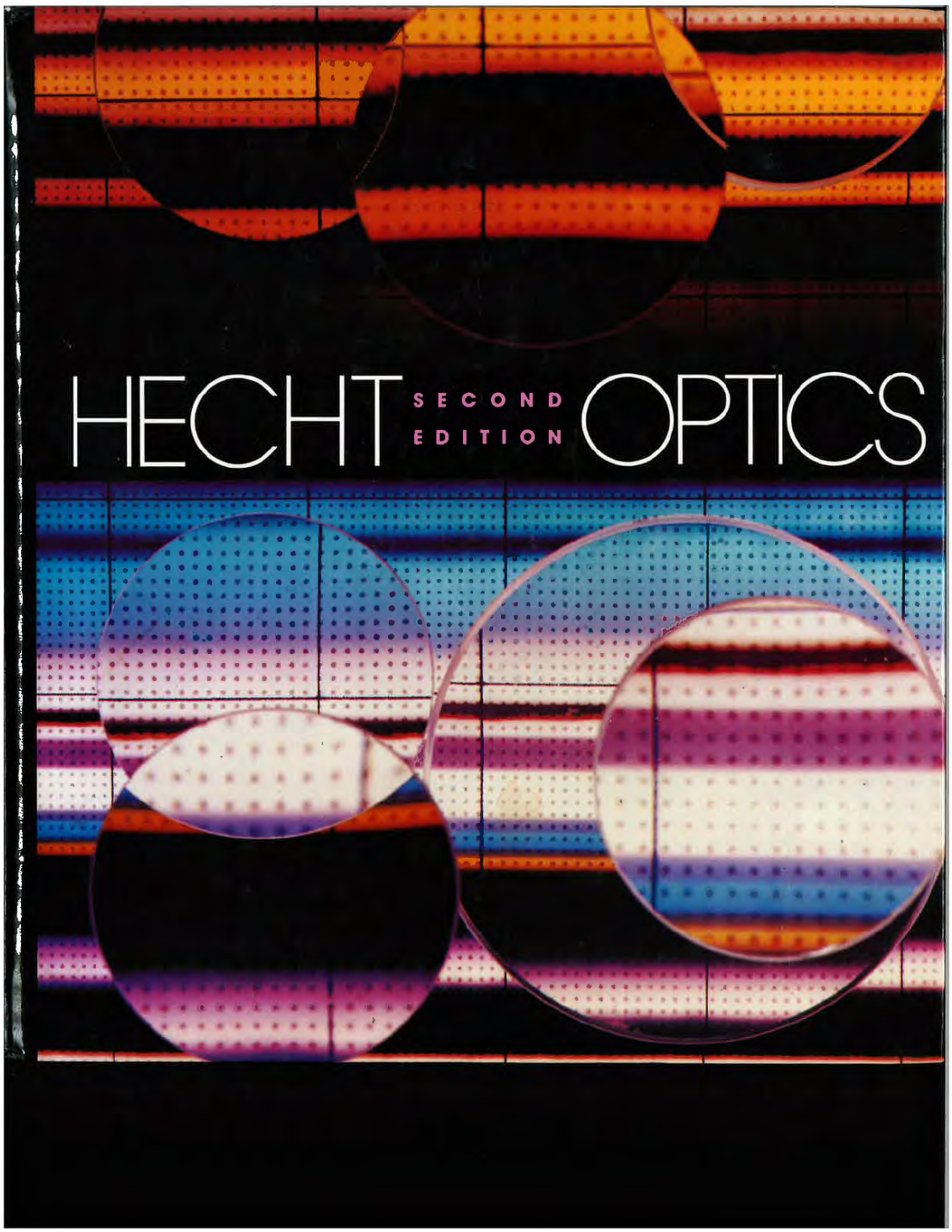


OPTICS

SECOND EDITION

EUGENE HECHT
Adelphi University

With Contributions by Alfred Zajac

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California • Don Mills, Ontario
Wokingham, England • Amsterdam • Sydney • Singapore
Tokyo • Madrid • Bogatá • Santiago • San Juan

Sponsoring editor: Bruce Spatz
Production supervisors: Margaret Pinette and Lorraine Ferrier
Text designer: Joyce Weston
Illustrators: Oxford Illustrators
Art consultant: Loretta Bailey
Manufacturing supervisor: Ann DeLacey

Library of Congress Cataloging-in-Publication Data

Hecht, Eugene.
Optics.

Bibliography: p.
Includes indexes.
1. Optics. I. Zajac, Alfred. II. Title.
QC355.2.H42 1987 535 86-14067
ISBN 0-201-11609-X

Reprinted with corrections May, 1990.

11 12 13 14 15 MA 96959493

Contents

4 THE PROPAGATION OF LIGHT

4.1 INTRODUCTION

We now consider a number of phenomena related to the propagation of light and its interaction with material media. In particular, we shall study the characteristics of lightwaves as they progress through various substances, crossing interfaces, and being reflected and refracted in the process. For the most part, we shall envision light as a classical electromagnetic wave whose velocity through any medium is dependent upon that material's electric and magnetic properties. It is an intriguing fact that many of the basic principles of optics are predicated on the wave aspects of light but are completely independent of the exact nature of the wave. As we shall see, this accounts for the longevity of *Huygens's principle*, which has served in turn to describe mechanical aether waves, electromagnetic waves, and now, after three hundred years, applies to quantum optics.

Suppose, for the moment, that a wave impinges on the interface separating two different media (e.g., a piece of glass in air). As we know from our everyday experiences, a portion of the incident flux density will be diverted back in the form of a *reflected wave*, while the remainder will be transmitted across the boundary as a *refracted wave*. On a submicroscopic scale we can envision an assemblage of atoms that scatter the incident radiant energy. The manner in which these emitted light wavelets superimpose and combine with each other will depend on the spatial distribution of the scattering atoms. As we know from the previous chapter, the scattering process is responsible for the *index of refraction*, as well as the resultant *reflected* and *refracted* waves. This atomistic description is quite satisfying conceptually, even though it is not a simple matter to treat analytically. It should, however, be kept in mind even when applying macroscopic techniques, as indeed we shall later on.

We now seek to determine the general principles governing or at least describing the propagation, reflection, and refraction of light. In principle it should be possible to trace the progress of radiant energy through any system by applying Maxwell's equations and the associated boundary conditions. In practice, however, this is often an impractical if not an impossible task (see Section 10.1). So we shall take a somewhat different route, stopping, when appropriate, to verify that our results are in accord with electromagnetic theory.

4.2 THE LAWS OF REFLECTION AND REFRACTION

4.2.1 Huygens's Principle

Recall that a wavefront is a surface over which an optical disturbance has a constant phase. As an illustration, Fig. 4.1 shows a small portion of a spherical wavefront Σ emanating from a monochromatic point source S in a homogeneous medium. Clearly, if the radius of the wavefront as shown is r, at some later time t it will simply be $(r + vt)$, where v is the phase velocity of the wave.

But suppose instead that the light passes through a nonuniform sheet of glass, as in Fig. 4.2, so that the wavefront itself is distorted. How can we determine its new form Σ'? Or for that matter, what will Σ' look like at some later time, if it is allowed to continue unobstructed?

A preliminary step toward the solution of this problem appeared in print in 1690 in the work entitled *Traité de la Lumière*, which had been written 12 years earlier by the Dutch physicist Christiaan Huygens. It was there that he enunciated what has since become known as **Huygens's principle**, that *every point on a primary wavefront serves as the source of spherical secondary wavelets, such that the primary wavefront at some later time is the envelope of these wavelets. Moreover, the wavelets advance with a speed and frequency equal to those of the primary wave at each point in space.* If the medium is homogeneous, the wavelets may be constructed with finite radii, whereas if it is inhomogeneous, the wavelets must have infinitesimal radii. Figure 4.3 should make this fairly clear; it shows a view of a wavefront Σ, as well as a number of spherical secondary wavelets, which, after a time t, have propagated out to a radius of vt. The envelope of all these wavelets is then asserted to correspond to the advanced primary wave Σ'. It is easy to visualize the process in terms of mechanical vibrations of an elastic medium. Indeed this is the way that Huygens envisioned it within the context of an all-pervading aether, as is evident from this comment by him:

> We have still to consider, in studying the spreading out of these waves, that each particle of matter in which a wave proceeds not only communicates its motion to the next particle to it, which is on the straight line drawn from the luminous point, but that it also necessarily gives a motion to all the others which touch it and which oppose its motion. The result is that around each particle there arises a wave of which this particle is a center.

We can make use of these ideas in two different ways. On one level, a mathematical representation of the wavelets will serve as the basis for a valuable analytical technique in treating diffraction theory. One can trace the progress of a primary wave past all sorts of apertures and obstacles by summing up the wavelet contributions

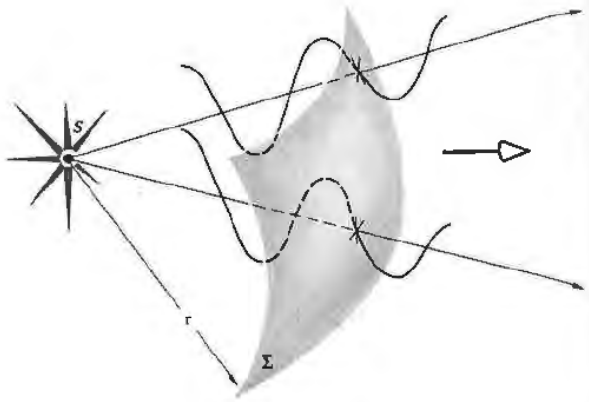


Figure 4.1 A segment of a spherical wave.

mathematically. On another level, Fig. 4.3 represents a graphical application of the essential ideas and as such is known as *Huygens's construction.*

Thus far we have merely stated Huygens's principle, without any justification or proof of its validity. As we shall see (Chapter 10), Fresnel successfully modified Huygens's principle somewhat in the 1800s. A little later on, Kirchhoff showed that the *Huygens–Fresnel principle* was a direct consequence of the differential wave equation (2.59), thereby putting it on a firm mathematical base. That there was a need for a reformulation

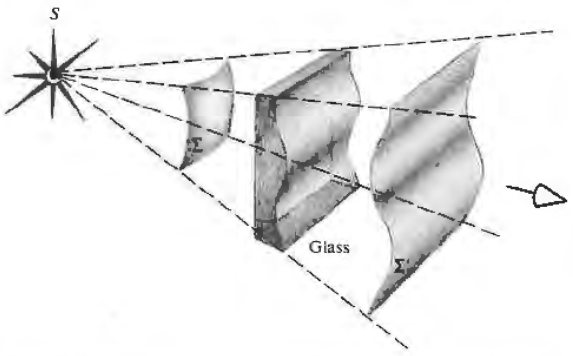


Figure 4.2 Distortion of a portion of a wavefront on passing through a material of nonuniform thickness.

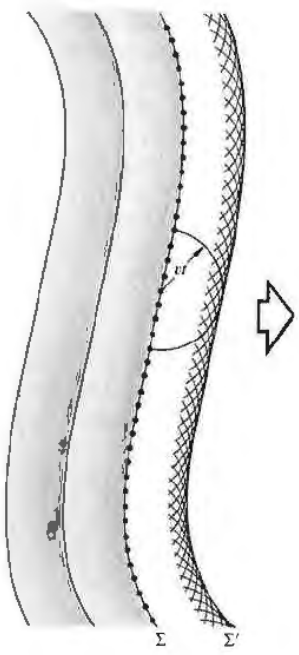


Figure 4.3 The propagation of a wavefront via Huygens's principle.

of the principle is evident from Fig. 4.3, where we deceptively only drew hemispherical wavelets.* Had we drawn them as spheres, there would have been a *back-wave* moving toward the source—something that is not observed. Since this difficulty was taken care of theoretically by Fresnel and Kirchhoff, we need not be disturbed by it. In fact, we shall overlook it completely when applying Huygens's construction, which, in the end, is best thought of as a highly useful fiction.

Still, Huygens's principle fits in rather nicely with our earlier discussion of the atomic scattering of radiant energy. Each atom of a material substance that interacts with an incident primary wavefront can be regarded as a point source of scattered secondary wavelets. Things are not quite as clear when we apply the principle to the propagation of light through a vacuum. It is helpful, however, to keep in mind that at any point in empty space on the primary wavefront there exists both a time-varying **E**-field and a time-varying **B**-field. These in turn create new fields that move out from the point. In this sense each point on the wavefront is analogous to a physical scattering center.

* See E. Hecht, *Phys. Teach.* 18, 149 (1980).

4.2.2 Snell's Law and the Law of Reflection

The fundamental laws of reflection and refraction can be derived in several different ways; the first approach to be used here is based on Huygens's principle. It should be said, however, that our intention at the moment is as much to elaborate on the use of the method as to arrive at the end results. Huygens's principle will provide a highly useful and fairly simple means of analyzing and visualizing some complex propagation problems, for example, those involving anisotropic media (p. 287) or diffraction (p. 392). Consequently, it is to our advantage to gain some practice in using the technique, even if it is not the most elegant procedure for deriving the desired laws.

Figure 4.4 shows a monochromatic plane wave impinging normally down onto the smooth interface separating two homogeneous transparent media. When an incident wave comes into contact with the interface, it can be imagined as split into two: we observe one wave reflected upward and another transmitted downward. If we consider an incident wavefront Σ_i coincident with the interface splitting into Σ_r and Σ_t, both also congruent with the interface, we can utilize Huygens's construction (neglecting the back-waves). Every point on Σ_r serves as a source of secondary wavelets, which travel more or less upward into the incident medium at a speed v_i. At a time t later, the front will advance a distance $v_i t$ and appear as Σ_r'. Similarly, every point on the downward-moving front Σ_t will serve as a source for wavelets essentially heading down with a speed v_t. After a time t the transmitted front will appear a distance $v_t t$ below as Σ_t'.

The process is ongoing, repeating itself with the frequency of the incident wave.* The media are

* This assumes the use of light whose flux density is not so extraordinarily high that the fields are gigantic. With this assumption the medium will behave linearly, as is most often the case. In contrast, observable harmonics can be generated if the fields are made large enough (Section 14.4).

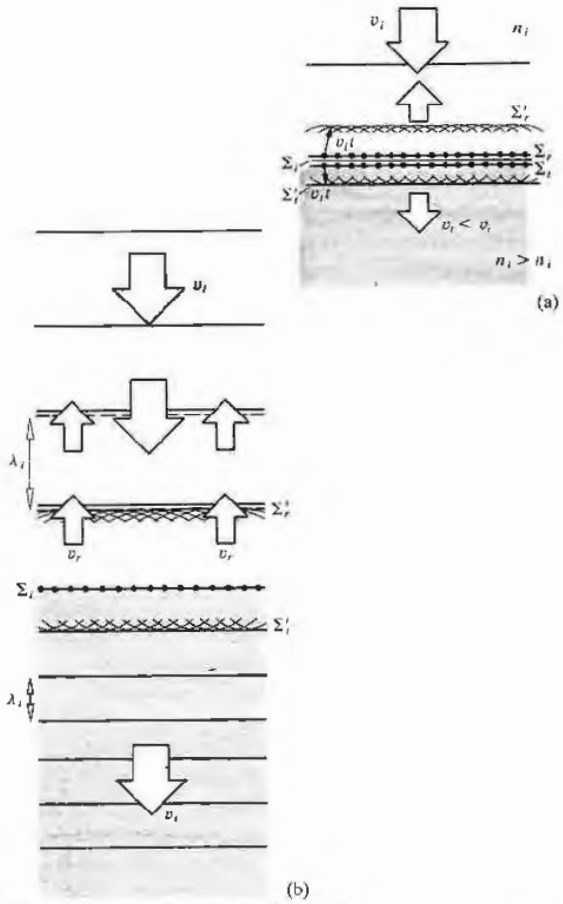


Figure 4.4 A monochromatic plane wave impinging down onto a homogeneous, isotropic medium of index n_t. Σ_i, Σ_r, and Σ_t should actually overlap.

assumed to respond linearly, so the reflected and transmitted waves have that same frequency (and period), as do all the secondary wavelets. Taking $n_t > n_i$, it follows that $c/v_t > c/v_i$, thus $v_t < v_i$, and the wavelengths (the distances between wavefronts drawn in consecutive intervals of τ) will be such that $\lambda_i > \lambda_t$ and $\lambda_i = \lambda_r$, as shown in Fig. 4.4(b). The incoming plane wave is perpendicular to the interface, and symmetry produces both reflected and transmitted plane waves that also travel out from the interface perpendicularly.

Now suppose the incident wave comes in at some other angle, as indicated in Fig. 4.5. Clearly, it sweeps across the interface again, essentially splitting into two waves: one reflected and one refracted. Let's follow the progress of a typical front in Fig. 4.6, envisioning the diagram as if it were a series of snapshots taken in successive intervals of time τ. Start when Σ_i makes contact with the interface at point a. At that point, both the reflected and transmitted wavefronts begin, so a, which lies on both fronts, can be taken as a source of both an upwardly emitted wavelet traveling at a speed v_i and a downwardly emitted wavelet traveling at a speed v_t. Now focus on another point, say, b on Σ_i.

After a time t_1 the plane Σ_i will have moved a distance in the incident medium of $v_i t_1$, so that b then corresponds to b'. Presumably, two wavelets will then propagate out from b' into the incident and transmitting media, contributing to the reflected, Σ_r', and transmitted, Σ_t', wavefronts. These wavelets are shown here after a time t_2, where $\tau = t_1 + t_2$. The rest of the diagram

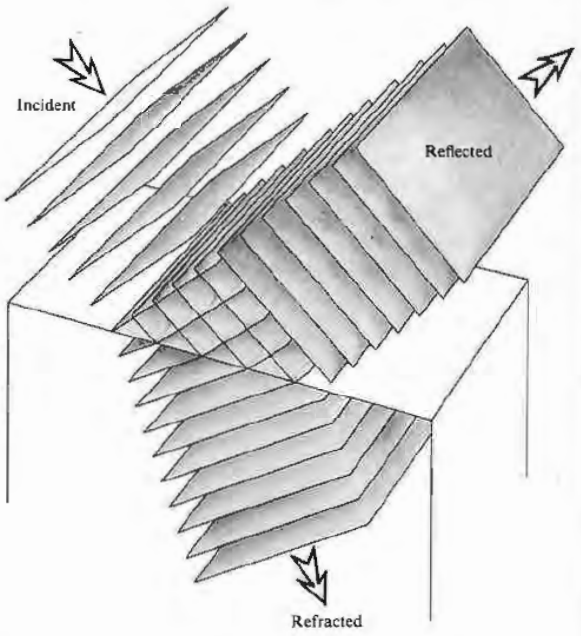


Figure 4.5 Reflection and transmission of plane waves.

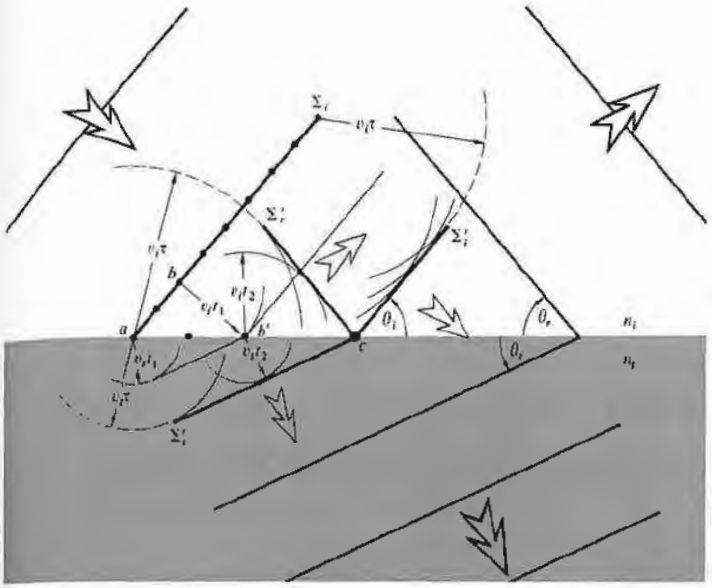


Figure 4.6 Reflection and transmission at an interface via Huygens's principle.

should be self-explanatory. Figure 4.7 is a somewhat simplified version in which θ_i, θ_r, and θ_t, as before, are the angles of *incidence*, *reflection*, and *transmission* (or *refraction*), respectively. Notice that

$$\frac{\sin\theta_i}{\overline{BD}} = \frac{\sin\theta_r}{\overline{AC}} = \frac{\sin\theta_t}{\overline{AE}} = \frac{1}{\overline{AD}}. \tag{4.1}$$

By comparison with Fig. 4.6, it should be evident that

$$\overline{BD} = v_i t, \qquad \overline{AC} = v_i t, \qquad \overline{AE} = v_t t,$$

so substituting into Eq. (4.1) and canceling t, we have

$$\frac{\sin\theta_i}{v_i} = \frac{\sin\theta_r}{v_i} = \frac{\sin\theta_t}{v_t}. \tag{4.2}$$

It follows from the first two terms that **the angle of incidence equals the angle of reflection**, that is,

$$\theta_i = \theta_r. \tag{4.3}$$

Known as the **law of reflection**, it first appeared in the book entitled *Catoptrics*, which was purported to have been written by Euclid.

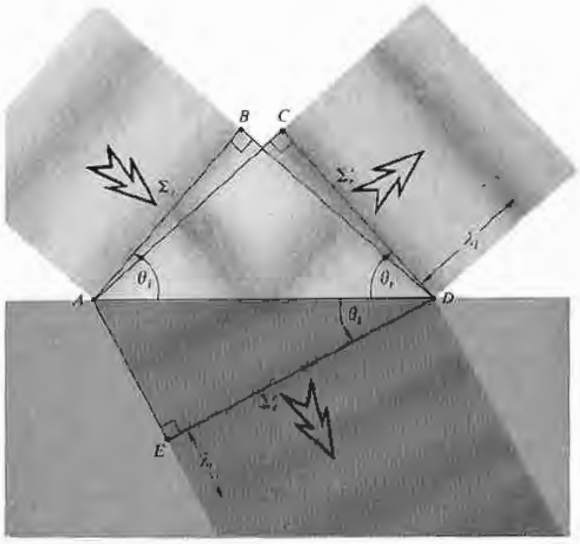


Figure 4.7 Reflected and transmitted wavefronts at a given instant.

The first and last terms of Eq. (4.2) yield

$$\frac{\sin \theta_i}{\sin \theta_t} = \frac{v_i}{v_t}, \tag{4.4}$$

or since $v_i/v_t = n_t/n_i$,

$$n_i \sin \theta_i = n_t \sin \theta_t. \tag{4.5}$$

This is the very important **law of refraction**, the physical consequences of which have been studied, at least on record, for over eighteen hundred years. On the basis of some fine observations, Claudius Ptolemy of Alexandria attempted unsuccessfully to divine the expression. Kepler nearly succeeded in deriving the law of refraction in his book *Supplements to Vitello* in 1604. Unfortunately he was misled by some erroneous data compiled earlier by Vitello (ca. 1270). The correct relationship seems to have been arrived at first by Snell* at the University of Leyden and then by the French mathematician Descartes.† In English-speaking countries Eq. (4.5) is generally referred to as **Snell's law**. Notice that it can be rewritten in the form

$$\frac{\sin \theta_i}{\sin \theta_t} = n_{ti}, \tag{4.6}$$

where $n_{ti} \equiv n_t/n_i$ is the *ratio of the absolute indices of refraction*. In other words, it is the *relative index of refraction of the two media*. It is evident in Fig. 4.6, where $n_{ti} > 1$ (i.e., $n_t > n_i$ and $v_i > v_t$), that $\lambda_i > \lambda_t$, whereas the opposite would be true if $n_{ti} < 1$.

One feature of the above treatment merits some further discussion. It was reasonably assumed that each point on the interface, such as c in Fig. 4.6, coincides with a particular point on each of the incident, reflected, and transmitted waves. In other words, there is a fixed phase relationship between each of the waves at points a, b, c, and so forth. As the incident front sweeps across the interface, every point on it in contact with the interface is also a point on both a corresponding reflected front and a corresponding transmitted front. This situation is known as *wavefront continuity*, and it will be

* This is the common spelling, although Snel is probably more accurate.

† For a more detailed history, see Max Herzberger, "Optics from Euclid to Huygens," *Appl. Opt.* 5, 1383 (1966).

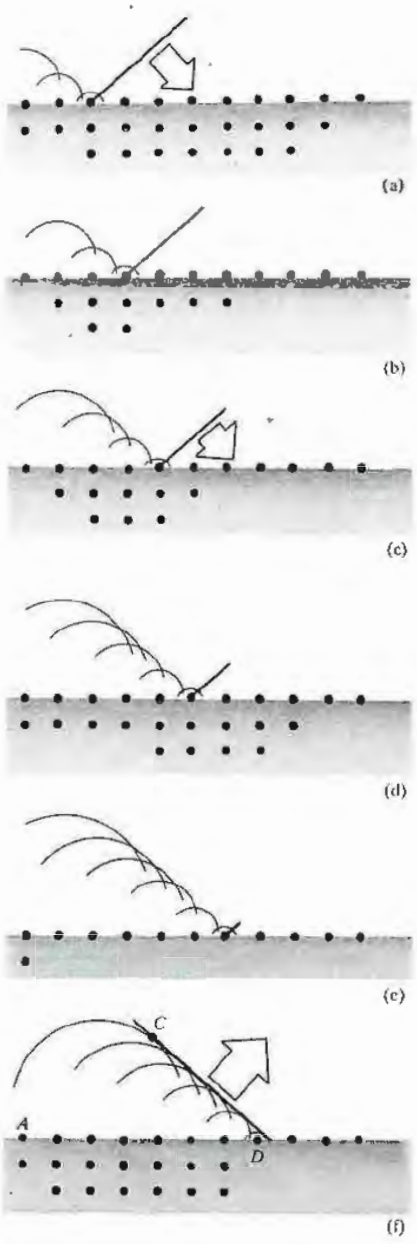


Figure 4.8 The reflection of a wave as the result of scattering.

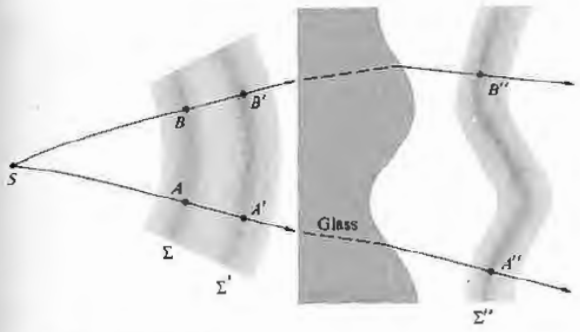


Figure 4.9 Wavefronts and rays.

justified in a more mathematically rigorous treatment in Section 4.3.1. Interestingly, Sommerfeld* has shown that the laws of reflection and refraction (independent of the kind of wave involved) can be derived directly from the requirement of wavefront continuity without any recourse to Huygens's principle, and the solution to Problem 4.9 demonstrates as much.

A far more physically appealing view of the whole process is depicted in Fig. 4.8. An electromagnetic disturbance whose wavelength (λ) is several thousand times larger than the spacing between the atoms ($d \approx 0.1$ nm) sweeps across an interface. Each atom is driven successively and scatters a wavelet. The tilt of the incident wave determines the phase delay between the scattering of each atom in turn (see Section 10.1.3 for the details). The front running from C to D is composed of wavelets that arrive in phase, superimpose, and interfere constructively. Since every point on the incident front (ranging from A to B in Fig. 4.7) has the same phase, if $\overline{AC} = \overline{BD}$, the distances traveled and therefore the phases of the wavelets arriving at C and D will be equal, as indeed they will be all across the front. From the geometry, this can happen only for a reflected wavefront propagating in the one direction such that $\theta_i = \theta_r$. This picture of scattered interfering wavelets is essentially an atomic version of the Huygens–Fresnel principle.

Although theoretically all the dipoles throughout the medium contribute to the reflected wave, the dominant effect is due to a surface layer only about $\frac{1}{2}\lambda$ thick, which is nonetheless typically several thousand atoms deep. Furthermore, the condition that only one beam is reflected is true provided that $\lambda \gg d$; it would not be the case with x-rays where $\lambda \approx d$, and there several scattered beams actually result; nor is it the case with a diffraction grating, where the separation between scatterers is again comparable to λ, and several reflected and transmitted beams are produced. A similar argument can be made for the scattering process giving rise to the transmitted wave and Snell's law, as Problem 4.11 establishes.

4.2.3 Light Rays

The concept of a light ray is one that will be of interest to us throughout our study of optics. *A* **ray** *is a line drawn in space corresponding to the direction of flow of radiant energy.* As such, it is a mathematical device rather than a physical entity. In practice one can produce very narrow *beams* or *pencils* of light (e.g., a laserbeam), and we might imagine a ray to be the unattainable limit on the narrowness of such a beam. Bear in mind that in an *isotropic medium* (i.e., one whose properties are the same in all directions) *rays are orthogonal trajectories of the wavefronts.* That is to say, *they are lines normal to the wavefronts at every point of intersection.* Evidently, *in such a medium a ray is parallel to the propagation vector* **k**. As you might suspect, this is not true in *anisotropic* substances, which we will consider later (see Section 8.4.1). *Within homogeneous isotropic materials, rays will be straight lines,* since by symmetry they cannot bend in any preferred direction, there being none. Moreover, because the speed of propagation is identical in all directions within a given medium, the spatial separation between two wavefronts, measured along rays, must be the same everywhere.* Points where a single ray intersects a set of wavefronts are called *corresponding points,* for example, A, A', and A'' in Fig. 4.9. *Evidently the separation in time between any two corresponding points on any two*

* A. Sommerfeld, *Optics*, p. 151. See also J. J. Stein, *Am. J. Phys.* **50**, 180 (1982).

* When the material is inhomogeneous or when there is more than one medium involved, it will be the *optical path length* (see Section 4.2.4) between the two wavefronts that is the same.

sequential wavefronts is identical. In other words, if wavefront Σ is transformed into Σ'' after a time t'', the distance between corresponding points on any and all rays will be traversed in that same time t''. This will be true even if the wavefronts pass from one homogeneous isotropic medium into another. This just means that each point on Σ can be imagined as following the path of a ray to arrive at Σ'' in the time t''.

If a group of rays is such that we can find a surface that is orthogonal to each and every one of them, they are said to form a *normal congruence.* For example, the rays emanating from a point source are perpendicular to a sphere centered at the source and consequently form a normal congruence.

We can now briefly consider an alternative to Huygens's principle that will also allow us to follow the progress of light through various isotropic media. The basis for this approach is the *theorem of Malus and Dupin* (introduced in 1808 by E. Malus and modified in 1816 by C. Dupin), according to which *a group of rays will preserve its normal congruence after any number of reflections and refractions* (as in Fig. 4.9). From our present vantage point of the wave theory, this is equivalent to the statement that rays remain orthogonal to wavefronts throughout all propagation processes in isotropic media. As shown in Problem 4.12, the theorem can be used to derive the law of reflection as well as Snell's law. It is often most convenient to carry out a ray trace through an optical system using the laws of reflection and refraction and then reconstruct the wavefronts. The latter can be accomplished in accord with the above considerations of equal transit times between corresponding points and the orthogonality of the rays and wavefronts.

Figure 4.10 depicts the parallel ray formation concomitant with a plane wave, where θ_i, θ_r, and θ_t, which have the exact same meanings as before, are now measured from the normal to the interface. The incident ray and the normal determine a plane known as the **plane of incidence**. Because of the symmetry of the situation, we must anticipate that both the reflected and transmitted rays will be undeflected from that plane. In other words, the respective unit propagation vectors $\hat{\mathbf{k}}_i$, $\hat{\mathbf{k}}_r$, and $\hat{\mathbf{k}}_t$ are coplanar.

In summary, then, the three basic laws of reflection and refraction are:

1. The incident, reflected, and refracted rays all lie in the plane of incidence.
2. $\theta_i = \theta_r$. [4.3]
3. $n_i \sin \theta_i = n_t \sin \theta_t$. [4.5]

These are illustrated rather nicely with a narrow light beam in the photographs of Fig. 4.11. Here, the incident medium is air ($n_i \approx 1.0$), and the transmitting medium is glass ($n_t \approx 1.5$). Consequently, $n_i < n_t$, and it follows

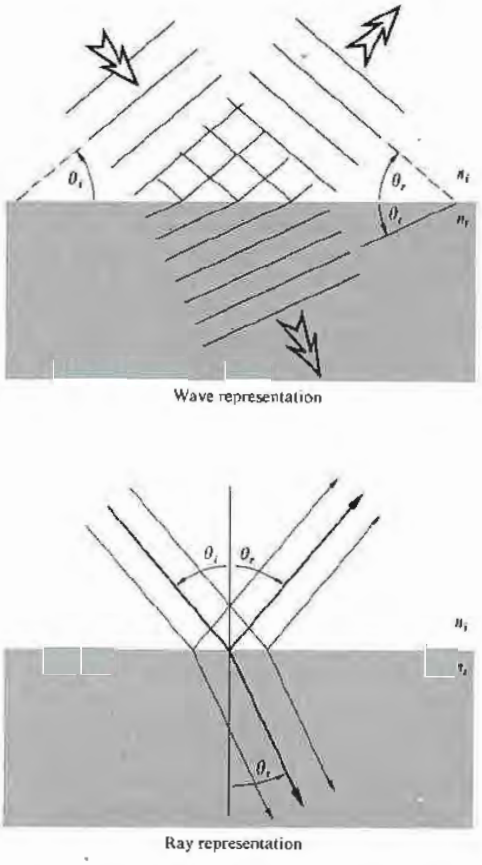


Figure 4.10 The wave and ray representations of an incident, reflected, and transmitted beam.

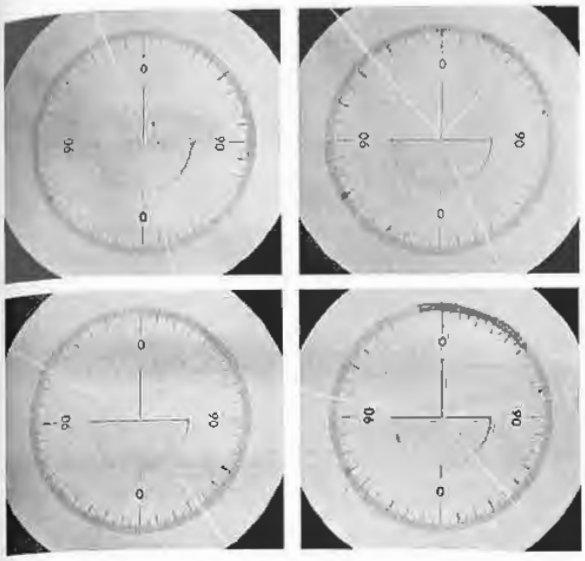


Figure 4.11 Refraction at various angles of incidence. (Photos courtesy *PSSC College Physics*, D. C. Heath & Co., 1968.)

from Snell's law that $\sin\theta_i > \sin\theta_t$. Since both angles, θ_i and θ_t, vary between 0° and 90°, a region over which the sine function is smoothly rising, it can be concluded that $\theta_i > \theta_t$. *Rays entering a higher-index medium from a lower one refract toward the normal* and vice versa. This much is evident in the figure. Notice that the bottom surface is cut circular so that the transmitted beam within the glass always lies along a radius and is therefore normal to the lower surface in every case. If a ray is normal to an interface, $\theta_i = 0 = \theta_t$, and it sails right through with no bending.

The incident beam in each portion of Fig. 4.11 is narrow and sharp, and the reflected beam is equally well defined. Accordingly, the process is known as **specular reflection** (from the word for a common mirror alloy in ancient times, speculum). In this case, as in Fig. 4.12(a), the reflecting surface is smooth, or more precisely, any irregularities in it are small compared with a wavelength.* In contrast, the **diffuse reflection** in Fig. 4.12(b) occurs when the surface is relatively rough. For example, "nonreflecting" glass used to cover pictures is actually glass whose surface is roughened so that it reflects diffusely. The law of reflection holds exactly over any region that is small enough to be considered smooth. These two forms of reflection are extremes; a whole range of intermediate behavior is possible. Thus, although the paper of this page was manufactured deliberately to be a fairly diffuse scatterer, the cover of the book reflects in a manner that is somewhere between diffuse and specular.

