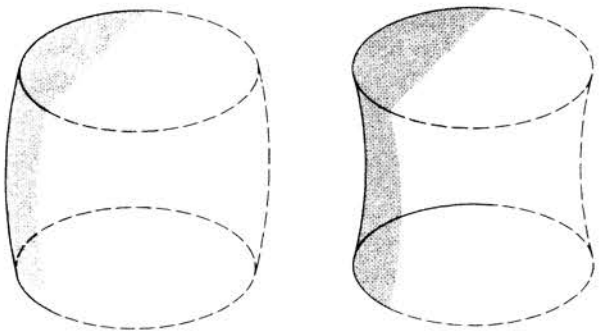


Figure 5.89 Toric surfaces.

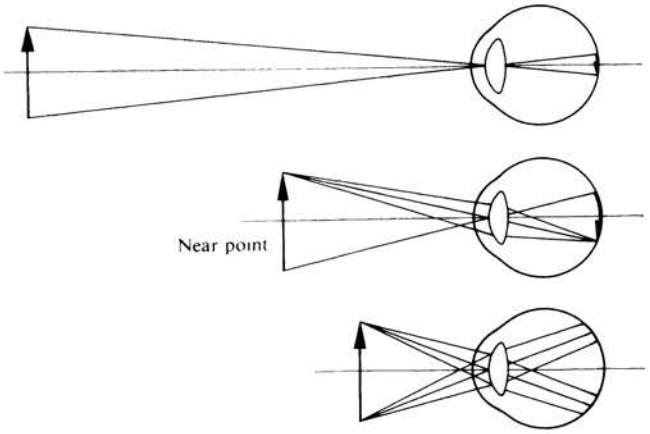


Figure 5.90 Images in relation to the near point.

motion pictures, where an extra-large horizontal field of view is compacted onto the regular film format. When shown through a special lens, the distorted picture spreads out again. On occasion a television station will show short excerpts without the special lens—you may have seen the weirdly elongated result.

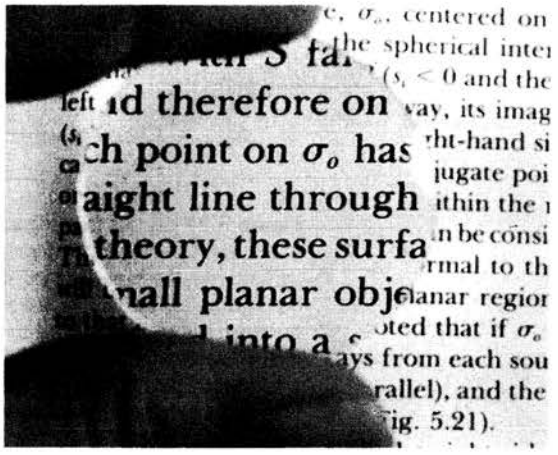
5.7.3 The Magnifying Glass

An observer can cause an object to appear larger, for the purpose of examining it in detail, by simply bringing it closer to her eye. As the object is brought nearer and nearer, its retinal image increases, remaining in focus until the crystalline lens can no longer provide adequate accommodation. Should the object come closer than this *near point*, the image will blur (Fig. 5.90). A single positive lens can be used, in effect to add refractive power to the eye, so that the object can be brought still closer and yet be in focus. The lens so used is referred to variously as a **magnifying glass**, a *simple magnifier*, or a *simple microscope*. In any event, its function is *to provide an image of a nearby object that is larger than the image seen by the unaided eye* (Fig. 5.91). Devices of this sort have been around for a long time. In fact, a quartz convex lens ($f \approx 10$ cm), which may have served as a magnifier, was unearthed in 1885 among the ruins of the palace of King Sennacherib (705–681 B.C.) of Assyria.

Evidently, it would be desirable for the lens to form a magnified, erect image. Furthermore, the rays entering the normal eye should not be converging. Table 5.3 (p. 163) immediately suggests placing the object within the focal length (i.e., $s_o < f$). The result is shown in Fig. 5.91. Because of the relatively tiny size of the eye's pupil, it will almost certainly always be the aperture stop, and as in Fig. 5.34 (p. 172), it will also be the exit pupil.

The **magnifying power**, *MP*, or equivalently, the **angular magnification**, M_A, of a visual instrument is defined as *the ratio of the size of the retinal image as seen through the instrument over the size of the retinal image as seen by the unaided eye at normal viewing distance*. The latter is generally taken as the distance to the near point, d_o. The ratio of angles α_a and α_u (which are made by chief rays from the top of the object in the instance of the aided and unaided eye, respectively) is equivalent to MP, that is,

$$\mathrm{MP} = \frac{\alpha_a}{\alpha_u} \tag{5.75}$$



A positive lens used as a magnifying glass. (Photo by E.H.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

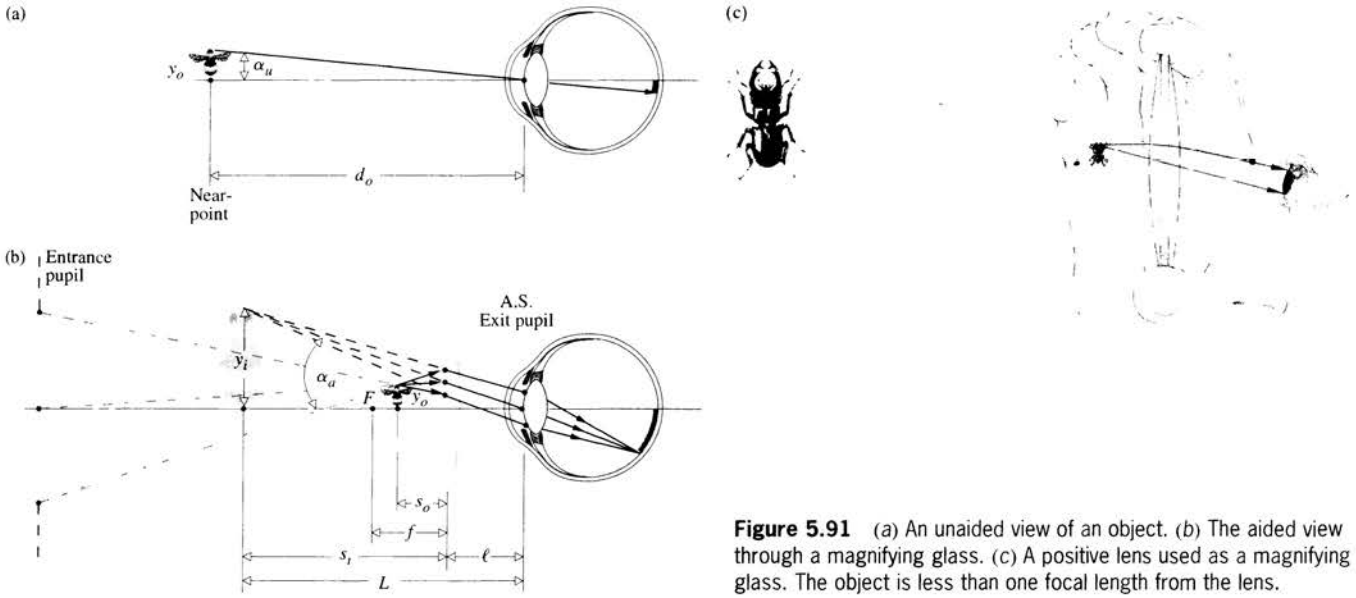


Figure 5.91 (*a*) An unaided view of an object. (*b*) The aided view through a magnifying glass. (*c*) A positive lens used as a magnifying glass. The object is less than one focal length from the lens.

Keeping in mind that we are restricted to the paraxial region, $\tan \alpha_a = y_i/L \approx \alpha_a$ and $\tan \alpha_u = y_o/d_o \approx \alpha_u$, so

$$\text{MP} = \frac{y_i d_o}{y_o L}$$

wherein y_i and y_o are above the axis and positive. If we make d_o and L positive quantities, MP will be positive, which is quite reasonable. When we use Eqs. (5.24) and (5.25) for M_T along with the Gaussian Lens Formula, the expression becomes

$$\text{MP} = -\frac{s_i d_o}{s_o L} = \left(1 - \frac{s_i}{f}\right)\frac{d_o}{L}$$

Inasmuch as the image distance is negative, $s_i = -(L - \ell)$, and

$$\text{MP} = \frac{d_o}{L}[1 + \mathcal{D}(L - \ell)] \tag{5.76}$$

$\mathcal{D}$ of course being the power of the magnifier $(1/f)$. There are three situations of particular interest: (1) When $\ell = f$ the magnifying power equals $d_o\mathcal{D}$. (2) When ℓ is effectively zero,

$$[\text{MP}]_{\ell=0} = d_o\left(\frac{1}{L} + \mathcal{D}\right)$$

In that case the largest value of MP corresponds to the smallest value of L, which, if vision is to be clear, must equal d_o. Thus

$$[\text{MP}]_{\substack{\ell=0 \\ L=d_o}} = d_o\mathcal{D} + 1 \tag{5.77}$$

Taking $d_o = 0.25$ m for the standard observer, we have

$$[\text{MP}]_{\substack{\ell=0 \\ L=d_o}} = 0.25\mathcal{D} + 1 \tag{5.78}$$

As L increases, MP decreases, and similarly as ℓ increases, MP decreases. If the eye is very far from the lens, the retinal image will indeed be small. (3) This last is perhaps the most common situation. Here we position the object at the focal point ($s_o = f$), in which case the virtual image is at infinity ($L = \infty$). Thus from Eq. (5.76)

$$[\text{MP}]_{L=\infty} = d_o\mathcal{D} \tag{5.79}$$

for all practical values of ℓ. Because the rays are parallel, the eye views the scene in a relaxed, unaccommodated configuration, a highly desirable feature. Notice that $M_T = -s_i/s_o$ approaches infinity as $s_o \to f$, whereas in marked contrast, M_A merely decreases by 1 under the same circumstances.

A magnifier with a power of 10 D has a focal length $(1/\mathcal{D})$ of 0.1 m and a MP equal to 2.5 when $L = \infty$. This is conventionally denoted as 2.5×, which means that the retinal image

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

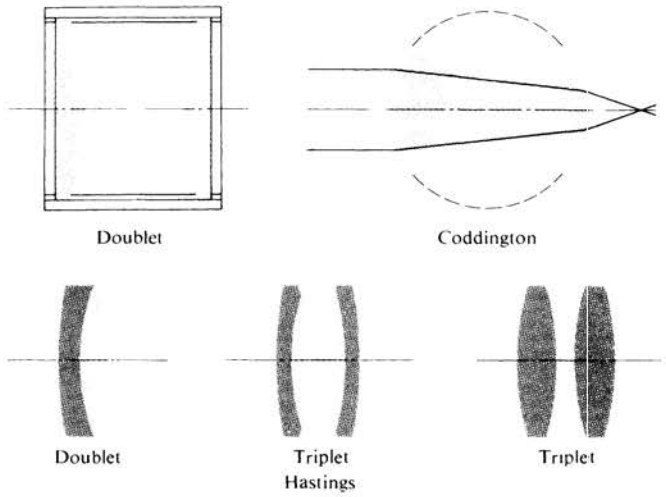


Figure 5.92 Magnifiers.

is 2.5 times larger with the object at the focal length of the lens than it would be were the object at the near point of the unaided eye (where the largest clear image is possible). The simplest single-lens magnifiers are limited by aberrations to roughly 2× or 3×. A large field of view generally implies a large lens; for practical reasons, this usually dictates a fairly small curvature of the surfaces. The radii are large, as is f, and therefore MP is small. The reading glass, the kind Sherlock Holmes made famous, is a typical example. The watchmaker's eye loupe is frequently a single-element lens, also of about 2× or 3×. Figure 5.92 shows a few more complicated magnifiers designed to operate in the range from roughly 10× to 20×. The double lens is quite common in a number of configurations. Although not particularly good, they perform satisfactorily, for example, in high-powered loupes. The Coddington is essentially a sphere with a slot cut in it to allow an aperture smaller than the pupil of the eye. A clear marble (any small sphere of glass qualifies) will also greatly magnify—but not without a good deal of distortion.

The relative refractive index of a lens and the medium in which it is immersed, n_{lm}, is wavelength dependent. But since the focal length of a simple lens varies with $n_{lm}(\lambda)$, this means that f is a function of wavelength, and the constituent colors of white light will focus at different points in space. The resultant defect is known as *chromatic aberration*. In order that the image be free of this coloration, positive and negative lenses made of different glasses are combined to form *achromates* (see Section 6.3.2). Achromatic, cemented, doublet, and triplet lenses are comparatively expensive and are usually found in small, highly corrected, high-power magnifiers.

5.7.4 Eyepieces

The **eyepiece**, or **ocular**, is a visual optical instrument. Fundamentally a magnifier, it views not an actual object but the intermediate image of that object as formed by a preceding lens system. In effect, the eye looks into the ocular, and the ocular looks into the optical system—be it a spotting scope, compound microscope, telescope, or binocular. A single lens could serve the purpose, but poorly. If the retinal image is to be more satisfactory, the ocular cannot have extensive aberrations. The eyepiece of a special instrument, however, might be designed as part of the complete system, so that its lenses can be utilized in the overall scheme to balance out aberrations. Even so, standard eyepieces are used interchangeably on most telescopes and compound microscopes. Moreover, eyepieces are quite difficult to design, and the usual, and perhaps most fruitful, approach is to incorporate or slightly modify one of the existing designs.