Let $\hat{\mathbf{u}}_n$ be a unit vector normal to the interface pointing in the direction from the incident to the transmitting medium (Fig. 4.13). As you will have the opportunity to prove in Problem 4.13, the first and third basic laws can be combined in the form of a *vector refraction equation:*

$$n_i(\hat{\mathbf{k}}_i \times \hat{\mathbf{u}}_n) = n_t(\hat{\mathbf{k}}_t \times \hat{\mathbf{u}}_n) \tag{4.7}$$

or, alternatively,

$$n_t\hat{\mathbf{k}}_t - n_i\hat{\mathbf{k}}_i = (n_t\cos\theta_t - n_i\cos\theta_i)\hat{\mathbf{u}}_n. \tag{4.8}$$

4.2.4 Fermat's Principle

The laws of reflection and refraction, and indeed the manner in which light propagates in general, can be viewed from an entirely different and intriguing perspective afforded us by **Fermat's principle**. The ideas that will unfold presently have had a tremendous influence on the development of physical thought in and beyond the study of classical optics. Apart from its implications in quantum optics (Section 13.6, p. 552), Fermat's principle provides us with an insightful and highly useful way of appreciating and anticipating the behavior of light.

Hero of Alexandria, who lived some time between 150 B.C. and 250 A.D., was the first to set forth what has since become known as a *variational principle.* In his formulation of the law of reflection, he asserted that *the path actually taken by light in going from some point S to a point P via a reflecting surface was the shortest possible one.* This can be seen rather easily in Fig. 4.14, which depicts a point source S emitting a number of rays that are

*If the surface ridges and valleys are small compared with λ, the scattered wavelets will still interfere constructively in only one direction ($\theta_i = \theta_r$).

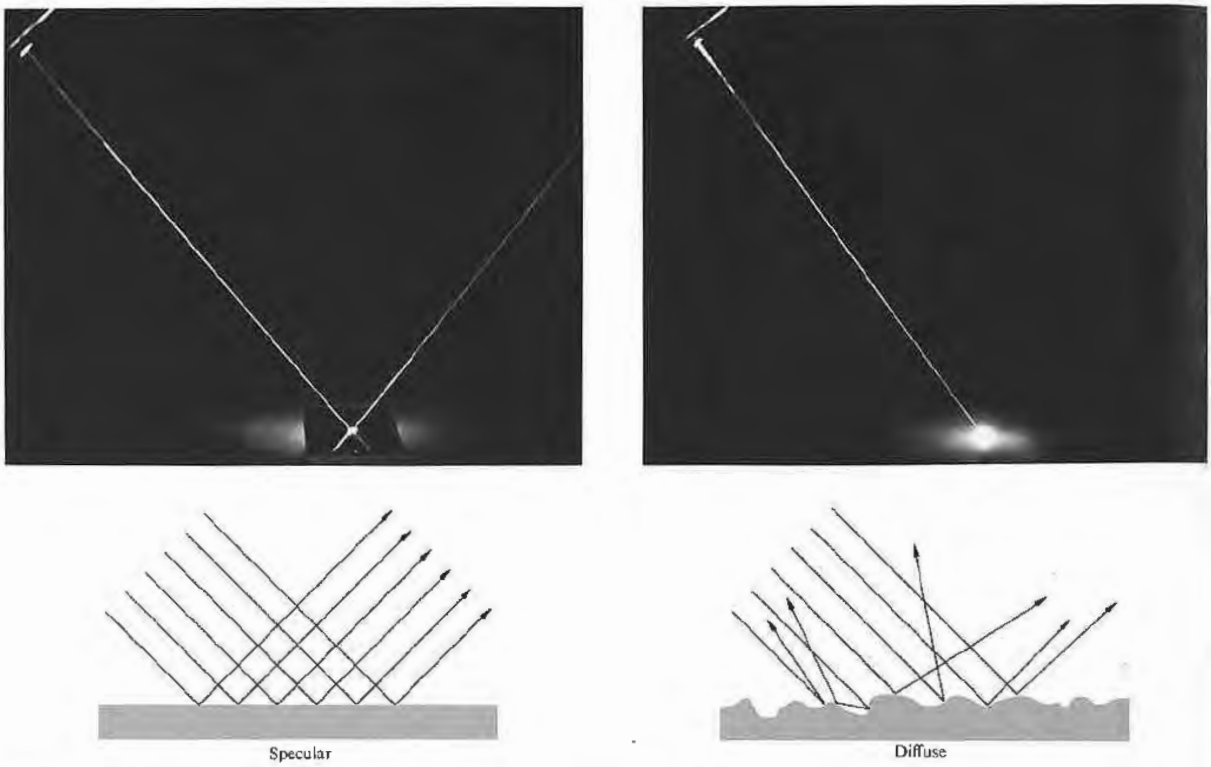


Figure 4.12 (a) Specular reflection. (b) Diffuse reflection. (Photos courtesy Donald Dunitz.)

then "reflected" toward P. Of course, only one of these paths will have any physical reality. If we simply draw the rays as if they emanated from S' (the image of S), none of the distances to P will have been altered (i.e., $SAP = S'AP$, $SBP = S'BP$, etc.). But obviously the straight-line path $S'BP$, which corresponds to $\theta_i = \theta_r$, is the shortest possible one. The same kind of reasoning (Problem 4.15) makes it evident that points S, B, and P must lie in what has previously been defined as the plane of incidence. For over fifteen hundred years Hero's curious observation stood alone, until in 1657 Fermat propounded his celebrated *principle of least time*, which encompassed both reflection and refraction. Obviously, a beam of light traversing an interface does not take a straight line or *minimum spatial path* between a point in the incident medium and one in the transmitting medium. Fermat consequently reformulated Hero's statement to read: *the actual path between two points taken by a beam of light is the one that is traversed in the least time.* As we shall see, even this form of the statement is somewhat incomplete and a bit erroneous at that. For the moment then, let us embrace it but not passionately.

As an example of the application of the principle to the case of refraction, refer to Fig. 4.15, where we minimize t, the transit time from S to P, with respect to the variable x. In other words, changing x shifts point O, thereby changing the ray from S to P. The smallest

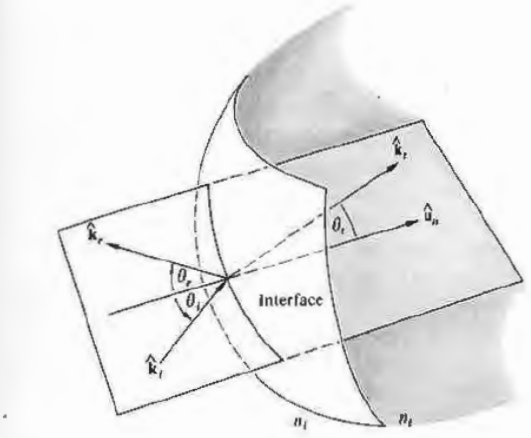


Figure 4.13 The ray geometry.

transit time will then presumably coincide with the actual path. Hence

$$t = \frac{\overline{SO}}{v_i} + \frac{\overline{OP}}{v_t}$$

or

$$t = \frac{(h^2 + x^2)^{1/2}}{v_i} + \frac{[b^2 + (a-x)^2]^{1/2}}{v_t}.$$

To minimize $t(x)$ with respect to variations in x, we set $dt/dx = 0$, that is,

$$\frac{dt}{dx} = \frac{x}{v_i(h^2 + x^2)^{1/2}} + \frac{-(a-x)}{v_t[b^2 + (a-x)^2]^{1/2}} = 0.$$

Using the diagram, we can rewrite the expression as

$$\frac{\sin\theta_i}{v_i} = \frac{\sin\theta_t}{v_t},$$

which is of course no less than Snell's law (Eq. 4.4). Thus if a beam of light is to advance from S to P in the least possible time, it must comply with the empirical law of refraction.

Suppose that we have a stratified material composed of m layers, each having a different index of refraction, as in Fig. 4.16. The transit time from S to P will then be

$$t = \frac{s_1}{v_1} + \frac{s_2}{v_2} + \cdots + \frac{s_m}{v_m}$$

or

$$t = \sum_{i=1}^{m} s_i/v_i,$$

where s_i and v_i are the path length and speed, respectively, associated with the ith contribution. Thus

$$t = \frac{1}{c}\sum_{i=1}^{m} n_i s_i, \tag{4.9}$$

in which the summation is known as the **optical path length** (OPL) traversed by the ray, in contrast to the spatial path length $\sum_{i=1}^{m} s_i$. Clearly, for an inhomogeneous medium where n is a function of position, the summation must be changed to an integral:

$$(\text{OPL}) = \int_S^P n(s)\, ds. \tag{4.10}$$

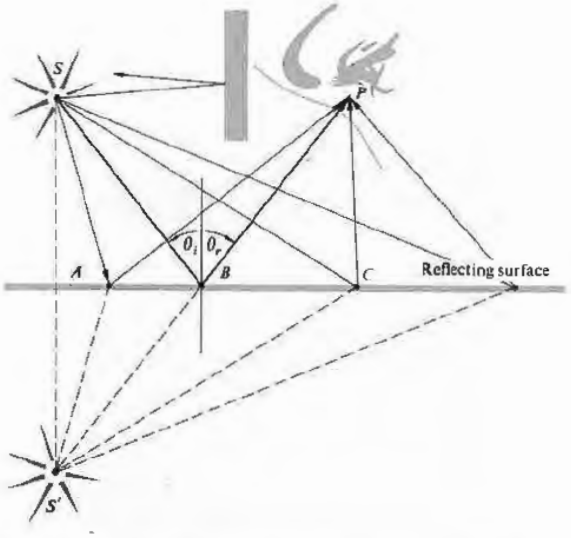


Figure 4.14 Minimum path from the source S to the observer's eye at P.

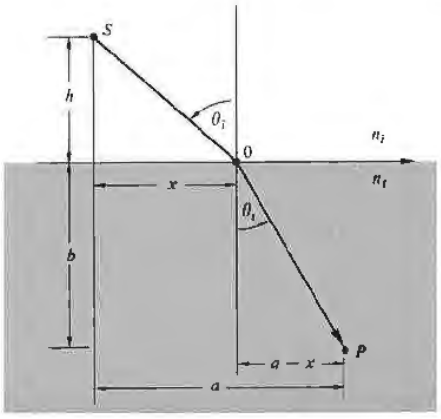


Figure 4.15 Fermat's principle applied to refraction.

Inasmuch as $t = (\text{OPL})/c$, we can restate Fermat's principle: *light, in going from points S to P, traverses the route having the smallest optical path length.* Accordingly, when light rays from the Sun pass through the inhomogeneous atmosphere of the Earth, as shown in Fig. 4.17(a), they bend so as to traverse the lower, denser regions as abruptly as possible, thus minimizing the OPL. Ergo, one can still see the Sun after it has actually passed below the horizon. In the same way, a road viewed at a glancing angle, as in Fig. 4.17(b), will appear to reflect the environs as if it were covered with a sheet of water. The air near the roadway will be warmer and less dense than that farther above it. Rays will bend upward, taking the shortest optical path, and in so doing they will appear to be reflected from a mirrored surface. The effect is particularly easy to see on long modern highways. The only requirement is that you look at the road at near glancing incidence, because the rays bend very gradually.

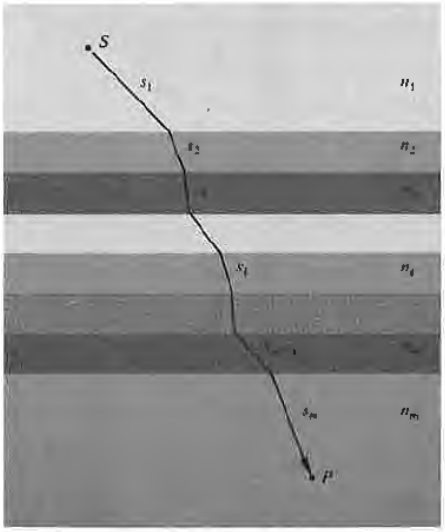


Figure 4.16 A ray propagating through a layered material.

The original statement of Fermat's *principle of least time* has some serious failings and is, as we shall see, in need of alteration. To that end, recall that if we have a function, say $f(x)$, we can determine the specific value of the variable x that causes $f(x)$ to have a *stationary* value by setting $df/dx = 0$ and solving for x. By a stationary value we mean one for which the slope of $f(x)$ versus x is zero or equivalently where the function has a maximum ⌒, minimum ⌣, or a point of inflection with a horizontal tangent —.

Fermat's principle in its modern form reads: *a light ray in going from point S to point P must traverse an optical path length that is stationary with respect to variations of that path.* In other words, the OPL for the true trajectory will equal, to a first approximation, the OPL of paths immediately adjacent to it.* Thus there will be many curves neighboring the actual one, which would take nearly the same time for the light to traverse. This latter point makes it possible to begin to understand how light manages to be so clever in its meanderings. Suppose that we have a beam of light advancing through a homogeneous isotropic medium so that a ray passes from points S to P. Atoms within the material are driven by the incident disturbance, and they reradiate in all directions. Generally, wavelets originating in the immediate vicinity of a stationary path will arrive at P by routes that differ only slightly and will therefore

*The first derivative of the OPL vanishes in its Taylor series expansion, since the path is stationary.

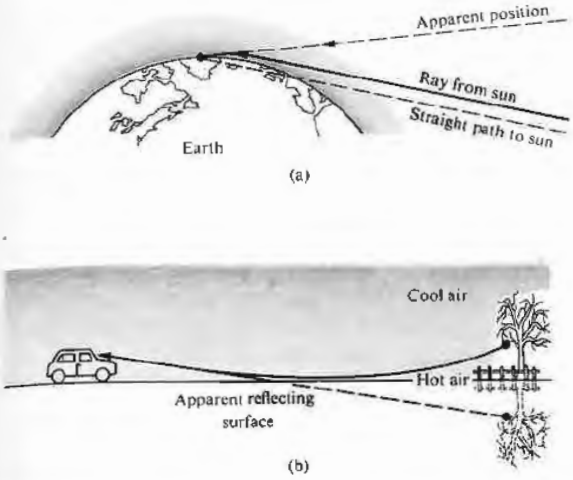


Figure 4.17 The bending of rays through inhomogeneous media.

arrive nearly in phase and reinforce each other (see Section 7.1). Wavelets taking other paths will arrive at P out of phase and will therefore tend to cancel each other. That being the case, energy will effectively propagate along that ray from S to P that satisfies Fermat's principle.

To show that the OPL for a ray need not always be a minimum, examine Fig. 4.18, which depicts a segment of a hollow three-dimensional ellipsoidal mirror. If the source S and the observer P are at the foci of the ellipsoid, then by definition the length SQP will be constant, regardless of where on the perimeter Q happens to be. It is also a geometrical property of the ellipse that $\theta_i = \theta_r$ for any location of Q. All optical paths from S to P via a reflection are therefore precisely equal—none is a minimum, and the OPL is clearly stationary with respect to variations. Rays leaving S and striking the mirror will arrive at the focus P. From another viewpoint we can say that radiant energy emitted by S will be scattered by electrons in the mirrored surface such that the wavelets will substantially reinforce each other only at P, where they have traveled the same distance and have the same phase. In any case, if a plane mirror was tangent to the ellipse at Q, the exact same

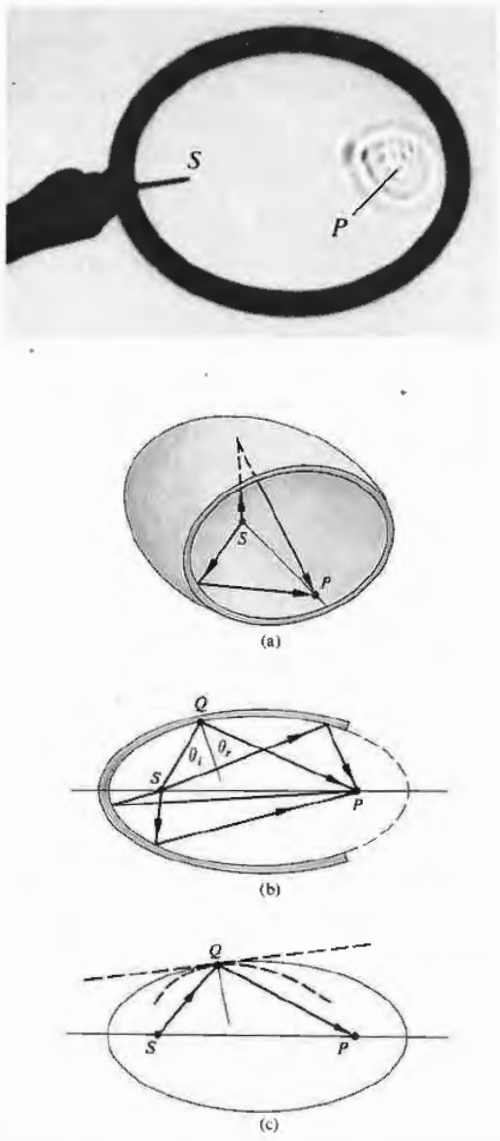


Figure 4.18 Reflection off an ellipsoidal surface. Observe the reflection of waves using a frying pan filled with water. Even though these are usually circular it is well worth playing with. (Photo courtesy *PSSC College Physics*, D. C. Heath & Co., 1968.)

path *SQP* traversed by a ray would then be a relative minimum. At the other extreme, if the mirrored surface conformed to a curve lying within the ellipse, like the dashed one shown, that same ray along *SQP* would now negotiate a relative maximum OPL. This is true even though other unused paths (where $\theta_i \neq \theta_r$) would actually be shorter (i.e., apart from inadmissible curved paths). Thus in all cases the rays travel a stationary OPL in accord with the reformulated Fermat's principle. Note that since the principle speaks only about the path and not the direction along it, a ray going from P to S will trace the same route as one from S to P. This is the very useful *principle of reversibility.*

Fermat's achievement stimulated a great deal of effort to supersede Newton's laws of mechanics with a similar variational formulation. The work of many men, notably Pierre de Maupertuis (1698–1759) and Leonhard Euler, finally led to the mechanics of Joseph Louis Lagrange (1736–1813) and hence to the *principle of least action,* formulated by William Rowan Hamilton (1805–1865). The striking similarity between the principles of Fermat and Hamilton played an important part in Schrödinger's development of quantum mechanics. In 1942 Richard Phillips Feynman (b. 1918) showed that quantum mechanics can be fashioned in an alternative way using a variational approach. The continuing evolution of variational principles brings us back to optics via the modern formalism of quantum optics (see Chapter 13).

Fermat's principle is not so much a computational device as it is a concise way of thinking about the propagation of light. It is a statement about the grand scheme of things without any concern for the contributing mechanisms, and as such it will yield insights under a myriad of circumstances.

4.3 THE ELECTROMAGNETIC APPROACH

Thus far we have been able to deduce the laws of reflection and refraction using three different approaches: *Huygens's principle,* the *theorem of Malus and Dupin,* and *Fermat's principle.* Each yields a distinctive and valuable point of view. Yet another and even more powerful approach is provided by the electromagnetic theory of light. Unlike the previous techniques, which say nothing about the incident, reflected, and transmitted radiant flux densities (i.e., I_i, I_r, I_t, respectively), the electromagnetic theory treats these within the framework of a far more complete description.

The body of information that forms the subject of optics has accrued over many centuries. As our knowledge of the physical universe becomes more extensive, the concomitant theoretical descriptions must become ever more encompassing. This, quite generally, brings with it an increased complexity. And so, rather than using the formidable mathematical machinery of the quantum theory of light, we will often avail ourselves of the simpler insights of simpler times (e.g., Huygens's and Fermat's principles). Thus even though we are now going to develop another and more extensive description of reflection and refraction, we will not put aside those earlier methods. In fact, throughout this study we shall use the simplest technique that can yield sufficiently accurate results for our particular purposes.

4.3.1 Waves at an Interface

Suppose that the incident monochromatic lightwave is planar, so that it has the form

$$\mathbf{E}_i = \mathbf{E}_{0i} \exp[i(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t)] \tag{4.11}$$

or, more simply,

$$\mathbf{E}_i = \mathbf{E}_{0i} \cos(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t). \tag{4.12}$$

Assume that $\mathbf{E}_{0i}$ is constant in time, that is, the wave is linearly or plane polarized. We'll find in Chapter 8 that any form of light can be represented by two orthogonal linearly polarized waves, so that this doesn't actually represent a restriction. Note that just as the origin in time, $t = 0$, is arbitrary, so too is the origin O in space, where $\mathbf{r} = 0$. Thus, making no assumptions about their directions, frequencies, wavelengths, phases, or amplitudes, we can write the reflected and transmitted waves as

$$\mathbf{E}_r = \mathbf{E}_{0r} \cos(\mathbf{k}_r \cdot \mathbf{r} - \omega_r t + \varepsilon_r) \tag{4.13}$$

and

$$\mathbf{E}_t = \mathbf{E}_{0t} \cos(\mathbf{k}_t \cdot \mathbf{r} - \omega_t t + \varepsilon_t). \tag{4.14}$$

Here ε_r and ε_t are *phase constants* relative to $\mathbf{E}_i$ and are introduced because the position of the origin is not unique. Figure 4.19 depicts the waves in the vicinity of the planar interface between two homogeneous lossless dielectric media of indices n_i and n_t.

The laws of electromagnetic theory (Section 3.1) lead to certain requirements that must be met by the fields, and these are referred to as the boundary conditions. Specifically, one of these is that the component of the electric field intensity $\mathbf{E}$ that is tangent to the interface must be continuous across it (the same is true for $\mathbf{H}$). In other words, the total tangential component of $\mathbf{E}$ on one side of the surface must equal that on the other (Problem 4.22). Thus since $\hat{\mathbf{u}}_n$ is the unit vector normal to the interface, regardless of the direction of the electric field within the wavefront, the cross-product of it with $\hat{\mathbf{u}}_n$ will be perpendicular to $\hat{\mathbf{u}}_n$ and therefore tangent to the interface. Hence

$$\hat{\mathbf{u}}_n \times \mathbf{E}_i + \hat{\mathbf{u}}_n \times \mathbf{E}_r = \hat{\mathbf{u}}_n \times \mathbf{E}_t \tag{4.15}$$

or

$$\begin{aligned}\hat{\mathbf{u}}_n \times \mathbf{E}_{0i} \cos(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t) \\ + \hat{\mathbf{u}}_n \times \mathbf{E}_{0r} \cos(\mathbf{k}_r \cdot \mathbf{r} - \omega_r t + \varepsilon_r) \\ = \hat{\mathbf{u}}_n \times \mathbf{E}_{0t} \cos(\mathbf{k}_t \cdot \mathbf{r} - \omega_t t + \varepsilon_t).\end{aligned} \tag{4.16}$$

This relationship must obtain at any instant in time and at any point on the interface ($y = b$). Consequently, $\mathbf{E}_i$, $\mathbf{E}_r$, and $\mathbf{E}_t$ must have precisely the same functional dependence on the variables t and r, which means that

$$\begin{aligned}(\mathbf{k}_i \cdot \mathbf{r} - \omega_i t)|_{y=b} &= (\mathbf{k}_r \cdot \mathbf{r} - \omega_r t + \varepsilon_r)|_{y=b} \\ &= (\mathbf{k}_t \cdot \mathbf{r} - \omega_t t + \varepsilon_t)|_{y=b}.\end{aligned} \tag{4.17}$$

With this as the case, the cosines in Eq. (4.16) cancel, leaving an expression independent of t and r, as indeed it must be. Inasmuch as this has to be true for all values of time, the coefficients of t must be equal, to wit

$$\omega_i = \omega_r = \omega_t. \tag{4.18}$$

Recall that the electrons within the media are undergoing (linear) forced vibrations at the frequency of the incident wave. Clearly, whatever light is scattered has that same frequency. Furthermore,

$$\begin{aligned}(\mathbf{k}_i \cdot \mathbf{r})|_{y=b} &= (\mathbf{k}_r \cdot \mathbf{r} + \varepsilon_r)|_{y=b} \\ &= (\mathbf{k}_t \cdot \mathbf{r} + \varepsilon_t)|_{y=b},\end{aligned} \tag{4.19}$$

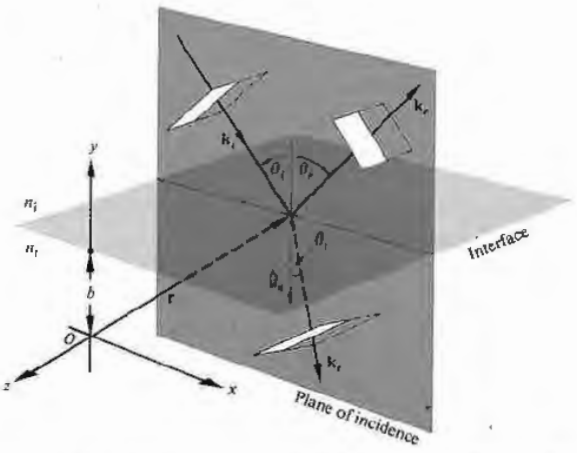


Figure 4.19 Plane waves incident on the boundary between two homogeneous, isotropic, lossless dielectric media.

wherein $\mathbf{r}$ terminates on the interface. The values of ε_r and ε_t correspond to a given position of O, and thus they allow the relation to be valid regardless of that location. (For example, the origin might be chosen such that $\mathbf{r}$ was perpendicular to $\mathbf{k}_i$ but not to $\mathbf{k}_r$ or $\mathbf{k}_t$.) From the first two terms we obtain

$$[(\mathbf{k}_i - \mathbf{k}_r) \cdot \mathbf{r}]_{y=b} = \varepsilon_r. \tag{4.20}$$

Recalling Eq. (2.42), this expression simply says that the endpoint of $\mathbf{r}$ sweeps out a plane (which is of course the interface) perpendicular to the vector $(\mathbf{k}_i - \mathbf{k}_r)$. To phrase it slightly differently, $(\mathbf{k}_i - \mathbf{k}_r)$ is parallel to $\hat{\mathbf{u}}_n$. Notice, however, that since the incident and reflected waves are in the same medium, $k_i = k_r$. From the fact that $(\mathbf{k}_i - \mathbf{k}_r)$ has no component in the plane of the interface, that is, $\hat{\mathbf{u}}_n \times (\mathbf{k}_i - \mathbf{k}_r) = 0$, we conclude that

$$k_i \sin\theta_i = k_r \sin\theta_r;$$

hence we have the law of reflection, that is,

$$\theta_i = \theta_r.$$

Furthermore, since $(\mathbf{k}_i - \mathbf{k}_r)$ is parallel to $\hat{\mathbf{u}}_n$ all three vectors, $\mathbf{k}_i$, $\mathbf{k}_r$, and $\hat{\mathbf{u}}_n$, are in the same plane, the plane of incidence. Again, from Eq. (4.19) we obtain

$$[(\mathbf{k}_i - \mathbf{k}_t) \cdot \mathbf{r}]_{y=b} = \varepsilon_t, \tag{4.21}$$

and therefore $(\mathbf{k}_i - \mathbf{k}_t)$ is also normal to the interface.

Thus $\mathbf{k}_i$, $\mathbf{k}_r$, $\mathbf{k}_t$, and $\hat{\mathbf{u}}_n$ are all coplanar. As before, the tangential components of $\mathbf{k}_i$ and $\mathbf{k}_t$ must be equal, and consequently

$$k_i \sin \theta_i = k_t \sin \theta_t. \tag{4.22}$$

But because $\omega_i = \omega_t$, we can multiply both sides by c/ω_i to get

$$n_i \sin \theta_i = n_t \sin \theta_t,$$

which is Snell's law. Finally, if we had chosen the origin O to be in the interface, it is evident from Eqs. (4.20) and (4.21) that ε_r and ϵ_t would both have been zero. That arrangement, although not as instructive, is certainly simpler, and we'll use it from here on.

4.3.2 Derivation of the Fresnel Equations

We have just found the relationship that exists among the phases of $\mathbf{E}_i(\mathbf{r}, t)$, $\mathbf{E}_r(\mathbf{r}, t)$, and $\mathbf{E}_t(\mathbf{r}, t)$ at the boundary. There is still an interdependence shared by the amplitudes $\mathbf{E}_{0i}$, $\mathbf{E}_{0r}$, and $\mathbf{E}_{0t}$, which can now be evaluated. To that end, suppose that a plane monochromatic wave is incident on the planar surface separating two isotropic media. Whatever the polarization of the wave, we shall resolve its **E**- and **B**-fields into components parallel and perpendicular to the plane of incidence and treat these constituents separately.

Case 1: E perpendicular to the plane of incidence. We now assume that **E** is perpendicular to the plane of incidence and that **B** is parallel to it (Fig. 4.20). Recall that $E = vB$, so that

$$\hat{\mathbf{k}} \times \mathbf{E} = v\mathbf{B} \tag{4.23}$$

and, of course,

$$\hat{\mathbf{k}} \cdot \mathbf{E} = 0 \tag{4.24}$$

(i.e., **E**, **B**, and the unit propagation vector $\hat{\mathbf{k}}$ form a right-handed system). Again making use of the continuity of the tangential components of the **E**-field, we have at the boundary at any time and any point

$$\mathbf{E}_{0i} + \mathbf{E}_{0r} = \mathbf{E}_{0t}, \tag{4.25}$$

where the cosines cancel. Realize that the field vectors as shown really ought to be envisioned at $y = 0$ (i.e., at the surface), from which they have been displaced for the sake of clarity. Note too that although $\mathbf{E}_r$ and $\mathbf{E}_t$ must be normal to the plane of incidence by symmetry, *we are guessing that they point outward* at the interface when $\mathbf{E}_i$ does. The directions of the **B**-fields then follow from Eq. (4.23).

We will need to invoke another of the boundary conditions in order to get one more equation. The presence of material substances that become electrically polarized by the wave has a definite effect on the field configuration. Thus, although the tangential component of **E** is continuous across the boundary, its normal component is not. Instead the normal component of the product $\epsilon\mathbf{E}$ is the same on either side of the

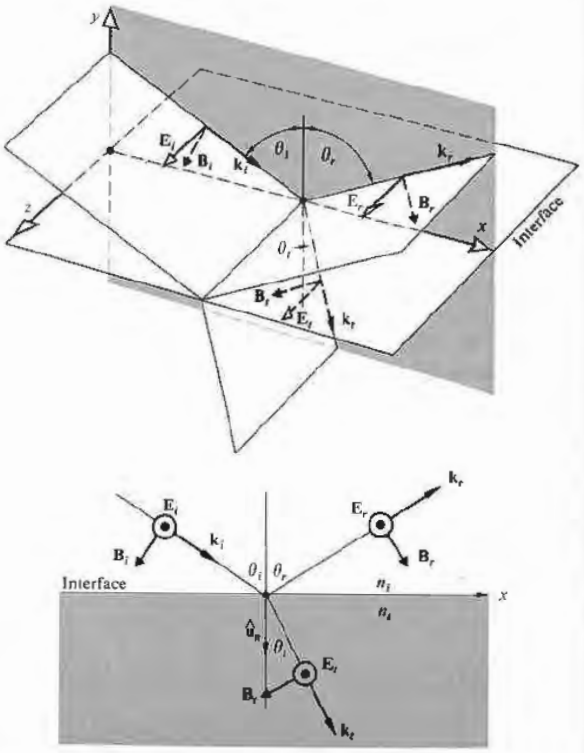


Figure 4.20 An incoming wave whose **E**-field is normal to the plane of incidence.

interface. Similarly, the normal component of **B** is continuous, as is the tangential component of $\mu^{-1}\mathbf{B}$. Here the effect of the two media appears via their permeabilities μ_i and μ_t. This boundary condition will be the simplest to use, particularly as applied to reflection from the surface of a conductor.* Thus the continuity of the tangential component of $\mathbf{B}/\mu$ requires that

$$-\frac{\mathbf{B}_i}{\mu_i}\cos\theta_i + \frac{\mathbf{B}_r}{\mu_r}\cos\theta_r = -\frac{\mathbf{B}_t}{\mu_t}\cos\theta_t, \tag{4.26}$$

where the left and right sides are the total magnitudes of $\mathbf{B}/\mu$ parallel to the interface in the incident and transmitting media, respectively. The positive direction is that of increasing x, so that the components of $\mathbf{B}_i$ and $\mathbf{B}_t$ appear with minus signs. From Eq. (4.23) we have

$$B_i = E_i/v_i, \tag{4.27}$$

$$B_r = E_r/v_r, \tag{4.28}$$

and

$$B_t = E_t/v_t. \tag{4.29}$$

Thus since $v_i = v_r$ and $\theta_i = \theta_r$, Eq. (4.26) can be written as

$$\frac{1}{\mu_i v_i}(E_i - E_r)\cos\theta_i = \frac{1}{\mu_t v_t}E_t\cos\theta_t. \tag{4.30}$$

Making use of Eqs. (4.12), (4.13), and (4.14) and remembering that the cosines therein equal one another $y = 0$, we obtain

$$\frac{n_i}{\mu_i}(E_{0i} - E_{0r})\cos\theta_i = \frac{n_t}{\mu_t}E_{0t}\cos\theta_t. \tag{4.31}$$

Combined with Eq. (4.25), this yields

$$\left(\frac{E_{0r}}{E_{0i}}\right)_\perp = \frac{\dfrac{n_i}{\mu_i}\cos\theta_i - \dfrac{n_t}{\mu_t}\cos\theta_t}{\dfrac{n_i}{\mu_i}\cos\theta_i + \dfrac{n_t}{\mu_t}\cos\theta_t} \tag{4.32}$$

and

$$\left(\frac{E_{0t}}{E_{0i}}\right)_\perp = \frac{2\dfrac{n_i}{\mu_i}\cos\theta_i}{\dfrac{n_i}{\mu_i}\cos\theta_i + \dfrac{n_t}{\mu_t}\cos\theta_t}. \tag{4.33}$$

The $\perp$ subscript serves as a reminder that we are dealing with the case in which **E** is perpendicular to the plane of incidence. These two expressions, *which are completely general statements applying to any linear, isotropic, homogeneous media,* are two of the **Fresnel equations**. Quite often one deals with dielectrics for which $\mu_i \approx \mu_t \approx \mu_0$; consequently the most common form of these equations is simply

$$r_\perp \equiv \left(\frac{E_{0r}}{E_{0i}}\right)_\perp = \frac{n_i\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t} \tag{4.34}$$

and

$$t_\perp \equiv \left(\frac{E_{0t}}{E_{0i}}\right)_\perp = \frac{2n_i\cos\theta_i}{n_i\cos\theta_i + n_t\cos\theta_t}. \tag{4.35}$$

Here $r_\perp$ denotes the **amplitude reflection coefficient**, and $t_\perp$ is the **amplitude transmission coefficient**.