The ocular must provide a virtual image (of the intermediate image), most often located at or near infinity, so that it can be comfortably viewed by a normal, relaxed eye. Furthermore, it must position the center of the exit pupil or *eye point* at which the observer's eye is placed at some convenient location, preferably at least 10 mm or so from the last surface. As before, ocular magnification is the product $d_o\mathcal{D}$, or as it is often written, MP = (250 mm)/f.

The **Huygens** ocular, which dates back over 250 years, is still in wide use today (Fig. 5.93), particularly in microscopy. The lens adjacent to the eye is known as the **eye-lens**, and the first lens in the ocular is the **field-lens**. The distance from the eye-lens to the eye point is known as the **eye relief**, and for the

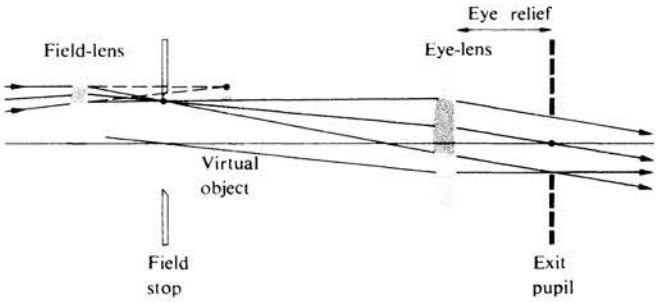


Figure 5.93 The Huygens eyepiece.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

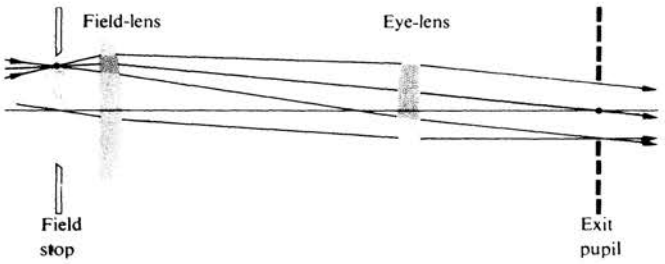


Figure 5.94 The Ramsden eyepiece.

Huygens ocular, it's only an uncomfortable 3 mm or so. Notice that this ocular requires the incoming rays to be converging so as to form a virtual object for the eye-lens. Clearly then, the Huygens eyepiece cannot be used as an ordinary magnifier. Its contemporary appeal rests in its low purchase price (see Section 6.3.2). Another old standby is the **Ramsden** eyepiece (Fig. 5.94). This time the principal focus is in front of the field-lens, so the intermediate image will appear there in easy access. This is where you would place a *reticle* (or *reticule*), which might contain a set of cross hairs, precision scales, or angularly divided circular grids. (When these are formed on a transparent plate, they are often called *graticules*.) Since the reticle and intermediate image are in the same plane, both will be in focus at the same time. The roughly 12-mm eye relief is an advantage over the previous ocular. The Ramsden is relatively popular and fairly inexpensive (see Problem 6.2). The **Kellner** eyepiece represents a definite increase in image quality, although eye relief is between that of the previous two devices. The Kellner is essentially an achromatized Ramsden (Fig. 5.95). It is most commonly used in moderately wide-field telescopic instruments. The **orthoscopic** eyepiece (Fig. 5.96) has a wide field, high magnification, and long eye relief (≈20 mm). The **symmetrical (Plössl)** eyepiece (Fig. 5.97) has characteristics similar to those of the orthoscopic ocular but is generally somewhat superior to it. The **Erfle** (Fig. 5.98) is

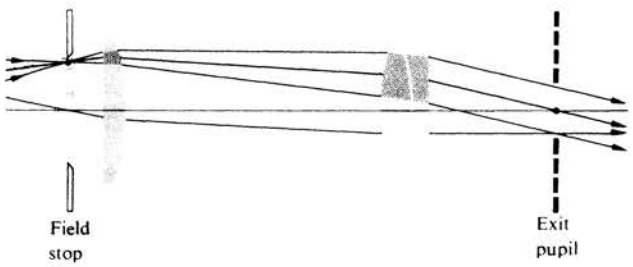


Figure 5.95 The Kellner eyepiece.

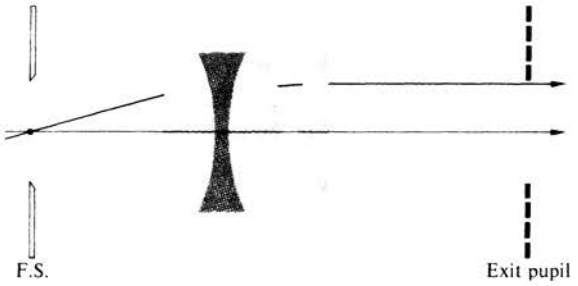


Figure 5.96 The orthoscopic eyepiece.

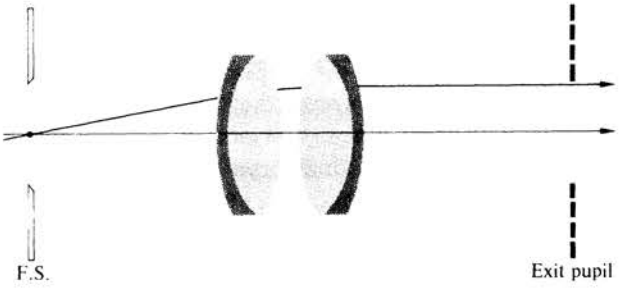


Figure 5.97 The symmetrical (Plössl) eyepiece.

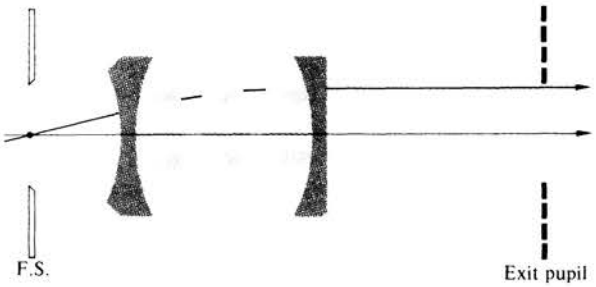


Figure 5.98 The Erfle eyepiece.

probably the most common wide-field (roughly ±30°) eyepiece. It is well corrected for all aberrations and comparatively expensive.*

Although there are many other eyepieces, including variable-power *zoom* devices and ones with aspherical surfaces, those discussed here are representative. They are the devices you will ordinarily find on telescopes and microscopes and on long lists in the commercial catalogs.

*Detailed designs of these and other oculars can be found in the *Military Standardization Handbook—Optical Design*, MIL-HDBK-141.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

5.7.5 The Compound Microscope

The compound microscope goes a step beyond the simple magnifier by providing higher angular magnification (greater than about 30×) of *nearby* objects. Its invention, which may have occurred as early as 1590, is generally attributed to a Dutch spectacle maker, Zacharias Janssen of Middleburg. Galileo runs a close second, having announced his invention of a compound microscope in 1610. A simple version, which is closer to these earliest devices than it is to a modern laboratory microscope, is depicted in Fig. 5.99.

The lens system, here a singlet, closest to the object is referred to as the **objective**. It forms a real, inverted, magnified image of the object. This image resides in space on the plane of the field stop of the eyepiece and has to be small enough to fit inside the barrel of the device. Rays diverging from each point of this image will emerge from the eye-lens (which in this simple case is the eyepiece itself) parallel to each other, as noted in the previous section. The ocular magnifies the intermediate image still further. Thus the magnifying power of the entire system is the product of the transverse linear magnification of the objective, M_{To}, and the angular magnification of the eyepiece, M_{Ae}, that is,

$$\text{MP} = M_{To}M_{Ae} \tag{5.80}$$

The objective magnifies the object and brings it up in the form of a real image, where it can be examined as if through a magnifying glass.

Recall that $M_T = -x_i/f$, Eq. (5.26). With this in mind most, but not all, manufacturers design their microscopes such that the distance (corresponding to x_i) from the second focus of the objective to the first focus of the eyepiece is standardized at 160 mm. This distance, known as the **tube length**, is denoted by L in the figure. (Some authors define tube length as the image distance of the objective.) Hence, with the final image at infinity [Eq. (5.79)] and the standard near point taken as 254 mm (10 inches),

$$\text{MP} = \left(-\frac{160}{f_o}\right)\left(\frac{254}{f_e}\right) \tag{5.81}$$

Here the focal lengths are in millimeters, and the image is inverted (MP < 0). Accordingly, the barrel of an objective with a focal length f_o of, say, 32 mm will be engraved with the marking 5× (or ×5), indicating a power of 5. Combined with a 10× eyepiece (f_e = 1 inch), the microscope MP would then be 50×.

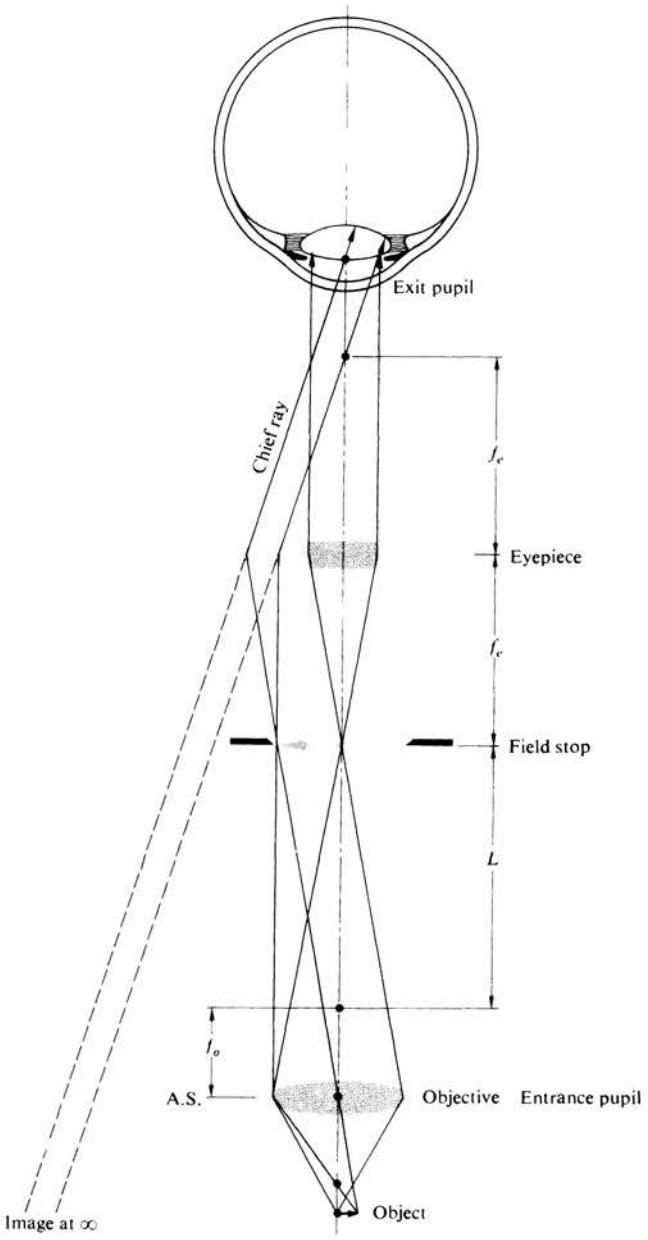


Figure 5.99 A rudimentary compound microscope. The objective forms a real image of a nearby object. The eyepiece, functioning like a magnifying glass, enlarges this intermediate image. The final virtual image can be bigger than the barrel of the device since it needn't fit inside.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

To maintain the distance relationships among the objective, field stop, and ocular, while a focused intermediate image of the object is positioned in the first focal plane of the eyepiece, all three elements are moved as a single unit.

The objective itself functions as the aperture stop and entrance pupil. Its image, formed by the eyepiece, is the exit pupil into which the eye is positioned. The field stop, which limits the extent of the largest object that can be viewed, is fabricated as part of the ocular. The image of the field stop formed by the optical elements following it is called the *exit window*, and the image formed by the optical elements preceding it is the *entrance window*. The cone angle subtended at the center of the exit pupil by the periphery of the exit window is said to be the *angular field of view in image space*.

A modern microscope objective can be roughly classified as one of three different kinds. It might be designed to work best with the object positioned below a cover glass, with no cover glass (metallurgical instruments), or with the object immersed in a liquid that is in contact with the objective. In some cases, the distinction is not critical, and the objective may be used with or without a cover glass. Four representative objectives are shown in Fig. 5.100 (see Section 6.3.1). In addition, the ordinary low-power (about 5×) cemented doublet achromate is quite common. Relatively inexpensive medium-power (10× or 20×) achromatic objectives, because of their short focal lengths, can conveniently be used when expanding and spatially filtering laserbeams.

There is one other characteristic quantity of importance that must be mentioned here, even if only briefly. The brightness of the image is, in part, dependent on the amount of light gathered in by the objective. The *f-number* is a useful parameter for describing this quantity, particularly when the object is a distant one (see Section 5.3.3). However, for an instrument working at *finite conjugates* (s_i and s_o both finite), the numerical aperture, NA, is more appropriate (see Section 5.6). In the present instance

$$\text{NA} = n_i \sin \theta_{\max} \tag{5.82}$$

where n_i is the refractive index of the immersing medium (air, oil, water, etc.) adjacent to the objective lens, and $\theta_{\max}$ is the half-angle of the maximum cone of light picked up by that lens (Fig. 5.100*b*). In other words, $\theta_{\max}$ is the angle made by a marginal ray with the axis. The numerical aperture is usually the second number etched in the barrel of the objective. It ranges from about 0.07 for low-power objectives to 1.4 or so for high-power (100×) ones. Of course, if the object is in the air, the numerical aperture cannot be greater than 1.0. Incidentally, Ernst Abbe (1840–1905), while working in the Carl Zeiss microscope workshop, introduced the concept of the numerical aperture. It was he who recognized that the minimum transverse distance between two object points that can be resolved in the image, that is, the *resolving power*, varied directly as λ and inversely as the NA.

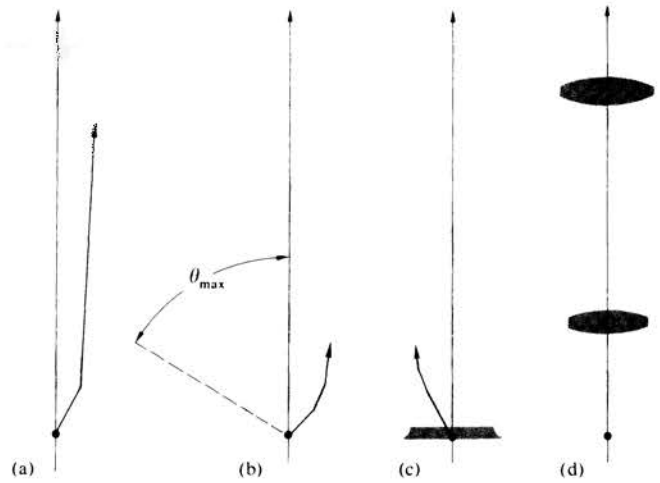


Figure 5.100 Microscope objectives: (*a*) Lister objective, 10×, NA = 0.25, f = 16 mm (two cemented achromates). (*b*) Amici objective, from 20×, NA = 0.5, f = 8 mm to 40×, NA = 0.8, f = 4 mm. (*c*) Oil-immersion objective, 100×, NA = 1.3, f = 1.6 mm (see Figure 6.18). (*d*) Apochromatic objective, 55×, NA = 0.95, f = 3.2 (contains two fluorite lenses).

In summary then, the microscope is a device for enlarging the image of a tiny nearby object. It does this by capturing as much of the emitted light as possible using a short focal length objective (so the magnification will be large) that is held close to the object. That produces a real image, which is further magnified by an eyepiece functioning as a magnifying glass.

5.7.6 The Camera

The prototype of the modern photographic camera* was a device known as the *camera obscura*, the earliest form of which was simply a dark room with a small hole in one wall. Light entering the hole cast an inverted image of the sunlit outside scene on an inside screen. The principle was known to

*See W. H. Price, "The Photographic Lens," *Sci. Am.* **72** (August 1976).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Aristotle, and his observations were preserved by Arab scholars throughout Europe's long Dark Ages. Alhazen utilized it to examine solar eclipses indirectly over eight hundred years ago. The notebooks of Leonardo da Vinci contain several descriptions of the obscura, but the first detailed treatment appears in *Magia naturalis* (*Natural Magic*) by Giovanni della Porta. He recommended it as a drawing aid, a function to which it was soon quite popularly put. Johannes Kepler, the renowned astronomer, had a portable tent version, which he used while surveying in Austria. By the latter part of the 1600s, the small hand-held camera obscura was commonplace. Note that the eye of the Nautilus, a little cuttlefish, is literally an open pinhole obscura that simply fills with seawater on immersion.

By replacing the viewing screen with a photosensitive surface, such as a film plate, the obscura becomes a camera in the modern sense of the word. The first permanent photograph was made in 1826 by Joseph Nicéphore Niépce (1765–1833), who used a box camera with a small convex lens, a sensitized pewter plate, and roughly an eight-hour exposure.

The lensless pinhole camera (Fig. 5.101) is by far the least complicated device for the purpose, yet it has several endearing and, indeed, remarkable virtues. It can form a well-defined, practically undistorted image of objects across an extremely wide angular field (due to great depth of focus) and over a large range of distances (great depth of field). If initially the entrance pupil is very large, no image results. As it is decreased in diameter, the image forms and grows sharper. After a point, further reduction in the hole size causes the image to blur again, and one quickly finds that the aperture size for maximum sharpness is proportional to its distance from the image plane. (A hole with a 0.5-mm diameter at 0.25 m from the film plate is convenient and works well.) There is no focusing of the rays at all, so no defects in that mechanism are responsible for the drop-off in clarity. The problem is actually one of diffraction, as we shall see later on (Section 10.2.5). In most practical situations, the pinhole camera's one overriding drawback is that it is insufferably slow (roughly $f/500$). This means that exposure times will generally be far too long, even with the most sensitive films. The obvious exception is a stationary subject, such as a building (see photo), for which the pinhole camera excels.

Figure 5.102 depicts the essential components of a popular and representative modern camera—the single-lens reflex, or SLR. Light traversing the first few elements of the lens then passes through an iris diaphragm, used in part to control the exposure time or, equivalently, the *f-number*; it is in effect a

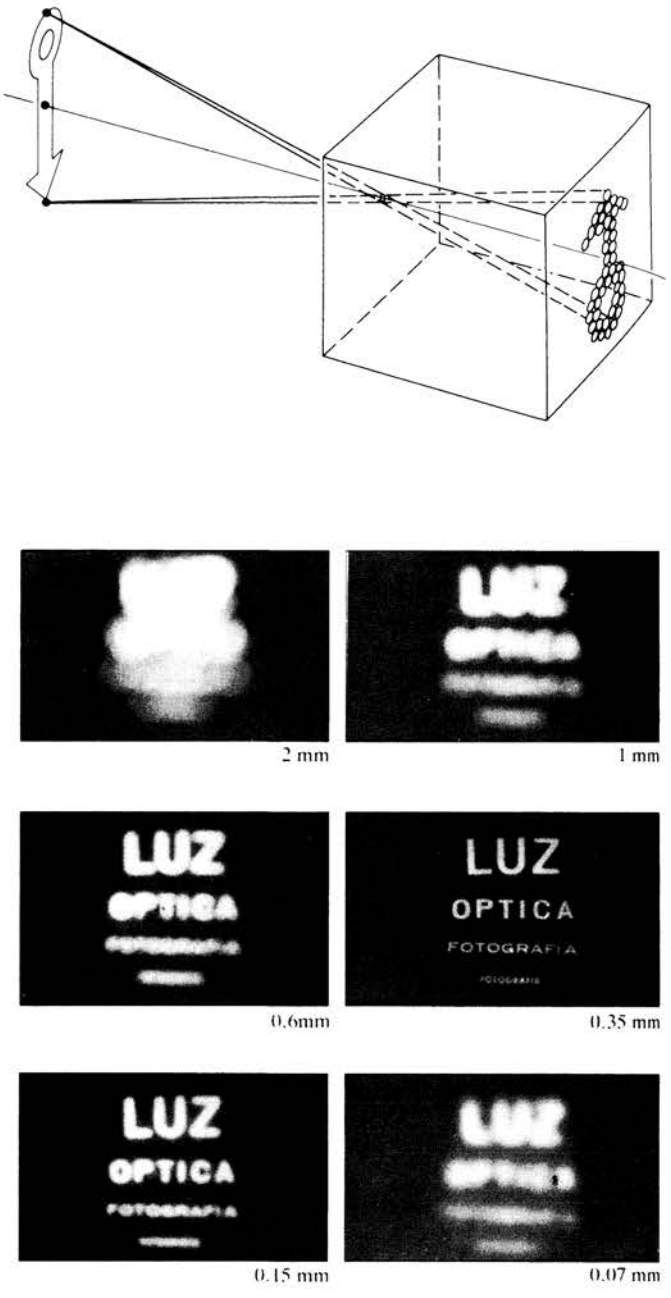
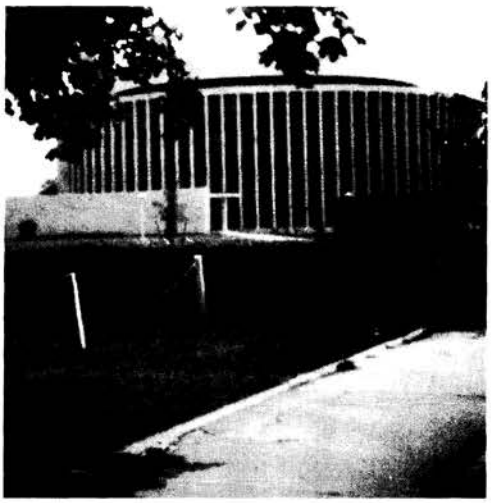


Figure 5.101 The pinhole camera. Note the variation in image clarity as the hole diameter decreases. (Photos courtesy Dr. N. Joel, UNESCO.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



Photograph taken with a pinhole camera (Science Building, Adelphi University). Hole diameter 0.5 mm, film plane distance 24 cm, A.S.A. 3000, shutter speed 0.25 s. Note depth of field. (Photo by E.H.)

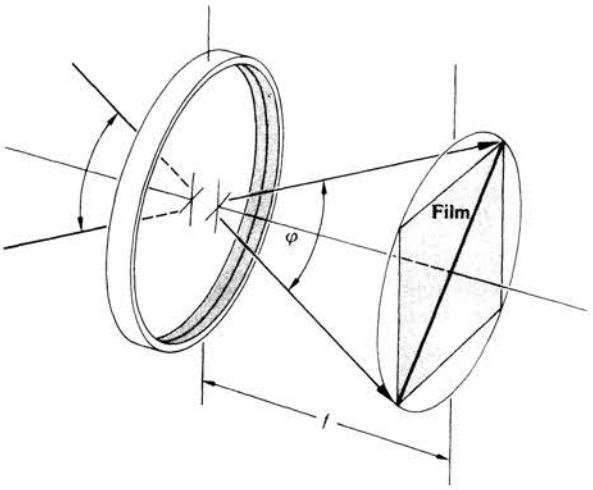


Figure 5.103 Angular field of view when focused at infinity.

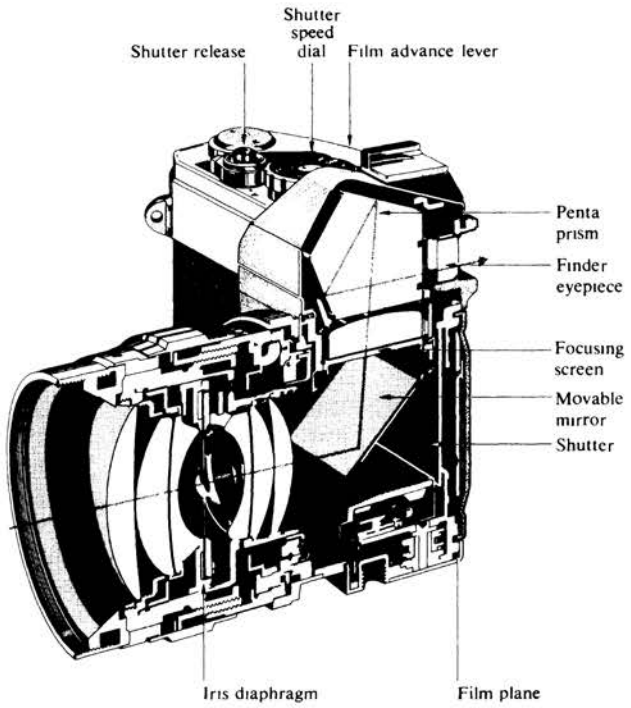


Figure 5.102 A single-lens reflex camera.

variable-aperture stop. On emerging from the lens, light strikes a movable mirror tilted at 45°, then goes up through the focusing screen to the penta prism and out the finder eyepiece. When the shutter release is pressed, the diaphragm closes down to a preset value, the mirror swings up out of the way, and the focal-plane shutter opens, exposing the film. The shutter then closes, the diaphragm opens fully, and the mirror drops back in place. Nowadays SLR systems have any one of a number of built-in light-meter arrangements, which are automatically coupled to the diaphragm and shutter, but those components are excluded from the diagram for the sake of simplicity.