Case 2: E parallel to the plane of incidence. A similar pair of equations can be derived when the incoming **E**-field lies in the plane of incidence, as shown in Fig. 4.21. Continuity of the tangential components of **E** on either side of the boundary leads to

$$E_{0i}\cos\theta_i - E_{0r}\cos\theta_r = E_{0t}\cos\theta_t. \tag{4.36}$$

In much the same way as before, continuity of the tangential components of $\mathbf{B}/\mu$ yields

$$\frac{1}{\mu_i v_i}E_{0i} + \frac{1}{\mu_r v_r}E_{0r} = \frac{1}{\mu_t v_t}E_{0t}. \tag{4.37}$$

Using the fact that $\mu_i = \mu_r$ and $\theta_i = \theta_r$, we can combine these formulas to obtain two more of the *Fresnel equations:*

$$r_\parallel \equiv \left(\frac{E_{0r}}{E_{0i}}\right)_\parallel = \frac{\dfrac{n_t}{\mu_t}\cos\theta_i - \dfrac{n_i}{\mu_i}\cos\theta_t}{\dfrac{n_i}{\mu_i}\cos\theta_t + \dfrac{n_t}{\mu_t}\cos\theta_i} \tag{4.38}$$

* In keeping with our intent to use only the **E**- and **B**-fields, at least in the early part of this exposition, we have avoided the usual statements in terms of **H**, where

$$\mathbf{H} = \mu^{-1}\mathbf{B}. \tag{A1.14}$$

and

$$t_{\parallel} \equiv \left(\frac{E_{0t}}{E_{0i}}\right)_{\parallel} = \frac{2\dfrac{n_i}{\mu_i}\cos\theta_i}{\dfrac{n_i}{\mu_i}\cos\theta_t + \dfrac{n_t}{\mu_t}\cos\theta_i}. \tag{4.39}$$

When both media forming the interface are dielectrics, the amplitude coefficients become

$$r_{\parallel} = \frac{n_t\cos\theta_i - n_i\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t} \tag{4.40}$$

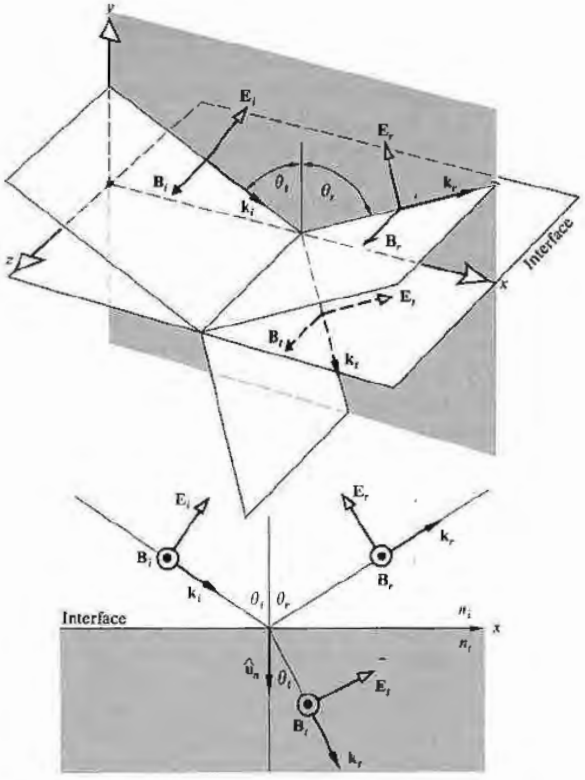


Figure 4.21 An incoming wave whose E-field is in the plane of incidence.

and

$$t_{\parallel} = \frac{2n_i\cos\theta_i}{n_i\cos\theta_t + n_t\cos\theta_i}. \tag{4.41}$$

One further notational simplification can be made by availing ourselves of Snell's law, whereupon the Fresnel equations for dielectric media become (Problem 4.23)

$$r_{\perp} = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)} \tag{4.42}$$

$$r_{\parallel} = +\frac{\tan(\theta_i - \theta_t)}{\tan(\theta_i + \theta_t)} \tag{4.43}$$

$$t_{\perp} = +\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i + \theta_t)} \tag{4.44}$$

$$t_{\parallel} = +\frac{2\sin\theta_t\cos\theta_i}{\sin(\theta_i + \theta_t)\cos(\theta_i - \theta_t)}. \tag{4.45}$$

A note of caution must be introduced before we move on to examine the considerable significance of the preceding calculation. Bear in mind that the directions (or more precisely, the phases) of the fields in Figs. 4.20 and 4.21 were selected rather arbitrarily. For example, in Fig. 4.20 we could have assumed that $\mathbf{E}_r$ pointed inward, whereupon $\mathbf{B}_r$ would have had to be reversed as well. Had we done that, the sign of $r_{\perp}$ would have turned out to be positive, leaving the other amplitude coefficients unchanged. The signs appearing in Eqs. (4.42) through (4.45), in this case positive, except for the first, correspond to the particular set of field directions selected. The minus sign, as we will see, just means that we didn't guess correctly concerning $\mathbf{E}_r$ in Fig. 4.20. Nonetheless, be aware that the literature is not standardized, and all possible sign variations have been labeled *Fresnel equations*—to avoid confusion *they must be related to the specific field directions from which they were derived.*

4.3.3 Interpretation of the Fresnel Equations

This section is devoted to an examination of the physical implications of the Fresnel equations. In particular we are interested in determining the fractional amplitudes and flux densities that are reflected and refracted. In

addition we shall be concerned with any possible phase shifts that might be incurred in the process.

i) Amplitude Coefficients

Let's briefly examine the form of the amplitude coefficients over the entire range of θ_i values. At nearly normal incidence ($\theta_i \approx 0$) the tangents in Eq. (4.43) are essentially equal to sines, in which case

$$[r_\parallel]_{\theta_i=0} = [-r_\perp]_{\theta_i=0} = \left[\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}\right]_{\theta_i=0}.$$

We will come back to the physical significance of the minus sign presently. After we have expanded the sines and used Snell's law, this expression becomes

$$[r_\parallel]_{\theta_i=0} = [-r_\perp]_{\theta_i=0} = \left[\frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_t \cos\theta_i + n_i \cos\theta_t}\right]_{\theta_i=0}, \tag{4.46}$$

which follows as well from Eqs. (4.34) and (4.40). In the limit, as θ_i goes to 0, $\cos\theta_i$ and $\cos\theta_t$ both approach one, and consequently

$$[r_\parallel]_{\theta_i=0} = [-r_\perp]_{\theta_i=0} = \frac{n_t - n_i}{n_t + n_i}. \tag{4.47}$$

Thus, for example, at an air ($n_i = 1$) glass ($n_t = 1.5$) interface at nearly normal incidence, the reflection coefficients equal ± 0.2.

When $n_t > n_i$ it follows from Snell's law that $\theta_i > \theta_t$, and $r_\perp$ is negative for all values of θ_i (Fig. 4.22). In contrast, $r_\parallel$ starts out positive at $\theta_i = 0$ and decreases gradually until it equals zero when $(\theta_i + \theta_t) = 90°$, since $\tan \pi/2$ is infinite. The particular value of the incident angle for which this occurs is denoted by θ_p and is referred to as the **polarization angle** (see Section 8.6.1). As θ_i increases beyond θ_p, $r_\parallel$ becomes progressively more negative, reaching -1.0 at 90°. If you place a single sheet of glass, a microscope slide, on this page and look straight down into it ($\theta_i = 0$), the region beneath the glass will seem decidedly grayer than the rest of the paper, because the slide will reflect at both its interfaces, and the light reaching and returning from the paper will be diminished appreciably. Now hold the slide near your eye and again view the page through it as you tilt it, increasing θ_i. The amount of light reflected will increase, and it will become more difficult to see

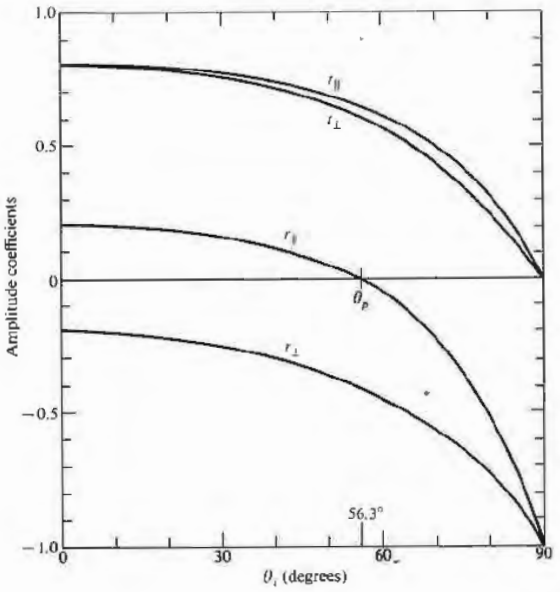


Figure 4.22 The amplitude coefficients of reflection and transmission as a function of incident angle. These correspond to external reflection $n_t > n_i$ at an air–glass interface ($n_{ig} = 1.5$).

the page through the glass. When $\theta_i \approx 90°$ the slide will look like a perfect mirror as the reflection coefficients (Fig. 4.22) go to -1.0. Even a rather poor surface, such as the cover of this book, will be mirrorlike at glancing incidence. Hold the book horizontally at the level of the middle of your eye and face a bright light; you will see the source reflected rather nicely in the cover. This suggests that even x-rays could be mirror-reflected at glancing incidence (p. 210), and modern x-ray telescopes are based on that very fact.

At normal incidence Eqs. (4.35) and (4.41) lead rather straightforwardly to

$$[t_\parallel]_{\theta_i=0} = [t_\perp]_{\theta_i=0} = \frac{2n_i}{n_i + n_t}. \tag{4.48}$$

It will be shown in Problem 4.24 that the expression

$$t_\perp + (-r_\perp) = 1 \tag{4.49}$$

holds for all θ_i, whereas

$$t_\parallel + r_\parallel = 1 \qquad (4.50)$$

is true only at normal incidence.

The foregoing discussion, for the most part, was restricted to the case of **external reflection** (i.e., $n_t > n_i$). The opposite situation of **internal reflection**, in which the incident medium is the more dense ($n_i > n_t$), is of interest as well. In that instance $\theta_t > \theta_i$, and $r_\perp$, as described by Eq. (4.42), will always be positive. Figure 4.23 shows that $r_\perp$ increases from its initial value (4.47) at $\theta_i = 0$, reaching +1 at what is called the **critical angle**, θ_c. Specifically, θ_c is the special value of the incident angle for which $\theta_t = \pi/2$. Likewise, $r_\parallel$ starts off negatively (4.47) at $\theta_i = 0$ and thereafter increases, reaching +1 at $\theta_i = \theta_c$, as is evident from the Fresnel equation (4.40). Again, $r_\parallel$ passes through zero at the *polarization angle* θ_p'. It is left for Problem 4.34 to show that the polarization angles θ_p' and θ_p for internal and external reflection at the interface between the same media are simply the complements of each other. We will return to internal reflection in Section 4.3.4, where it will be shown that $r_\perp$ and $r_\parallel$ are complex quantities for $\theta_i > \theta_c$.

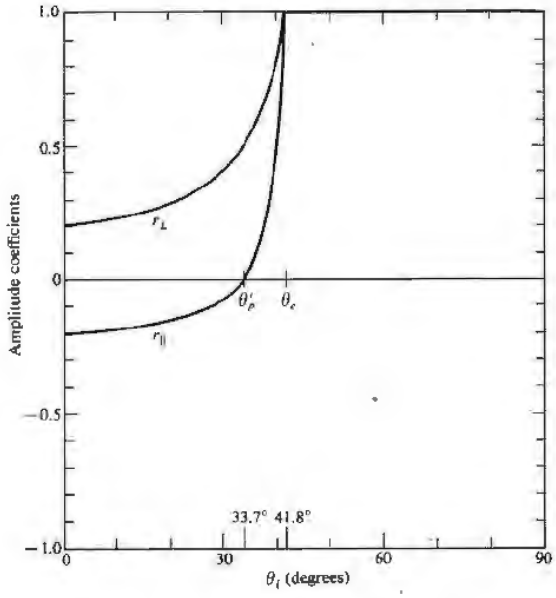


Figure 4.23 The amplitude coefficients of reflection as a function of incident angle. These correspond to internal reflection $n_i < n_t$ at an air–glass interface ($n_{ti} = 1/1.5$).

ii) Phase Shifts

It should be evident from Eq. (4.42) that $r_\perp$ is negative regardless of θ_i when $n_t > n_i$. Yet we saw earlier that had we chosen $[\mathbf{E}_r]_\perp$ in Fig. 4.20 to be in the opposite direction, the first Fresnel equation (4.42) would have changed signs, causing $r_\perp$ to become a positive quantity. Thus the sign of $r_\perp$ is associated with the relative directions of $[\mathbf{E}_{0i}]_\perp$ and $[\mathbf{E}_{0r}]_\perp$. Bear in mind that a reversal of $[\mathbf{E}_{0r}]_\perp$ is tantamount to introducing a phase shift, $\Delta\varphi_\perp$, of π radians into $[\mathbf{E}_r]_\perp$. Hence at the boundary $[\mathbf{E}_i]_\perp$ and $[\mathbf{E}_r]_\perp$ will be antiparallel and therefore π out of phase with each other, as indicated by the negative value of $r_\perp$. When we consider components normal to the plane of incidence, there is no confusion as to whether two fields are in phase or π radians out of phase: if parallel, they're in phase; if antiparallel, they're π out of phase. In summary, then, *the component of the electric field normal to the plane of incidence undergoes a phase shift of π radians upon reflection when the incident medium has a lower index than the transmitting medium.* Similarly, $t_\perp$ and $t_\parallel$ are always positive and $\Delta\varphi = 0$. *Furthermore, when $n_i > n_t$ no phase shift in the normal component results on reflection, that is, $\Delta\varphi_\perp = 0$ so long as $\theta_i < \theta_c$.*

Things are a bit less obvious when we deal with $[\mathbf{E}_i]_\parallel$, $[\mathbf{E}_r]_\parallel$, and $[\mathbf{E}_t]_\parallel$. It now becomes necessary to define more explicitly what is meant by *in phase*, since the field vectors are coplanar but generally not colinear. The field directions were chosen in Figs. 4.20 and 4.21 such that if you looked down any one of the propagation vectors toward the direction from which the light was coming, **E**, **B**, and **k** would appear to have the same relative orientation whether the ray was incident, reflected, or transmitted. We can use this as the required condition for two **E**-fields to be in phase. Equivalently, but more simply, *two fields in the incident plane are in phase if their y-components are parallel and are out of phase*

if the components are antiparallel. Notice that when two E-fields are out of phase so too are their associated B-fields and vice versa. With this definition we need only look at the vectors normal to the plane of incidence, whether they be **E** or **B**, to determine the relative phase of the accompanying fields in the incident plane. Thus in Fig. 4.24(a) $\mathbf{E}_i$ and $\mathbf{E}_t$ are in phase, as are $\mathbf{B}_i$ and $\mathbf{B}_t$, whereas $\mathbf{E}_i$ and $\mathbf{E}_r$ are out of phase, along with $\mathbf{B}_i$ and $\mathbf{B}_r$. Similarly, in Fig. 4.24(b) $\mathbf{E}_i$, $\mathbf{E}_r$, and $\mathbf{E}_t$ are in phase, as are $\mathbf{B}_i$, $\mathbf{B}_r$, and $\mathbf{B}_t$.

Now, the amplitude reflection coefficient for the parallel component is given by

$$r_\parallel = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_t \cos\theta_i + n_i \cos\theta_t},$$

which is positive ($\Delta\varphi_\parallel = 0$) as long as

$$n_t \cos\theta_i - n_i \cos\theta_i > 0,$$

that is, if

$$\sin\theta_i \cos\theta_i - \cos\theta_t \sin\theta_t > 0$$

or equivalently

$$\sin(\theta_i - \theta_t)\cos(\theta_i + \theta_t) > 0. \tag{4.51}$$

This will be the case for $n_i < n_t$ if

$$(\theta_i + \theta_t) < \pi/2 \tag{4.52}$$

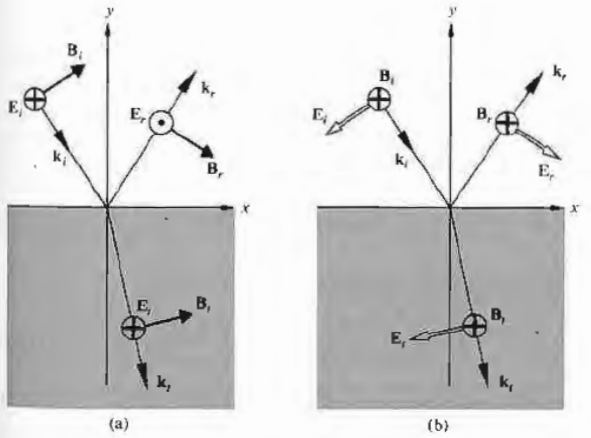


Figure 4.24 Field orientations and phase shifts.

and for $n_i > n_t$ when

$$(\theta_i + \theta_t) > \pi/2. \tag{4.53}$$

Thus when $n_i < n_t$, $[\mathbf{E}_{0r}]_\parallel$ and $[\mathbf{E}_{0i}]_\parallel$ will be in phase ($\Delta\varphi_\parallel = 0$) until $\theta_i = \theta_p$ and out of phase by π radians thereafter. The transition is not actually discontinuous, since $[\mathbf{E}_{0r}]_\parallel$ goes to zero at θ_p. In contrast, for internal reflection $r_\parallel$ is negative until θ_p', which means that $\Delta\varphi_\parallel = \pi$. From θ_p' to θ_c, $r_\parallel$ is positive and $\Delta\varphi_\parallel = 0$. Beyond θ_c, $r_\parallel$ becomes complex, and $\Delta\varphi_\parallel$ gradually increases to π at $\theta_i = 90°$.

Figure 4.25, which summarizes these conclusions, will be of continued use to us. The actual functional form of $\Delta\varphi_\parallel$ and $\Delta\varphi_\perp$ for internal reflection in the region where $\theta_i > \theta_c$ can be found in the literature,* but the curves depicted here will suffice for our purposes. Figure 4.25(e) is a plot of the relative phase shift between the parallel and perpendicular components, that is, $\Delta\varphi_\parallel - \Delta\varphi_\perp$. It is included here because it will be useful later on (e.g., when we consider polarization effects). Finally, many of the essential features of this discussion are illustrated in Figs. 4.26 and 4.27. The amplitudes of the reflected vectors are in accord with those of Figs. 4.22 and 4.23 (for an air–glass interface), and the phase shifts agree with those of Fig. 4.25.

Many of these conclusions can be verified with the simplest experimental equipment, namely, two linear polarizers, a piece of glass, and a small source, such as a flashlight or high-intensity lamp. By placing one polarizer in front of the source (at 45° to the plane of incidence), you can easily duplicate the conditions of Fig. 4.26. For example, when $\theta_i = \theta_p$ [Fig. 4.26(b)] no light will pass through the second polarizer if its transmission axis is parallel to the plane of incidence. In comparison, at near-glancing incidence the reflected beam will vanish when the axes of the two polarizers are almost normal to each other.

iii) Reflectance and Transmittance

Consider a circular beam of light incident on a surface, as shown in Fig. 4.28, such that there is an illuminated spot of area A. Recall that the power per unit area

* Born and Wolf, *Principles of Optics*, p. 49.

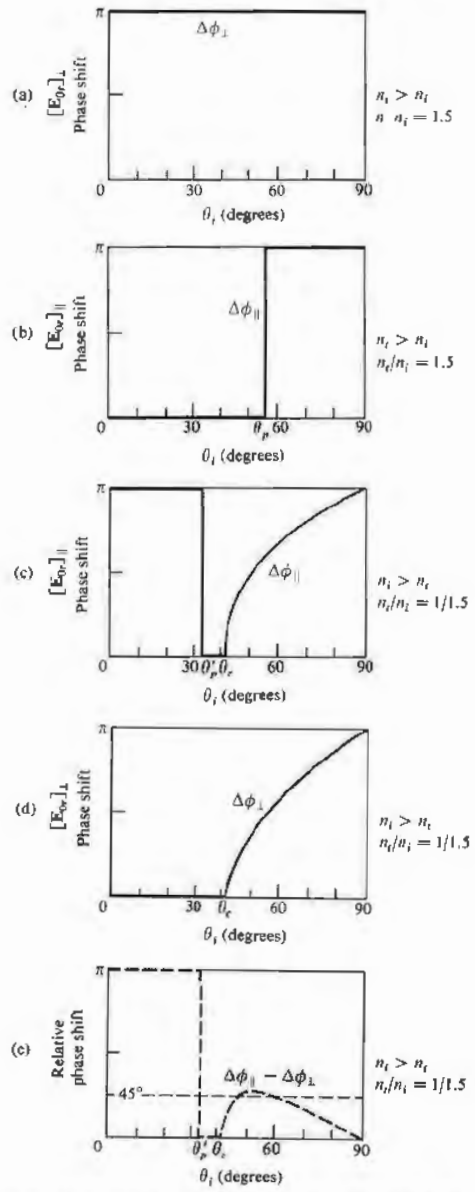


Figure 4.25 Phase shifts for the parallel and perpendicular components of the **E**-field corresponding to internal and external reflection.

crossing a surface in vacuum whose normal is parallel to **S**, the Poynting vector, is given by

$$\mathbf{S} = c^2\epsilon_0\mathbf{E} \times \mathbf{B}. \tag{3.48}$$

Furthermore, the radiant flux density (W/m^2) or irradiance is

$$I = \langle S \rangle = \frac{c\epsilon_0}{2} E_0^2. \tag{3.52}$$

This is the average energy per unit time crossing a unit area normal to **S** (in isotropic media **S** is parallel to **k**). In the case at hand (Fig. 4.28), let I_i, I_r, and I_t be the incident, reflected, and transmitted flux densities, respectively. The cross-sectional areas of the incident, reflected, and transmitted beams are, respectively, $A\cos\theta_i$, $A\cos\theta_r$, and $A\cos\theta_t$. Accordingly, the incident power is $I_iA\cos\theta_i$; this is the energy per unit time flowing in the incident beam and it's therefore the power arriving on the surface over A. Similarly, $I_rA\cos\theta_r$ is the power in the reflected beam, and $I_tA\cos\theta_t$ is the power being transmitted through A. We define the **reflectance** R to be the ratio of the reflected power (or flux) to the incident power:

$$R \equiv \frac{I_r\cos\theta_r}{I_i\cos\theta_i} = \frac{I_r}{I_i}. \tag{4.54}$$

In the same way, the **transmittance** T is defined as the ratio of the transmitted to the incident flux and is given by

$$T \equiv \frac{I_t\cos\theta_t}{I_i\cos\theta_i}. \tag{4.55}$$

The quotient I_r/I_i equals $(v_r\epsilon_r E_{0r}^2/2)/(v_i\epsilon_i E_{0i}^2/2)$, and since the incident and reflected waves are in the same medium, $v_r = v_i$, $\epsilon_r = \epsilon_i$, and

$$R = \left(\frac{E_{0r}}{E_{0i}}\right)^2 = r^2. \tag{4.56}$$

In like fashion (assuming $\mu_i = \mu_t = \mu_0$),

$$T = \frac{n_t\cos\theta_t}{n_i\cos\theta_i}\left(\frac{E_{0t}}{E_{0i}}\right)^2 = \left(\frac{n_t\cos\theta_t}{n_i\cos\theta_i}\right)t^2, \tag{4.57}$$

where use was made of the fact that $\mu_0\epsilon_t = 1/v_t^2$ and $\mu_0 v_t\epsilon_t = n_t/c$. Notice that at normal incidence, which is a situation of great practical interest, $\theta_t = \theta_i = 0$, and

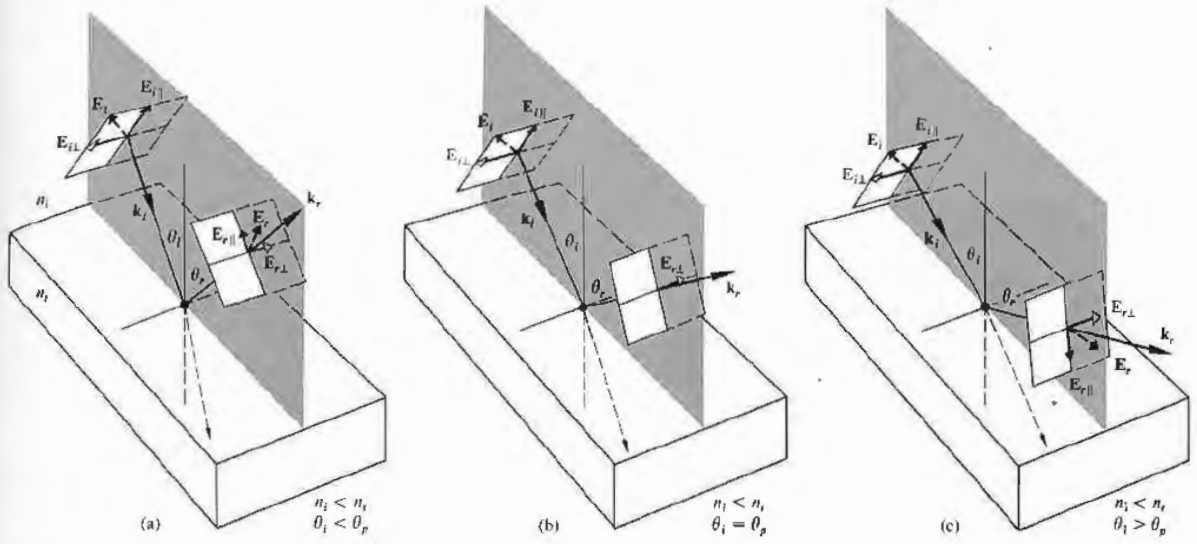


Figure 4.26 The reflected **E**-field at various angles concomitant with external reflection.

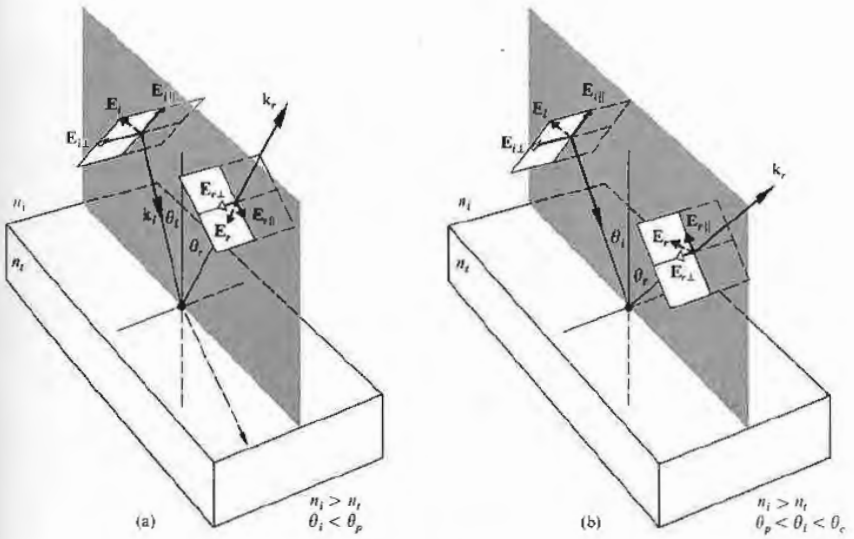


Figure 4.27 The reflected **E**-field at various angles concomitant with internal reflection.

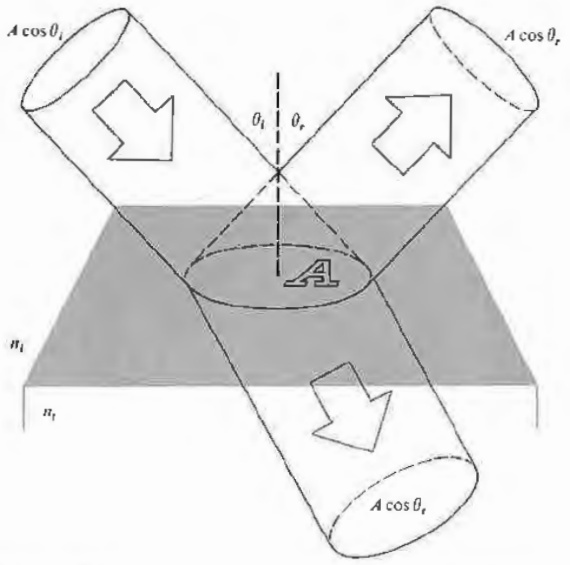


Figure 4.28 Reflection and transmission of an incident beam.

the transmittance [Eq. (4.55)], like the reflectance [Eq. (4.54)], is then simply the ratio of the appropriate irradiances. Since $R = r^2$, we need not worry about the sign of r in any particular formulation, and that makes reflectance a convenient notion. Observe that in Eq. (4.57) T is not simply equal to t^2, for two reasons. First, the ratio of the indices of refraction must be there, since the speeds at which energy is transported into and out of the interface are different, in other words, $I \propto v$, from Eq. (3.47). Second, the cross-sectional areas of the incident and reflected beams are different, and so the energy flow per unit area is affected accordingly, and that manifests itself in the presence of the ratio of the cosine terms.

Let's now write an expression representing the conservation of energy for the configuration depicted in Fig. 4.26. In other words, the total energy flowing into area A per unit time must equal the energy flowing outward from it per unit time:

$$I_i A \cos\theta_i = I_r A \cos\theta_r + I_t A \cos\theta_t. \tag{4.58}$$

When both sides are multiplied by c this expression becomes

$$n_i E_{0i}^2 \cos\theta_i = n_i E_{0r}^2 \cos\theta_i + n_t E_{0t}^2 \cos\theta_t$$

or

$$1 = \left(\frac{E_{0r}}{E_{0i}}\right)^2 + \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right)\left(\frac{E_{0t}}{E_{0i}}\right)^2. \tag{4.59}$$

But this is simply

$$R + T = 1, \tag{4.60}$$

where there was no absorption. It is convenient to use the component forms, that is,

$$R_\perp = r_\perp^2 \tag{4.61}$$

$$R_\parallel = r_\parallel^2 \tag{4.62}$$

$$T_\perp = \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right) t_\perp^2 \tag{4.63}$$

and

$$T_\parallel = \left(\frac{n_t \cos\theta_t}{n_i \cos\theta_i}\right) t_\parallel^2, \tag{4.64}$$

which are illustrated in Fig. 4.29. Furthermore, it can be shown (Problem 4.39) that

$$R_\parallel + T_\parallel = 1 \tag{4.65}$$

and

$$R_\perp + T_\perp = 1. \tag{4.66}$$

When $\theta_i = 0$ the incident plane becomes undefined, and any distinction between the parallel and perpendicular components of R and T vanishes. In this case Eqs. (4.61) through (4.64), along with (4.47) and (4.48), lead to

$$R = R_\parallel = R_\perp = \left(\frac{n_t - n_i}{n_t + n_i}\right)^2 \tag{4.67}$$

and

$$T = T_\parallel = T_\perp = \frac{4 n_t n_i}{(n_t + n_i)^2}. \tag{4.68}$$

Thus 4% of the light incident normally on an air–glass interface will be reflected back, whether internally, $n_i > n_t$, or externally, $n_i < n_t$ (Problem 4.40). This will

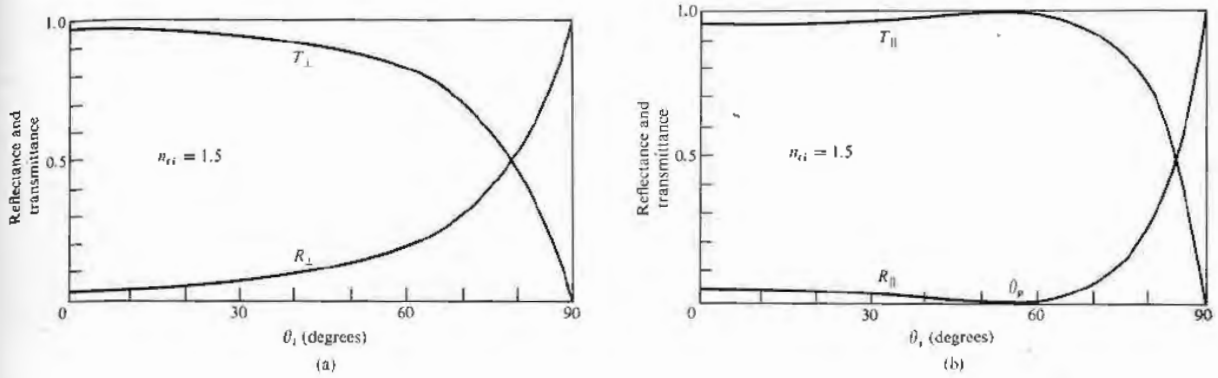


Figure 4.29 Reflectance and transmittance versus incident angle.

obviously be of great concern to anyone who is working with a complicated lens system, which might have 10 or 20 such air–glass boundaries. Indeed, if you look perpendicularly into a stack of about 50 microscope slides (cover-glass slides are much thinner and easier to handle in large quantities), most of the light will be reflected. The stack will look very much like a mirror (Fig. 4.30). Figure 4.31 is a plot of the reflectance at a single interface, assuming normal incidence for various transmitting media in air. Figure 4.32 depicts the corresponding dependence of the transmittance at normal incidence on the number of interfaces and the index of the medium. Of course, this is why you can't see through a roll of "clear" smooth-surfaced plastic tape,

Figure 4.30 Near normal reflection off a stack of microscope slides. You can see the image of the camera that took the picture. (Photo by E.H.)

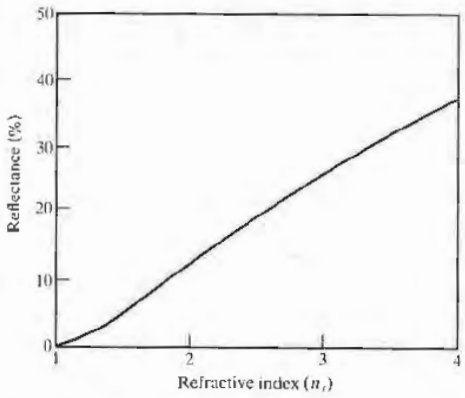


Figure 4.31 Reflectance at normal incidence in air ($n_i = 1.0$) at a single interface.

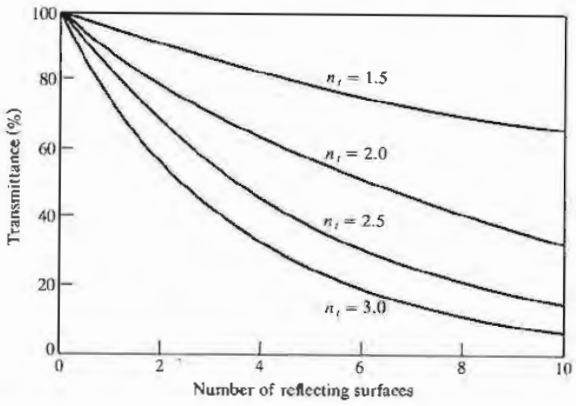


Figure 4.32 Transmittance through a number of surfaces in air ($n_i = 1.0$) at normal incidence.

and it's also why the many elements in a periscope must be coated with antireflection films (Section 9.9.2).

4.3.4 Total Internal Reflection

In the previous section it was evident that something rather interesting was happening in the case of internal reflection ($n_i > n_t$) when θ_i was equal to or greater than θ_c, the so-called **critical angle**. Let's now return to that situation for a somewhat closer look. Suppose that we have a source imbedded in an optically dense medium, and we allow θ_i to increase gradually, as indicated in Fig. 4.33. We know from the preceding section (Fig. 4.23) that $r_\parallel$ and $r_\perp$ increase with increasing θ_i, and therefore $t_\parallel$ and $t_\perp$ both decrease. Moreover $\theta_t > \theta_i$, since

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

and $n_i > n_t$, in which case $n_{ti} < 1$. Thus as θ_i becomes larger, the transmitted ray gradually approaches tangency with the boundary, and as it does so more and more of the available energy appears in the reflected beam. Finally, when $\theta_t = 90°$, $\sin \theta_t = 1$ and

$$\sin \theta_c = n_{ti}. \qquad (4.69)$$

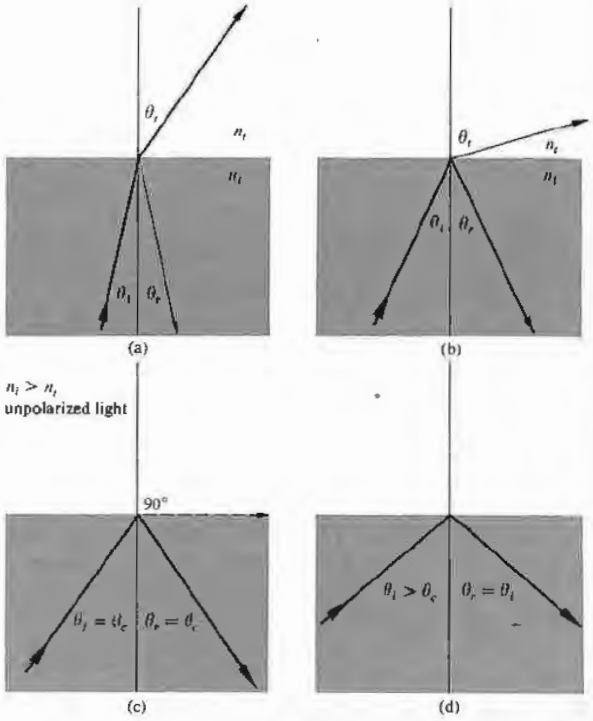


Figure 4.33 Internal reflection and the critical angle. (Photo courtesy of Educational Service, Inc.)

As noted earlier, *the critical angle is that special value of* θ_i *for which* $\theta_t = 90°$. For incident angles greater than or equal to θ_c, all the incoming energy is reflected back into the incident medium in the process known as **total internal reflection**. It should be stressed that the transition from the conditions depicted in Fig. 4.33(a) to those of 4.33(d) takes place without any discontinuities. That is to say, as θ_i becomes larger, the reflected beam grows stronger and stronger while the transmitted beam grows weaker, until the latter vanishes and the former carries off all the energy at $\theta_r = \theta_c$. It's an easy matter to observe the diminution of the transmitted beam as θ_i is made larger. Just place a glass microscope slide on a printed page, this time blocking out any specularly reflected light. At $\theta_i \approx 0$, θ_t is roughly zero, and the page as seen through the glass is fairly bright and clear. But if you move your head, allowing θ_t (the angle at which you view the interface) to increase, the region of the printed page covered by the glass will appear darker and darker, indicating that T has indeed been markedly reduced.