To focus the camera, the entire lens is moved toward or away from the film plane. Since its focal length is fixed, as s_o varies, so too must s_i. The *angular field of view* can loosely be thought of as relating to the fraction of the scene included in the photograph. It is furthermore required that the entire photograph surface correspond to a region of satisfactory image quality. More precisely, the angle subtended at the lens, by a circle encompassing the film area, is the angular field of view φ (Fig. 5.103). As a rough but reasonable approximation of a common arrangement, take the diagonal distance across the film to equal the focal length. Thus $\varphi/2 \approx \tan^{-1} \frac{1}{2}$; that is, $\varphi \approx 53°$. If the object comes in from infinity, s_i must increase. The lens is then backed away from the film plate to keep the image in focus, and the field of view, as recorded on the film whose periphery is the field stop, decreases. A **standard** SLR lens has a focal length in the range of about 50 to 58 mm and a

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

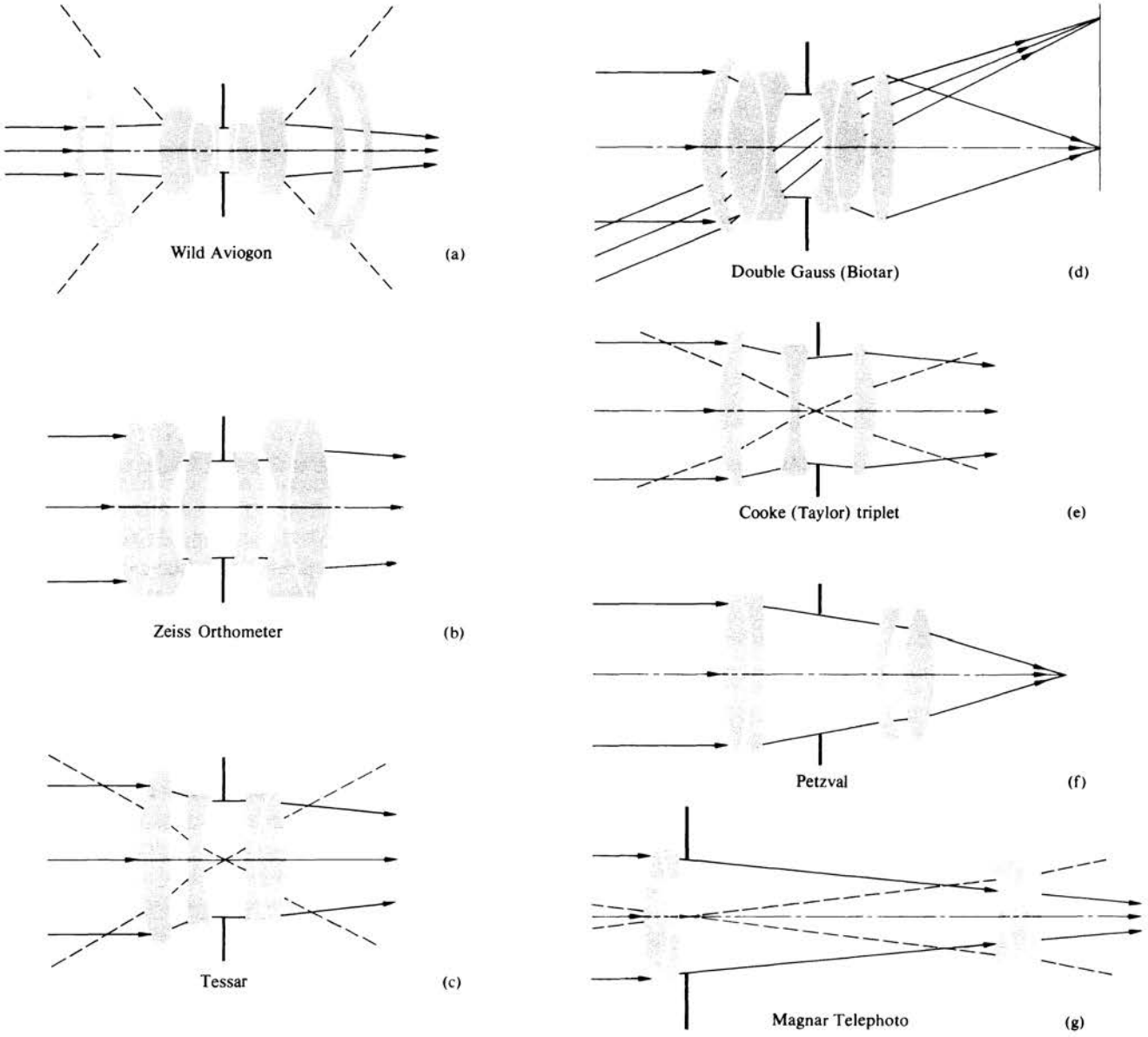


Figure 5.104 Camera lenses.

field of view of 40° to 50°. With the film size kept constant, a reduction of f results in a wider field angle. Accordingly, **wide-angle** SLR lenses range from $f \approx 40$ mm down to about 6 mm, and φ goes from about 50° to a remarkable 220° (the latter being a special-purpose lens wherein distortion is unavoidable). The *telephoto* has a long focal length, roughly 80 mm or more. Consequently, its field of view drops off rapidly, until it is only a few degrees at $f \approx 1000$ mm.

The standard photographic objective must have a large relative aperture, $1/(f/\#)$, to keep exposure times short. Moreover, the image is required to be flat and undistorted, and the lens should have a wide angular field of view as well. The evolution of a modern lens still begins with a creative insight that leads to a promising new form. In the past, these were laboriously perfected relying on intuition, experience, and, of course trial and error with a succession of developmental lenses.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Today, for the most part, the computer serves this function without the need of numerous prototypes.

Many contemporary photographic objectives are variations of well-known successful forms. Figure 5.104 illustrates the general configuration of several important lenses, roughly progressing from wide angle to telephoto. Particular specifications are not given, because variations are numerous. The *Aviogon* and *Zeiss Orthometer* are wide-angle lenses, whereas the *Tessar* and *Biotar* are often standard lenses. The *Cooke triplet*, described in 1893 by H. Dennis Taylor of Cooke and Sons, is still being made (note the similarity with the Tessar). It contains the smallest number of elements by which all seven third-order aberrations can essentially be made to vanish. Even earlier (ca. 1840), Josef Max Petzval designed what was then a rapid (portrait) lens for Voightländer and Son. Its modern offshoots are myriad.

5.7.7 The Telescope

It is not at all clear who actually invented the telescope. In point of fact, it was probably invented and reinvented many times. Recall that by the seventeenth century spectacle lenses had been in use in Europe for about three hundred years. During that long span of time, the fortuitous juxtapositioning of two appropriate lenses to form a telescope seems almost inevitable. In any event, it is most likely that a Dutch optician, possibly even the ubiquitous Zacharias Janssen of microscope fame, first constructed a telescope and in addition had inklings of the value of what he was peering into. The earliest indisputable evidence of the discovery, however, dates to October 2, 1608, when Hans Lippershey petitioned the States-General of Holland for a patent on a device for seeing at a distance (which is what *teleskopos* means in Greek). As you might have guessed, its military possibilities were immediately recognized. His patent was therefore not granted; instead the government purchased the rights to the instrument, and he received a commission to continue research. Galileo heard of this work, and by 1609 he had fashioned a telescope of his own, using two lenses and an organ pipe as a tube. It was not long before he had constructed a number of greatly improved instruments and was astounding the world with the astronomical discoveries for which he is famous.

Refracting Telescopes

A simple **astronomical telescope** is shown in Fig. 5.105. Unlike the compound microscope, which it closely resembles, its primary function is to enlarge the retinal image of a *distant*

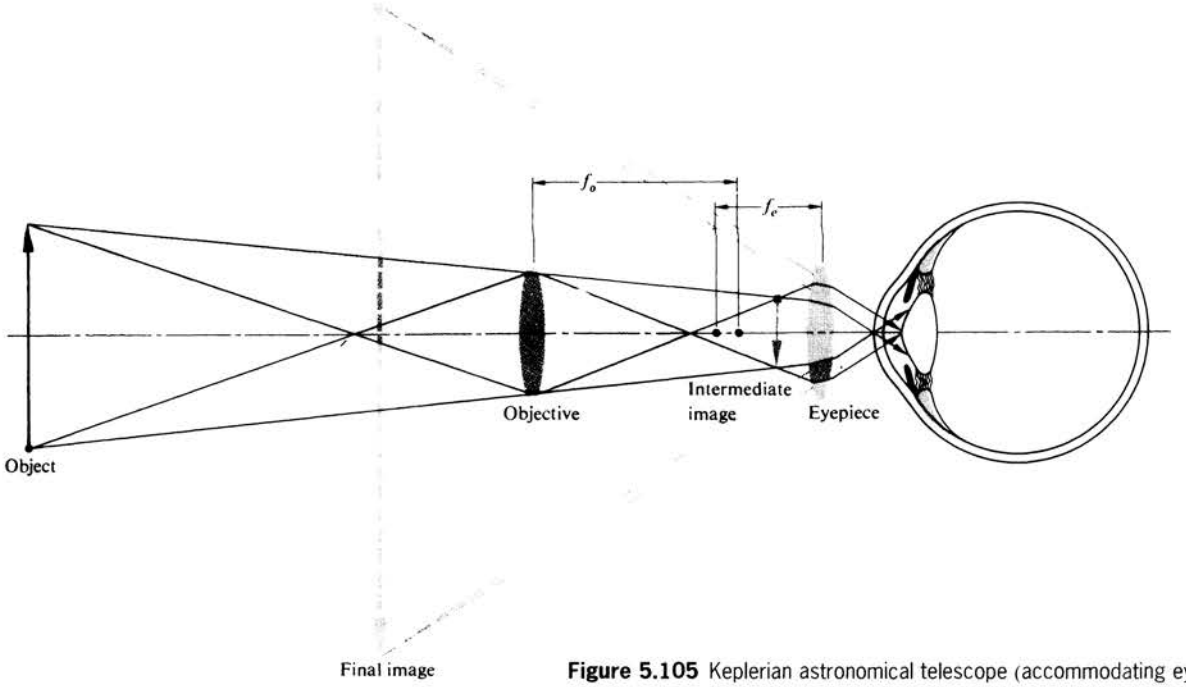


Figure 5.105 Keplerian astronomical telescope (accommodating eye).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

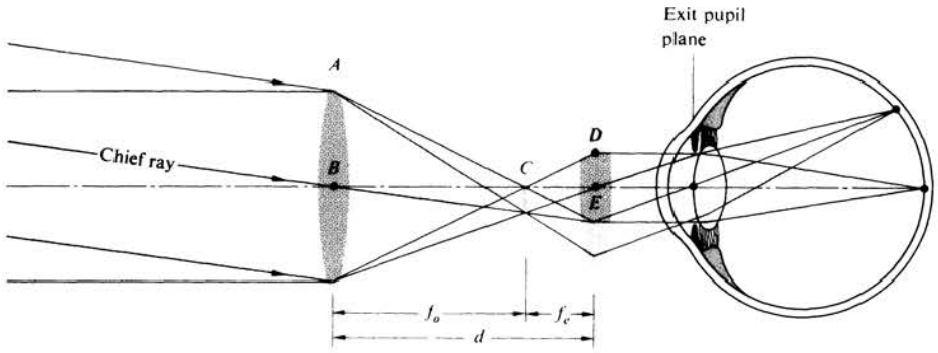


Figure 5.106 Astronomical telescope—infinite conjugates.

object. In the illustration, the object is at a finite far distance from the objective, so that the real intermediate image is formed just beyond its second focal point. This image will be the object for the next lens system, that is, the ocular. It follows from Table 5.3 (p. 163) that if the eyepiece is to form a virtual magnified final image (within the range of normal accommodation), the object distance must be less than or equal to the focal length, f_e. In practice, *the position of the intermediate image is fixed, and only the eyepiece is moved to focus the instrument.* Notice that *the final image is inverted*, but as long as the scope is used for astronomical observations, this is of little consequence, especially since most work is photographic.

At great object distances the incident rays are effectively parallel—the intermediate image resides at the second focus of the objective. Usually, the eyepiece is located so that its first focus overlaps the second focus of the objective, in which case rays diverging from a point on the intermediate image will leave the ocular parallel to each other. A normal viewing eye can then focus the rays in a relaxed configuration. If the eye is nearsighted or farsighted, the ocular can be moved in or out so that the rays diverge or converge a bit to compensate. (If you are astigmatic, you'll have to keep your glasses on when using ordinary visual instruments.) We saw earlier (Section 5.2.3) that both the back and front focal lengths of a thin-lens combination go to infinity when the two lenses are separated by a distance d equal to the sum of their focal lengths (Fig. 5.106). The astronomical telescope in this configuration of *infinite conjugates* is said to be *afocal*, that is, without a focal length. As a side note, if you shine a collimated (parallel rays, i.e., plane waves) narrow laserbeam into the back end of a scope focused at infinity, it will emerge still collimated but with an increased cross section. It is often desirable to have a broad, quasimonochromatic, plane-wave beam, and specific devices of this sort are now available commercially.

The periphery of the objective is the aperture stop, and it encompasses the entrance pupil as well, there being no lenses to the left of it. If the telescope is trained directly on some distant galaxy, the visual axis of the eye will presumably be colinear with the central axis of the scope. The entrance pupil of the eye should then coincide in space with the exit pupil of the scope. However, the eye is not immobile. It will move about scanning the entire field of view, which quite often contains many points of interest. In effect, the eye examines different regions of the field by rotating so that rays from a particular area fall on the fovea centralis. The direction established by the chief ray through the center of the entrance pupil to the fovea centralis is the *primary line-of-sight*. The axial point, fixed in reference to the head, through which the primary line-of-sight always passes, regardless of the orientation of the eyeball, is called the *sighting intersect*. When it is desirable to have the eye surveying the field, the sighting intersect should be positioned at the center of the telescope's exit pupil. In that case, the primary line-of-sight will always correspond to a chief ray through the center of the exit pupil, however the eye moves.

Suppose that the margin of the visible object subtends a half-angle of α at the objective (Fig. 5.107). This is essentially the same as the angle α_u, which would be subtended at the unaided eye. As in previous sections, the angular magnification is

$$\mathrm{MP} = \frac{\alpha_a}{\alpha_u} \qquad [5.75]$$

Here α_u and α_a are measures of the field of view in object and image space, respectively. The first is the half-angle of the actual cone of rays collected, and the second relates to the apparent cone of rays. If a ray arrives at the objective with a negative slope, it will enter the eye with a positive slope and vice versa. To make the sign of MP positive for erect images, and therefore consistent with previous usage (Fig. 5.91), either

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

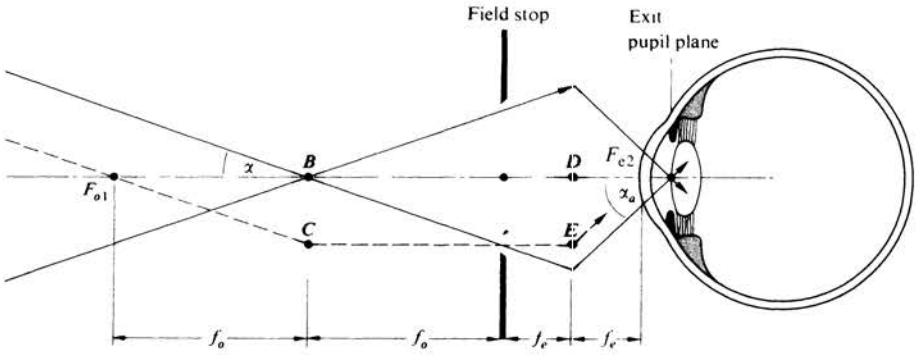


Figure 5.107 Ray angles for a telescope.

α_u or α_a must be taken to be negative—we choose the former because the ray has a negative slope. Observe that the ray passing through the first focus of the objective passes through the second focus of the eyepiece; that is, F_{o1} and F_{e2} are conjugate points. In the paraxial approximation $\alpha \approx \alpha_u \approx \tan \alpha_u$ and $\alpha_a \approx \tan \alpha_a$. The image fills the region of the field stop, and half its extent equals the distance $\overline{BC} = \overline{DE}$. Thus, from triangles $F_{o1}BC$ and $F_{e2}DE$, the ratio of the tangents yields

$$\text{MP} = -\frac{f_o}{f_e} \tag{5.83}$$

Another convenient expression for the MP comes from considering the transverse magnification of the ocular. Inasmuch as the exit pupil is the image of the objective (Fig. 5.107), we have

$$M_{Te} = -\frac{f_e}{x_o} = -\frac{f_e}{f_o}$$

Furthermore, if D_o is the *diameter of the objective* and D_{ep} is the *diameter of its image, the exit pupil*, then $M_{Te} = D_{ep}/D_o$. These two expressions for M_{Te} compared with Eq. (5.83) yield

$$\text{MP} = \frac{D_o}{D_{ep}} \tag{5.84}$$

Here D_{ep} is actually a negative quantity, since the image is inverted. It is an easy matter to build a simple refracting scope by holding a lens with a long focal length in front of one with a short focal length and making sure that $d = f_o + f_e$. But again, well-corrected telescopic instruments generally have multi-element objectives, usually doublets or triplets.

To be useful when the orientation of the object is of importance, a scope must contain an additional **erecting system**; such an arrangement is known as a **terrestrial telescope**. A single erecting lens or lens system is usually located between the ocular and objective, with the result that the image is right-side-up. Figure 5.108 shows one with a cemented doublet objective and a Kellner eyepiece. It will obviously have to have a long draw tube, the picturesque kind that comes to mind when you think of wooden ships and cannonballs.

For that reason, **binoculars** (binocular telescopes) generally utilize erecting prisms, which accomplish the same thing in less space and also increase the separation of the objectives, thereby enhancing the stereoscopic effect. Most often these are double Porro prisms, as in Fig. 5.109. (Notice the involved modified Erfle eyepiece, the wide field stop, and the achromatic doublet objective.) Binoculars customarily bear several numerical markings, for example, 6×30, 7×50, or 20×50.

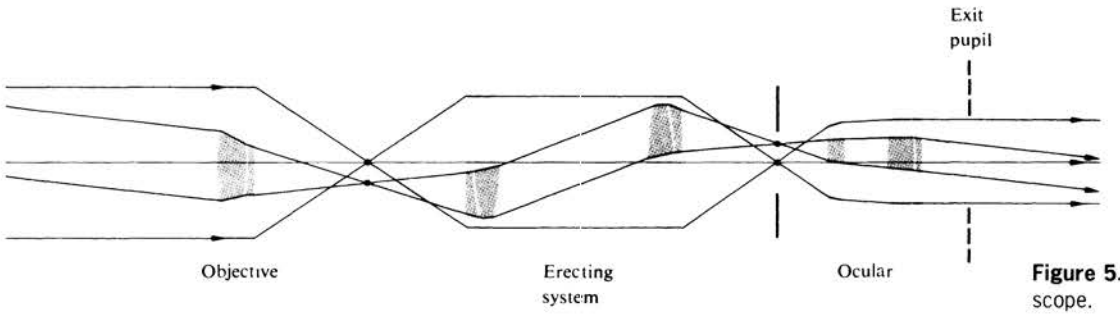


Figure 5.108 A terrestrial telescope.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

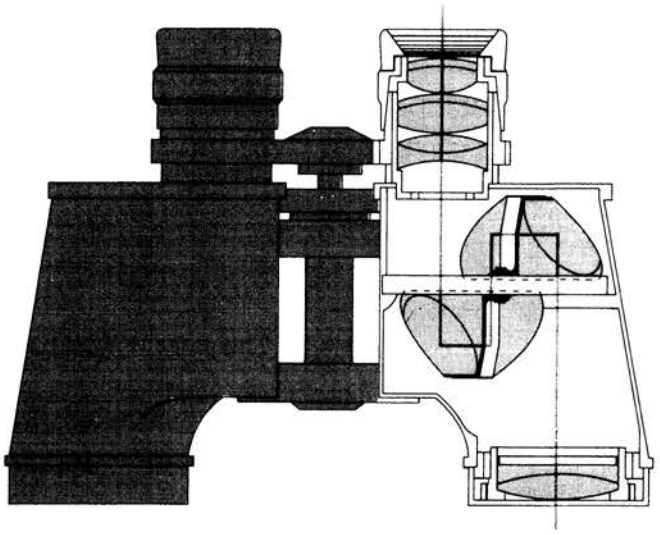


Figure 5.109 A binocular.

The initial number is the magnification, here 6×, 7×, or 20×. The second number is the entrance-pupil diameter or, equivalently, the clear aperture of the objective, expressed in millimeters. It follows from Eq. (5.84) that the exit-pupil diameter will be the second number divided by the first, or in this case 5, 7.1, and 2.5, all in millimeters. You can hold the instrument away from your eye and see the bright circular exit pupil surrounded by blackness. To measure it, focus the device at infinity, point it at the sky, and observe the emerging sharp disk of light, using a piece of paper as a screen. Determine the eye relief while you're at it.

By the way, as long as $d = f_o + f_e$, the scope will be afocal, even if the eyepiece is negative (i.e., $f_e < 0$). The telescope built by Galileo (Fig. 5.110) had just such a negative lens as an eyepiece and therefore formed an erect image [$f_e < 0$ and MP > 0 in Eq. (5.83)]. As a telescope, the system is now mainly of historical and pedagogical interest, although one can still purchase two such scopes mounted side by side to form a Galilean field glass. It is quite useful, however, as a laserbeam expander because it has no internal focal points where a high-power beam would otherwise ionize the surrounding air.

Reflecting Telescopes

Put rather simply, a telescope should allow us to see things *clearly* that are far away, and often extremely *faint*. We need to be able to resolve fine details, that is, to distinguish separate features that are small and close together, such as the two stars in a binary system. A spy satellite that can spot people walking around is highly desirable, but one that will identify their military service from the markings on their uniforms is even better. The measure of that ability is the **resolution**, and it increases with the diameter (D) of the aperture admitting light into the system. All other things being equal (under ideal seeing conditions), a large-diameter telescope will have better resolution than a small-diameter telescope. There's another even more compelling reason to increase the size of the aperture: to improve the **light-gathering power**. A telescope with a large aperture will be able to collect more light and see fainter, more distant objects than an otherwise identical but smaller one.

The difficulties inherent in making large lenses are underscored by the fact that the largest refracting instrument is the 40-inch Yerkes telescope in Williams Bay, Wisconsin, whereas the reflector on Mount Palomar in southwestern California is 200 inches in diameter. The problems are evident; a lens must be transparent and free of internal flaws

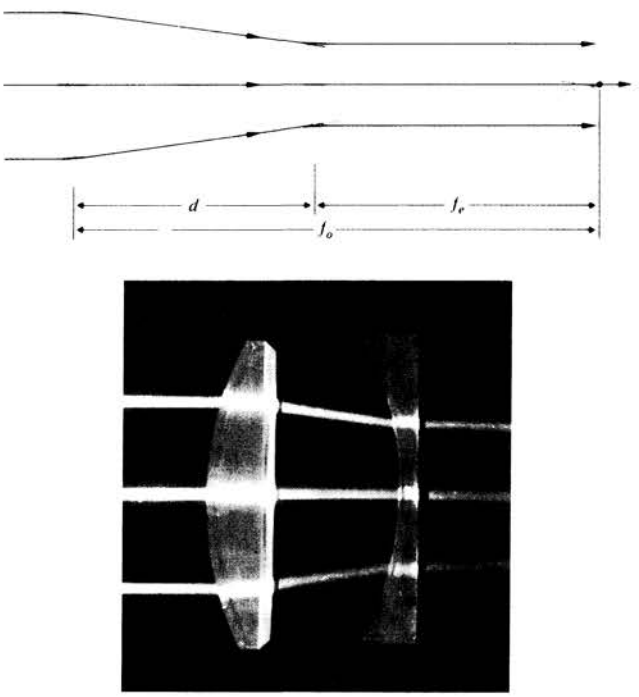


Figure 5.110 The Galilean telescope. Galileo's first scope had a planar-convex objective (5.6 cm in diameter, f = 1.7 m, R = 93.5 cm) and a planar-concave eyepiece, both of which he ground himself. It was 3× in contrast to his last scope, which was 32×. (Photo by E.H.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

such as bubbles. A front-surfaced mirror obviously need not be; indeed, it need not even be transparent. A lens can be supported only by its rim and may sag under its own weight; a mirror can be supported by its rim and back as well. Furthermore, since there is no refraction and therefore no effect on the focal length due to the wavelength dependence of the index, mirrors suffer no chromatic aberration. For these and other reasons (e.g., their frequency response), reflectors predominate in large telescopes.

Invented by the Scotsman James Gregory (1638–1675) in 1661, the reflecting telescope was first successfully constructed by Newton in 1668 and only became an important research tool in the hands of William Herschel a century later. Figure 5.111 depicts a number of reflector arrangements, each having a concave paraboloidal primary mirror. The venerable 200-inch Hale telescope is so large that a little enclosure, where an observer can sit, is positioned at the prime focus (Fig. 5.111*a*). In the Newtonian version (Fig. 5.111*b*) a plane mirror or prism brings the beam out at right angles to the axis of the scope, where it can be photographed, viewed, spectrally analyzed, or photoelectrically processed. In the classical Gregorian arrangement (Fig. 5.111*c*), which is not particularly popular, a concave ellipsoidal secondary mirror reinverts the image, returning the beam through a hole in the primary. The classical Cassegrain system (Fig. 5.111*d*) utilizes a convex hyperboloidal secondary mirror to increase the effective focal length (refer back to Fig. 5.47, p. 180). It functions as if the primary mirror had the same aperture but a larger focal length or radius of curvature.

The simple single-mirror paraboloidal telescope (Fig. 5.111*a*) was designed to function with rays entering along its optical axis. But there will always be objects of interest elsewhere in the field of view other than at its direct center. When a parallel bundle of off-axis rays are reflected by a paraboloid, they do not all meet at the same point. The image of a distant off-axis point (e.g., a star) is an off-axis asymmetric blur caused by the combined aberrations of *coma* (p. 258) and *astigmatism* (p. 261). This blurring becomes unacceptable rather quickly as the object moves off-axis; that's especially true for the contribution due to coma, and it ends up limiting the acceptable field of view to something quite narrow. Even for a slow $f/10$ system, the angular radius of the acceptable field of view is only about 9 arc-minutes off-axis, and it drops to a mere 1.4 arc-minutes at $f/4$. The classical two-mirror telescopes (Figs. 5.111*b*, *c*, and *d*) are similarly severely limited in their fields of view by coma.

Incidentally, if we put a liquid such as mercury in a shallow horizontal basin and continuously rotate it about a vertical axis at a constant rate ω, the equilibrium configuration of the surface will be parabolic. The elevation (z) above the lowest point in the liquid at any location on the surface is given by

$$z = \frac{\omega^2 r^2}{2g} \tag{5.85}$$

Large (upwards of 3 meters in diameter), robust, diffraction limited, liquid mirrors have been produced. The main advantage of a liquid telescope mirror over a glass one is that it's

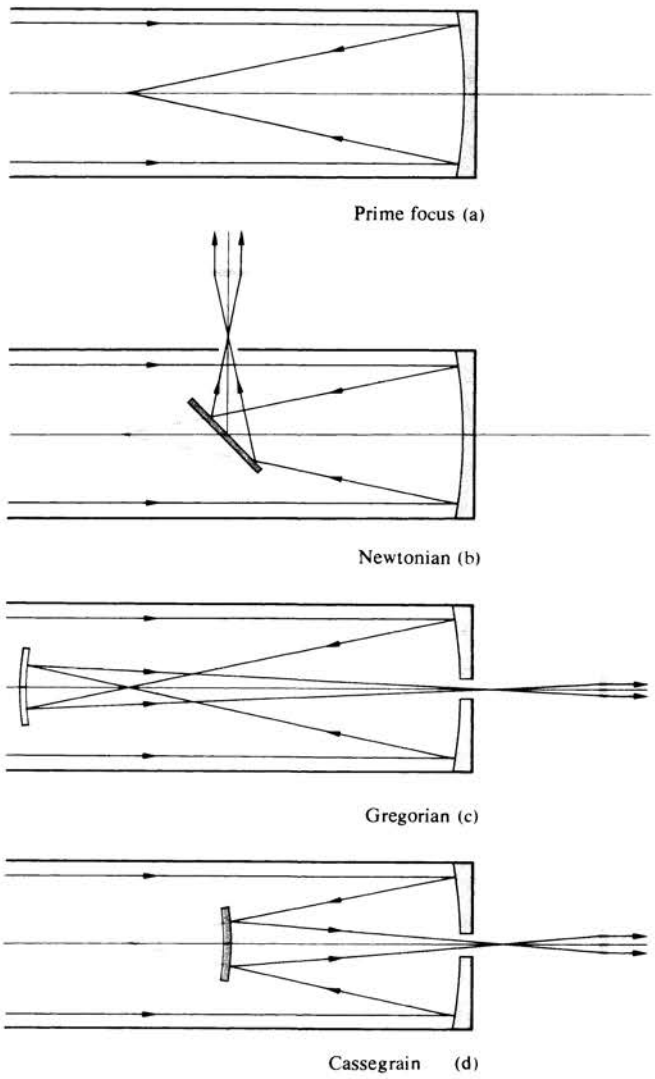
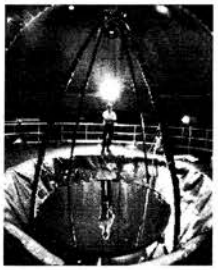


Figure 5.111 Reflecting telescopes.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



The rotating 3-meter diameter Liquid Mirror Telescope in New Mexico is used by NASA to detect chunks of low-Earth-orbit space debris as small as 5 cm.

very much less expensive. The main disadvantage is that it can only look straight up (see photo).