The critical angle for our air–glass interface is roughly 42° (see Table 4.1). Consequently, a ray incident normally on the left face of either of the prisms in Fig. 4.34

Table 4.1 Critical angles.

n_{ti}	θ_c (degrees)	θ_c (radians)	n_{ti}	θ_c (degrees)	θ_c (radians)
1.30	50.2849	0.8776	1.50	41.8103	0.7297
1.31	49.7612	0.8685	1.51	41.4718	0.7238
1.32	49.2509	0.8596	1.52	41.1395	0.7180
1.33	48.7535	0.8509	1.53	40.8132	0.7123
1.34	48.2682	0.8424	1.54	40.4927	0.7067
1.35	47.7946	0.8342	1.55	40.1778	0.7012
1.36	47.3321	0.8261	1.56	39.8683	0.6958
1.37	46.8803	0.8182	1.57	39.5642	0.6905
1.38	46.4387	0.8105	1.58	39.2652	0.6853
1.39	46.0070	0.8030	1.59	38.9713	0.6802
1.40	45.5847	0.7956	1.60	38.6822	0.6751
1.41	45.1715	0.7884	1.61	38.3978	0.6702
1.42	44.7670	0.7813	1.62	38.1181	0.6653
1.43	44.3709	0.7744	1.63	37.8428	0.6605
1.44	43.9830	0.7676	1.64	37.5719	0.6558
1.45	43.6028	0.7610	1.65	37.3052	0.6511
1.46	43.2302	0.7545	1.66	37.0427	0.6465
1.47	42.8649	0.7481	1.67	36.7842	0.6420
1.48	42.5066	0.7419	1.68	36.5296	0.6376
1.49	42.1552	0.7357	1.69	36.2789	0.6332

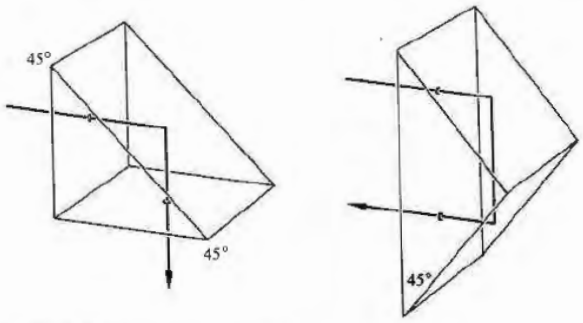


Figure 4.34 Total internal reflection.

will have a $\theta_i > 42°$ and therefore be internally reflected. This is a convenient way to reflect nearly 100% of the incident light without having to worry about the deterioration that can occur with metallic surfaces.

Another useful way to view the situation is shown in Fig. 4.35, which can be thought of as either a Huygens construction or a simplified representation of scattering off atomic oscillators. We know that the net effect of the presence of the homogeneous isotropic media is to alter the speed of the light from c to v_i and v_t, respectively (p. 63). This is equivalent mathematically (via Huygens's principle) to saying that the resultant wave is the superposition of these wavelets propagating at the appropriate speeds. In Fig. 4.35(a) an incident wave results in the emission of wavelets successively from scattering centers A and B. These overlap to form the transmitted wave. The reflected wave, which comes back down into the incident medium as usual ($\theta_i = \theta_r$), is not shown. In a time t the incident front travels a distance $v_i t = \overline{CB}$, while the transmitted front moves a distance $v_t t = \overline{AD} > \overline{CB}$. Since one wave moves from A to E in the same time that the other moves from C to B, and since they have the same frequency and period, they must change phase by the same amount in the process. Thus the disturbance at point E must be in phase with that at point B; both of these points must be on the same transmitted wavefront.

It can be seen that the greater v_t is in comparison to v_i, the more tilted the transmitted front will be (i.e., the larger θ_t will be). That much is depicted in Fig. 4.35(b), where n_{ti} has been taken to be smaller by assuming n_t

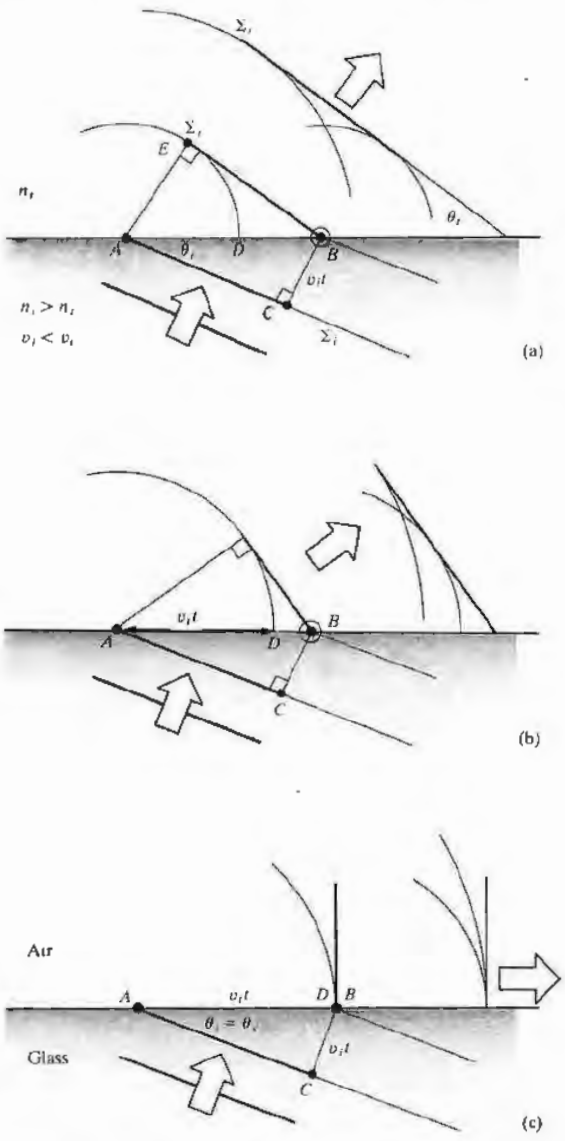


Figure 4.35 An examination of the transmitted wave in the process of total internal reflection from a scattering perspective. Here we keep θ_t and n_i constant and in successive parts of the diagram decrease n_t, thereby increasing v_t. The reflected wave ($\theta_r = \theta_i$) is not drawn.

to be smaller. The result is a higher speed v_t, increasing $\overline{AD}$ and causing a greater transmission angle. In Fig. 4.35(c) a special case is reached: $\overline{AD} = \overline{AB} = v_t t$, and the wavelets will overlap in phase *only along the line of the interface*, $\theta_t = 90°$. From triangle ABC, $\sin\theta_i = v_i t/v_t t = n_t/n_i$, which is Eq. (4.69). For the two given media (i.e., for the particular value of n_{ti}), the direction in which the scattered wavelets will add constructively in the transmitting medium is along the interface. The resulting disturbance ($\theta_t = 90°$) is known as a *surface wave*.

If we assume that there is no transmitted wave, it becomes impossible to satisfy the boundary conditions using only the incident and reflected waves—things are not at all as simple as they might seem. Furthermore, we can reformulate Eqs. (4.34) and (4.40) (Problem 4.43) such that

$$r_\perp = \frac{\cos\theta_i - (n_{ti}^2 - \sin^2\theta_i)^{1/2}}{\cos\theta_i + (n_{ti}^2 - \sin^2\theta_i)^{1/2}} \tag{4.70}$$

and

$$r_\parallel = \frac{n_{ti}^2\cos\theta_i - (n_{ti}^2 - \sin^2\theta_i)^{1/2}}{n_{ti}^2\cos\theta_i + (n_{ti}^2 - \sin^2\theta_i)^{1/2}}. \tag{4.71}$$

Clearly then, since $\sin\theta_c = n_{ti}$ when $\theta_i > \theta_c$, $\sin\theta_i > n_{ti}$, and both $r_\perp$ and $r_\parallel$ become complex quantities. Despite this (Problem 4.44), $r_\perp r_\perp^* = r_\parallel r_\parallel^* = 1$ and $R = 1$, which means that $I_r = I_i$ and $I_t = 0$. Thus, although there must be a transmitted wave, it cannot, on the average, carry energy across the boundary. We shall not perform the complete and rather lengthy computation needed to derive expressions for all the reflected and transmitted fields, but we can get an appreciation of what's happening in the following way. The wave function for the transmitted electric field is

$$\mathbf{E}_t = \mathbf{E}_{0t}\exp i(\mathbf{k}_t\cdot\mathbf{r} - \omega t),$$

where

$$\mathbf{k}_t\cdot\mathbf{r} = k_{tx}x + k_{ty}y,$$

there being no z-component of $\mathbf{k}$. But

$$k_{tx} = k_t\sin\theta_t$$

and

$$k_{ty} = k_t\cos\theta_t,$$

as seen in Fig. 4.36. Once again using Snell's law, we find that

$$k_t \cos \theta_t = \pm k_t \left(1 - \frac{\sin^2 \theta_i}{n_{ti}^2}\right)^{1/2} \tag{4.72}$$

or, since we are concerned with the case where $\sin \theta_i > n_{ti}$,

$$k_{ty} = \pm i k_t \left(\frac{\sin^2 \theta_i}{n_{ti}^2} - 1\right)^{1/2} \equiv \pm i\beta$$

and

$$k_{tx} = \frac{k_t}{n_{ti}} \sin \theta_i.$$

Hence

$$\mathbf{E}_t = \mathbf{E}_{0t} e^{\mp \beta y} e^{i(k_t x \sin \theta_i / n_{ti} - \omega t)}. \tag{4.73}$$

Neglecting the positive exponential, which is physically untenable, we have a wave whose amplitude drops off exponentially as it penetrates the less dense medium. The disturbance advances in the x-direction as a surface or **evanescent wave.** Notice that the wavefronts or surfaces of constant phase (parallel to the yz-plane) are perpendicular to the surfaces of constant amplitude (parallel to the xz-plane), and as such the wave is *inhomogeneous* (see Section 2.5). Its amplitude decays rapidly in the y-direction, becoming negligible at a distance into the second medium of only a few wavelengths.

If you are still concerned about the conservation of energy, a more extensive treatment would have shown that energy actually circulates back and forth across the interface, resulting on the average in a zero net flow through the boundary into the second medium. Yet one puzzling point remains, inasmuch as there is still a bit of energy to be accounted for, namely, that associated with the evanescent wave that moves along the boundary in the plane of incidence. Since this energy could not have penetrated into the less dense medium under the present circumstances (so long as $\theta_i \geq \theta_c$), we must look elsewhere for its source. Under actual experimental conditions the incident beam would have a finite cross section and therefore would obviously differ from a true plane wave. This deviation gives rise (via diffraction) to a slight transmission of energy across the interface, which is manifested in the evanescent wave.

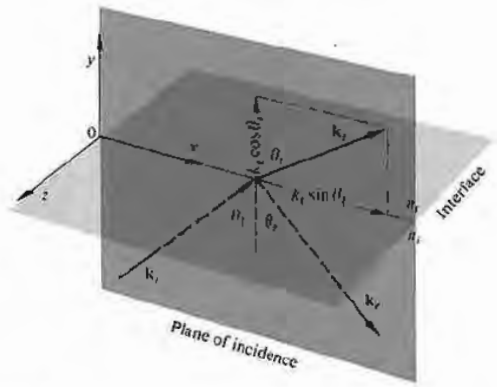


Figure 4.36 Propagation vectors for internal reflection.

Incidentally, it is clear from (c) and (d) in Fig. 4.25 that the incident and reflected waves (except at $\theta_i = 90°$) do not differ in phase by π and cannot therefore cancel each other. It follows from the continuity of the tangential component of **E** that there must be an oscillatory field in the less dense medium with a component parallel to the interface having a frequency ω (i.e., the evanescent wave).

The exponential decay of the surface wave, or *boundary wave*, as it is also sometimes called, has been confirmed experimentally at optical frequencies.*

Imagine that a beam of light traveling within a block of glass is internally reflected at a boundary. Presumably, if you pressed another piece of glass against the first, the air–glass interface could be made to vanish, and the beam would then propagate onward undisturbed. Furthermore, you might expect this transition from total to no reflection to occur gradually as the air film thinned out. In much the same way, if you hold a drinking glass or a prism, you can see the ridges of your fingerprints in a region that, because of total internal reflection, is otherwise mirrorlike. In more general terms, if the evanescent wave extends with appreciable amplitude across the rare medium into a nearby region occupied by a higher-index material, energy may flow through the gap in what is known as **frustrated total**

* Take a look at the fascinating article by K. H. Drexhage, "Monomolecular Layers and Light." *Sci. Am.* **222**, 108 (1970).

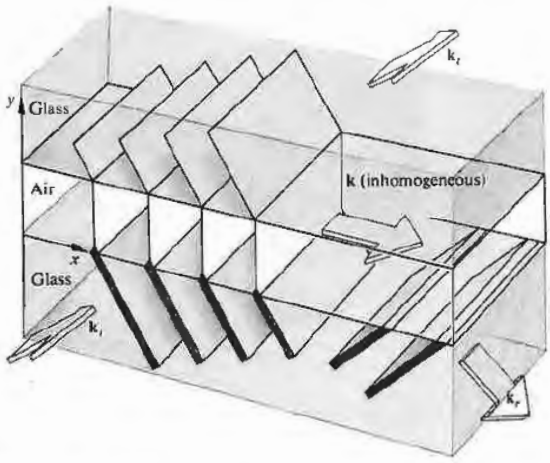


Figure 4.37 Frustrated total internal reflection.

internal reflection (FTIR). In other words, if the evanescent wave, having traversed the gap, is still strong enough to drive electrons in the "frustrating" medium, they in turn will generate a wave that significantly alters the field configuration, thereby permitting energy to flow. Figure 4.37 is a schematic representation of FTIR. The width of the lines depicting the wavefronts decreases across the gap as a reminder that the amplitude of the field behaves in the same way. The process as a whole is remarkably similar to the quantum-mechanical phenomenon of *barrier penetration* or *tunneling*, which has numerous applications in contemporary physics.

One can demonstrate FTIR with the prism arrangement of Fig. 4.38 in a manner that is fairly self-evident. Moreover, if the hypotenuse faces of both prisms are made planar and parallel, they can be positioned so as to transmit and reflect any desired fraction of the incident flux density. Devices that perform this function are known as *beam-splitters*. A *beam-splitter cube* can be made rather conveniently by using a thin, low-index transparent film as a precision spacer. Low-loss reflectors whose transmittance can be controlled by frustrating internal reflection are of considerable practical interest. FTIR can also be observed in other regions of the electromagnetic spectrum. Three-centimeter microwaves are particularly easy to work with, inasmuch as the evanescent wave will extend roughly 10^5 times farther than it would at optical frequencies. One can duplicate the above optical experiments with solid prisms made of paraffin or hollow ones of acrylic plastic filled with kerosene or motor oil. Any one of these would have an index of about 1.5 for 3-cm waves. It then becomes an easy matter to measure the dependence of the field amplitude on y.

4.3.5 Optical Properties of Metals

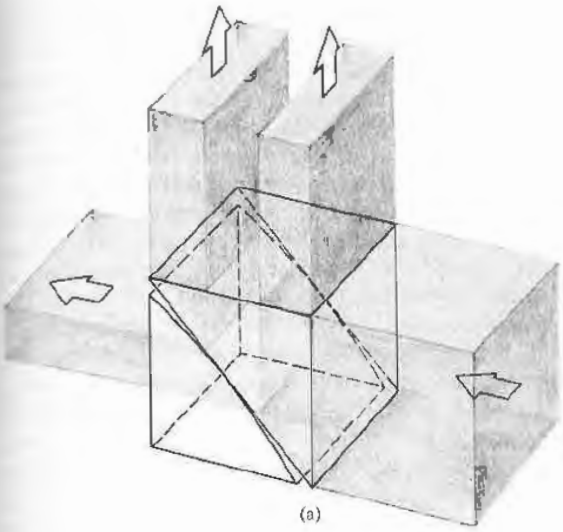
The characteristic feature of conducting media is the presence of a number of free electric charges (free in the sense of being unbound, i.e., able to circulate within the material). For metals these charges are of course electrons, and their motion constitutes a current. The current per unit area resulting from the application of a field $\mathbf{E}$ is related by means of Eq. (A1.15) to the conductivity of the medium σ. For a dielectric there are no free or conduction electrons and $\sigma = 0$, whereas for actual metals σ is nonzero and finite. In contrast, an idealized "perfect" conductor would have an infinite conductivity. This is equivalent to saying that the electrons, driven into oscillation by a harmonic wave, would simply follow the field's alternations. There would he no restoring force, no natural frequencies, and no absorption, only reemission. In real metals the conduction electrons undergo collisions with the thermally agitated lattice or with imperfections and in so doing irreversibly convert electromagnetic energy into joule heat. Evidently the absorption of radiant energy by a material is a function of its conductivity.

i) Waves in a Metal

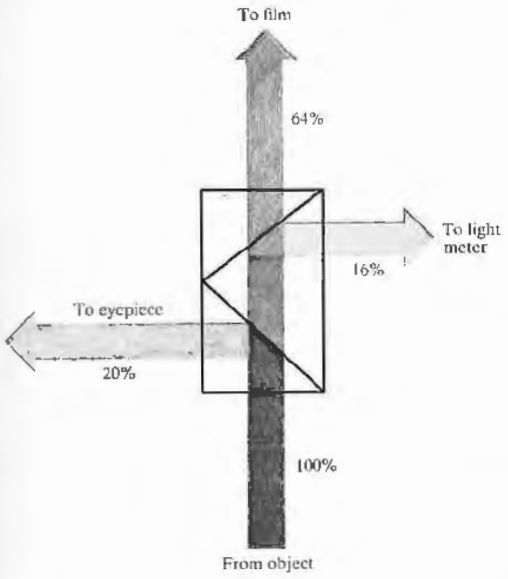
If we visualize the medium as continuous, Maxwell's equations lead to

$$\frac{\partial^2 \mathbf{E}}{\partial x^2} + \frac{\partial^2 \mathbf{E}}{\partial y^2} + \frac{\partial^2 \mathbf{E}}{\partial z^2} = \mu\epsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} + \mu\sigma \frac{\partial \mathbf{E}}{\partial t}, \qquad (4.74)$$

which is Eq. (A1.21) in Cartesian coordinates. The last term, $\mu\sigma\, \partial\mathbf{E}/\partial t$, is a first-order time derivative, like the damping force in the oscillator model discussed in Sec-



(a)



(b)

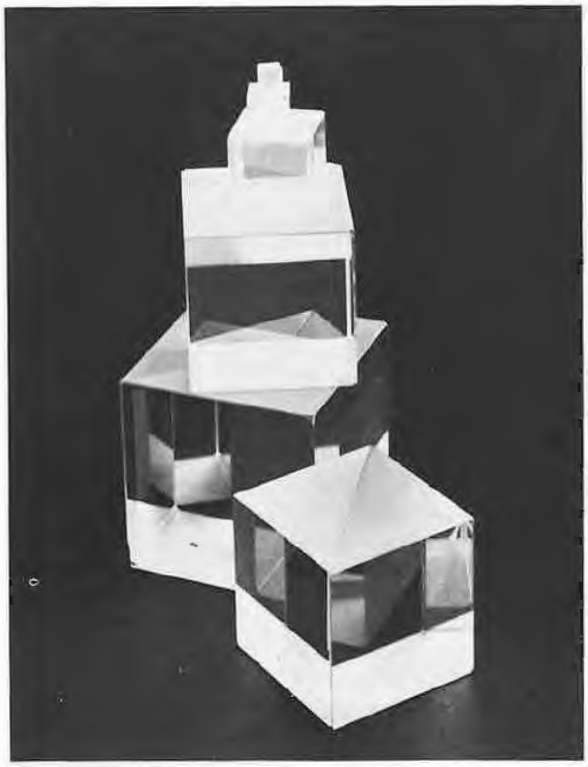


Figure 4.38 (a) A beam-splitter utilizing FTIR. (b) A typical modern application of FTIR: a conventional beam-splitter arrangement used to take photographs through a microscope. (c) Beam-splitter cubes. (Photo courtesy Melles Griot.)

tion 3.5.1. The time rate of change of **E** generates a voltage, currents circulate, and since the material is resistive, light is converted to heat—ergo absorption. This expression can be reduced to the unattenuated wave equation, if the permittivity is reformulated as a complex quantity. This in turn leads to a complex index of refraction, which, as we saw earlier (Section 3.5.1), is tantamount to absorption. We then need only substitute the complex index

$$n_c = n_R - in_I \tag{4.75}$$

(where the real and imaginary indices n_R and n_I are both real numbers) into the corresponding solution for a nonconducting medium. Alternatively, we can utilize the wave equation and appropriate boundary conditions to yield a specific solution. In either event, we can find a simple sinusoidal plane-wave solution applicable within the conductor. Such a wave propagating in the y-direction is ordinarily written as

$$\mathbf{E} = \mathbf{E}_0 \cos(\omega t - ky)$$

or as a function of n,

$$\mathbf{E} = \mathbf{E}_0 \cos \omega(t - ny/c),$$

but here the refractive index must be taken as complex. Accordingly, writing the wave as an exponential and using Eq. (4.75), we obtain

$$\mathbf{E} = \mathbf{E}_0 e^{(-\omega n_I y/c)} e^{i\omega(t - n_R y/c)} \tag{4.76}$$

or

$$\mathbf{E} = \mathbf{E}_0 e^{-\omega n_I y/c} \cos \omega(t - n_R y/c). \tag{4.77}$$

The disturbance advances in the y-direction with a speed c/n_R, precisely as if n_R were the more usual index of refraction. As the wave progresses into the conductor, its amplitude, $\mathbf{E}_0 \exp(-\omega n_I y/c)$, is exponentially attenuated. Inasmuch as irradiance is proportional to the square of the amplitude, we have

$$I(y) = I_0 e^{-\alpha y}, \tag{4.78}$$

where $I_0 = I(0)$, that is, I_0 is the irradiance at $y = 0$ (the interface), and $\alpha = 2\omega n_I/c$ is called the *absorption coefficient* or (even better) the **attenuation coefficient**. The flux density will drop by a factor of $e^{-1} = 1/2.7 \approx \frac{1}{3}$ after the wave has propagated a distance $y = 1/\alpha$, known as the *skin* or *penetration depth*. For a material to be transparent the penetration depth must be large in comparison to its thickness. The penetration depth for metals, however, is exceedingly small. For example, copper at ultraviolet wavelengths ($\lambda_0 \approx 100$ nm) has a miniscule penetration depth, about 0.6 nm, while it is still only about 6 nm in the infrared ($\lambda_0 \approx 10{,}000$ nm). This accounts for the generally observed opacity of metals, which nonetheless can become partly transparent when formed into extremely thin films (e.g., in the case of partially silvered two-way mirrors). The familiar metallic sheen of conductors corresponds to a high reflectance, which arises from the fact that the incident wave cannot effectively penetrate the material. Relatively few electrons in the metal "see" the transmitted wave, and therefore, although each absorbs strongly, little total energy is dissipated by them. Instead, most of the incoming energy reappears as the reflected wave. The majority of metals, including the less common ones (e.g., sodium, potassium, cesium, vanadium, niobium, gadolinium, holmium, yttrium, scandium, and osmium) have a silvery gray appearance like that of aluminum, tin, or steel. They reflect almost all the incident light regardless of wavelengths and are therefore essentially colorless.

Equation (4.77) is certainly reminiscent of Eq. (4.73) and FTIR. In both cases there is an exponential decay of the amplitude. Moreover, a complete analysis would show that the transmitted waves are not strictly transverse, there being a component of the field in the direction of propagation in both instances.

The representation of metal as a continuous medium works fairly well in the low-frequency, long-wavelength domain of the infrared. Yet we certainly might expect that as the wavelength of the incident beam decreased the actual granular nature of matter would have to be reckoned with. Indeed, the continuum model shows large discrepancies from experimental results at optical frequencies. And so we again turn to the classical atomistic picture initially formulated by Hendrik Lorentz, Paul Karl Ludwig Drude (1863–1906), and others. This simple approach will provide qualitative agreement with the experimental data, but the ultimate treatment nonetheless requires quantum theory.

ii) The Dispersion Equation

Envision the conductor as an assemblage of driven, damped oscillators. Some correspond to free electrons and will therefore have zero restoring force, whereas others are bound to the atom, much like those in the dielectric media of Section 3.5.1. The conduction electrons are, however, the predominant contributors to the optical properties of metals. Recall that the displacement of a vibrating electron was given by

$$x(t) = \frac{q_e/m_e}{(\omega_0^2 - \omega^2)} E(t). \qquad [3.65]$$

With no restoring force, $\omega_0 = 0$, the displacement is opposite in sign to the driving force $q_e E(t)$ and therefore 180° out of phase with it. This is unlike the situation for transparent dielectrics, where the resonance frequencies are above the visible and the electrons oscillate in phase with the driving force (Fig. 4.39). Free electrons oscillating out of phase with the incident light will reradiate wavelets that tend to cancel the incoming disturbance. The effect, as we have already seen, is a rapidly decaying refracted wave.

Assuming that the average field experienced by an electron moving about within a conductor is just the applied field $\mathbf{E}(t)$, we can extend the dispersion equation of a rare medium (3.71) to read

$$n^2(\omega) = 1 + \frac{Nq_e^2}{\epsilon_0 m_e}\left[\frac{f_e}{-\omega^2 + i\gamma_e\omega} + \sum_j \frac{f_j}{\omega_{0j}^2 - \omega^2 + i\gamma_j\omega}\right]. \qquad (4.79)$$

The first bracketed term is the contribution from the free electrons, wherein N is the number of atoms per unit volume. Each of these has f_e conduction electrons, which have no natural frequencies. The second term arises from the bound electrons and is identical to Eq. (3.71). It should be noted that if a metal has a particular color, it indicates that the atoms are partaking of selective absorption by way of the bound electrons, in addition to the general absorption characteristic of the free electrons. Recall that a medium that is very strongly absorbing at a given fequency doesn't actually absorb much of the incident light at that frequency but rather *selectively reflects* it. Gold and copper are reddish yellow

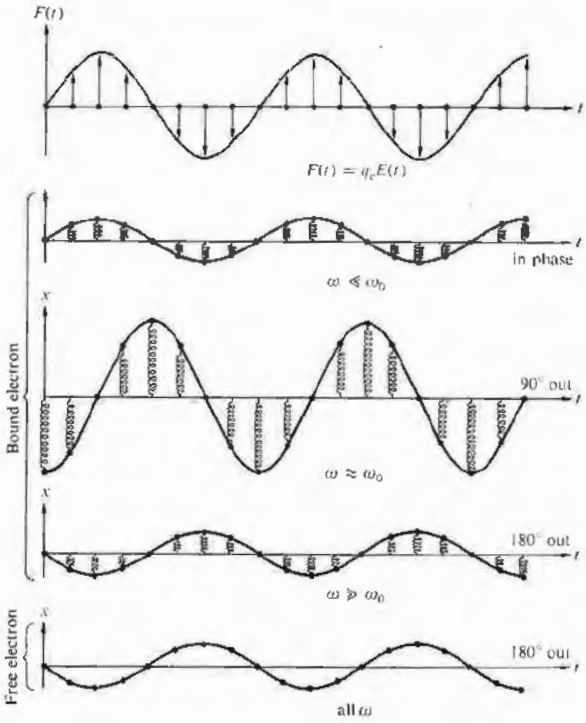


Figure 4.39 Oscillations of bound and free electrons.

because n_I increases with wavelength, and the larger values of λ are reflected more strongly. Thus, for example, gold should be fairly opaque to the longer visible wavelengths. Consequently, under white light, a gold foil less than roughly 10^{-6} m thick will indeed transmit predominantly greenish blue light.

We can get a rough idea of the response of metals to light by making a few simplifying assumptions. Accordingly, we neglect the bound electron contribution and assume that γ_e is also negligible for very large ω, whereupon

$$n^2(\omega) = 1 - \frac{Nq_e^2}{\epsilon_0 m_e \omega^2}. \qquad (4.80)$$

The latter assumption is based on the fact that at high frequencies the electrons will undergo a great many oscillations between each collision. Free electrons and positive ions within a metal may be thought of as a plasma whose density oscillates at a natural frequency ω_p, the *plasma frequency*. This in turn can be shown to equal $(Nq_e^2/\epsilon_0 m_e)^{1/2}$, and so

$$n^2(\omega) = 1 - (\omega_p/\omega)^2. \tag{4.81}$$

The plasma frequency serves as a critical value below which the index is complex and the penetrating wave drops off exponentially (4.77) from the boundary; at frequencies above ω_p, n is real, absorption is small, and the conductor is transparent. In the latter circumstance n is less than 1, as it was for dielectrics at very high frequencies. Hence we can expect metals in general to be fairly transparent to x-rays. Table 4.2 lists the plasma frequencies for some of the alkali metals that are transparent even to ultraviolet.

The index of refraction for a metal will usually be complex, and the impinging wave will suffer absorption in an amount that is frequency dependent. For example, the outer visors on the Apollo space suits were overlaid with a very thin film of gold (Fig. 4.40). The coating reflected about 70% of the incident light and was used under bright conditions, such as low and forward sun angles. It was designed to decrease the thermal load on the cooling system by strongly reflecting radiant energy in the infrared while still transmitting adequately in the visible. Inexpensive metal-coated sunglasses which are quite similar in principle are also available commercially and they're well worth having just to experiment with.

The ionized upper atmosphere of the Earth contains a distribution of free electrons that behave very much like those confined within a metal. The index of refraction of such a medium will be real and less than 1 for frequencies above ω_p. In July of 1965 the *Mariner IV* spacecraft made use of this effect to examine the ionosphere of the planet Mars, 216 million kilometers from Earth.*

If we wish to communicate between two distant terrestrial points, we might bounce low-frequency waves off the Earth's ionosphere. To speak to someone on the Moon, however, we should use high-frequency signals, to which the ionosphere would be transparent.

* R. Von Eshelman, *Sci. Am.* **220**, 78 (1969).

Table 4.2 Critical wavelengths and frequencies for some alkali metals.

Metal	λ_p (observed) nm	λ_p (calculated) nm	$\nu_p = c/\lambda_p$ (observed) Hz
Lithium (Li)	155	155	1.94×10^{15}
Sodium (Na)	210	209	1.43×10^{15}
Potassium (K)	315	287	0.95×10^{15}
Rubidium (Rb)	340	322	0.88×10^{15}

iii) Reflection From a Metal

Imagine that a plane wave initially in air impinges on a conducting surface. The transmitted wave advancing at some angle to the normal will be inhomogeneous. But if the conductivity of the medium is increased, the

Figure 4.40 Edwin Aldrin Jr. at Tranquility Base on the Moon. The photographer, Neil Armstrong, is reflected in the gold-coated visor. (Photo courtesy NASA.)

wavefronts will become aligned with the surfaces of constant amplitude, whereupon $\mathbf{k}_t$ and $\hat{\mathbf{u}}_n$ will approach parallelism. In other words, in a good conductor the transmitted wave propagates in a direction normal to the interface regardless of θ_i.

Let's now compute the reflectance, $R = I_r/I_i$, for the simplest case of normal incidence on a metal. Taking $n_i = 1$ and $n_t = n_c$ (i.e., the complex index), we have from Eq. (4.47) that

$$R = \left(\frac{n_c - 1}{n_c + 1}\right)\left(\frac{n_c - 1}{n_c + 1}\right)^*, \tag{4.82}$$

and therefore, since $n_c = n_R - in_I$,

$$R = \frac{(n_R - 1)^2 + n_I^2}{(n_R + 1)^2 + n_I^2}. \tag{4.83}$$

If the conductivity of the material goes to zero, we have the case of a dielectric, whereupon in principle the index is real ($n_I = 0$), and the attenuation coefficient, α, is zero. Under those circumstances, the index of the transmitting medium n_t is n_R, and the reflectance (4.83) becomes identical with that of Eq. (4.67). If instead n_I is large while n_R is comparatively small, R in turn becomes large (Problem 4.49). In the unattainable limit where n_c is purely imaginary, 100% of the incident flux density would be reflected ($R = 1$). Notice that it is possible for the reflectance of one metal to be greater than that of another even though its n_I is smaller. For example, at $\lambda_0 = 589.3$ nm the parameters associated with solid sodium are roughly $n_R = 0.04$, $n_I = 2.4$, and $R = 0.9$; and those for bulk tin are $n_R = 1.5$, $n_I = 5.3$, and $R = 0.8$; whereas for a gallium single crystal $n_R = 3.7$, $n_I = 5.4$, and $R = 0.7$.

The curves of $R_\parallel$ and $R_\perp$ for oblique incidence shown in Fig. 4.41 are somewhat typical of absorbing media. Thus, although R at $\theta_i = 0$ is about 0.5 for gold, as opposed to nearly 0.9 for silver in white light, the two metals have reflectances that are quite similar in shape, approaching 1.0 at $\theta_i = 90°$. Just as with dielectrics (Fig. 4.29), $R_\parallel$ drops to a minimum at what is now called the *principle angle of incidence*, but here that minimum is nonzero. Figure 4.42 illustrates the spectral reflectance at normal incidence for a number of evaporated metal films under ideal conditions. Observe that although gold transmits fairly well in and below the green region of the spectrum, silver, which is highly reflective across the visible, becomes transparent in the ultraviolet at about 316 nm.

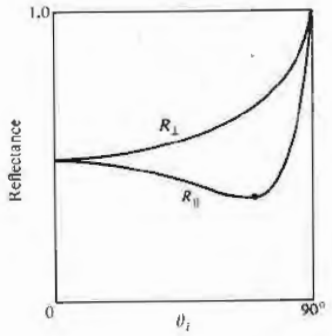


Figure 4.41 Typical reflectance for a linearly polarized beam of white light incident on an absorbing medium.

Phase shifts arising from reflection off a metal occur in both components of the field (i.e., parallel and perpendicular to the plane of incidence). These are generally neither 0 nor π, with a notable exception at $\theta_i = 90°$, where, just as with a dielectric, both components shift phase by 180° on reflection.

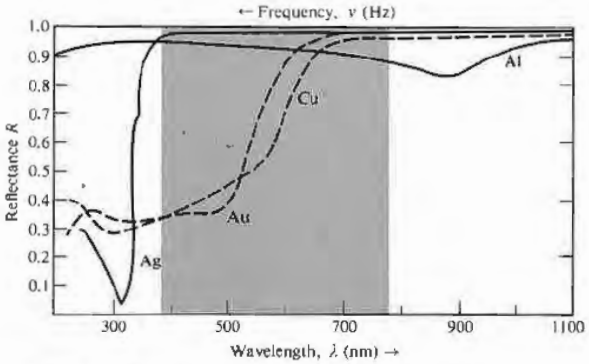


Figure 4.42 Reflectance versns wavelength for silver, gold, copper, and aluminum.

4.4 FAMILIAR ASPECTS OF THE INTERACTION OF LIGHT AND MATTER

Let's now examine some of the phenomena that paint the everyday world in a marvel of myriad colors.

As we saw earlier (p. 72), light that contains a roughly equal amount of every frequency in the visible region of the spectrum is perceived as white. Thus a broad source of white light (whether natural or artificial) is one for which every point on its surface can be imagined as sending out, more or less in all directions, a stream of light of every visible frequency. Similarly, a reflecting surface that accomplishes essentially the same thing will also appear white: a highly reflecting, frequency-independent, *diffusely* scattering object will be perceived as white under white light.

Although water is essentially transparent, water vapor appears white, as does grouud glass. The reason is simple enough—if the grain size is small but much larger than the wavelengths involved, light will enter each transparent particle, be reflected and refracted several times, and emerge. There will be no distinction among any of the frequency components, so the reflected light reaching the observer will be white. This is the mechanism accountable for the whiteness of things like sugar, salt, paper, clouds, talcum powder, snow, and paint, each grain of which is actually transparent. Similarly, a wadded-up piece of crumpled clear plastic wrap will appear whitish, as will an ordinarily transparent material filled with small air bubbles (e.g., beaten egg white). Even though we usually think of paper, talcum powder, and sugar as each consisting of some sort of opaque white substance, it's an easy matter to dispel that misconception. Cover a printed page with a few of these materials (a sheet of white paper, some grains of sugar, or talcum) and illuminate it from behind. You'll have little difficulty in seeing through them. In the case of white paint, one simply suspends colorless transparent particles, such as the oxides of zinc, titanium, or lead, in an equally transparent vehicle, for example, linseed oil or the newer acrylics. Obviously, if the particles and vehicle have the same index of refraction, there will not be any reflections at the grain boundaries. The particles will simply disappear into the conglomeration, which itself remains clear. In contrast, if the indices are markedly different, there will be a good deal of reflection at all wavelengths (Problem 4.42), and the paint will appear white and opaque [take another look at Eq. (4.67)]. To color paint one need only dye the particles so that they absorb all frequencies except the desired range.

Carrying the logic in the reverse direction, if we reduce the relative index, n_{ti}, at the grain or fiber boundaries, the particles of material will reflect less, thereby decreasing the overall whiteness of the object. Consequently, a wet white tissue will have a grayish, more transparent look. Wet talcum powder loses its sparkling whiteness, becoming a dull gray, as does wet white cloth. In the same way, a piece of dyed fabric soaked in a clear liquid (e.g., water, gin, or benzene) will lose its whitish haze and become much darker, the colors then being deep and rich like those of a still-wet water-color painting.

A diffusely reflecting surface that absorbs somewhat—uniformly right across the spectrum—will reflect a bit less than a white surface and so appear mat gray. The less it reflects, the darker the gray, until it absorbs almost all the light and appears black. A surface that reflects perhaps 70% or 80% or more, but does so specularly, will appear the familiar shiny gray of a typical metal. Metals possess tremendous numbers of free electrons (p. 111) that scatter light very effectively, independent of frequency: they are not bound to the atoms and have no associated resonances. Moreover, the amplitudes of the vibrations are an order of magnitude larger than they were for the bound electrons. The incident light cannot penetrate into the metal any more than a fraction of a wavelength or so before it's canceled completely. There is little or no refracted light; most of the energy is reflected out, and only the small remainder is absorbed. Note that the primary difference between a gray surface and a mirrored surface is one of diffuse versus specular reflection. An artist paints a picture of a polished "white" metal, such as silver or aluminum, by "reflecting" images of things in the room on top of a gray surface.

When the distribution of energy in a beam of light is not effectively uniform across the spectrum, the light appears colored. Figure 4.43 depicts typical frequency

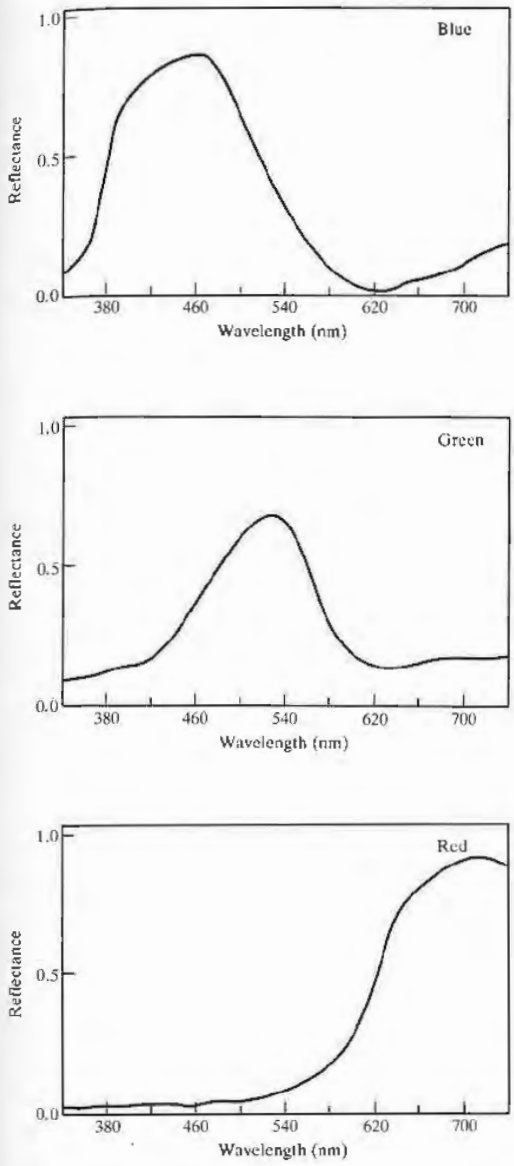


Figure 4.43 Reflection curves for blue, green, and red pigments. These are typical, but there is a great deal of possible variation among the colors.

distributions for what would be perceived as red, green, and blue light. These curves show the predominant frequency regions, but there can be a great deal of variation in the distributions, and they will still provoke the responses of red, green, and blue. In the early 1800s Thomas Young showed that a broad range of colors could be generated by mixing three beams of light, provided their frequencies were widely separated. When three such beams combine to produce white light they are called **primary colors**. There is no single unique set of these primaries, nor do they have to be quasimonochromatic. Since a wide range of colors can be created by mixing red (R), green (G), and blue (B), these tend to be used most frequently. They are the three components (emitted by three phosphors) that generate the whole gamut of hues seen on a color television set.

Figure 4.44 summarizes the results when beams of these three primaries are overlapped in a number of different combinations: Red plus blue is seen as *magenta* (M), a reddish purple; blue plus green is seen as *cyan* (C), a bluish green or turquoise; and perhaps most surprising, red plus green is seen as *yellow* (Y). The sum of all three primaries is white:

$$R + B + G = W,$$

$$M + G = W, \text{ since } R + B = M,$$

$$C + R = W, \text{ since } B + G = C,$$

$$Y + B = W, \text{ since } R + G = Y.$$

Any two colors that together produce white are said to be **complementary**, and the last three symbolic state-

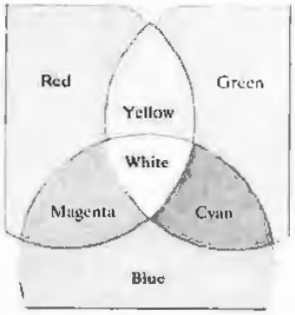


Figure 4.44 Three overlapping beams of colored light. A color television set uses these same three primary light sources—red, green, and blue.

ments exemplify that situation. Now suppose we overlap a beam of magenta and a beam of yellow:

$$M + Y = (R + B) + (R + G) = W + R;$$

the result is a combination of red and white, or pink. That raises another point: we say a color **is saturated,** that it is deep and intense, when it does not contain any white light. As Fig. 4.45 shows, pink is unsaturated red—red superimposed on a background of white.

The mechanism responsible for the yellowish red hue of gold and copper is, in some respects, similar to the process that causes the sky to appear blue. Putting it rather succinctly (see Section 8.5 for a further discussion of scattering in the atmosphere), the molecules of air have resonances in the ultraviolet and will therefore be driven into larger-amplitude oscillations as the frequency of the incident light increases toward the ultraviolet. Consequently, they will effectively take energy from and reemit (i.e., scatter) the blue component of sunlight in all directions, transmitting the complementary red end of the spectrum with little alteration. This is analogous to the selective reflection or scattering of yellow-red light that takes place at the surface of a gold film and the concomitant transmission of blue-green light. In contradistinction, the characteristic colors of most substances have their origin in the phenomenon of *selective* or *preferential absorption.* For example, water has a very light green-blue tint because of its absorption of red light. That is, the H_2O molecules have a broad resonance in the infrared, which extends somewhat into the visible. The absorption isn't very strong, so there is no accentuated reflection of red light at the surface. Instead it is transmitted and gradually absorbed out until at a depth of about 30 m of sea water, red is almost completely removed from sunlight. This same process of selective absorption is responsible for the colors of brown eyes and butterflies, of birds and bees and cabbages and kings. Indeed the great majority of objects in nature appear to have characteristic colors as the result of preferential absorption by pigment molecules. In contrast with most atoms and molecules, which have resonances in the ultraviolet and infrared, the pigment molecules must obviously have resonances in the visible. Yet visible photons have energies of roughly 1.6 eV to 3.2 eV, which, as you

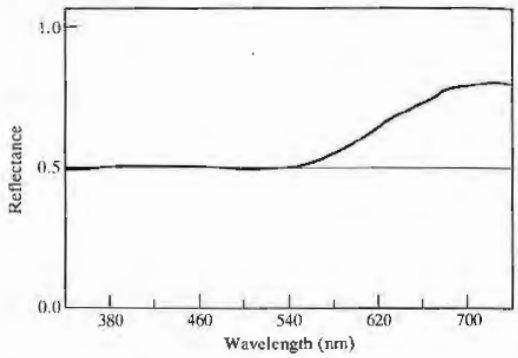


Figure 4.45 Spectral reflection of a pink pigment.

might expect, are on the low side for ordinary electron excitation and on the high side for excitation via molecular vibration. Despite this, there are atoms where the bound electrons form incomplete shells (gold, for example) and variations in the configuration of these shells provide a mode for low-energy excitation. In addition, there is the large group of organic dye molecules, which evidently also have resonances in the visible. All such substances, whether natural or synthetic, consist of long-chain molecules made up of regularly alternating single and double bonds in what is called a conjugated system. This structure is typified by the carotene molecule $C_{40}H_{56}$ (Fig. 4.46). The carotenoids range in color from yellow to red and are found in carrots, tomatoes, daffodils, dandelions, autumn leaves, and people. The chlorophylls are another group of familiar natural pigments, but here a portion of the long chain is turned around on itself to form a ring. In any event, conjugated systems of this sort contain a number of particularly mobile electrons known as *pi electrons.* They are not bound to specific atomic sites but instead can range over the relatively large dimensions of the molecular chain or ring. In the phraseology of quantum mechanics, we would say that these are long-wavelength, low-frequency, and therefore low-energy, electron states. The energy required to raise a pi electron to an excited state is accordingly comparatively low, corresponding to that of visible photons. In effect, we can imagine the molecule as an

oscillator having a resonance frequency in the visible.

The energy levels of an individual atom are precisely defined, that is, the resonances are very sharp. With solids and liquids, however, the proximity of the atoms results in a broadening of the energy levels into wide bands. In other words, the resonances spread over a broad range of frequencies. Consequently, we can expect that a dye will not absorb just a narrow portion of the spectrum; indeed if it did, it would reflect most frequencies and appear nearly white.

Imagine a piece of stained glass with a resonance in the blue where it strongly absorbs. If you look through it at a white-light source composed of red, green, and blue, the glass will absorb blue, passing red and green, which is yellow (Fig. 4.47). The glass looks yellow: yellow cloth, paper, dye, paint, and ink all selectively absorb blue. If you peer at something that is a pure blue through a yellow filter, one that passes yellow and absorbs blue, the object will appear black. Here the filter colors the light yellow by removing blue, and we speak

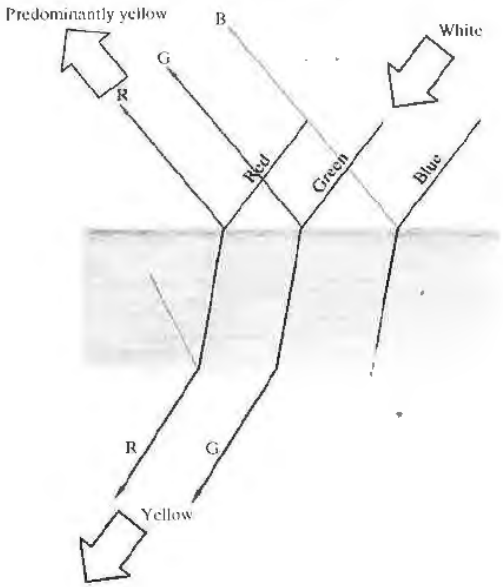


Figure 4.47 Yellow stained glass.

Figure 4.46 The carotene molecule.

of the process as *subtractive* coloration, as opposed to *additive* coloration, which results from overlapping beams of light.

In the same way, fibers of a sample of white cloth or paper are essentially transparent, but when dyed each fiber behaves as if it were a chip of colored glass. The incident light penetrates the paper, emerging for the most part as a reflected beam only after undergoing numerous reflections and refractions within the dyed fibers. The exiting light will be colored to the extent that it lacks the frequency component absorbed by the dye. This is precisely why a leaf appears green, or a banana yellow.

A bottle of ordinary blue ink looks blue in either reflected or transmitted light. But if the ink is painted on a glass slide and the solvent evaporates, something rather interesting happens. The concentrated pigment absorbs so effectively that it preferentially reflects at the resonant frequency, and we are back to the idea that a strong absorber (large n_I) is a strong reflector. Thus,

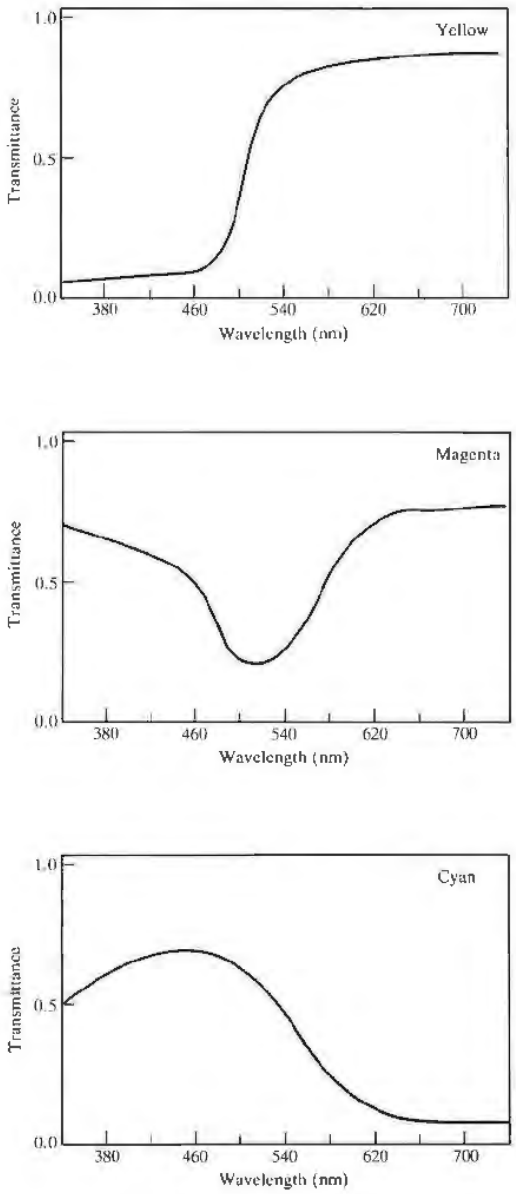


Figure 4.48 Transmission curves for colored filters.

concentrated blue-green ink reflects red, whereas red-blue ink reflects green. Try it with a felt marking pen, but you must use reflected light, being careful not to inundate the sample with unwanted light from below. (Wipe the ink to obtain a thin layer and then place the slide on a piece of black paper.)

The whole range of colors (including red, green, and blue) can be produced by passing light through various combinations of magenta, cyan, and yellow filters (Fig. 4.48). These are the primary colors of subtractive mixing, the primaries of the paint box, although they are often mistakenly spoken of as red, blue, and yellow. They are the basic colors of the dyes used to make photographs and the inks used to print them. Ideally, if you mix all the subtractive primaries together (either by combining paints or by stacking filters), you get no color, no light—black. Each removes a region of the spectrum, and together they absorb it all.

If the range of frequencies being absorbed spreads across the visible, the object will appear black. That is not to say that there is no reflection at all—you obviously can see a reflected image in a piece of black patent leather, and a rough black surface reflects also, only diffusely. If you still have those red and blue inks, mix them, add some green, and you'll get black.

In addition to the above processes specifically related to reflection, refraction, and absorption, there are many other color-generating mechanisms, which we shall explore later on. For example, the scarabaeid beetles mantle themselves in the brilliant colors produced by diffraction gratings on their wing cases, and wavelength-dependent interference effects contribute to the color patterns seen on oil slicks, mother-of-pearl, soap bubbles, peacocks, and hummingbirds.

4.5 THE STOKES TREATMENT OF REFLECTION AND REFRACTION

A rather elegant and novel way of looking at reflection and transmission at a boundary was developed by the British physicist Sir George Gabriel Stokes (1819–1903). Since we will often make use of his results in future chapters, let's now examine that derivation. Suppose that we have an incident wave of amplitude E_{0i} imping-

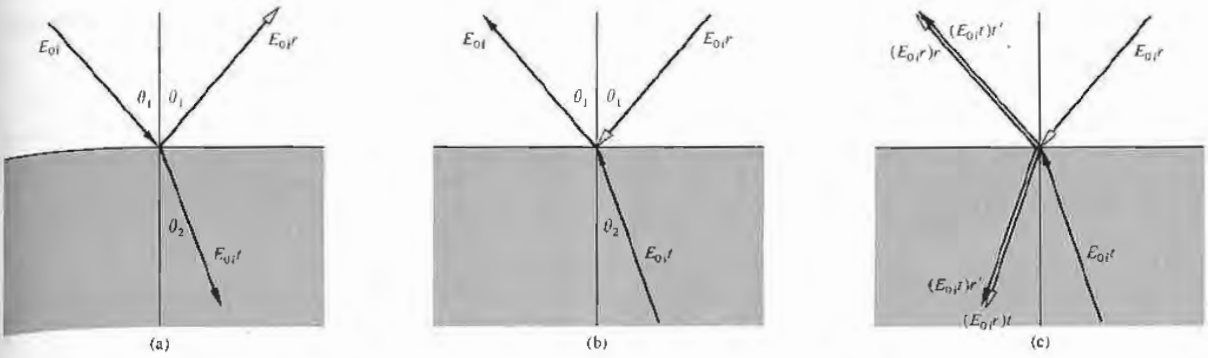


Figure 4.49 Reflection and refraction via the Stokes treatment.

ing on the planar interface separating two dielectric media, as in Fig. 4.49(a). As we saw earlier in this chapter, since r and t are the fractional amplitudes reflected and transmitted, respectively (where $n_i = n_1$ and $n_t = n_2$), then $E_{0r} = rE_{0i}$ and $E_{0t} = tE_{0i}$. Again we are reminded of the fact that Fermat's principle led to the principle of reversibility, which implies that the situation depicted in Fig. 4.49(b), where all the ray directions are reversed, must also be physically possible. With the one proviso that there be no energy dissipation (no absorption), a wave's meanderings must be reversible. Equivalently, in the idiom of modern physics one speaks of *time-reversal invariance*, that is, if a process occurs, the reverse process can also occur. Thus if we take a hypothetical motion picture of the wave incident on, reflecting from, and transmitting through the interface, the behavior depicted when the film is run backward must also be physically realizable. Accordingly, examine Fig. 4.49(c), where there are now two incident waves of amplitudes $E_{0i}r$ and $E_{0i}t$. A portion of the wave whose amplitude is $E_{0i}t$ is both reflected and transmitted at the interface. Without making any assumptions, let r' and t' be the amplitude reflection and transmission coefficients, respectively, for a wave incident from below (i.e., $n_i = n_2$, $n_t = n_1$). Consequently, the reflected portion is $E_{0i}tr'$, and the transmitted portion is $E_{0i}tt'$. Similarly, the incoming wave whose amplitude is $E_{0i}r$ splits into segments of amplitude $E_{0i}rr$ and $E_{0i}rt$. If the configuration in Fig. 4.49(c) is to be identical with that in Fig. 4.49(b), then obviously

$$E_{0i}tt' + E_{0i}rr = E_{0i} \tag{4.84}$$

and

$$E_{0i}rt + E_{0i}tr' = 0. \tag{4.85}$$

Hence

$$tt' = 1 - r^2 \tag{4.86}$$

and

$$r' = -r, \tag{4.87}$$

the latter two equations being known as the Stokes relations. Actually this discussion calls for a bit more caution than is usually granted it. It must be pointed out that *the amplitude coefficients are functions of the incident angles*, and therefore the Stokes relations might better be written as

$$t(\theta_1)t'(\theta_2) = 1 - r^2(\theta_1) \tag{4.88}$$

and

$$r'(\theta_2) = -r(\theta_1), \tag{4.89}$$

where $n_1 \sin \theta_1 = n_2 \sin \theta_2$. The second equation indicates, by virtue of the minus sign, that *there is a* 180° *phase difference between the waves internally and externally reflected.* It is most important to keep in mind that here θ_1 and θ_2 are pairs of angles that are related by way of Snell's law. Note as well that we never did say whether n_1 was greater or less than n_2, so Eqs. (4.88) and (4.89)

apply in either case. Let's return for a moment to one of the Fresnel equations:

$$r_\perp = -\frac{\sin(\theta_i - \theta_t)}{\sin(\theta_i + \theta_t)}. \qquad [4.42]$$

If a ray enters from above, as in Fig. 4.49(a), and we assume $n_2 > n_1$, $r_\perp$ is computed by setting $\theta_i = \theta_1$ and $\theta_t = \theta_2$ (external reflection), the latter being derived from Snell's law. If, on the other hand, the wave is incident at that same angle from below (in this instance internal reflection), $\theta_i = \theta_1$ and we again substitute in Eq. (4.42), but here θ_t is not θ_2, as before. The values of $r_\perp$ for internal and external reflection *at the same incident angle* are obviously different. Now suppose, in this case of internal reflection, that $\theta_i = \theta_2$. Then $\theta_t = \theta_1$, the ray directions are the reverse of those in the first situation, and Eq. (4.42) yields

$$r'_\perp(\theta_2) = -\frac{\sin(\theta_2 - \theta_1)}{\sin(\theta_2 + \theta_1)}.$$

Although it may be unnecessary we once again point out that this is just the negative of what was determined for $\theta_i = \theta_1$ and external reflection, that is,

$$r'_\perp(\theta_2) = -r_\perp(\theta_1). \qquad (4.90)$$

The use of primed and unprimed symbols to denote the amplitude coefficients should serve as a reminder that we are once more dealing with angles related by Snell's law. In the same way, interchanging θ_i and θ_t in Eq. (4.43) leads to

$$r'_\parallel(\theta_2) = -r_\parallel(\theta_1). \qquad (4.91)$$

The 180° phase difference between each pair of components is evident in Fig. 4.25, but do keep in mind that when $\theta_i = \theta_p$, $\theta_t = \theta'_p$ and vice versa (Problem 4.46). Beyond $\theta_i = \theta_c$ there is no transmitted wave, Eq. (4.89) is not applicable, and as we have seen, the phase difference is no longer 180°.

It is common to conclude that both the parallel and perpendicular components of the externally reflected beam change phase by π radians while the internally reflected beam undergoes no phase shift at all. By now, within the particular convention we've established, this should be recognized as incorrect, or at least almost obviously [compare Figs. 4.26(a) and 4.27(a)].

4.6 PHOTONS AND THE LAWS OF REFLECTION AND REFRACTION

Suppose that light consists of a stream of photons and that one such photon strikes the interface between two dielectric media at an angle θ_i and is subsequently transmitted across it at an angle θ_t. We know that if this were just one of billions of such quanta in a narrow laserbeam, it would obediently conform to Snell's law. To appreciate this behavior let's examine the dynamics associated with the odyssey of our single photon. Recall that

$$\mathbf{p} = \hbar\mathbf{k}, \qquad [3.53]$$

and consequently the incident and transmitted momenta are $\mathbf{p}_i = \hbar\mathbf{k}_i$ and $\mathbf{p}_t = \hbar\mathbf{k}_t$, respectively. We assume (without much justification) that although the material in the vicinity of the interface affects the y-component of momentum, it leaves the x-component unchanged. Indeed we know experimentally that linear momentum can be transferred to a medium from a light beam (see Section 3.3.2). The statement of conservation of the component of momentum parallel to the interface takes the form

$$p_{ix} = p_{tx} \qquad (4.92)$$

or

$$p_i \sin\theta_i = p_t \sin\theta_t.$$

If we use Eq. (3.53), this becomes

$$k_i \sin\theta_i = k_t \sin\theta_t$$

and hence

$$\frac{1}{\lambda_i}\sin\theta_i = \frac{1}{\lambda_t}\sin\theta_t.$$

Multiplying both sides by c/ν, we have

$$n_i \sin\theta_i = n_t \sin\theta_t,$$

which of course is Snell's law. In exactly the same way, if the photon reflects off the interface instead of being transmitted, Eq. (4.92) leads to

$$k_i \sin\theta_i = k_r \sin\theta_r,$$

and since $\lambda_i = \lambda_r$, $\theta_i = \theta_r$. It is interesting to note that

$$n_{ii} = \frac{p_t}{p_i}, \tag{4.99}$$

and so if $n_{ii} > 1$, $p_t > p_i$. Experiments dating back as far as 1850, to those of Foucault, have shown that when $n_{ii} > 1$ the speed of propagation is actually reduced in the transmitting media, even though the momentum apparently increases.†

Do keep in mind that we have been dealing with a very simple representation that leaves much to be desired. For example, it says nothing about the atomic structure of the media or about the probability that a photon will traverse a given path. Even though this treatment is obviously simplistic, it is appealing pedagogically (see Chapter 13).

† This suggests an increase in the photon's effective mass. See F. R. Tangherlini, "On Snell's Law and the Gravitational Deflection of Light." *Am. J. Phys.* **36**, 1001 (1968). Take a *cautious* look at R. A. Houstoun, "Nature of Light." *J. Opt. Soc. Am.* **55**, 1186 (1965).

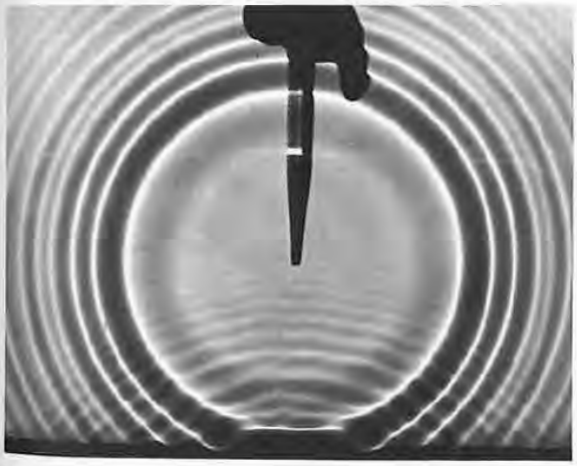


Figure 4.50 (Photos courtesy *Physics*, Boston, D. C. Heath & Co., 1960.)

PROBLEMS

4.1 Calculate the transmission angle for a ray incident in air at 30° on a block of crown glass ($n_g = 1.52$).

4.2* A ray of yellow light from a sodium discharge lamp falls on the surface of a diamond in air at 45°. If at that frequency $n_d = 2.42$, compute the angular deviation suffered upon transmission.

4.3 Use Huygens's construction to create a wavefront diagram showing the form a spherical wave will have after reflection from a planar surface, as in the ripple tank photos of Fig. 4.50. Draw the ray diagram as well.

4.4* Given an interface between water ($n_w = 1.33$) and glass ($n_g = 1.50$), compute the transmission angle for a beam incident in the water at 45°. If the transmitted beam is reversed so that it impinges on the interface, show that $\theta_t = 45°$.

4.5 A beam of 12-cm planar microwaves strikes the surface of a dielectric at 45°. If $n_{ti} = \frac{4}{3}$, compute (a) the wavelength in the transmitting medium, and (b) the angle θ_t.

4.6* Light of wavelength 600 nm in vacuum enters a block of glass where $n_g = 1.5$. Compute its wavelength in the glass. What color would it appear to someone imbedded in the glass (see Table 3.2)?

4.7 Figure 4.51 shows a bundle of rays entering and emerging from a glass disk (a lens). From the configuration of the rays, determine the shape of the wavefronts at various points. Draw a diagram in profile.

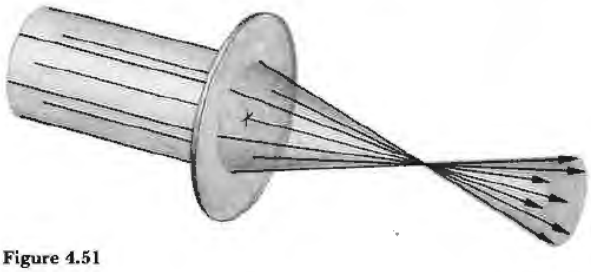


Figure 4.51

4.8 Make a plot of θ_i versus θ_t for an air–glass boundary where $n_{ga} = 1.5$.

4.9 In Fig. 4.52 the wavefronts in the incident medium match the fronts in the transmitting medium everywhere on the interface—a concept known as *wavefront continuity*. Write expressions for the number of waves per unit length along the interface in terms of θ_i and λ_i in one case and θ_t and λ_t in the other. Use these to derive Snell's law. Do you think Snell's law applies to sound waves? Explain.

4.10* With the previous problem in mind, return to Eq. (4.19) and take the origin of the coordinate system in the plane of incidence and on the interface (Fig. 4.20). Show that that equation is then equivalent to equating the x-components of the various propagation vectors. Show that it is also equivalent to the notion of wavefront continuity.

4.11* Figure 4.53 depicts a wavefront at $\overline{AB}$ that subsequently sweeps across the interface, driving atoms along it, which in turn radiate transmitted wavelets. Since the refracted wave travels at a speed v_t, assume the transmitted wavelets also propagate at v_t. These wavelets then overlap and interfere (which is essentially the Huygens–Fresnel principle) to form the refracted wave. Show that the transmitted wavelets will arrive in phase along $\overline{DC}$, provided Snell's law obtains.

4.12 Making use of the ideas of equal transit times between corresponding points and the orthogonality of

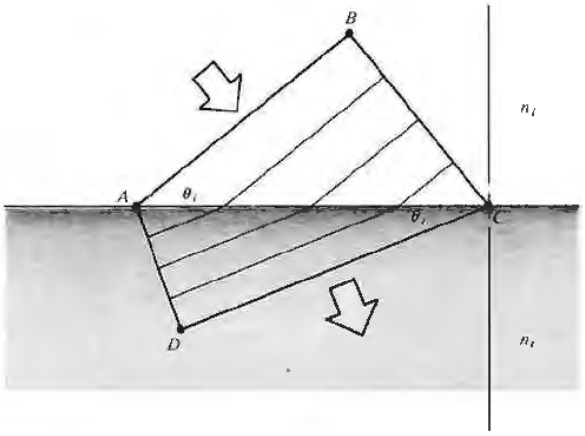


Figure 4.52

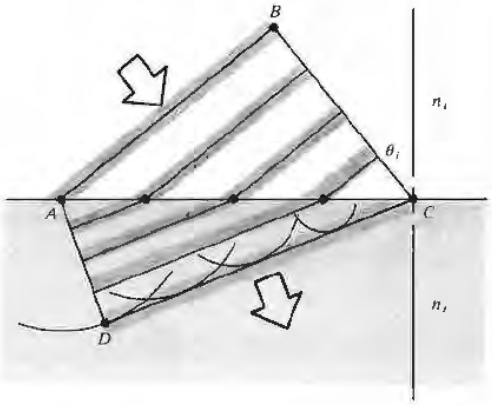


Figure 4.53

rays and wavefronts, derive the law of reflection and Snell's law. The ray diagram of Fig. 4.54 should be helpful.

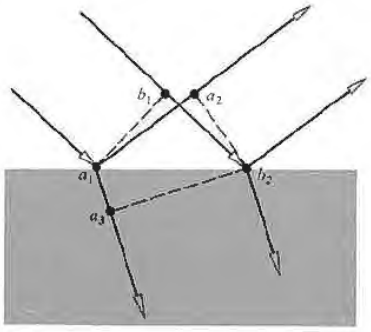


Figure 4.54

4.13 Starting with Snell's law, prove that the vector refraction equation has the form

$$n_t\hat{\mathbf{k}}_t - n_i\hat{\mathbf{k}}_i = (n_t \cos\theta_t - n_i \cos\theta_i)\hat{\mathbf{u}}_n. \qquad [4.8]$$

4.14 Derive a vector expression equivalent to the law of reflection. As before, let the normal go from the incident to the transmitting medium, even though it obviously doesn't really matter.

4.15 In the case of reflection from a planar surface, use Fermat's principle to prove that the incident and reflected rays share a common plane with the normal $\hat{\mathbf{u}}_n$, namely, the plane of incidence.

4.16* Derive the law of reflection, $\theta_i = \theta_r$, by using the calculus to minimize the transit time, as required by Fermat's principle.

4.17* According to the mathematician Hermann Schwarz, there is one triangle that can be inscribed within an acute triangle such that it has a minimal perimeter. Using two planar mirrors, a laserbeam, and Fermat's principle, explain how you can show that this inscribed triangle has its vertices at the points where the altitudes of the acute triangle intersect its corresponding sides.

4.18 Show analytically that a beam entering a planar transparent plate, as in Fig. 4.55, emerges parallel to its initial direction. Derive an expression for the lateral displacement of the beam. Incidentally, the incoming and outgoing rays would be parallel even for a stack of plates of different material.

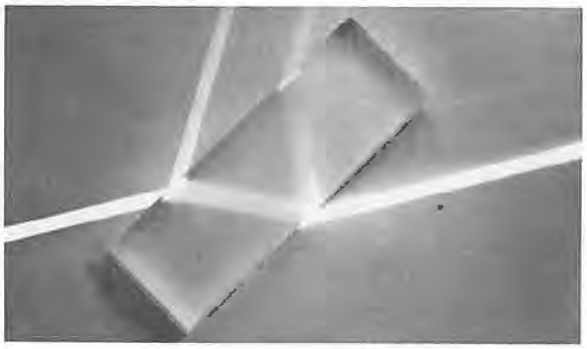


Figure 4.55 (Source unknown.)

4.19* Show that the two rays that enter the system in Fig. 4.56 parallel to each other emerge from it being parallel.

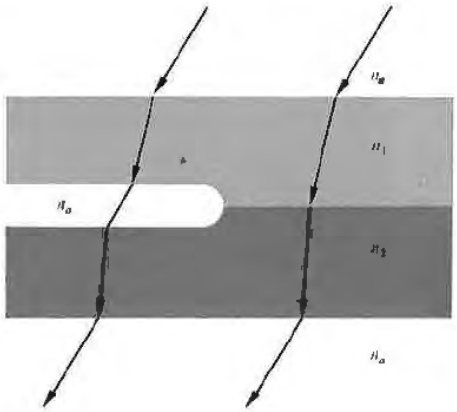


Figure 4.56

4.20 Discuss the results of Problem 4.18 in the light of Fermat's principle, that is, how does the relative index n_{21} affect things? To see the lateral displacement, look

at a broad source through a thick piece of glass ($\approx\frac{1}{4}$ inch) or a stack (four will do) of microscope slides *held at an angle.* There will be an obvious shift between the region of the source seen directly and the region viewed through the glass.

4.21 Suppose a lightwave that is linearly polarized in the plane of incidence impinges at 30° on a crown-glass ($n_g = 1.52$) plate in air. Compute the appropriate amplitude reflection and transmission coefficients at the interface. Compare your results with Fig. 4.22.

4.22 Show that even in the nonstatic case the tangential component of the electric field intensity **E** is continuous across an interface. [*Hint:* using Fig. 4.57 and Eq. (3.5), shrink sides *FB* and *CD*, thereby letting the area bounded go to zero.]

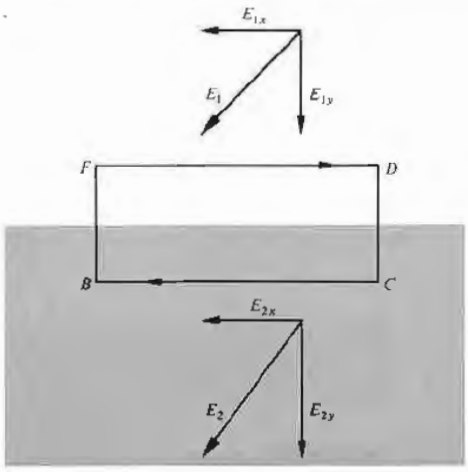


Figure 4.57

4.23 Derive Eqs. (4.42) through (4.45) for $r_\perp$, $r_\parallel$, $t_\perp$, and $t_\parallel$.

4.24 Prove that

$$t_\perp + (-r_\perp) = 1 \qquad [4.49]$$

for all θ_i, first from the boundary conditions and then from the Fresnel equations.

4.25* Verify that

$$t_\perp + (-r_\perp) = 1 \qquad [4.49]$$

for $\theta_i = 30°$ at a crown glass and air interface ($n_{ti} \approx 1.52$).

4.26* Calculate the critical angle beyond which there is total internal reflection at an air–glass ($n_g = 1.5$) interface. Compare this result with that of Problem 4.8.

4.27 Derive an expression for the speed of the evanescent wave in the case of internal reflection. Write it in terms of c, n_i, and θ_i.

4.28 Light having a vacuum wavelength of 600 nm, traveling in a glass ($n_g = 1.50$) block, is incident at 45° on a glass–air interface. It is then totally internally reflected. Determine the distance into the air at which the amplitude of the evanescent wave has dropped to a value of $1/e$ of its maximum value at the interface.