Aplanatic Reflectors

An optical system that has negligible amounts of both spherical aberration (p. 253) and coma is called an **aplanat**, and there are aplanatic versions of both the Cassegrain and the Gregorian scopes. The Ritchey-Chrétien telescope is an aplanatic Cassegrain having a hyperboloidal primary and secondary. In recent times, this configuration has become the leading choice for devices with apertures of 2 m or more. Perhaps the best known example of its kind is the 2.4-m Hubble Space Telescope (HST), pictured in Fig. 5.112. Only telescopes in space (i.e., above the absorbing atmosphere) can work efficiently in the ultraviolet—which, for example, is where one would like to examine hot young stars. With its newly installed (1993) charge-coupled-device (CCD) arrays, the HST can "see" from about 1 μm in the IR to 121.6 nm in the UV. This complements ground-based telescopes that can provide diffraction-limited imaging in the wavelength range greater than 10 μm. (Incidentally, CCDs have a sensitivity about 50 times greater than otherwise comparable photographic film; the era of dropping film packs out of spy satellites is over.)

With little or no coma, the field of view of the Ritchey-Chrétien is limited by astigmatism. Thus an $f/10$ instrument will have an acceptable angular field radius of about 18 arcminutes, twice that of an equivalent paraboloidal telescope. In comparison to the aplanatic Gregorian, the Ritchey-Chrétien has a smaller secondary and therefore blocks less light, and is substantially shorter in length; both features make it much more desirable.

Because an instrument can only collect a portion of the incident wavefront to be reformed into an image, there will always be diffraction: the light will deviate from straight-line propagation and spread out somewhat in the image plane. When an optical system with a circular aperture receives plane waves, rather than there being an image "point" (whatever that means), the light actually spreads out into a tiny circular spot (called an Airy disk, containing about 84% of the energy) surrounded by very faint rings. The radius of the Airy disk determines the overlapping of neighboring images and therefore the resolution. That's why an imaging system that is as perfect as possible is referred to as **diffraction limited**.

For a perfect instrument, the ideal theoretical angular resolution is given by Eq. (10.59), namely, the radius of the Airy disk, $1.22\lambda/D$ radians. Here D is the diameter of the instrument in the same units as λ. Another way to present the angular resolution is in arc-seconds, in which case it equals $2.52 \times 10^5 \lambda/D$. Because of atmospheric distortions, ground-based telescopes, regardless of their size, seldom have angular resolutions better than about 1 arc-second. That is, the images of two stars separated by an angle of less than 1 arc-second blend into an undecipherable blur. By comparison, the HST, high above the atmosphere, for which $D = 2.4$ m, has a diffraction-limited angular resolution at $\lambda = 500 \times 10^{-9}$ m, of about 0.05 arc-second.

The world's two largest telescopes are the twin Keck aplanatic Cassegrains. Separated by 85 m, these two giant telescopes are perched atop the extinct volcano Mauna Kea in Hawaii, at an altitude of 13 600 feet. Each has a 10-m hyperboloidal primary composed of 36 hexagonal elements. These are deeply curved so that the $f/1.75$ system has a focal length of only 17.5 m. This is indicative of the new generation of large telescopes that tend to have fast mirrors (less than $f/2$) with relatively small focal lengths. Short telescopes are more economical to build and house, and are more stable and accurately steered.

The technology now exists for interferometrically combining the images from several separate optical telescopes, thereby tremendously increasing the overall effective aperture. A new generation of ground-based optical telescope arrays is destined to contribute significantly to the way we see the Universe.

Catadioptric Telescopes

A combination of reflecting (*catoptric*) and refracting (*dioptric*) elements is called a *catadioptric* system. The best known of these, although not the first, is the classic *Schmidt optical system*. We must treat it here, even if only briefly, because it represents an important approach to the design of large-aperture, extended-field reflecting systems. As seen in Fig. 5.113, bundles of parallel rays reflecting off a spherical mirror will form images, let's say of a field of stars, on a spherical image

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

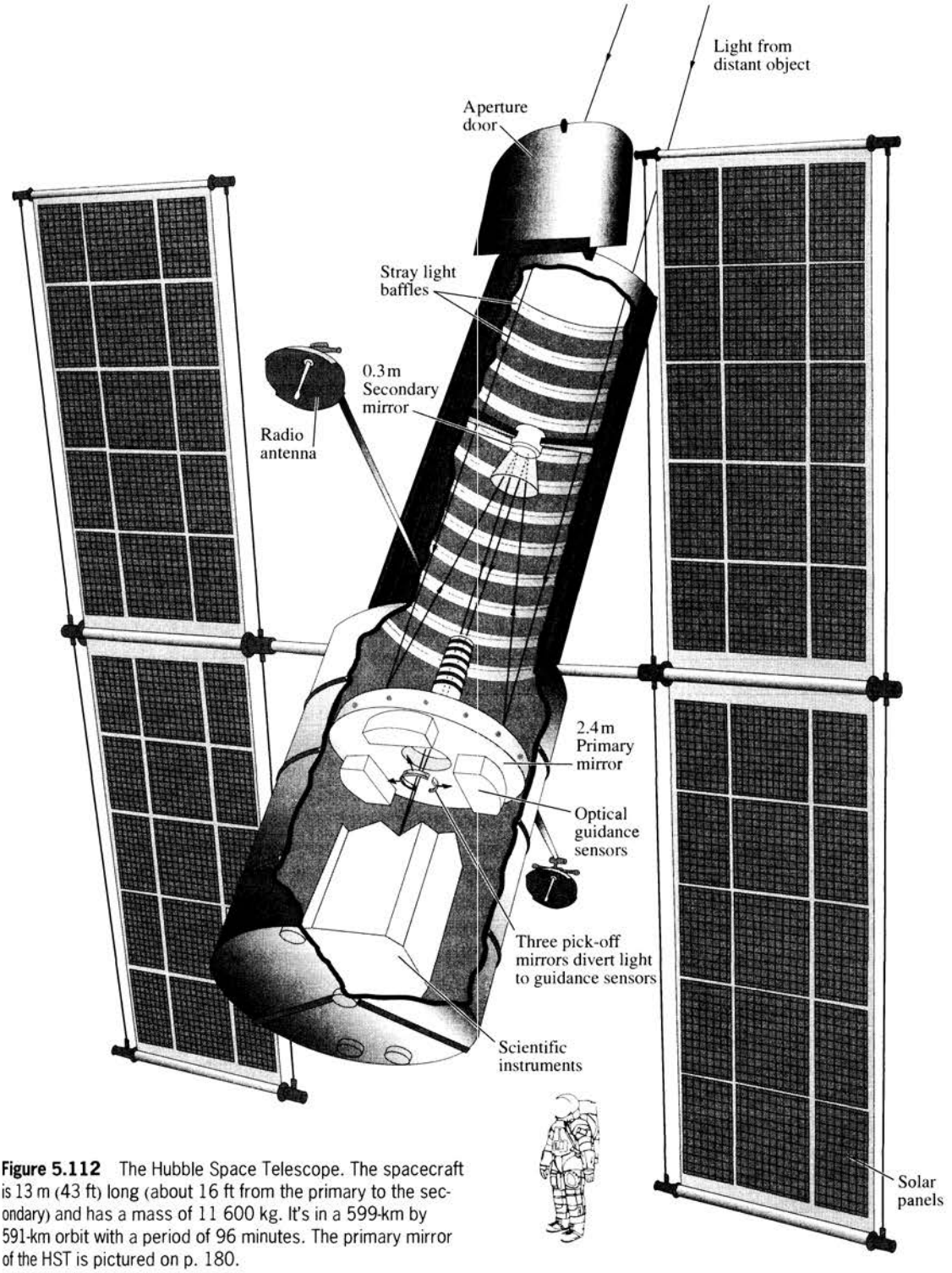


Figure 5.112 The Hubble Space Telescope. The spacecraft is 13 m (43 ft) long (about 16 ft from the primary to the secondary) and has a mass of 11 600 kg. It's in a 599-km by 591-km orbit with a period of 96 minutes. The primary mirror of the HST is pictured on p. 180.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

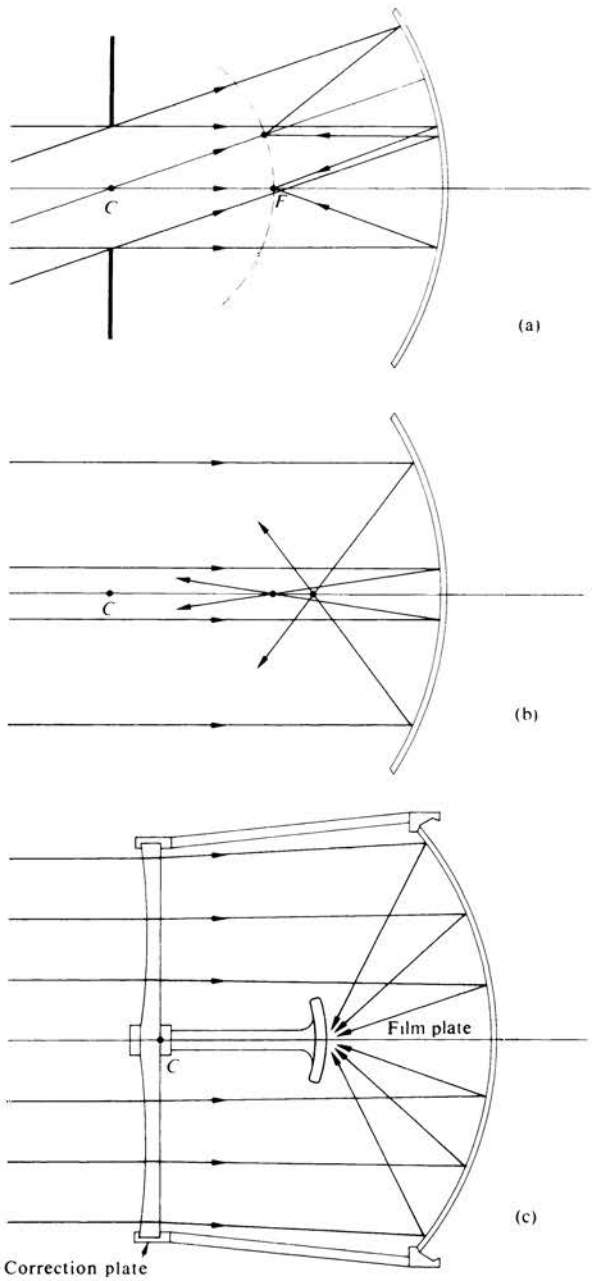


Figure 5.113 The Schmidt optical system.

surface, the latter being a curved film plate in practice. The only problem with such a scheme is that although it is free of other aberrations (astigmatism and coma, see Section 6.3.1), we know that rays reflected from the outer regions of the mirror will not arrive at the same focus as those from the paraxial region. In other words, the mirror is a sphere, not a paraboloid, and it suffers *spherical aberration* (Fig. 5.113*b*). If this could be corrected, the system (in theory at least) would be capable of perfect imagery over a wide field of view. Since there is no one central axis, there are, in effect, no off-axis points. Recall that the paraboloid forms perfect images only at axial points, the image deteriorating rapidly off axis.

One evening in 1929, while sailing on the Indian Ocean (returning from an eclipse expedition to the Philippines), Bernhard Voldemar Schmidt (1879–1935) showed a colleague a sketch of a system he had designed to cope with the spherical aberration of a spherical mirror. He would use a thin glass corrector plate on whose surface would be ground a very shallow toroidal curve (Fig. 5.113*c*). Light rays traversing the outer regions would be deviated by just the amount needed to be sharply focused on the image sphere. The corrector must overcome one defect without introducing appreciable amounts of other aberrations. This first system was built in 1930, and in 1949 the famous 48-inch Schmidt telescope of the Palomar Observatory was completed. It is a fast ($f/2.5$), wide-field device, ideal for surveying the night sky. A single photograph could encompass a region the size of the bowl of the Big Dipper—this compared with roughly 400 photographs by the 200-inch reflector to cover the same area.

Major advances in the design of catadioptric instrumentation have occurred since the introduction of the original Schmidt system. There are now catadioptric satellite and missile tracking instruments, meteor cameras, compact commercial telescopes, telephoto objectives, and missile-homing guidance systems. Innumerable variations on the theme exist; some replace the correcting plate with concentric meniscus lens arrangements (Bouwers–Maksutov), and others use solid thick mirrors. One highly successful approach utilizes a triplet aspheric lens array (Baker).