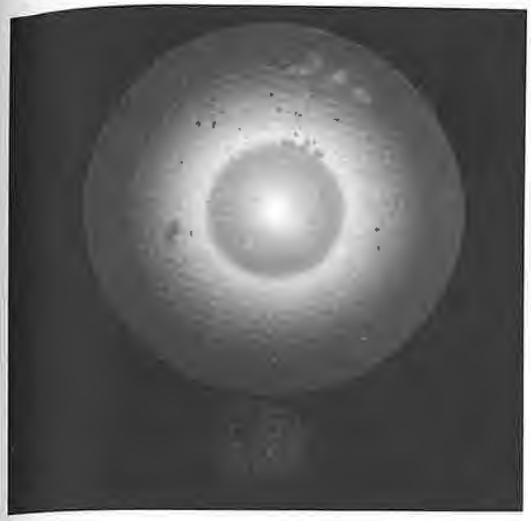
4.29 Figure 4.58 shows a laserbeam incident on a wet piece of filter paper atop a sheet of glass whose index of refraction is to be measured—the photograph shows the resulting light pattern. Explain what is happening and derive an expression for n_i in terms of R and d.

4.30 Consider the common mirage associated with an inhomogeneous distribution of air situated above a warm roadway. Envision the bending of the rays as if it were instead a problem in total internal reflection. If an observer, at whose head $n_a = 1.00029$, sees an apparent wet spot at $\theta_i \geq 88.7°$ down the road, find the index of the air immediately above the road.

4.31* Use the Fresnel equations to prove that light incident at $\theta_p = \frac{1}{2}\pi - \theta_t$ results in a reflected beam that is indeed polarized.

4.32 Show that $\tan\theta_p = n_t/n_i$ and calculate the polarization angle for external incidence on a plate of crown glass ($n_g = 1.52$) in air.

4.33* Beginning with Eq. (4.38), show that for



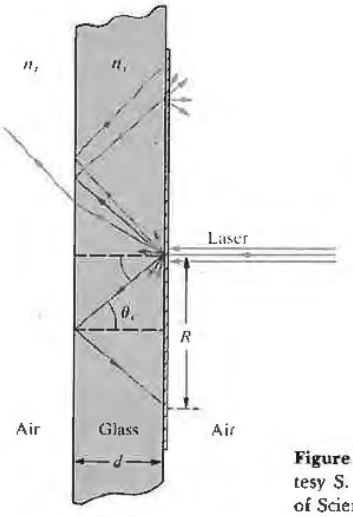


Figure 4.58 (Photo and diagram courtesy S. Reich, The Weizmann Institute of Science, Israel.)

two dielectric media, in general $\tan \theta_p = [\epsilon_t(\epsilon_t\mu_i - \epsilon_i\mu_t)/\epsilon_i(\epsilon_i\mu_i - \epsilon_t\mu_t)]^{1/2}$.

4.34 Show that the polarization angles for internal and external reflection at a given interface are complementary, that is, $\theta_p + \theta_p' = 90°$ (see Problem 4.32).

4.35 It is often useful to work with the *azimuthal angle* γ, which is defined as the angle between the plane of vibration and the plane of incidence. Thus for linearly polarized light,

$$\tan \gamma_i = [E_{0i}]_\perp/[E_{0i}]_\parallel \tag{4.94}$$

$$\tan \gamma_t = [E_{0t}]_\perp/[E_{0t}]_\parallel \tag{4.95}$$

and

$$\tan \gamma_r = [E_{0r}]_\perp/[E_{0r}]_\parallel. \tag{4.96}$$

Figure 4.59 is a plot of γ_r versus θ_i for internal and external reflection at an air–glass interface ($n_{ga} = 1.51$), where $\gamma_i = 45°$. Verify a few of the points on the curves and in addition show that

$$\tan \gamma_r = -\frac{\cos(\theta_i - \theta_t)}{\cos(\theta_i + \theta_t)} \tan \gamma_i. \tag{4.97}$$

4.36* Making use of the definitions of the azimuthal angles in Problem 4.35, show that

$$R = R_\parallel \cos^2 \gamma_i + R_\perp \sin^2 \gamma_i \tag{4.98}$$

and

$$T = T_\parallel \cos^2 \gamma_i + T_\perp \sin^2 \gamma_i. \tag{4.99}$$

4.37 Make a sketch of $R_\perp$ and $R_\parallel$ for $n_i = 1.5$ and $n_t = 1$ (i.e., internal reflection).

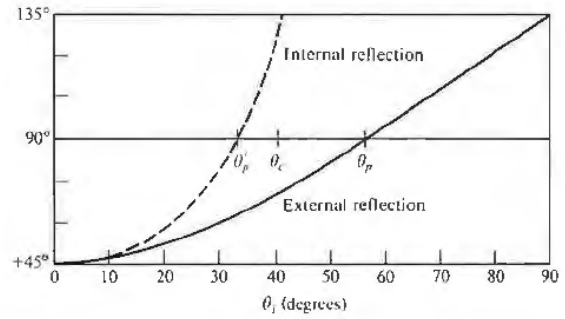


Figure 4.59

4.38 Show that

$$T_{\parallel} = \frac{\sin 2\theta_i \sin 2\theta_t}{\sin^2(\theta_i + \theta_t)\cos^2(\theta_i - \theta_t)} \tag{4.100}$$

and

$$T_{\perp} = \frac{\sin 2\theta_i \sin 2\theta_t}{\sin^2(\theta_i + \theta_t)}. \tag{4.101}$$

4.39* Using the results of Problem 4.38, that is, Eqs. (4.100) and (4.101), show that

$$R_{\parallel} + T_{\parallel} = 1 \tag{4.65}$$

and

$$R_{\perp} + T_{\perp} = 1. \tag{4.66}$$

4.40 Suppose that we look at a source perpendicularly through a stack of N microscope slides. The source seen through even a dozen slides will be noticeably darker. Assuming negligible absorption, show that the total transmittance of the stack is given by

$$T_i = (1 - R)^{2N}$$

and evaluate T_i for three slides in air.

4.41 Making use of the expression

$$I(y) = I_0 e^{-\alpha y} \tag{4.78}$$

for an absorbing medium, we define a quantity called the *unit transmittance* T_1. At normal incidence (4.55) $T = I_t/I_i$, and thus when $y = 1$, $T_1 \equiv I(1)/I_0$. If the total thickness of the slides in the previous problem is d and if they now have a transmittance per unit length T_1, show that

$$T_i = (1 - R)^{2N}(T_1)^d.$$

4.42 Show that at normal incidence on the boundary between two dielectrics, as $n_{ti} \to 1$, $R \to 0$, and $T \to 1$. Moreover, prove that as $n_{ti} \to 1$, $R_{\parallel} \to 0$, $R_{\perp} \to 0$, $T_{\parallel} \to 1$, and $T_{\perp} \to 1$ for all θ_i. Thus as the two media take on more similar indices of refraction, less and less energy is carried off in the reflected wave. It should be obvious that when $n_{ti} = 1$ there will be no interface and no reflection.

4.43* Derive the expressions for $r_{\perp}$ and $r_{\parallel}$ given by Eqs. (4.70) and (4.71).

4.44 Show that when $\theta_i > \theta_c$ at a dielectric interface, $r_{\parallel}$ and $r_{\perp}$ are complex and $r_{\perp} r_{\perp}^* = r_{\parallel} r_{\parallel}^* = 1$.

4.45 Figure 4.60 depicts a ray being multiply reflected by a transparent dielectric plate (the amplitudes of the resulting fragments are indicated). As in Section 4.5, we use the primed coefficient notation, because the angles are related by Snell's law.

a) Finish labeling the amplitudes of the last four rays.
b) Show, using the Fresnel equations, that

$$t_{\parallel} t'_{\parallel} = T_{\parallel} \tag{4.102}$$

$$t_{\perp} t'_{\perp} = T_{\perp} \tag{4.103}$$

$$r_{\parallel}^2 = r_{\parallel}'^2 = R_{\parallel} \tag{4.104}$$

and

$$r_{\perp}^2 = r_{\perp}'^2 = R_{\perp}. \tag{4.105}$$

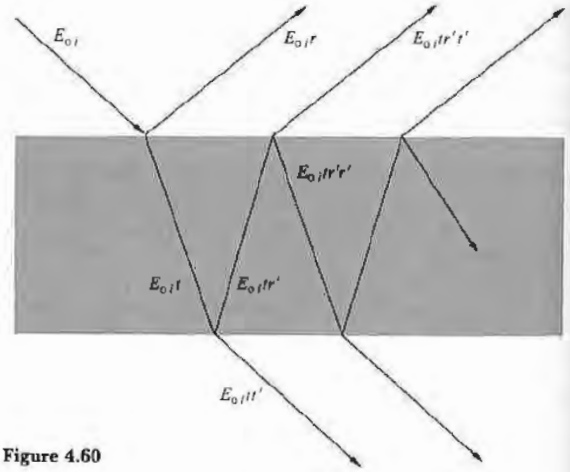


Figure 4.60

4.46* A wave, linearly polarized in the plane of incidence, impinges on the interface between two dielectric media. If $n_i > n_t$ and $\theta_i = \theta'_p$, there is no reflected wave, that is, $r'_{\parallel}(\theta'_p) = 0$. Using Stokes's tech-

nique, start from scratch to show that $t_{\parallel}(\theta_p)t'_{\parallel}(\theta'_p) = 1$, $r_{\parallel}(\theta_p) = 0$, and $\theta_i = \theta_p$ (Problem 4.34). How does this compare with Eq. (4.102)?

4.47 Making use of the Fresnel equations, show that $t_{\parallel}(\theta_p)t'_{\parallel}(\theta'_p) = 1$, as in the previous problem.

4.48 Figure 4.61 depicts a glass cube surrounded by four glass prisms in very close proximity to its sides. Sketch in the paths that will be taken by the two rays shown and discuss a possible application for the device.

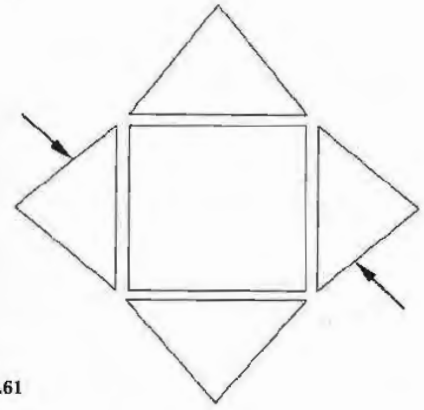


Figure 4.61

4.49 Figure 4.62 is a plot of n_I and n_R versus λ for a common metal. Identify the metal by comparing its characteristics with those considered in the chapter and discuss its optical properties.

4.50 Figure 4.63 shows a prism-coupler arrangement developed at the Bell Telephone Laboratories. Its function is to feed a laserbeam into a thin (0.00001-inch) transparent film, which then serves as a sort of waveguide. One application is that of thin-film laserbeam circuitry—a kind of integrated optics. How do you think it works?

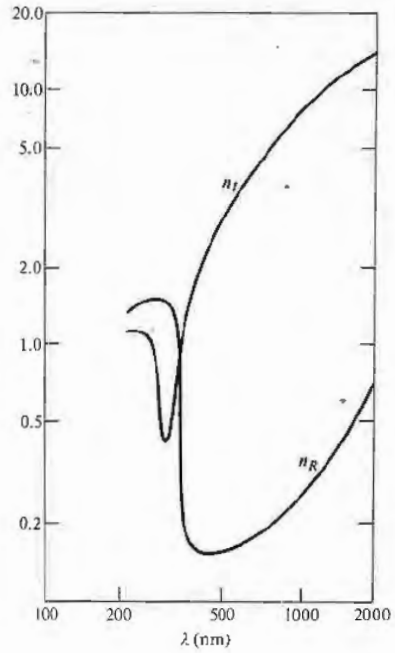


Figure 4.62

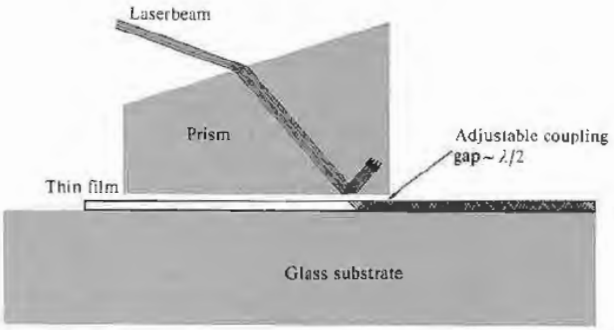


Figure 4.63

5 GEOMETRICAL OPTICS—PARAXIAL THEORY

5.1 INTRODUCTORY REMARKS

Suppose we have an object that is either self-luminous or externally illuminated, and imagine its surface as consisting of a large number of point sources. Each of these emits spherical waves, that is, rays emanate radially in the direction of energy flow or, if you like, in the direction of the Poynting vector (Fig. 4.1). In this case, the rays *diverge* from a given point source S, whereas if the spherical wave were collapsing to a point, the rays would of course be *converging*. Generally one deals only with a small portion of a wavefront. *A point from which a portion of a spherical wave diverges, or one toward which the wave segment converges, is known as a focal point of the bundle of rays.*

Now envision the situation in which we have a point source in the vicinity of some arrangement of reflecting and refracting surfaces representing an *optical system*. Of the infinity of rays emanating from S, generally speaking, only one will pass through an arbitrary point in space. Even so, it is possible to arrange for an infinite number of rays to arrive at a certain point P, as in Fig. 5.1. Thus, if for a cone of rays coming from S there is a corresponding cone of rays passing through P, the system is said to be *stigmatic* for these two points. The energy in the cone (apart from some inadvertent losses due to reflection, scattering, and absorption) reaches P, which is then referred to as a *perfect image* of S. The wave could conceivably arrive to form a finite patch of light, or *blur spot*, about P; it would still be an image of S but no longer a perfect one.

It follows from the principle of reversibility (see Section 4.2.4) that a point source placed at P would be equally well imaged at S, and accordingly the two are spoken of as *conjugate points*. In an *ideal optical system* every point of a three-dimensional region will be perfectly (or stigmatically) imaged in another region, the former being the *object space*, the latter the *image space*.

Most commonly, the function of an optical device is to collect and reshape a portion of the incident wavefront, often with the ultimate purpose of forming an image of an object. Notice that inherent in realizable systems is the limitation of being unable to collect all the emitted light; the system accepts only a segment of the wavefront. As a result, there will always be an

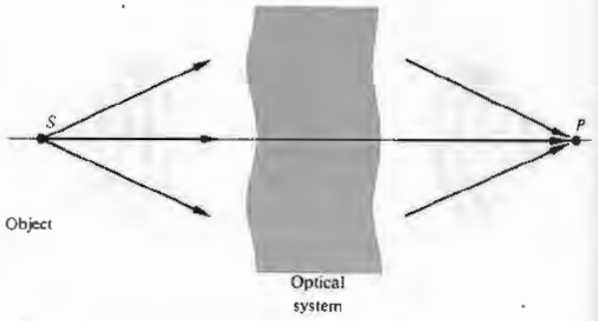


Figure 5.1 Converging and diverging waves.

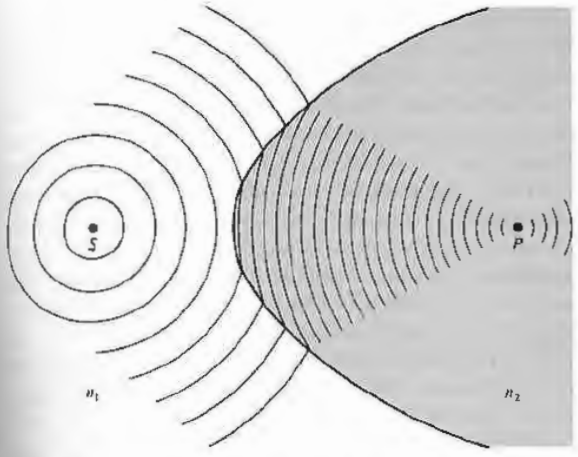


Figure 5.2 Reshaping a spherical wave at a refracting interface ($n_1 < n_2$).

apparent deviation from rectilinear propagation even in homogeneous media—the waves will be *diffracted.* The attainable degree of perfection in the imaging capability of a real optical system will therefore be **diffraction-limited** (there will always be a blur spot). As the wavelength of the radiant energy decreases in comparison to the physical dimensions of the optical system, the effects of diffraction become less significant. In the conceptual limit as $\lambda_0 \to 0$, rectilinear propagation obtains in homogeneous media, and we have the idealized domain of *geometrical optics.** Behavior that is specifically attributable to the wave nature of light (e.g., interference and diffraction) would no longer be observable. There are many situations in which the great simplicity arising from the approximation of geometrical optics more than compensates for its inaccuracies. In short, *the subject treats the controlled manipulation of wavefronts (or rays) by means of the interpositioning of reflecting and/or refracting bodies, neglecting any diffraction effects.*

Physical optics deals with situations in which the nonzero wavelength of light must be reckoned with. Analogously, when the de Broglie wavelength of a material object is negligible, we have *classical mechanics;* when it is not, we have the domain of *quantum mechanics* (see Chapter 13).

5.2 LENSES

No doubt the most widely used optical device is the lens, and that notwithstanding the fact that we see the world through a pair of them. Lenses date back to the burning glasses of antiquity, and indeed who can say when people first peered through the liquid lens formed by a droplet of water?

As an initial step toward an understanding of what lenses do and how they manage to do it, let's examine what happens when light impinges on the curved surface of a transparent dielectric medium.

5.2.1 Refraction at Aspherical Surfaces

Imagine that we have a point source S whose spherical waves arrive at a boundary between two transparent media, as shown in Fig. 5.2. We would like to determine the shape that the interface must have for the wave traveling within the second medium to converge at a point P, there forming a perfect image of S. Practical reasons for wanting to focus a diverging wave to a point will become evident as we proceed.

The time it takes for each and every portion of a wavefront leaving S to converge at P must be identical, if a perfect image is to be formed—that much was implied by Huygens in 1678. Or as we saw in Section

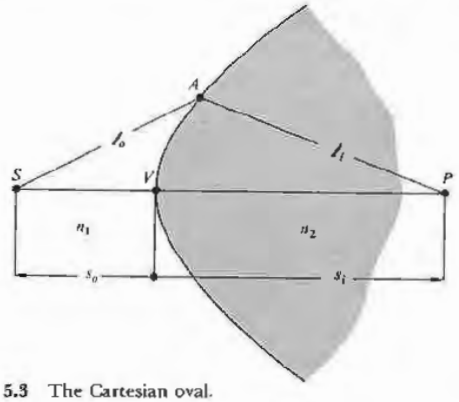


Figure 5.3 The Cartesian oval.

4.2.3, the distance between corresponding points on any and all rays will be traversed in that same time. Another way to say essentially the same thing from the perspective of Fermat's principle is that if a great many different rays are to go from S to P (i.e., if point A in Fig. 5.3 can be anywhere on the interface), each ray must traverse the same optical path length. Thus, for example, if S is in a medium of index n_1 and P is in an optically more dense medium of index n_2,

$$\ell_o n_1 + \ell_i n_2 = s_o n_1 + s_i n_2, \tag{5.1}$$

where s_o and s_i are the **object** and **image distances** measured from the *vertex* or *pole* V, respectively. Once we choose s_o and s_i, the right-hand side of this equation becomes fixed, and so

$$\ell_o n_1 + \ell_i n_2 = \text{constant}. \tag{5.2}$$

This is the equation of a *Cartesian oval* whose significance in optics was studied extensively by René Descartes in the early 1600s (Problem 5.1). Hence, when the boundary between two media has the shape of a Cartesian oval of revolution about the $\overline{SP}$, or **optical axis**, S and P will be conjugate points, that is, a point source at either location will be perfectly imaged at the other. What's actually occurring physically is rather easy to comprehend. Since $n_2 > n_1$, those regions of the wavefront traveling in the optically more dense medium move slower than those regions traversing the rarer material. Consequently, as the wave begins to pass through the vertex of the oval, the segment immediately about the optical axis is slowed down from c/n_1 to c/n_2. Regions of the same wavefront remote from the axis are still in the first medium traveling with a greater speed, c/n_1. Thus the wavefronts bend, and if the boundary is properly configured (in the form of a Cartesian ovoid), the wavefronts will be inverted from diverging to converging spherical segments.

In addition to focusing a spherical wave, we would like to be able to perform a few other reshaping operations using refracting interfaces; some of these are illustrated in Fig. 5.4. We shall consider them only briefly and more for pedagogical than practical reasons. The surfaces in Fig. 5.4(a) and (b) are ellipsoidal, whereas those in (c) and (d) are hyperboloidal. Notice

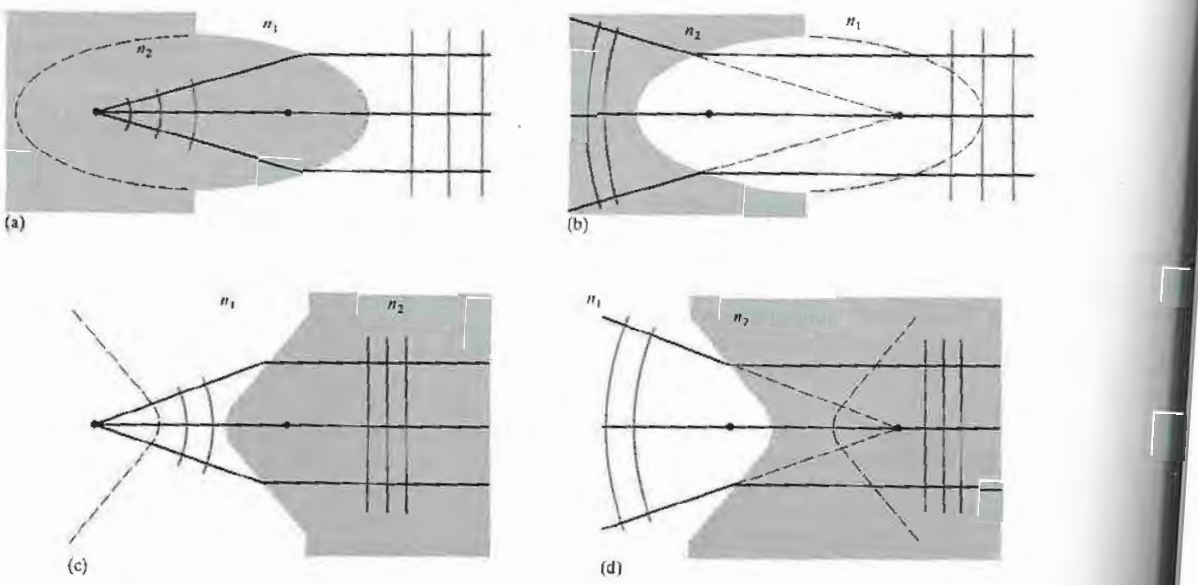


Figure 5.4 Ellipsoidal and hyperboloidal refracting surfaces ($n_2 > n_1$).

that in all cases, the rays either diverge from or converge toward the foci. The arrowheads have been omitted to indicate that the rays can go either way. In other words, an incident plane wave will converge to the farthest focus of an ellipsoid just as a spherical wave emitted from that focus will emerge as a plane wave. Furthermore, as you might expect, if we let the point S in Fig. 5.2 move out to infinity, the ovoid would gradually metamorphose into an ellipsoid.

Rather than deriving expressions for these surfaces, let's just justify the above remarks. To that end, examine **Fig.** 5.5, which relates back to Fig. 5.4(a). The optical path lengths from any point D on the planar wavefront Σ to the focus F_1 must all be equal to the same constant C, that is,

$$(\overline{F_1A})n_2 + (\overline{AD})n_1 = C$$

or

$$(\overline{F_1A}) + (\overline{AD})n_{12} = C/n_2. \tag{5.3}$$

To see that this relationship is indeed satisfied by an ellipsoid of revolution, recall that if Σ corresponds to the directrix of the ellipse, $(\overline{F_2A}) = e(\overline{AD})$, where e is the eccentricity. Thus if $e = n_{12}$, the left-hand side of Eq. (5.3) becomes $(\overline{F_1A}) + (\overline{F_2A})$, which is certainly constant for an ellipse. Here the eccentricity is less than 1 ($e = n_1/n_2$) and it is left for Problem 5.2 to show that had it been greater than 1 (i.e., $n_1 > n_2$), the curve would have been a hyperbola instead [compare (a) with (c) and (b) with (d) in Fig. 5.4]. If all this brings back memories of analytic geometry, you might keep in mind that that subject was originated by Descartes. Interestingly, it was Kepler who first (1611) suggested using conic sections for mirrors and lenses.

The knowledge we have at hand now may be used to construct lenses such that both the object and image points can be in the same medium, which is usually air. The first such device to be considered [**Fig.** 5.6(a)] is a *double convex hyperbolic* lens, which utilizes the response characterized in Fig. 5.4(c). A diverging spherical wave becomes planar after traversing the first hyperbolic surface and then spherically converging on leaving the lens. Alternatively, if the second surface is made planar so that we have a *hyperbolic planar convex* lens, as in Fig. 5.6(b), the plane waves within the lens will strike the back surface perpendicularly and emerge unaltered.

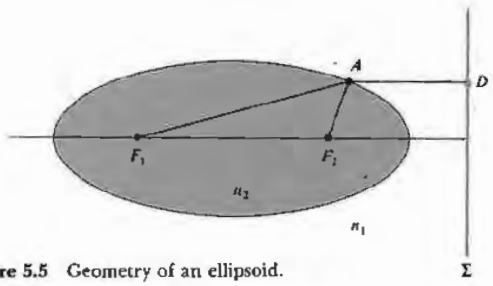
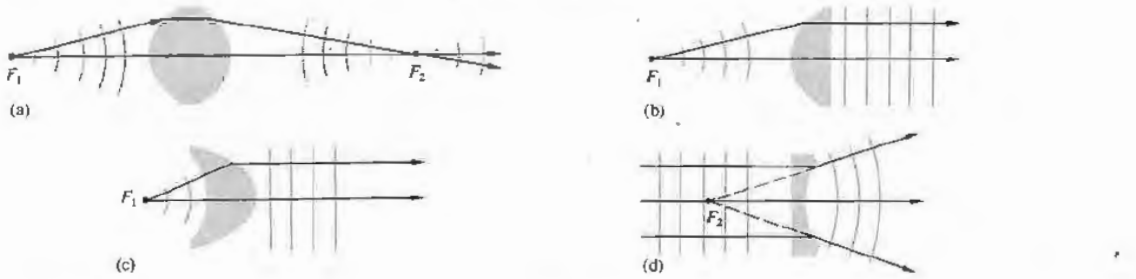


Figure 5.5 Geometry of an ellipsoid.

Another arrangement that will convert diverging spherical waves into plane waves is illustrated in Fig. 5.6(c). This is a *sphero-elliptic convex* lens, where F_1 is simultaneously at the center of the spherical surface and at the focus of the ellipsoid. Rays from F_1 strike the first surface perpendicularly and are therefore undeviated by it. As in Fig. 5.4(a), the exiting wavefronts are planar. All the elements thus far examined have been thicker at their midpoints than at their edges and are for that reason said to be *convex* (from the Latin *convexus*, meaning arched). In contrast, the *planar hyperbolic concave* lens (from the Latin *concavus*, meaning hollow, and easily remembered because it contains the word cave) is thinner at the middle than at the edges, as is evident in Fig. 5.6(d). A number of other arrangements are possible, and a few will be considered in the problems (5.3). Note that each of these lenses will work just as well in reverse: the waves shown emerging can instead be thought of as entering from the right.

If a point source is positioned on the optical axis at the point F_1 of the lens in Fig. 5.6(a), rays will *converge* to the conjugate point F_2. A luminous image of the source would appear on a screen placed at F_2, an image that is therefore said to be **real**. On the other hand, in Fig. 5.6(d) the point source is at infinity, and the rays emerging from the system this time are *diverging*. They appear to come from a point F_2, but no actual luminous image would appear on a screen at that location. The image here is spoken of as **virtual**, as is the familiar image generated by a plane mirror.

Optical elements (lenses and mirrors) of the sort we have talked about, with one or both surfaces neither planar nor spherical, are referred to as *aspherics*. Although their operation is easy to understand and they perform certain tasks exceedingly well, they are still



(e)

Figure 5.6 (a) A double hyperbolic lens. (b) A hyperbolic planar convex lens. (c) A sphero-elliptic lens. (d) A planar hyperbolic lens. (e) Photo courtesy Melles Griot.

difficult to manufacture with great accuracy. Nonetheless, where the costs are justifiable or the required precision is not restrictive or the volume produced is large enough, aspherics are being used extensively and will surely have an increasingly important role. The first quality glass aspheric to be manufactured in great quantities (tens of millions) was a lens for the Kodak disk camera (1982). And the small-scale production of diffraction-limited molded-glass aspheric lenses has been reported in recent times. Today aspherical lenses are frequently used as an elegant means of correcting imaging errors in complicated optical systems.

A new generation of computer-controlled machines, aspheric generators, is producing elements with tolerances (i.e., departures from the desired surface) of better than 0.5 μm (0.000020 inch). This is still about a factor of 10 away from the generally required tolerance of $\lambda/4$ for quality optics, but that will surely come in time. Nowadays aspherics made in plastic and glass can be found in all kinds of instruments across the whole range of quality, including telescopes, projectors, cameras, and reconnaissance devices.

5.2.2 Refraction at Spherical Surfaces

Imagine that we have two pieces of material, one with a concave and the other a convex spherical surface, both having the same radius. It is a unique property of the sphere that such pieces will fit together in intimate contact regardless of their mutual orientation. Thus if we take two roughly spherical objects of suitable cur-

Figure 5.7 Polishing a spherical lens. (Photo courtesy Optical Society of America.)

vature, one a grinding tool and the other a disk of glass, separate them with some abrasive, and then randomly move them with respect to each other, we can anticipate that any high spots on either object will wear away. As they wear, both pieces will gradually become more spherical (Fig. 5.7). Such surfaces are now commonly generated in batches by automatic grinding and polishing machines. In contrast, high-quality aspherical shapes require considerably more effort to produce.

It should therefore come as no surprise that the vast majority of quality lenses in use today have spherical surfaces. Our intent here is to establish techniques for using such surfaces whereby a great many object points can be satisfactorily imaged simultaneously in light composed of a broad frequency range. Image errors, known as *aberrations*, will occur, but it is possible with the present technology to construct high-quality spherical lens systems whose aberrations are so well controlled that image fidelity is limited only by diffraction.

Now that we know why and where we are going, let's move on. Figure 5.8 depicts a wave from the point source S impinging on a spherical interface of radius R centered at C. The ray $(\overline{SA})$ will be refracted at the interface toward the local normal ($n_2 > n_1$) and therefore toward the optical axis. Assume that at some point P it will cross the axis, as will all other rays incident at the same angle θ_i (Fig. 5.9). Fermat's principle maintains that the optical path length (OPL) will be stationary, that is, its derivative with respect to the position variable will be zero. For the ray in question,

$$(\text{OPL}) = n_1\ell_o + n_2\ell_i. \tag{5.4}$$

Using the law of cosines in triangles SAC and ACP along with the fact that $\cos\varphi = -\cos(180 - \varphi)$, we get

$$\ell_o = [R^2 + (s_o + R)^2 - 2R(s_o + R)\cos\varphi]^{1/2}$$

and

$$\ell_i = [R^2 + (s_i - R)^2 + 2R(s_i - R)\cos\varphi]^{1/2}.$$

The OPL can be rewritten as

$$(\text{OPL}) = n_1[R^2 + (s_o + R)^2 - 2R(s_o + R)\cos\varphi]^{1/2} + n_2[R^2 + (s_i - R)^2 + 2R(s_i - R)\cos\varphi]^{1/2}.$$

All the quantities in the diagram (s_i, s_o, R, etc.) are positive numbers, and these form the basis of a *sign convention* which is gradually unfolding and to which

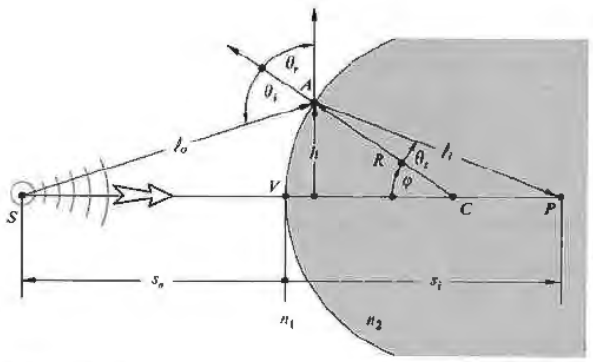


Figure 5.8 Refraction at a spherical interface.

we shall return time and again (see Table 5.1). Inasmuch as the point A moves at the end of a fixed radius (i.e., R = constant), φ is the position variable, and thus setting $d(\text{OPL})/d\varphi = 0$, via Fermat's principle we have

$$\frac{n_1R(s_o + R)\sin\varphi}{2\ell_o} - \frac{n_2R(s_i - R)\sin\varphi}{2\ell_i} = 0,$$

from which it follows that

$$\frac{n_1}{\ell_o} + \frac{n_2}{\ell_i} = \frac{1}{R}\left(\frac{n_2s_i}{\ell_i} - \frac{n_1s_o}{\ell_o}\right). \tag{5.5}$$

This is the relationship that must hold among the parameters for a ray going from S to P by way of refraction at the spherical interface. Although this expression is exact, it is rather complicated. We already know that if A is moved to a new location by changing φ, the new ray will not intercept the optical axis at P—this is not a

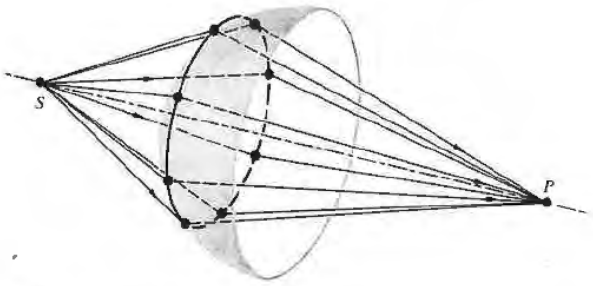


Figure 5.9 Rays incident at the same angle.

Table 5.1 Sign convention for spherical refracting surfaces and thin lenses* (light entering from the left).

s_o, f_o	+ left of V
x_o	+ left of F_o
s_i, f_i	+ right of V
x_i	+ right of F_i
R	+ if C is right of V
y_o, y_i	+ above optical axis

* This table anticipates the imminent introduction of a few quantities not yet spoken of.

Cartesian oval. The approximations that are used to represent ℓ_o and ℓ_i, and thereby simplify Eq. (5.5), are crucial in all that is to follow. Recall that

$$\cos\varphi = 1 - \frac{\varphi^2}{2!} + \frac{\varphi^4}{4!} - \frac{\varphi^6}{6!} + \cdots \tag{5.6}$$

and

$$\sin\varphi = \varphi - \frac{\varphi^3}{3!} + \frac{\varphi^5}{5!} - \frac{\varphi^7}{7!} + \cdots. \tag{5.7}$$

If we assume small values of φ (i.e., A close to V), $\cos\varphi \approx 1$. Consequently, the expressions for ℓ_o and ℓ_i yield $\ell_o \approx s_o$, $\ell_i \approx s_i$, and to that approximation

$$\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}. \tag{5.8}$$

We could have begun this derivation with Snell's law rather than Fermat's principle (Problem 5.4), in which case small values of φ would have led to $\sin\varphi \approx \varphi$ and Eq. (5.8) once again. This approximation delineates the domain of what is called *first-order theory*—we'll examine *third-order theory* ($\sin\varphi \approx \varphi - \varphi^3/3!$) in the next chapter. Rays that arrive at shallow angles with respect to the optical axis (such that φ and h are appropriately small) are known as **paraxial rays**. The *emerging wavefront segment corresponding to these paraxial rays is essentially spherical and will form a "perfect" image at its center P located at* s_i. Notice that Eq. (5.8) is independent of the location of A over a small area about the symmetry axis, namely, the *paraxial region*. Gauss, in 1841, was the first to give a systematic exposition of the formation of images under the above approximation, and the result is variously known as *first-order, paraxial,* or **Gaussian optics.** It soon became the basic theoretical tool by which lenses would be designed for several decades to come. If the optical system is well corrected, an incident spherical wave will emerge in a form very closely resembling a spherical wave. Consequently, as the perfection of the system increases, it more closely approaches first-order theory. Deviations from that of paraxial analysis will provide a convenient measure of the quality of an actual optical device.

If the point F_o in Fig. 5.10 is imaged at infinity ($s_i = \infty$), we have

$$\frac{n_1}{s_o} + \frac{n_2}{\infty} = \frac{n_2 - n_1}{R}.$$

That special object distance is defined as the *first focal length* or the *object focal length*, $s_o = f_o$, so that

$$f_o = \frac{n_1}{n_2 - n_1} R. \tag{5.9}$$

The point F_o is known as the *first* or *object focus*. Similarly the *second or image focus* is the axial point F_i, where the image is formed when $s_o = \infty$, that is,

$$\frac{n_1}{\infty} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}.$$

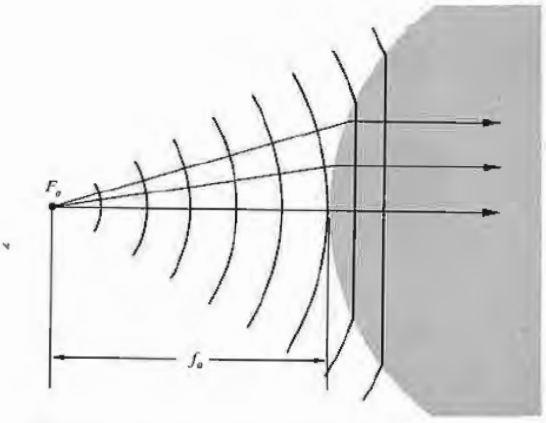


Figure 5.10 Plane waves propagating beyond a spherical interface—the object focus.

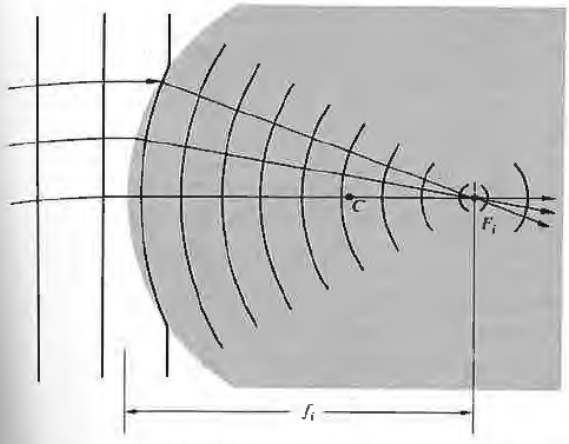


Figure 5.11 The reshaping of plane into spherical waves at a spherical interface—the image focus.

Defining the *second or image focal length* f_i as equal to s_i in this special case (Fig. 5.11), we have

$$f_i = \frac{n_2}{n_2 - n_1} R. \tag{5.10}$$

Recall that an image is virtual when the rays diverge from it (Fig. 5.12). Analogously, *an object is virtual when the rays converge toward it* (Fig. 5.13). Observe that the virtual object is now on the right-hand side of the vertex, and therefore s_o will be a negative quantity. Moreover, the surface is concave, and its radius will also be negative, as required by Eq. (5.9), since f_o would be negative. In the same way the virtual image distance appearing to the left of V is negative.

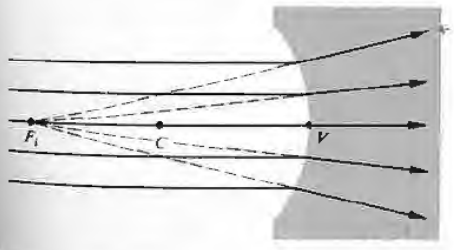


Figure 5.12 A virtual image point.

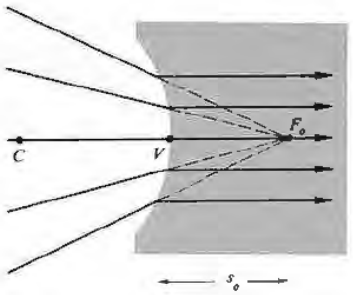


Figure 5.13 A virtual object point.

5.2.3 Thin Lenses

Lenses are made in a wide range of forms; for example, there are acoustic and microwave lenses; some of the latter are made of glass or wax in easily recognizable shapes, whereas others are far more subtle in appearance (Fig. 5.14). In the traditional sense, *a lens is an optical system consisting of two or more refracting interfaces, at least one of which is curved.* Generally the nonplanar surfaces are centered on a common axis. These surfaces are most frequently spherical segments and are often coated with thin dielectric films to control their transmission properties (see Section 9.9). A lens that consists of one element (i.e., it has only two refracting surfaces) is a *simple lens.* The presence of more than one element makes it a *compound lens.* A lens is also classified as to whether it is *thin* or *thick*, that is, whether its thickness is effectively negligible or not. We will limit ourselves, for the most part, to *centered systems* (for which all surfaces are rotationally symmetric about a common axis) of spherical surfaces. Under these restrictions, the simple lens can take the diverse forms shown in Fig. 5.15. Lenses that are variously known as *convex, converging,* or *positive* are thicker at the center and so tend to decrease the radius of curvature of the wavefronts. In other words, the wave converges more as it traverses the lens, assuming, of course, that the index of the lens is greater than that of the media in which it is immersed. *Concave, diverging,* or *negative* lenses, on the other hand,

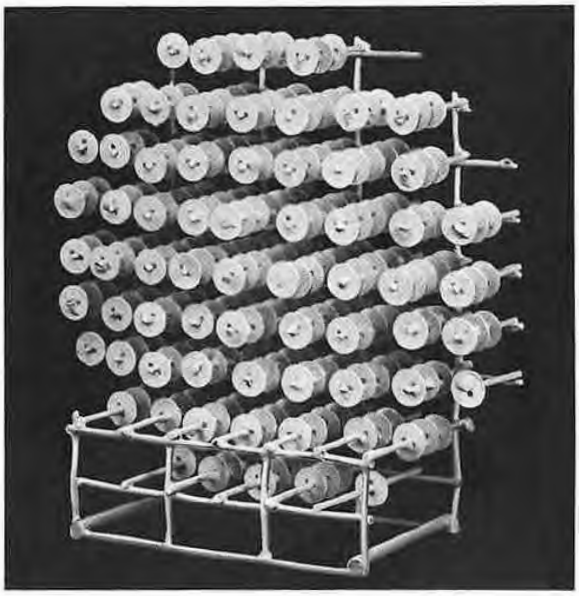
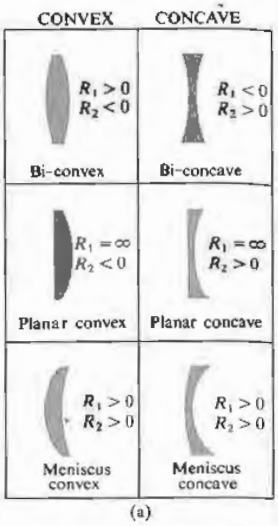


Figure 5.14 A lens for short-wavelength radiowaves. The disks serve to refract these waves much as rows of atoms refract light. (Photo courtesy Optical Society of America.)



(a)

Figure 5.15 Cross sections of various centered spherical simple lenses. The surface on the left is #1 since it is encountered first. Its radius is R_1. (Photo courtesy of Melles Griot.)

(b)

are thinner at the center and tend to advance that portion of the wavefront, causing it to diverge more then it did upon entry.

In the broadest sense, *a lens is a refracting device that is used to reshape wavefronts in a controlled manner.* Although this is usually done by passing the wave through at least one specially shaped interface separating two different homogeneous media, it is not the only approach available. For example, it is also possible to reconfigure a wavefront by passing it through an inhomogeneous medium. A *gradient-index*, or GRIN, lens is one where the desired effect is accomplished by using a medium in which the index of refraction varies in a prescribed fashion. Different portions of the wave propagate at different speeds, and the front changes shape as it progresses. In the commercial GRIN material (available only since 1976) the index varies radially, decreasing parabolically out from the central axis.

Today GRIN lenses are still fabricated in quantity only in the form of small-diameter, parallel, flat-faced rods. Usually grouped together in large arrays, they have been used extensively in such equipment as facsimile machines and compact copiers. There are other unconventional lenses, including the holographic lens and even the gravitational lens (where, for example, the gravity of a galaxy bends light passing in its vicinity, thereby forming multiple images of distant celestial objects, such as quasars). We shall focus our attention in the remainder of this chapter on the more traditional types of lenses, even though you are actually reading these words through a GRIN lens (p.179).

Thin-Lens Equations

Return for a moment to the discussion of refraction at a single spherical interface, where the location of the conjugate points S and P is given by

$$\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}. \qquad [5.8]$$

When s_o is large for a fixed $(n_2 - n_1)/R$, s_i is relatively small. As s_o decreases, s_i moves away from the vertex, that is, both θ_i and θ_t increase until finally $s_o = f_o$ and $s_i = \infty$. At that point, $n_1/s_o = (n_2 - n_1)/R$, so that if s_o gets any smaller, s_i will have to be negative, if Eq. (5.8) is to hold. In other words, the image becomes virtual (Fig. 5.16). Let's now locate the conjugate points for the lens of index n_l surrounded by a medium of index n_m, as in Fig. 5.17, where another end has simply been ground on the piece in Fig. 5.16(c). This certainly isn't the most general set of circumstances, but it is the most common, and even more cogently, it is the simplest.* We know from Eq. (5.8) that the paraxial rays issuing from S at s_{o1} will meet at P', a distance, which we now call s_{i1}, from V_1, given by

$$\frac{n_m}{s_{o1}} + \frac{n_l}{s_{i1}} = \frac{n_l - n_m}{R_1}. \qquad (5.11)$$

Thus as far as the second surface is concerned, it "sees" rays coming toward it from P', which serves as its object

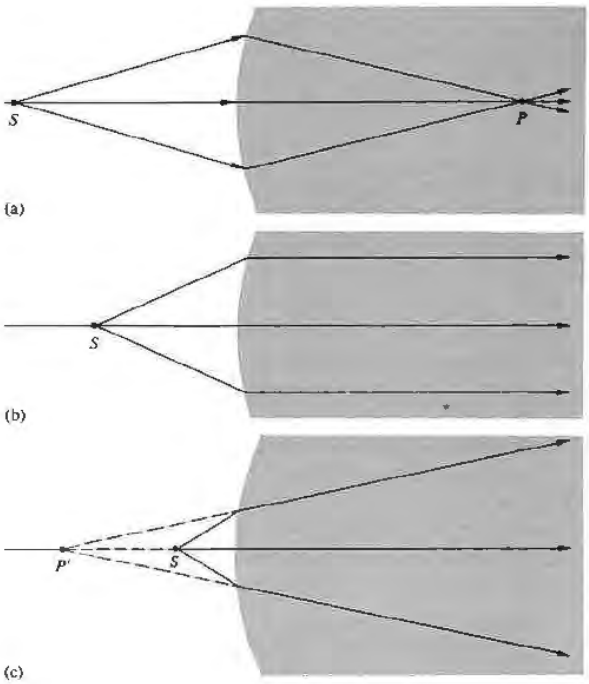


Figure 5.16 Refraction at a spherical interface.

point a distance s_{o2} away. Furthermore, the rays arriving at that second surface are in the medium of index n_l. Thus, the object space for the second interface that contains P' has an index n_l. Note that the rays from P' to that surface are indeed straight lines. Considering the fact that

$$|s_{o2}| = |s_{i1}| + d,$$

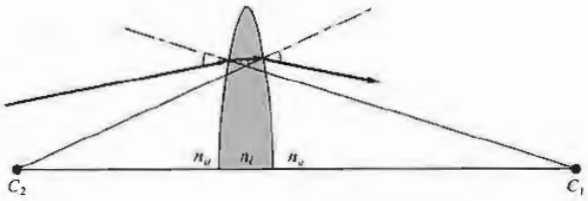
since s_{o2} is on the left and therefore positive, $s_{o2} = |s_{o2}|$, and s_{i1} is also on the left and therefore negative, $-s_{i1} = |s_{i1}|$, we have

$$s_{o2} = -s_{i1} + d. \qquad (5.12)$$

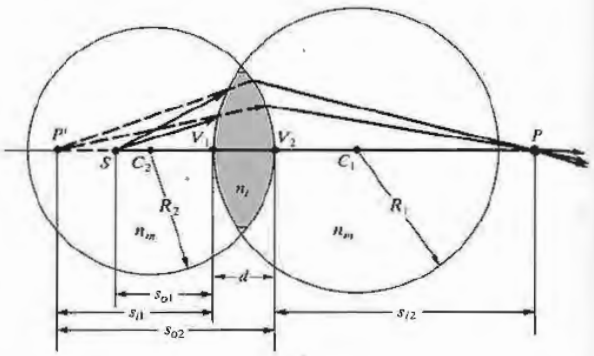
Thus at the second surface Eq. (5.8) yields

$$\frac{n_l}{(-s_{i1} + d)} + \frac{n_m}{s_{i2}} = \frac{n_m - n_l}{R_2}. \qquad (5.13)$$

*See Jenkins and White, *Fundamentals of Optics*, p. 57, for a derivation containing three different indices.



(a)



(b)

Figure 5.17 A spherical lens. (a) Refraction at the interfaces. The radius drawn from C_1 is normal to the first surface, and as the ray enters the lens it bends down *toward* that normal. The radius from C_2 is normal to the second surface; and as the ray emerges, since $n_l > n_a$, the ray bends down *away* from that normal. (b) The geometry.

Here $n_l > n_m$ and $R_2 < 0$, so that the right-hand side is positive. Adding Eqs. (5.11) and (5.13), we have

$$\frac{n_m}{s_{o1}} + \frac{n_m}{s_{i2}} = (n_l - n_m)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{n_l d}{(s_{i1} - d)s_{i1}}. \tag{5.14}$$

If the lens is thin enough ($d \to 0$), the last term on the right is effectively zero. As a further simplification, assume the surrounding medium to be air (i.e., $n_m \approx 1$). Accordingly, we have the very useful **thin-lens equation**, often referred to as the **lensmaker's formula**:

$$\frac{1}{s_o} + \frac{1}{s_i} = (n_l - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right), \tag{5.15}$$

where we let $s_{o1} = s_o$ and $s_{i2} = s_i$. The points V_1 and V_2 tend to coalesce as $d \to 0$, so that s_o and s_i can be measured from either the vertices or the lens center.

Just as in the case of the single spherical surface, if s_o is moved out to infinity, the image distance becomes the focal length f_i, or symbolically,

$$\lim_{s_o \to \infty} s_i = f_i.$$

Similarly

$$\lim_{s_i \to \infty} s_o = f_o.$$

It is evident from Eq. (5.15) that for a thin lens $f_i = f_o$, and consequently we drop the subscripts altogether. Thus

$$\frac{1}{f} = (n_l - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{5.16}$$

and

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}, \tag{5.17}$$

which is the famous **Gaussian lens formula**. As an example of how these expressions might be used, let's compute the focal length in air of a thin planar-convex lens having a radius of curvature of 50 mm and an index of 1.5. With light entering on the planar surface ($R_1 = \infty$, $R_2 = -50$),

$$\frac{1}{f} = (1.5 - 1)\left(\frac{1}{\infty} - \frac{1}{-50}\right),$$

whereas if instead it arrives at the curved surface ($R_1 = +50$, $R_2 = \infty$),

$$\frac{1}{f} = (1.5 - 1)\left(\frac{1}{+50} - \frac{1}{\infty}\right),$$

and in either case $f = 100$ mm. If an object is alternately

placed at distances 600 mm, 200 mm, 150 mm, 100 mm, and 50 mm from the lens on either side, we can find the image points from Eq. (5.17). Hence

$$\frac{1}{600}+\frac{1}{s_i}=\frac{1}{100}$$

and $s_i = 120$ mm. Similarly, the other image distances are 200 mm, 300 mm, ∞, and -100 mm, respectively. Interestingly enough, when $s_o = \infty$, $s_i = f$; as s_o decreases, s_i increases positively until $s_o = f$ and s_i is negative thereafter. You can qualitatively check this out with a simple convex lens and a small electric light—the high-intensity variety that uses auto lamps is probably the most convenient. Standing as far as you can from the source, project a clear image of it onto a white sheet of paper. You should be able to see the lamp quite clearly and not just as a blur. That image distance approximates f. Now move the lens in toward S, adjusting s_i to produce a clear image. It will surely increase. As $s_o \to f$, a clear image of the filament can be projected, but only on an increasingly distant screen. For $s_o < f$, there will just be a blur where the farthest wall intersects the diverging cone of rays—the image is virtual.

ii) Focal Points and Planes

Figure 5.18 summarizes pictorially some of the situations described analytically by Eq. 5.16. Observe that if a lens of index n_l is in a medium of index n_m,

$$\frac{1}{f}=(n_{lm}-1)\left(\frac{1}{R_1}-\frac{1}{R_2}\right). \qquad (5.18)$$

The focal lengths in (a) and (b) of Fig. 5.18 are equal, because the same medium exists on either side of the lens. Since $n_l > n_m$, it follows that $n_{lm} > 1$. In both cases $R_1 > 0$ and $R_2 < 0$, so that each focal length is positive. We have a real object in (a) and a real image in (b). In (c), $n_l < n_m$, and consequently f is negative. In (d) and (e), $n_{lm} > 1$ but $R_1 < 0$, whereas $R_2 > 0$, so f is again negative, and the object in one case and the image in

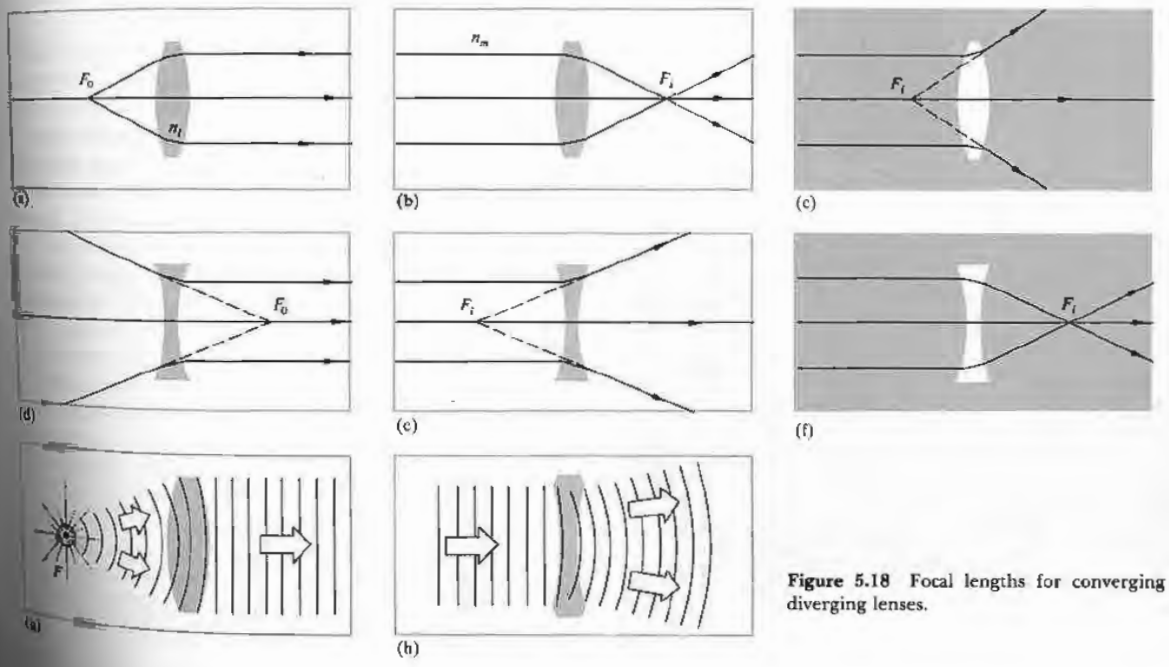
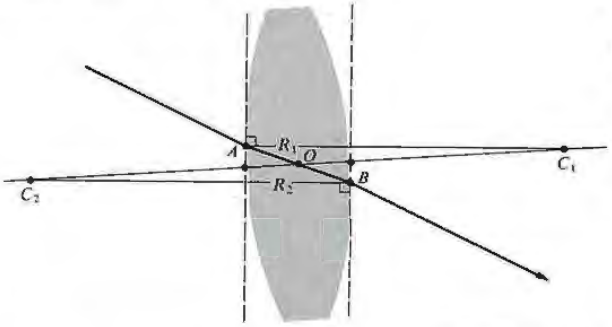


Figure 5.18 Focal lengths for converging and diverging lenses.

the other are virtual. The last situation shows $n_{1m} < 1$, yielding an $f > 0$.

Notice that in each instance it is particularly convenient to draw a ray through the center of the lens, which, because it is perpendicular to both surfaces, is undeviated. Suppose, however, that an off-axis paraxial ray emerges from the lens parallel to its incident direction, as in Fig. 5.19. We maintain that all such rays will pass through the point defined as the *optical center* of the lens O. To see this, draw two parallel planes, one on each side tangent to the lens at any pair of points A and B. This can easily be done by selecting A and B such that the radii $\overline{AC_1}$ and $\overline{BC_2}$ are themselves parallel. It is to be shown that the paraxial ray traversing $\overline{AB}$ enters and leaves the lens in the same direction. It is evident from the diagram that triangles AOC_1 and BOC_2 are similar, in the geometric sense, and therefore their sides are proportional. Hence, $|R_1|(\overline{OC_2}) = |R_2|(\overline{OC_1})$, and since the radii are constant, the location of O is constant, independent of A and B. As we saw earlier (Problem 4.19 and Fig. 4.55), a ray traversing a medium bounded by parallel planes will be displaced laterally but will suffer no angular deviation. This displacement is proportional to the thickness, which for a thin lens is negligible. *Rays passing through O may, accordingly, be drawn as straight lines.* It is customary when dealing with thin lenses simply to place O midway between the vertices.



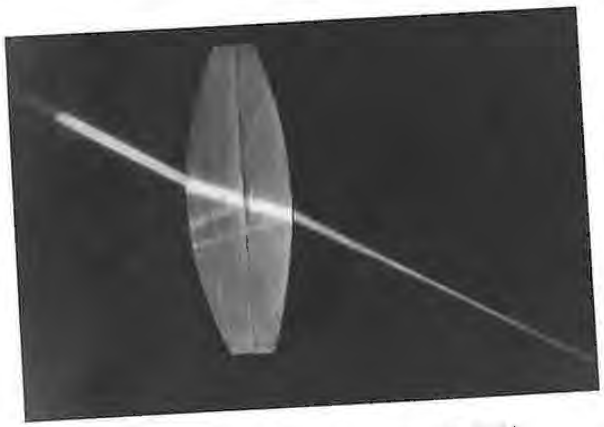


Figure 5.19 The optical center of a lens. (Photo by E.H.)

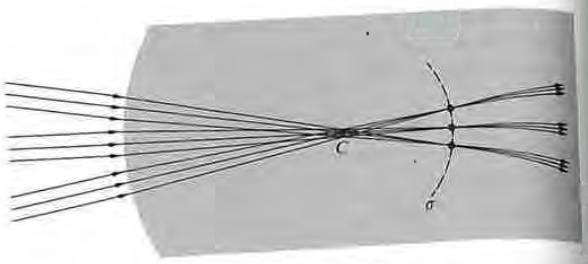


Figure 5.20 Focusing of several ray bundles.

Recall that a bundle of parallel paraxial rays incident on a spherical refracting surface comes to a focus at a point on the optical axis (Fig. 5.11). As shown in Fig. 5.20, this implies that several such bundles entering in a narrow cone will be focused on a spherical segment σ, also centered on C. The undeviated rays normal to the surface, and therefore passing through C, locate the foci on σ. Since the ray cone must indeed be narrow, σ can satisfactorily be represented as a plane normal to the symmetry axis and passing through the image focus. It is known as a **focal plane**. In the same way, limiting ourselves to paraxial theory, a lens will focus all incident parallel bundles of rays* onto a surface called the *second or back focal plane*, as in Fig. 5.21. Here each point on σ is located by the undeviated ray through O. Similarly, the *first* or *front focal plane* contains the object focus F_o.

* Perhaps the earliest literary reference to the focal properties of a lens appears in Aristophanes' play, *The Clouds*, which dates back to 423 B.C. In it Strepsiades plots to use a burning glass to focus the Sun's rays onto a wax tablet and thereby melt out the record of his gambling debt.

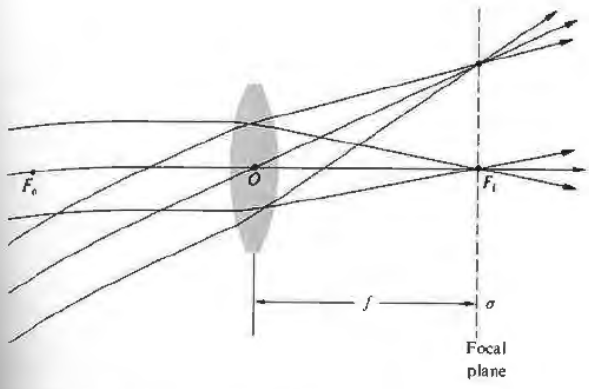


Figure 5.21 The focal plane of a lens.

iii) Finite Imagery

Thus far we've dealt with the mathematical abstraction of a single-point source, but now let's suppose that a great many such points combine to form a continuous finite object. For the moment, imagine the object to be a segment of a sphere, σ_o, centered on C, as in Fig. 5.22. If σ_o is close to the spherical interface, point S will have a virtual image P ($s_i < 0$ and therefore on the left of V). With S farther away, its image will be real ($s_i > 0$ and therefore on the right-hand side). In either case, each point on σ_o has a conjugate point on σ_i lying on a straight line through C. Within the restrictions of paraxial theory, these surfaces can be considered planar. Thus a small planar object normal to the optical axis will be imaged into a small planar region also normal to that axis. It should be noted that if σ_o is moved out to infinity, the cone of rays from each source point will become **collimated** (i.e., parallel), and the image points will lie on the focal plane (Fig. 5.21).

By cutting and polishing the right side of the piece depicted in Fig. 5.22, we can construct a thin lens, just as was done in Section (i). Once again, the image (σ_i in Fig. 5.22) formed by the first surface of the lens will serve as the object for the second surface, which in turn will generate a final image. Suppose then that σ_i in Fig. 5.22(a) is the object for the second surface, which is assumed to have a negative radius. We already know what will happen next—the situation is identical to Fig. 5.22(b) with the ray directions reversed. The *final image formed by a lens of a small planar object normal to the optical axis will itself be a small plane normal to that axis.*

The location, size, and orientation of an image produced by a lens can be determined, particularly simply, with ray diagrams. To find the image of the object in Fig. 5.23, we must locate the image point corresponding to each object point. Since all rays issuing from a source point in a paraxial cone will arrive at the image point, any two such rays will suffice to fix that point. Since we know the positions of the focal points, there are three rays that are especially easy to apply. Two of these make use of the fact that a ray passing through the focal point will emerge from the lens parallel to the optical axis and vice versa; the third is the undeviated ray through

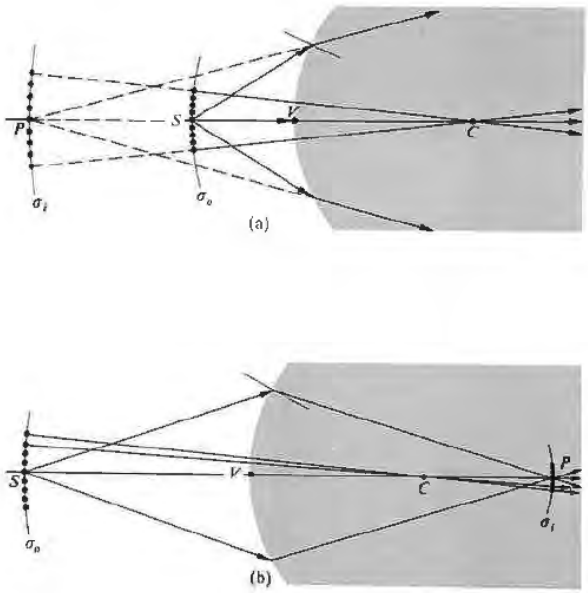


Figure 5.22 Finite imagery.

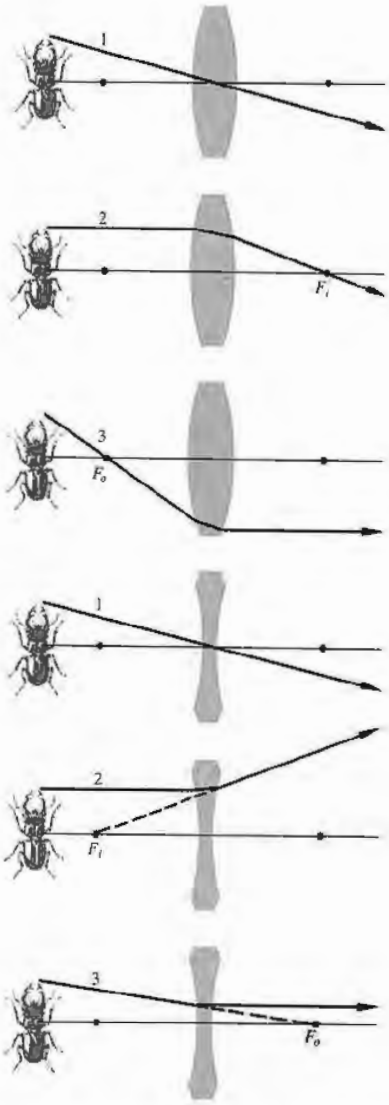


Figure 5.23 Tracing a few key rays through a positive and negative lens.

O. Figure 5.24 shows how any *two* of these three rays locate the image of a point on the object. Incidentally, this technique dates back to the work of Robert Smith as long ago as 1738.

This graphical procedure can be made even simpler by replacing the thin lens with a plane passing through its center (Fig. 5.25). Presumably, if we were to extend every incoming ray forward a little and every outgoing ray backward a bit, each pair would meet on this plane. Thus the total deviation of any ray can be envisaged as occurring all at once on that plane. This is equivalent to the actual process consisting of two separate angular shifts, one at each interface. (As we will see later, this is tantamount to saying that the two principal planes of a thin lens coincide.)

In accord with convention, transverse distances above the optical axis are taken as positive quantities, and those below the axis are given negative numerical values. Therefore in Fig. 5.25 $y_o > 0$ and $y_i < 0$. Here the image is said to be *inverted*, whereas if $y_i > 0$ when $y_o > 0$, it is *erect*. Observe that triangles AOF_i and $P_2P_1F_i$ are similar. Ergo

$$\frac{y_o}{|y_i|} = \frac{f}{(s_i - f)}. \tag{5.19}$$

Likewise, triangles S_2S_1O and P_2P_1O are similar and

$$\frac{y_o}{|y_i|} = \frac{s_o}{s_i}, \tag{5.20}$$

where all quantities other than y_i are positive. Hence

$$\frac{s_o}{s_i} = \frac{f}{(s_i - f)} \tag{5.21}$$

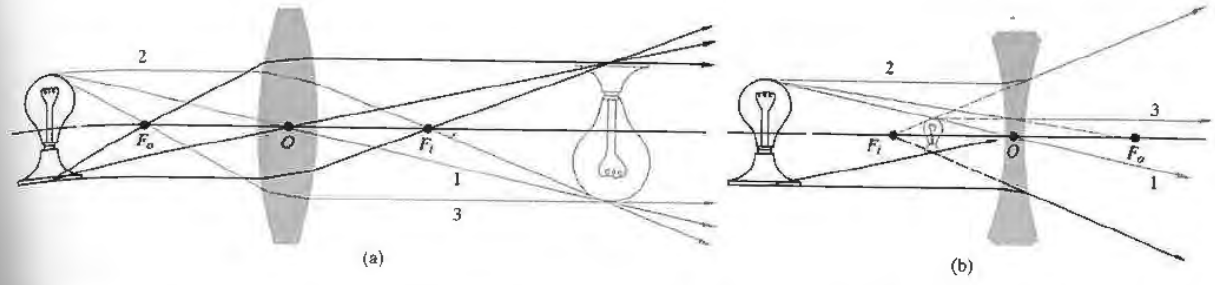
and

$$\frac{1}{f} = \frac{1}{s_o} + \frac{1}{s_i},$$

which is, of course, the Gaussian lens equation (5.17). Furthermore, triangles $S_2S_1F_o$ and BOF_o are similar and

$$\frac{f}{(s_o - f)} = \frac{|y_i|}{y_o}. \tag{5.22}$$

Using the distances measured from the focal points and



(c)

(d)

Figure 5.24 (a) A real object and a positive lens. (b) A real object and a negative lens. (c) A real image projected on the viewing screen much as the eye projects its image on the retina. (d) The minified, rightside-up, virtual image formed by a negative lens.

combining this information with Eq. (5.19), we have

$$x_o x_i = f^2. \tag{5.23}$$

This is the **Newtonian form** of the lens equation, the first statement of which appeared in Newton's *Opticks* in 1704. The signs of x_o and x_i are reckoned with respect to their concomitant foci. By convention x_o is taken to be positive left of F_o, whereas x_i is positive on the right of F_i. To be sure, it is evident from Eq. (5.23) that x_o and x_i have like signs, which means that *the object and image must be on opposite sides of their respective focal points.* This is a good thing for the neophyte to remember

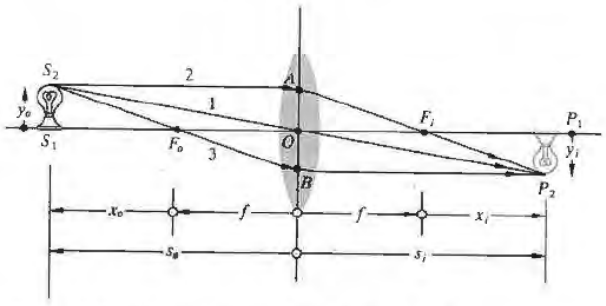


Figure 5.25 Object and image location for a thin lens.

6 MORE ON GEOMETRICAL OPTICS

The preceding chapter, for the most part, dealt with paraxial theory as applied to thin spherical lens systems. The two predominant approximations were, rather obviously, that we had *thin* lenses and that first-order theory was sufficient for their analysis. Neither of these assumptions can be maintained throughout the design of a precision optical system, but, taken together, they provide the basis for a first rough solution. This chapter will carry things a bit further by examining thick lenses and aberrations; even at that, it is only a beginning. The advent of computerized lens design requires a certain shift in emphasis—there is litttle need to do what a computer can do better. Moreover, the sheer wealth of existing material developed over centuries demands a bit of judicious pruning to avoid a plethora of pedantry.

6.1 THICK LENSES AND LENS SYSTEMS

Figure 6.1 depicts a thick lens (i.e., one whose thickness is by no means negligible). As we shall see, it could equally well be envisioned more generally as an optical system, allowing for the possibility that it consists of a number of simple lenses, not merely one. The first and second focal points, or if you like, the object and image foci, F_o and F_i, can conveniently be measured from the two (outermost) vertices. In that case we have the familiar front and back focal lengths denoted by f.f.l. and b.f.l. When extended, the incident and emerged

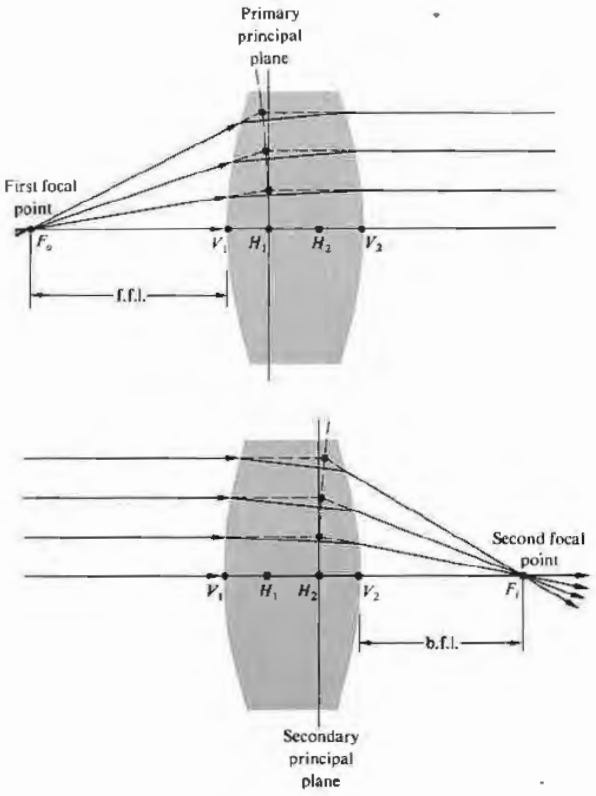


Figure 6.1 A thick lens.

rays will meet at points, the locus of which forms a curved surface that may or may not reside within the lens. The surface, approximating a plane in the paraxial region, is termed the **principal plane** (see Section 6.3.1). Points where the primary and secondary principal planes (as shown in Fig. 6.1) intersect the optical axis are known as the **first** and **second principal points**, H_1 and H_2, respectively. They provide a set of very useful references from which to measure several of the system parameters. We saw earlier (Fig. 5.19, p.140) that a ray traversing the lens through its optical center emerges parallel to the incident direction. Extending both the incoming and outgoing rays until they cross the optical axis locates what are called the **nodal points**, N_1 and N_2 in Fig. 6.2. *When the lens is surrounded on both sides by the same medium, generally air, the nodal and principal points will be coincident.* The six points, two focal, two principal, and two nodal, constitute the **cardinal points** of the system. As shown in Fig. 6.3, the principal planes can lie completely outside the lens system. Here, although differently configured, each lens in either group has the same power. Observe that in the symmetrical lens the principal planes are, quite reasonably, symmetrically located. In the case of either the planar-concave or planar-convex lens, one principal plane is tangent to the curved surface—as should be expected from the definition (applied to the paraxial region). In contrast, the principal points can be external for meniscus lenses. One often speaks of this succession of shapes with the same power as exemplifying *lens bending*. A

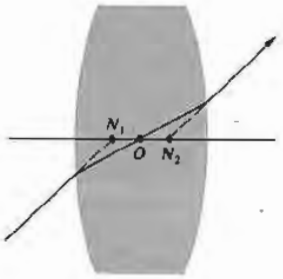


Figure 6.2 Nodal points.