5.8 Wavefront Shaping

This chapter has been about reshaping wavefronts in one way or another, but the changes introduced by traditional lenses and mirrors are global, affecting the whole processed portion of the wavefront in more-or-less the same way. By contrast, for the first time it is now possible to take an incoming wavefront and reconfigure every portion of it differently to fit specific needs.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Consider a plane wave passing either through some inhomogeneous medium of index $n(\mathbf{r})$ or through a medium of nonuniform thickness—a piece of shower-door glass will do (Fig. 5.114*a*). The wavefronts are essentially held back in proportion to the *OPL* and distort accordingly. When, for example, such a wrinkled wave reflects from an ordinary planar mirror, it goes off reversed in direction but otherwise unchanged (Fig. 5.114*b*). The leading and trailing wavefront regions remain leading and trailing, with only the direction of propagation reversed; the wavefront remains distorted. The scene beyond a crinkled-glass shower door is equally blurred whether you look at it directly or in a mirror.

If a more sophisticated mirror could be devised that could reshape the reflected wavefronts, we might be able to get rid of undesirable distortions that are unavoidably introduced in a variety of situations. This section explores two different state-of-the-art techniques for accomplishing just that.

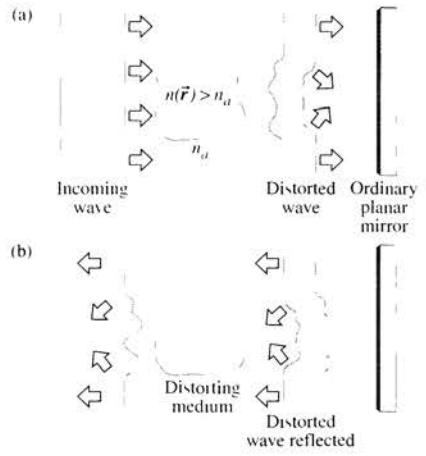


Figure 5.114 (*a*) A plane wave becomes distorted on passing through an inhomogeneous medium. (*b*) When such a wrinkled wave reflects off a traditional mirror, it changes direction. Regions that were originally leading or trailing remain that way as the wave, still wrinkled, moves off in a new direction. Passing through the inhomogeneous medium a second time increases the distortion.

5.8.1 Adaptive Optics

One of the most significant breakthroughs in telescope technology to occur in recent times is called **adaptive optics**, and it has already begun to provide a way to deal with the daunting problem of atmospheric distortion. As Newton put it, "If the Theory of making Telescopes could at length be fully brought into Practice, yet there would be certain Bounds beyond which telescopes could not perform. For the Air through which we look upon the Stars, is in perpetual Tremor; as may be seen by the tremulous Motion of Shadows cast from high Towers, and by the twinkling of the fix'd Stars." Adaptive optics is a methodology used to control that "perpetual Tremor"—first by measuring the turbulence-induced distortions of the incident light, and then by using that information to reconfigure the lightwave, bringing it back to a pristine condition as if it had never traversed the swirling tumult of the atmosphere (Fig. 5.115).

Driven by thermal energy from the Sun, the Earth's atmosphere is a shifting sea of turbulent air. Variations in density are accompanied by variations in the index of refraction and therefore in optical path length. Wavefronts streaming down from a point on a distant star arrive at the atmosphere almost precisely as plane waves (with a wavelength in the mid-visible of about 0.5 μm). As they sweep through the 100 miles or so of shifting air, path length differences of a few micrometers are introduced, and the wavefronts distort into a bumpy undulated surface. What reaches ground level is a succession of broadly wrinkled wavefronts shaped much like what you would get if you spread 10-cm tiles down on a floor that had first been randomly strewn with tiny tough beetles, that is, contiguous tiles, slightly tilted every which way. The turbulence changes unpredictably on a time scale of milliseconds, and the progression of wavefronts traversing it is continuously bent and buckled anew (as if the beetles were aimlessly walking around under the tiles, lifting and shifting them).

The tile imagery, however weird, is useful because in 1966 David L. Fried showed that the optical results of atmospheric turbulence could be modeled in a fairly simple way. In effect (because the speed of light is so great), one can assume that at any moment the atmosphere behaved as if it were compressed into a horizontal array of small, contiguous, wedge-shaped refracting regions or stable cells. At any given ground site, the local portion of a stellar wavefront is composed of many randomly tilted, small, fairly flat areas (each analogous to a single tile). In someone's backyard, these areas are typically about 10 cm across, although under the very best conditions (e.g., on a astronomical mountaintop) they might reach as much as 20 or occasionally 30 cm "when the seeing is good." Over each such **isoplanatic region**, the wavefront is fairly smooth and has little curvature: the difference between leading bumps and trailing depressions is about $\lambda/17$, and it's a rule-of-thumb that if wave distortions are less than $\lambda/10$ the image quality will be

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

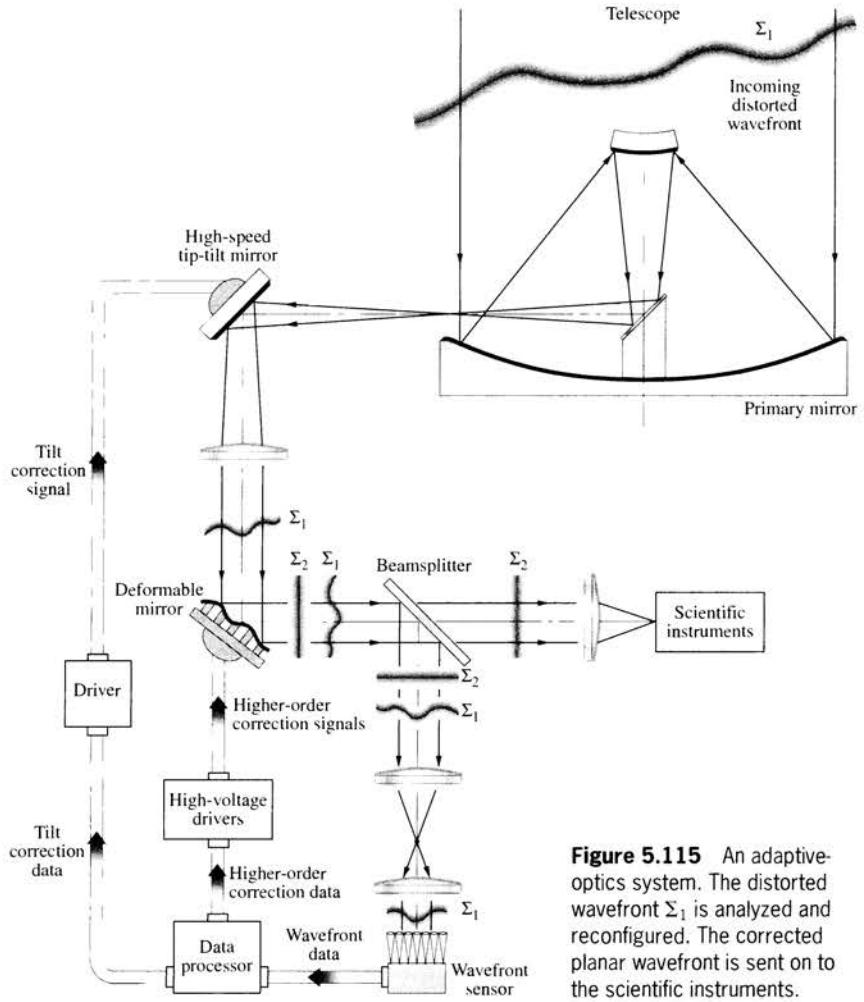


Figure 5.115 An adaptive-optics system. The distorted wavefront Σ_1 is analyzed and reconfigured. The corrected planar wavefront is sent on to the scientific instruments.

very good. The more turbulence there is, the smaller the stable cells are, and the smaller are the corresponding isoplanatic regions of the wavefront.

The effect of turbulence on the image formed by a telescope, one trained on a star, depends strongly on the size of its aperture. If the instrument has an aperture of only a few centimeters, the small admitted portion of a wavefront (having traversed only a part of a stable cell) will likely be quite flat. Turbulence will primarily alter the tilt of that otherwise planar incoming wavefront section. That means that a sharp Airy image can momentarily be formed via that section, but the Airy-image spot will wander around as the atmosphere changes and each successive planar wavefront section arrives at a different angle (our mythical beetles keep moving). By contrast, for a large-diameter telescope, several meters across, the large admitted wavefront section is a mosaic of many flat, tilted regions. The image is then a simultaneous superposition of numerous shifting Airy spots, and the result is a shimmering blur. Clearly, increasing the aperture will collect more light, but it will not proportionately improve the resolution.

The critical aperture size at which blurring becomes appreciable is a measure of



When looking through the atmosphere with a telescope the probability of experiencing a moment of clear viewing decreases exponentially with aperture diameter. Using a moderate-size objective (≈12-inch) in ordinary seeing conditions the odds are 1 in 100. This sequence of photos of a star taken at 1/60-second intervals shows how the image "twinkles." The right most picture was taken at an instant of very good seeing. With a diffraction-limited instrument the image should resemble the Airy disk pattern (p. 470) of a central bright spot surrounded by faint concentric rings.
(Photo courtesy Ron Dantowitz, Museum of Science, Boston.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

the turbulence. It's called the **Fried parameter**, and it's almost universally represented by r_0; this is an unfortunate choice of symbol since this is not a radius. Pronounced "*r naught*," it corresponds to the size of the region over which the incoming wavefront can be taken to be essentially planar. On those rare occasions when r_0 exceeds 30 cm, a very distant star will be "perfectly" imaged as an Airy disk. As the turbulence increases, r_0 decreases; moreover, as the wavelength increases, r_0 increases: $r_0 \propto \lambda^{1.2}$. It follows that the angular resolution of a large ground-based telescope is actually $1.22\lambda/r_0$ and since r_0 is rarely better than 20 cm, the most powerful Earth-bound instrument has little more resolution than a humble 6-inch telescope!

When there's a wind above a telescope, it, in effect, blows the isoplanatic regions past the aperture. A 5-m/s breeze will carry an $r_0 = 10$-cm isoplanatic region past in 20 ms. To monitor and ultimately respond to such atmospheric changes, an electro-optical-mechanical control system should operate 10 or 20 times faster, sampling the data at upwards of 1000 times per second.

Figure 5.115 is a schematic drawing of a typical astronomical adaptive optics system. In this simple arrangement, the telescope is pointed at a star that will serve both as the object of attention and as a beacon for correcting distortions. Before anything clever is done, the large beam from the primary mirror is reduced to several centimeters in diameter so that it can be handled more conveniently. In the process, each isoplanatic region at the primary becomes focused down to a correspondingly small region in the reduced beam.

The first step is to analyze the distorted wavefront, Σ_1, transmitted by the telescope and now present in miniature in the reduced beam. This is done with a **wavefront sensor**, of which there are several types. The one considered here is a Hartmann sensor (Fig. 5.116), which consists of a compact array of thousands of independent detectors tightly grouped side-by-side. Light incident on the sensor first encounters a sheet of closely packed tiny identical lenslets, at whose focal plane there is a CCD array (Fig. 5.116*a*). The device is located in the beam in such a way that a lenslet is about the size of an isoplanatic region. Each lenslet then forms a minute image of the star on a cluster of four CCD pixel elements grouped around its optical axis. If the overall wavefront were perfectly flat, that is, if every isoplanatic region had zero tilt and all were parallel, each lenslet would produce an Airy-image spot at a null position between its own four pixel elements (Fig. 5.116*b*). But when any isoplanatic region is tilted, the corresponding image spot shifts and the four CCD elements record an unbalanced signal that indicates the exact displacement (Fig. 5.116*c*). The output from all of these minute detectors is computer analyzed, Σ_1 is theoretically reconstructed, and the corrections necessary to flatten the wavefront are calculated.

If an overall tilt of the wavefront is detected, a signal is sent to the fast-steering flat mirror, which initially receives the light from the primary, and that tilt is counteracted. The now untilted, but still wrinkled, wavefront heads toward a "rubber mirror," a flexible reflector that can rapidly and precisely be deformed. It might, for example, be composed of a thin faceplate reflector mounted on hundreds of actuators that rapidly push and pull it into the desired shape. Driven by signals from the computer, the mirror is bent into an inverse configuration to that of the wavefront. In effect, wavefront bumps impinge on matching mirror depressions, and vice versa. The result is to reflect a beam of distortion-free wavefronts, Σ_2, that correspond to the condition of the starlight before it entered the atmosphere. A small fraction of the radiant energy goes back into the sensor-computer-mirror control loop to continuously maintain the correction process, while the remainder travels on to the scientific instruments.