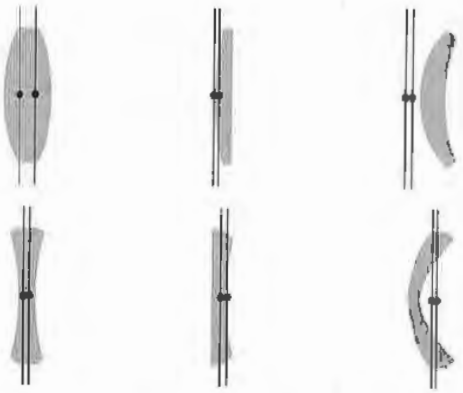


Figure 6.3 Lens bending.

rule of thumb for ordinary glass lenses in air is that the separation $\overline{H_1H_2}$ roughly equals one third the lens thickness $\overline{V_1V_2}$.

The thick lens can be treated as consisting of two spherical refracting surfaces separated by a distance d between their vertices, as in Section 5.2.3, where the thin-lens equation was derived. After a great deal of algebraic manipulation,* wherein d is not negligible, one arrives at a very interesting result for the thick lens immersed in air. The expression for the conjugate points once again can be put in the Gaussian form,

$$\frac{1}{s_o}+\frac{1}{s_i}=\frac{1}{f}, \tag{6.1}$$

provided that both these object and image distances are measured from the first and second principal planes, respectively. Moreover, the *effective focal length*, or simply the *focal length*, f, is also reckoned with respect to the principal planes and is given by

$$\frac{1}{f}=(n_l-1)\left[\frac{1}{R_1}-\frac{1}{R_2}+\frac{(n_l-1)d}{n_l R_1 R_2}\right]. \tag{6.2}$$

The principal planes are located at distances of $\overline{V_1H_1}=h_1$ and $\overline{V_2H_2}=h_2$, *which are positive when the planes lie to the right of their respective vertices.* Figure 6.4 illustrates

* For the complete derivation, see Morgan, *Introduction to Geometrical and Physical Optics*, p. 57.

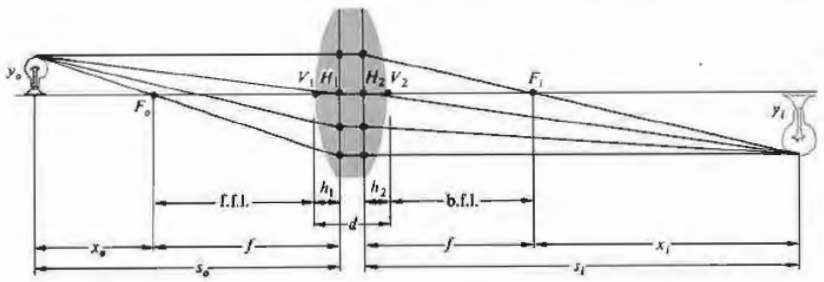


Figure 6.4 Thick lens geometry.

the arrangement of the various quantities. The values of h_1 and h_2 are given by

$$h_1 = -\frac{f(n_l - 1)d}{R_2 n_l} \tag{6.3}$$

and

$$h_2 = -\frac{f(n_l - 1)d}{R_1 n_l}. \tag{6.4}$$

In the same way the Newtonian form of the lens equation holds, as is evident from the similar triangles in Fig. 6.4. Thus

$$x_i x_o = f^2, \tag{6.5}$$

so long as f is given the present interpretation. And from the same triangles

$$M_T = \frac{y_i}{y_o} = -\frac{x_i}{f} = -\frac{f}{x_o}. \tag{6.6}$$

Obviously if $d \to 0$, Eqs. (6.1), (6.2), and (6.5) are transformed into the thin-lens expressions (5.17), (5.16), and (5.23). As a numerical example, let's find the image distance for an object positioned 30 cm from the vertex of a double convex lens having radii of 20 cm and 40 cm, a thickness of 1 cm, and an index of 1.5. From Eq. (6.2) the focal length (in centimeters) is

$$\frac{1}{f} = (1.5 - 1)\left[\frac{1}{20} - \frac{1}{-40} + \frac{(1.5 - 1)1}{1.5(20)(-40)}\right],$$

so $f = 26.8$ cm. Furthermore,

$$h_1 = -\frac{26.8(0.5)1}{-40(1.5)} = +0.22 \text{ cm}$$

and

$$h_2 = -\frac{26.8(0.5)1}{20(1.5)} = -0.44 \text{ cm},$$

which means that H_1 is to the right of V_1, and H_2 is to the left of V_2. Finally, $s_o = 30 + 0.22$, whence

$$\frac{1}{30.2} + \frac{1}{s_i} = \frac{1}{26.8},$$

and $s_i = 238$ cm, measured from H_2.

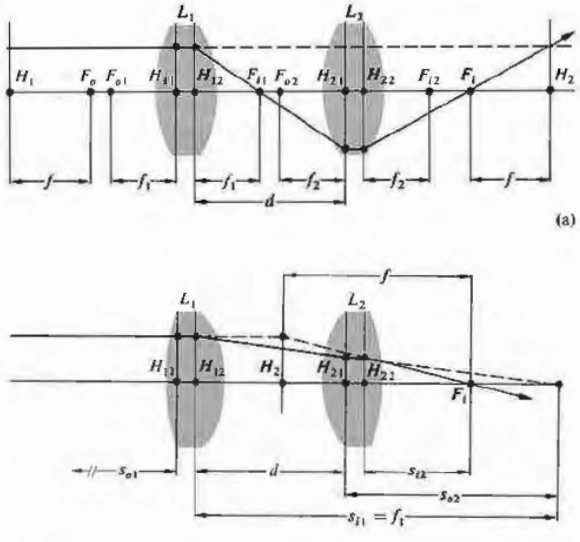


Figure 6.5 A compound thick lens.

The principal points are conjugate to each other. In other words, since $f = s_o s_i/(s_o + s_i)$, when $s_o = 0$, s_i must be zero, because f is finite and thus a point at H_1 is imaged at H_2. Furthermore, an object in the first principal plane ($x_o = -f$) is imaged in the second principal plane ($x_i = -f$) with unit magnification ($M_T = 1$). It is for this reason that they are sometimes spoken of as *unit planes*. Hence any ray directed toward a point on the first principal plane will emerge from the lens as if it originated at the corresponding point (the same distance above or below the axis) on the second principal plane.

Suppose we now have a compound lens consisting of two thick lenses, L_1 and L_2 (Fig. 6.5). Let s_{o1}, s_{i1}, and f_1 and s_{o2}, s_{i2}, and f_2 be the object and image distances and focal lengths for the two lenses, all measured with respect to their own principal planes. We know that the transverse magnification is the product of the magnifications of the individual lenses, that is,

$$M_T = \left(-\frac{s_{i1}}{s_{o1}}\right)\left(-\frac{s_{i2}}{s_{o2}}\right) = -\frac{s_i}{s_o}, \tag{6.7}$$

where s_o and s_i are the object and image distances for the combination as a whole. When s_o is equal to infinity $s_o = s_{o1}$, $s_{i1} = f_1$, $s_{o2} = -(s_{i1} - d)$, and $s_i = f$. Since

$$\frac{1}{s_{o2}} + \frac{1}{s_{i2}} = \frac{1}{f_2},$$

it follows (Problem 6.1), upon substituting into Eq. (6.7), that

$$-\frac{f_1 s_{i2}}{s_{o2}} = f$$

or

$$f = -\frac{f_1}{s_{o2}}\left(\frac{s_{o2} f_2}{s_{o2} - f_2}\right) = \frac{f_1 f_2}{s_{i1} - d + f_2}.$$

Hence

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2}. \tag{6.8}$$

This is the effective focal length of the combination of two thick lenses where all distances are measured from principal planes. The principal planes for the system as a whole are located using the expressions

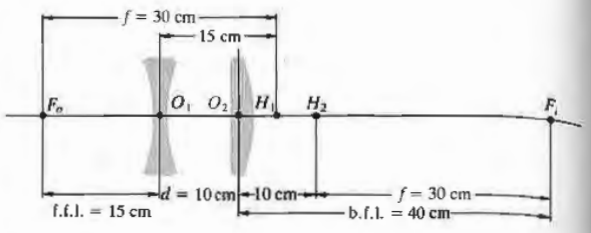


Figure 6.6 A compound lens.

$$\overline{H_{11}H_1} = \frac{fd}{f_2} \tag{6.9}$$

and

$$\overline{H_{22}H_2} = -\frac{fd}{f_1}, \tag{6.10}$$

which will not be derived here (see Section 6.2.1). We have in effect found an equivalent thick-lens representation of the compound lens. Note that if the component lenses are thin, the pairs of points H_{11}, H_{12} and H_{21}, H_{22} coalesce, whereupon d becomes the center-to-center lens separation, as in Section 5.2.3. For example, returning to the thin lenses of Fig. 5.31 where $f_1 = -30$, $f_2 = 20$, and $d = 10$, as in Fig. 6.6,

$$\frac{1}{f} = \frac{1}{-30} + \frac{1}{20} - \frac{10}{(-30)(20)},$$

so $f = 30$ cm. We found earlier (p.148) that b.f.l. = 40 cm and f.f.l. = 15 cm. Moreover, since these are thin lenses, Eqs. (6.9) and (6.10) can be written as

$$\overline{O_1H_1} = \frac{30(10)}{20} = +15 \text{ cm}$$

and

$$\overline{O_2H_2} = -\frac{30(10)}{-30} = +10 \text{ cm}.$$

Both are positive, and therefore the planes lie to the right of O_1 and O_2, respectively. Both computed values agree with the results depicted in the diagram. If light

enters from the right, the system resembles a telephoto lens that must be placed 15 cm from the film plane, yet has an effective focal length of 30 cm.

The same procedures can be extended to three, four, or more lenses. Thus

$$f = f_1\left(-\frac{s_{i2}}{s_{o2}}\right)\left(-\frac{s_{i3}}{s_{o3}}\right)\cdots. \qquad (6.11)$$

Equivalently, the first two lenses can be envisioned as combined to form a single thick lens whose principal points and focal length are calculated. It, in turn, is combined with the third lens, and so on with each successive element.

6.2 ANALYTICAL RAY TRACING

Ray tracing is unquestionably one of the designer's chief tools. Having formulated an optical system on paper, one can mathematically shine rays through it to evaluate its performance. Any ray, paraxial or otherwise, can be traced through the system exactly. Conceptually it's a simple matter of applying the refraction equation

$$n_i(\hat{\mathbf{k}}_i \times \hat{\mathbf{u}}_n) = n_t(\hat{\mathbf{k}}_t \times \hat{\mathbf{u}}_n) \qquad [4.7]$$

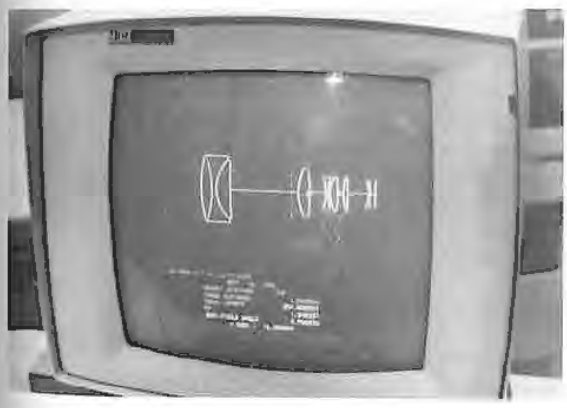
at the first surface, locating where the transmitted ray then strikes the second surface, applying the equation once again, and so on all the way through. At one time *meridional rays* (those in the plane of the optical axis) were traced almost exclusively, because nonmeridional or *skew rays* (which do not intersect the axis) are considerably more complicated to deal with mathematically. The distinction is of less importance to a high-speed electronic computer (Fig. 6.7) which simply takes a trifle longer to make the trace. Thus, whereas it would probably take 10 or 15 minutes for a skilled person with a desk calculator to evaluate the trajectory of a single skew ray through a single surface, a computer might require less than a thousandth of a second for the same job, and equally important, it would be ready for the next calculation with undiminished enthusiasm.

The simplest case that will serve to illustrate the ray-tracing process is that of a paraxial, meridional ray traversing a thick spherical lens. Applying Snell's law in Fig. 6.8 at point P_1 yields

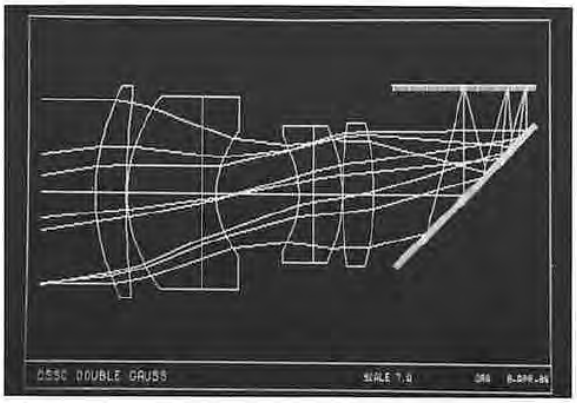
$$n_{i1}\theta_{i1} = n_{t1}\theta_{t1}$$

or

$$n_{i1}(\alpha_{i1} + \alpha_1) = n_{t1}(\alpha_{t1} + \alpha_1).$$



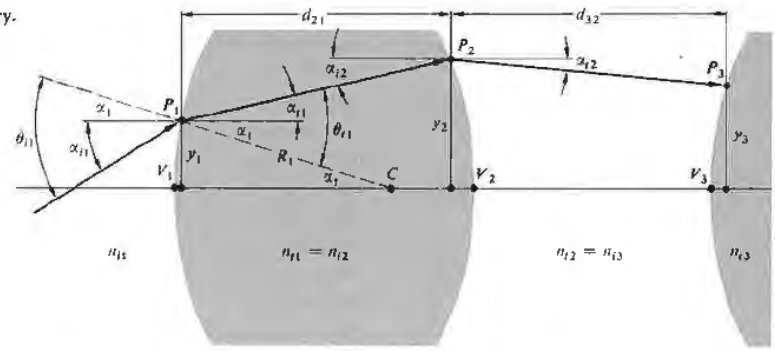
(a)



(b)

Figure 6.7 (a) Computer lens display. (Photo by E.H.) (b) Computer output. (Photo courtesy of Optical Research Associates.)

Figure 6.8 Ray geometry.



Inasmuch as $\alpha_1 = y_1/R_1$, this becomes

$$n_{i1}(\alpha_{i1} + y_1/R_1) = n_{t1}(\alpha_{t1} + y_1/R_1).$$

Rearranging terms, we get

$$n_{t1}\alpha_{t1} = n_{i1}\alpha_{i1} - \left(\frac{n_{t1} - n_{i1}}{R_1}\right) y_1,$$

but as we saw in Section 5.7.2, the power of a single refracting surface is

$$\mathcal{P}_1 = \frac{(n_{t1} - n_{i1})}{R_1}.$$

Hence

$$n_{t1}\alpha_{t1} = n_{i1}\alpha_{i1} - \mathcal{P}_1 y_1. \tag{6.12}$$

This is often called the *refraction equation* pertaining to the first interface. Having undergone refraction at point P_1, the ray advances through the homogeneous medium of the lens to point P_2 on the second interface. The height of P_2 can be expressed as

$$y_2 = y_1 + d_{21}\alpha_{t1}, \tag{6.13}$$

on the basis that $\tan \alpha_{t1} \approx \alpha_{t1}$. This is known as the *transfer equation*, because it allows us to follow the ray from P_1 to P_2. Recall that the angles are positive if the ray has a positive slope. Since we are dealing with the paraxial region $d_{21} \approx \overline{V_2 V_1}$ and y_2 is easily computed. Equations (6.11) and (6.12) are then used successively to trace a ray through the entire system. Of course, these are meridional rays and because of the lenses' symmetry about the optical axis, such a ray remains in the same meridional plane throughout its sojourn. The process is two-dimensional; there are two equations and two unknowns, α_{t1} and y_2. In contrast, a skew ray would have to be treated in three dimensions.

6.2.1 Matrix Methods

In the beginning of the 1930s, T. Smith formulated a rather interesting way of handling the ray-tracing equations. The simple linear form of the expressions and the repetitive manner in which they are applied suggested the use of matrices. The processes of refraction and transfer might then be performed mathematically by matrix operators. These initial insights were not widely appreciated for almost thirty years. However, the early 1960s saw a rebirth of interest in this approach, which is now flourishing.* We shall only outline some of the salient features of the method, leaving a more detailed study to the references.

Let's begin by writing the formulas

$$n_{t1}\alpha_{t1} = n_{i1}\alpha_{i1} - \mathcal{P}_1 y_{i1} \tag{6.14}$$

and

$$y_{t1} = 0 + y_{i1}, \tag{6.15}$$

* For further reading see K. Hallbach, "Matrix Representation of Gaussian Optics." *Am. J. Phys.* **32**, 90 (1964); W. Brouwer, *Matrix Methods in Optical Instrument Design*; E. L. O'Neill, *Introduction to Statistical Optics*; or A. Nussbaum, *Geometric Optics*.

which are not very insightful, since we merely replaced y_1 in Eq. (6.12) by the symbol y_{i1} and then let $y_{t1} = y_{i1}$. This last bit of business is for purely cosmetic purposes, as you will see in a moment. In effect, it simply says that the height of reference point P_1 above the axis in the incident medium (y_{i1}) equals its height in the transmitting medium (y_{t1})—which is obvious. But now the pair of equations can be recast in matrix form as

$$\begin{bmatrix} n_{t1}\alpha_{t1} \\ y_{t1} \end{bmatrix} = \begin{bmatrix} 1 & -\mathscr{D}_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} n_{i1}\alpha_{i1} \\ y_{i1} \end{bmatrix}. \tag{6.16}$$

This could equally well be written as

$$\begin{bmatrix} \alpha_{t1} \\ y_{t1} \end{bmatrix} = \begin{bmatrix} n_{i1}/n_{t1} & -\mathscr{D}_1/n_{t1} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{i1} \\ y_{i1} \end{bmatrix}, \tag{6.17}$$

so that the precise form of the 2×1 column matrices is actually a matter of preference. In any case, these can be envisioned as rays on either side of P_1, one before and the other after refraction. Accordingly, using $\mathscr{z}_{t1}$ and $\mathscr{z}_{i1}$ for the two rays, we can write

$$\mathscr{z}_{t1} \equiv \begin{bmatrix} n_{t1}\alpha_{t1} \\ y_{t1} \end{bmatrix} \quad \text{and} \quad \mathscr{z}_{i1} \equiv \begin{bmatrix} n_{i1}\alpha_{i1} \\ y_{i1} \end{bmatrix}. \tag{6.18}$$

The 2×2 matrix is the **refraction matrix**, denoted as

$$\mathscr{R}_1 \equiv \begin{bmatrix} 1 & -\mathscr{D}_1 \\ 0 & 1 \end{bmatrix}, \tag{6.19}$$

so Eq. (6.16) can be concisely stated as

$$\mathscr{z}_{t1} = \mathscr{R}_1 \mathscr{z}_{i1}, \tag{6.20}$$

which just says that $\mathscr{R}_1$ transforms the ray $\mathscr{z}_{i1}$ into the ray $\mathscr{z}_{t1}$ during refraction at the first interface. From Fig. 6.8 we have $n_{i2}\alpha_{i2} = n_{t1}\alpha_{t1}$, that is,

$$n_{i2}\alpha_{i2} = n_{t1}\alpha_{t1} + 0 \tag{6.21}$$

and

$$y_{i2} = d_{21}\alpha_{t1} + y_{t1}, \tag{6.22}$$

where $n_{i2} = n_{t1}$, $\alpha_{i2} = \alpha_{t1}$, and use was made of Eq. (6.13), with y_2 rewritten as y_{i2} to make things pretty. Thus

$$\begin{bmatrix} n_{i2}\alpha_{i2} \\ y_{i2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ d_{21}/n_{t1} & 1 \end{bmatrix} \begin{bmatrix} n_{t1}\alpha_{t1} \\ y_{t1} \end{bmatrix}. \tag{6.23}$$

The **transfer matrix**

$$\mathscr{T}_{21} \equiv \begin{bmatrix} 1 & 0 \\ d_{21}/n_{t1} & 1 \end{bmatrix} \tag{6.24}$$

takes the transmitted ray at P_1 (i.e., $\mathscr{z}_{t1}$) and transforms it into the incident ray at P_2:

$$\mathscr{z}_{i2} \equiv \begin{bmatrix} n_{i2}\alpha_{i2} \\ y_{i2} \end{bmatrix}.$$

Hence Eqs. (6.21) and (6.22) become simply

$$\mathscr{z}_{i2} = \mathscr{T}_{21}\mathscr{z}_{t1}. \tag{6.25}$$

If we make use of Eq. (6.20), this becomes

$$\mathscr{z}_{i2} = \mathscr{T}_{21}\mathscr{R}_1\mathscr{z}_{i1}. \tag{6.26}$$

The 2×2 matrix formed by the product of the transfer and refraction matrices $\mathscr{T}_{21}\mathscr{R}_1$ will carry the ray incident at P_1 into the ray incident at P_2. Notice that the determinant of $\mathscr{T}_{21}$, denoted by $|\mathscr{T}_{21}|$, equals 1, that is, $(1)(1) - (0)(d_{21}/n_{t1}) = 1$. Similarly $|\mathscr{R}_1| = 1$, and since the determinant of a matrix product equals the product of the individual determinants, $|\mathscr{T}_{21}\mathscr{R}_1| = 1$. This provides a quick check on the computations. Carrying the procedure through the second interface (Fig. 6.8) of the lens, which has a refraction matrix $\mathscr{R}_2$, it follows that

$$\mathscr{z}_{t2} = \mathscr{R}_2\mathscr{z}_{i2}, \tag{6.27}$$

or from Eq. (6.26)

$$\mathscr{z}_{i2} = \mathscr{R}_2\mathscr{T}_{21}\mathscr{R}_1\mathscr{z}_{i1}. \tag{6.28}$$

The **system matrix** $\mathscr{M}$ is defined as

$$\mathscr{M} = \mathscr{R}_2\mathscr{T}_1\mathscr{R}_1 \tag{6.29}$$

and has the form

$$\mathscr{M} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{6.30}$$

Since

$$\mathscr{M} = \begin{bmatrix} 1 & -\mathscr{D}_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ d_{21}/n_{t1} & 1 \end{bmatrix} \begin{bmatrix} 1 & -\mathscr{D}_1 \\ 0 & 1 \end{bmatrix}$$

or

$$\mathscr{M} = \begin{bmatrix} 1 & -\mathscr{D}_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -\mathscr{D}_1 \\ d_{21}/n_{t1} & -\mathscr{D}_1 d_{21}/n_{t1} + 1 \end{bmatrix},$$

we can write

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 - \mathcal{D}_2 d_{21}/n_{t1} & -\mathcal{D}_1 + (\mathcal{D}_2\mathcal{D}_1 d_{21}/n_{t1}) - \mathcal{D}_2 \\ d_{21}/n_{t1} & -\mathcal{D}_1 d_{21}/n_{t1} + 1 \end{bmatrix}, \tag{6.31}$$

and again $|\mathscr{A}| = 1$ (Problem 6.15). The value of each element in $\mathscr{A}$ is expressed in terms of the physical lens parameters, such as thickness, index, and radii (via $\mathcal{D}$). Thus the cardinal points that are properties of the lens, determined solely by its make-up, should be deducible from $\mathscr{A}$. The system matrix in this case (6.31) transforms an incident ray at the first surface to an emerging ray at the second surface; as a reminder we will write it as $\mathscr{A}_{21}$.

The concept of image formation enters rather directly (Fig. 6.9) after introduction of appropriate object and image planes. Consequently, the first operator $\mathscr{T}_{1O}$ transfers the reference point from the object (i.e., P_O to P_1). The next operator $\mathscr{A}_{21}$ then carries the ray through the lens, and a final transfer $\mathscr{T}_{I2}$ brings it to the image plane (i.e., P_I). Thus the ray at the image point ($\mathfrak{r}_I$) is given by

$$\mathfrak{r}_I = \mathscr{T}_{I2}\mathscr{A}_{21}\mathscr{T}_{1O}\mathfrak{r}_O, \tag{6.32}$$

where $\mathfrak{r}_O$ is the ray at P_O. In component form this is

$$\begin{bmatrix} n_I\alpha_I \\ y_I \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ d_{I2}/n_I & 1 \end{bmatrix}\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ d_{1O}/n_O & 1 \end{bmatrix}\begin{bmatrix} n_O\alpha_O \\ y_O \end{bmatrix}. \tag{6.33}$$

Notice that $\mathscr{T}_{1O}\mathfrak{r}_O = \mathfrak{r}_{i1}$ and that $\mathscr{A}_{21}\mathfrak{r}_{i1} = \mathfrak{r}_{t2}$, hence $\mathscr{T}_{I2}\mathfrak{r}_{t2} = \mathfrak{r}_I$. The subscripts $O, 1, 2, \ldots, I$ correspond to reference points P_O, P_1, P_2, and so on, and subscripts i and t denote the side of the reference point (i.e., whether incident or transmitted). Operation by a refraction matrix will change i to t but not the reference point designation. On the other hand, operation by a transfer matrix obviously does change the latter.

Ordinarily the physical significances of the components of $\mathscr{A}$ are found by expanding out Eq. (6.33), but this is too involved to do here. Instead, let's return

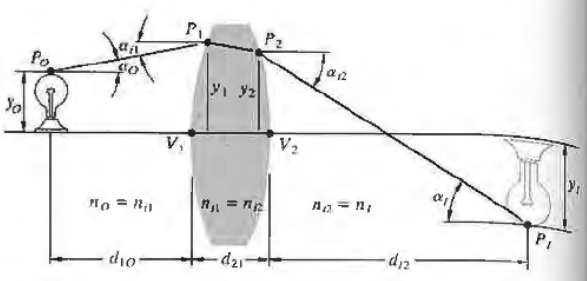


Figure 6.9 Image geometry.

to Eq. (6.31) and examine several of the terms. For example,

$$-a_{12} = \mathcal{D}_1 + \mathcal{D}_2 - \mathcal{D}_2\mathcal{D}_1 d_{21}/n_{t1}.$$

If we suppose, for the sake of simplicity, that the lens is in air, then

$$\mathcal{D}_1 = \frac{n_{t1} - 1}{R_1} \quad \text{and} \quad \mathcal{D}_2 = \frac{n_{t1} - 1}{-R_2},$$

as in Eqs. (5.70) and (5.71). Hence

$$-a_{12} = (n_{t1} - 1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n_{t1} - 1)d_{21}}{R_1R_2n_{t1}}\right].$$

But this is the expression for the focal length of a thick lens (6.2); in other words,

$$a_{12} = -1/f. \tag{6.34}$$

If the imbedding media were different on each side of the lens (Fig. 6.10), this would become

$$a_{12} = -\frac{n_{i1}}{f_o} = -\frac{n_{t2}}{f_i}. \tag{6.35}$$

Similarly it is left as a problem to verify that

$$\overline{V_1H_1} = \frac{n_{i1}(1 - a_{11})}{-a_{12}} \tag{6.36}$$

and

$$\overline{V_2H_2} = \frac{n_{t2}(a_{22} - 1)}{-a_{12}}, \tag{6.37}$$

which locate the principal points.

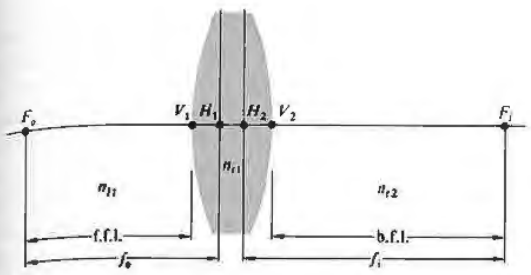


Figure 6.10 Principal planes and focal lengths.

As an example of how the technique can be used, let's apply it, at least in principle, to the Tessar lens* shown in Fig. 6.11. The system matrix has the form

$$\mathscr{A}_{71} = \mathscr{R}_7\mathscr{T}_{76}\mathscr{R}_6\mathscr{T}_{65}\mathscr{R}_5\mathscr{T}_{54}\mathscr{R}_4\mathscr{T}_{43}\mathscr{R}_3\mathscr{T}_{32}\mathscr{R}_2\mathscr{T}_{21}\mathscr{R}_1,$$

where

$$\mathscr{T}_{21} = \begin{bmatrix} 1 & 0 \\ \dfrac{0.357}{1.6116} & 1 \end{bmatrix}, \qquad \mathscr{T}_{32} = \begin{bmatrix} 1 & 0 \\ \dfrac{0.189}{1} & 1 \end{bmatrix},$$

$$\mathscr{T}_{43} = \begin{bmatrix} 1 & 0 \\ \dfrac{0.081}{1.6053} & 1 \end{bmatrix},$$

and so forth. Furthermore,

$$\mathscr{R}_1 = \begin{bmatrix} 1 & -\dfrac{1.6116-1}{1.628} \\ 0 & 1 \end{bmatrix}, \qquad \mathscr{R}_2 = \begin{bmatrix} 1 & -\dfrac{1-1.6116}{-27.57} \\ 0 & 1 \end{bmatrix},$$

$$\mathscr{R}_3 = \begin{bmatrix} 1 & -\dfrac{1.6053-1}{-3.457} \\ 0 & 1 \end{bmatrix},$$

and so on. Multiplying out the matrices, in what is obviously a horrendous although conceptually simple calculation, one presumably will get

$$\mathscr{A}_{71} = \begin{bmatrix} 0.848 & -0.198 \\ 1.338 & 0.867 \end{bmatrix},$$

and from that, $f = 5.06$, $\overline{V_1H_1} = 0.77$, and $\overline{V_7H_2} = -0.67$.

As a last point, it is often convenient to consider a system of thin lenses using the matrix representation. To that end, return to Eq. (6.31). It describes the system matrix for a single lens, and if we let $d_{21} \to 0$, it corresponds to a thin lens. This is equivalent to making $\mathscr{T}_{21}$ a unit matrix, thus

$$\mathscr{A} = \mathscr{R}_2\mathscr{R}_1 = \begin{bmatrix} 1 & -(\mathscr{D}_1 + \mathscr{D}_2) \\ 0 & 1 \end{bmatrix}. \tag{6.38}$$

But as we saw in Section 5.7.2, the power of a thin lens $\mathscr{D}$ is the sum of the powers of its surfaces. Hence

$$\mathscr{A} = \begin{bmatrix} 1 & -\mathscr{D} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1/f \\ 0 & 1 \end{bmatrix}. \tag{6.39}$$

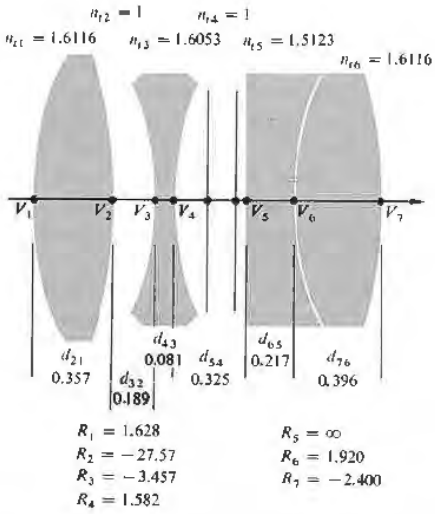


Figure 6.11 A Tessar.

*This particular example was chosen primarily because Nussbaum's book *Geometric Optics* contains a simple Fortran computer program specifically written for this lens. It would be almost silly to evaluate the system matrix by hand. Since Fortran is an easily mastered computer language, the program is well worth further study.

In addition, for two thin lenses separated by a distance d, in air, the system matrix is

$$\mathscr{A} = \begin{bmatrix} 1 & -1/f_2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ d & 1 \end{bmatrix} \begin{bmatrix} 1 & -1/f_1 \\ 0 & 1 \end{bmatrix}$$

or

$$\mathscr{A} = \begin{bmatrix} 1 - d/f_2 & -1/f_1 + d/f_1 f_2 - 1/f_2 \\ d & -d/f_1 + 1 \end{bmatrix}.$$

Clearly then,

$$-a_{12} = \frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2},$$

and from Eqs. (6.36) and (6.37)

$$\overline{O_1 H_1} = fd/f_2, \qquad \overline{O_2 H_2} = -fd/f_1,$$

all of which by now should be quite familiar. Note how easy it would be with this approach to find the focal length and principal points for a compound lens composed of three, four, or more thin lenses.

6.3 ABERRATIONS

To be sure, we already know that first-order theory is no more than a good approximation—an exact ray trace or even measurements performed on a prototype system would certainly reveal inconsistencies with the corresponding paraxial description. Such departures from the idealized conditions of Gaussian optics are known as **aberrations**. There are two main types: **chromatic aberrations** (which arise from the fact that n is actually a function of frequency or color) and **monochromatic aberrations**. The latter occur even with light that is highly monochromatic, and they in turn fall into two subgroupings. There are monochromatic aberrations that deteriorate the image, making it unclear, such as *spherical aberration*, *coma*, and *astigmatism*. In addition, there are aberrations that deform the image, for example, *Petzval field curvature* and *distortion*.

We have known all along that spherical surfaces in general would yield perfect imagery only in the paraxial region. Now we must determine the kind and extent of deviations that result simply from using those surfaces with finite apertures. By the judicious manipulation of a system's physical parameters (e.g., the powers, shapes, thicknesses, glass types, and separations of the lenses, as well as the locations of stops), these aberrations can indeed be minimized. In effect, one cancels out the most undesirable faults by a slight change in the shape of a lens here or a shift in the position of a stop there (very much like trimming up a circuit with small variable capacitors, coils, and pots). When it's all finished, the unwanted deformations of the wavefront incurred as it passes through one surface will, it is hoped, be negated as it traverses some other surfaces further down the line.

As early as 1950 ray-tracing programs were being developed for the new digital computers, and by 1954 efforts were already under way to create lens-designing software. In the early 1960s computerized lens design was a tool of the trade used by manufacturers worldwide. Today there are elaborate computer programs for "automatically" designing and analyzing the performance of all sorts of complicated optical systems. Broadly speaking, you give the computer a quality factor (or merit function) of some sort to aim for (i.e., you essentially tell it how much of each aberration you are willing to tolerate). Then you give it a roughly designed system (e.g., some Tessar configuration), which in the first approximation meets the particular requirements. Along with that, you feed in whatever parameters must be held constant, such as a given *f-number*, focal length, or lens diameter, the field of view, or magnification. The computer will then trace several rays through the system and evaluate the image errors. Having been given leave to vary, say, the curvatures and axial separations of the elements, it will calculate the optimum effect of such changes on the quality factor, make them, and then reevaluate. After a number of iterations, it will have changed the initial configuration so that it now meets the specified limits on aberrations. The final lens design will still be a Tessar, but not the original one. The result is, if you will, an *optimum configuration* but probably not *the* optimum. We can be fairly certain that all aberrations cannot be made exactly zero in any real system comprising spherical surfaces. Moreover, there is no currently known way to determine how close to zero we can actually come. A quality factor is somewhat like a crater-pocked surface in a multidimensional