Because many objects of interest to astronomers—planets, galaxies, nebulae, and so on—are imaged as extended bodies, using these as an adaptive-optics beacon is precluded. Still, if you wish to examine a galaxy, you could use a nearby star as a

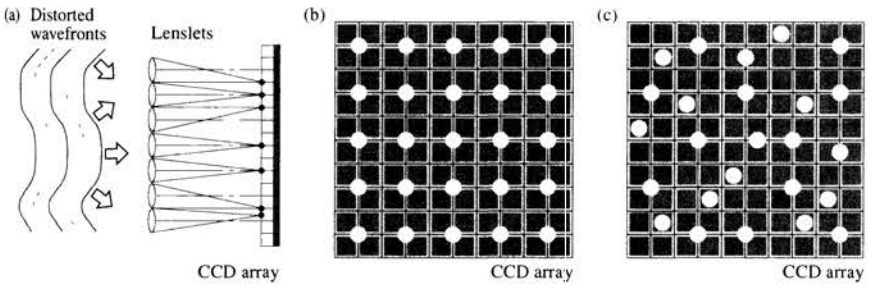
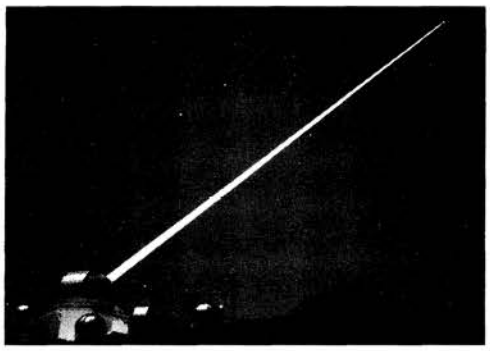
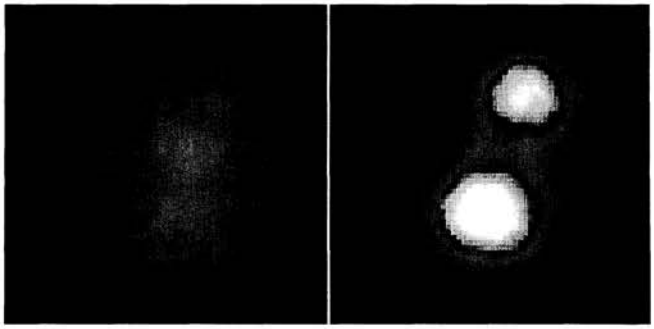


Figure 5.116 The Hartmann wavefront sensor. (*a*) Lenslets focus light down to a CCD array. Each square cluster of four CCD elements forms a detector. (*b*) When the incident wave is planar, Airy-image spots form at null points at the centers of each four-element detector. (*c*) When the wavefront is distorted, Airy-image spots are shifted from the null positions.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



The creation of a laser guide star at the Phillips Laboratory, Kirtland Air Force Base, New Mexico. (Photo courtesy of Phillips Laboratory, Department of the US Air Force.)



A 1-second exposure of 53ξ Ursa Major using a 1.5-m telescope at the Phillips Laboratory. (*a*) The ordinary uncompensated image is undecipherable. (*b*) Using adaptive optics, the image is improved dramatically. (Photos courtesy of Phillips Laboratory, Department of the US Air Force.)

beacon. Unfortunately, however, there will frequently not be any stars in the vicinity that are bright enough for the purpose. One way to get around this limitation is to use a laserbeam to create an artificial guide star (see photo). This has successfully been done in two different ways. In one, a laser pulse, focused at altitudes in the range from around 10 to 40 km, is projected up through the telescope. A portion of that light is backscattered downward from air molecules via Rayleigh Scattering. Alternatively, there is a layer of sodium atoms (probably deposited by meteors) at an altitude of 92 km, well above most of the atmospheric turbulence. A laser tuned to 589 nm can excite the sodium, thereby producing a small bright yellow beacon anywhere in the sky.

The results (see photo) have been so encouraging* that most of the world's major telescopes are already, or soon will be, using adaptive optics.

5.8.2 Phase Conjugation

Another important new technology for reshaping wavefronts is known as **phase conjugation**; here the wave is turned inside-out during a special kind of reflection.

Imagine a stream of plane waves traveling to the right in the positive z-direction impinging perpendicularly on an ordinary flat mirror. The incident wave is expressible as $E_i = E_0 \cos(kz - \omega t)$, or in complex form as $\tilde{E}_i = E_0 e^{i(kz-\omega t)} = E_0 e^{ikz} e^{-i\omega t} = \tilde{E}(z) e^{-i\omega t}$, where the space and time parts have been separated. For this simple geometry, the reflected waves ride right back over the incident waves; ***they are identical except for the direction of propagation***. The reflected wave is given by $E_r = E_0 \cos(-kz - \omega t)$, or $\tilde{E}_r = E_0 e^{-ikz} e^{-i\omega t} = E^*(z) e^{-i\omega t}$. Changing the sign of the space part of the phase changes the direction of the wave, and the same thing is accomplished by taking the complex conjugate in the exponential formulation. For this reason, the reflected wave is also called a **phase-conjugated wave** or just a **conjugate wave**. A situation of this sort is characterized by the fact that we could, in principle, take a motion picture of it, which when shown forwards or backwards, would be indistinguishable. Consequently, a phase-conjugated wave is said to be **time reversed**. For monochromatic waves, changing the sign of the time part (i.e., time reversal) is equivalent to reversing the direction of propagation: $\cos[kz - \omega(-t)] = \cos(kz + \omega t) = \cos(-kz - \omega t)$.

A very simple, phase-conjugated reflection occurs when there is a point source at the center of curvature of a concave spherical mirror. The waves flow, expanding out to the mirror, and on reflection, contract back on themselves to the source point. Presumably, a conventional reflecting surface could be made to exactly match any particular wavefront and thereby reflect a conjugate for that specific incoming wave (Fig. 5.117). That's a rather impractical approach, especially if you can't anticipate the shape of the wavefront or if it changes from moment to moment.

Fortunately, in 1972 a team of Russian scientists discov-

*See L. A. Thompson, "Adaptive Optics in Astronomy," *Phys. Today* **47**, 24 (1994); J. W. Hardy, "Adaptive Optics," *Sci. Am.* **60** (June 1994); R. Q. Fugate and W. J. Wild, "Untwinkling the Stars—Part I," *Sky & Telescope* **24** (May 1994); W. J. Wild and R. Q. Fugate, "Untwinkling the Stars—Part II," *Sky & Telescope* **20** (June 1994).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

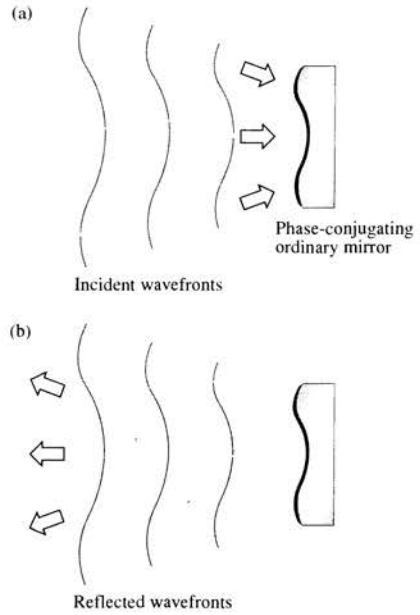


Figure 5.117 The operation of a rather limited phase-conjugating mirror. It only works for the incoming wavefronts shown in (*a*).

ered a method for producing phase conjugation for *any* incident wavefront using Brillouin Scattering. They directed an intense beam of laser light into a tube containing high-pressure methane gas. At power levels of about a million watts, pressure-density variations occur, and the medium becomes a remarkable kind of mirror, reflecting back almost all the incoming light. What surprised the investigators was that the beam scattered back out of the gas was phase conjugated. The medium, in this case the methane, adjusts itself to the presence of the electromagnetic field in just such a way as to turn the backscattered wave inside-out, so that regions that were originally leading were now trailing. Today there are several means to the same end, all using media that produce nonlinear optical effects. There are myriad potential applications from tracking satellites to improving laserbeam quality.*

As an example of the kinds of things that can be done, consider the following: If a beam that has been distorted by passing through an inhomogeneous medium (Fig. 5.114) is reflected from an ordinary mirror and made to retraverse that

*See D. M. Pepper, "Applications of Optical Phase Conjugation," *Sci. Am.* **74** (January 1986) and V. V. Shkunov and B. Ya. Zel'dovich, "Optical Phase Conjugation," *Sci. Am.* **54** (December 1985).

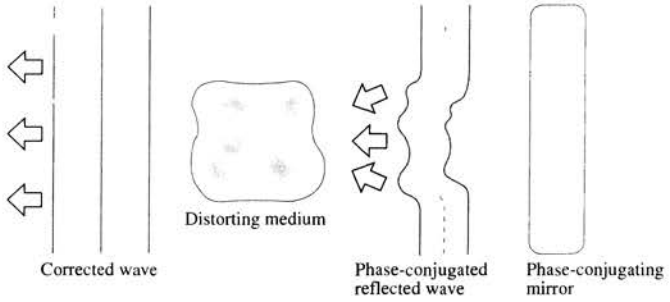


Figure 5.118 When the distorted wave in Fig. 5.113 is reflected by a phase-conjugating mirror, it's turned inside-out, or conjugated. Compare it to the conventionally reflected wave in Fig. 5.113*b*. On traversing the inhomogeneous medium a second time, regions of the wavefront that are now leading will be held back, and vice versa. The wave that emerges after a round-trip will be identical to the one that originally entered (Fig. 5.113*a*).

medium, the beam will become even more distorted. By contrast, if the same thing is done using a phase-conjugating mirror, on passing back through the distorting medium for a second time the beam will be restored to its pristine condition. Figure 5.118 illustrates the technique, and Fig. 5.119 shows the results of an actual experiment. The image of a cat was impressed on a collimated argon-ion laserbeam ($\lambda = 514.5$ nm) by simply passing the beam through a photographic transparency of the cat. As a reference standard, the image-carrying wave was sent, via a beamsplitter, to an ordinary mirror where it was reflected back through the beamsplitter and onto a ground-glass screen so it could be photographed (Fig. 5.119*a*). Next, a phase distorter (e.g., a piece of shower-door glass) was introduced between the beamsplitter and the mirror so that the wave traversed it twice. The image returned from the mirror was unrecognizable (Fig. 5.119*b*). Finally, the conventional mirror was removed and replaced by a phase conjugator. Even though the wave again passed twice through the distorter, the image was restored to its original clarity (Fig. 5.119*c*).

5.9 Gravitational Lensing

Among the most remarkable discoveries of the twentieth century—one that sprang directly from Einstein's General Theory of Relativity (1915)—was that matter gives rise to, or better yet corresponds to, a curvature of space-time. From either perspective, where there is a large concentration of mass there

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

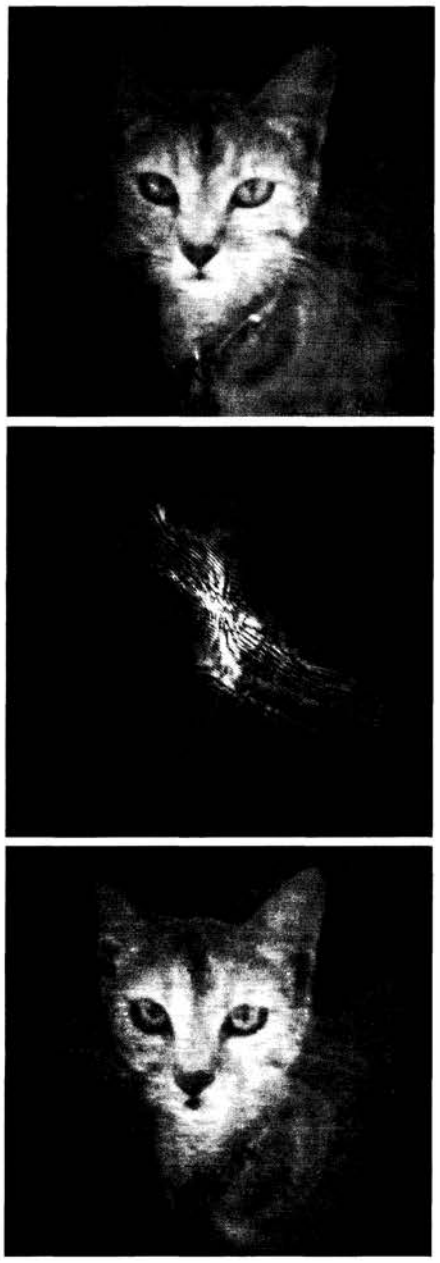


Figure 5.119 Using phase conjugation to remove distortion. (*a*) Image of a cat reflected from a mirror—no introduced distortion. (*b*) The same cat wave after twice traversing an inhomogeneous medium. (*c*) After passing through the inhomogeneous medium, the wave was phase conjugated and returned through the medium a second time. Most of the image distortion vanished. (Photos courtesy of Jack Feinberg, University of Southern California.)

will be an appreciable curvature of the local space-time. Relativity theory conceptually conjoins space and time, and gravity affects both. That means that a light beam traversing such a warped region will follow a curved path, bending inward toward the mass concentration. In other words, gravity alters the velocity of light—direction and speed. That shouldn't be altogether surprising since it slows time itself.

Up until now we've assumed that light in space propagates in straight lines at the fixed speed c. That understanding is in accord with Special Relativity, and it's true enough for any Earthbound experiment we might perform. But it's not true on the vast scale of stars and galaxies and black holes. Under the influence of a tremendous amount of mass, the gravitational potential (Φ_G) in the immediate surroundings can be immense. Light propagating through this kind of region behaves as if it were traversing an inhomogeneous medium having a position-dependent index of refraction, $n_G(\vec{r})$, greater than one. For that reason, and because the resulting effects are similar to those that can easily be produced by aspherical lenses, the phenomenon is called **gravitational lensing**. More fundamentally, the deviation of light from rectilinear propagation is the purview of diffraction, and the effect might better be called *gravitational diffraction*.

The geometry of the situation is straightforward: we need an observer, (e.g., someone with a telescope on Earth), a distant source of electromagnetic radiation (e.g., a quasar or a galaxy) that serves as the object being viewed, and a lensing mass (e.g., a quasar, galaxy, group of galaxies, or black hole) somewhere between the two, located on the source-observer axis.

A region of curved space-time functions much like a crude GRIN lens (Fig. 6.42) where the index of refraction drops off with distance from the central axis, just as Φ_G drops off. An even more rudimentary modeling approach would be to just match the index profile with a corresponding aspherical thickness profile. The lens in Fig. 5.120 might then correspond to a nice symmetrical galaxy, whereas a much more centrally pointed version would represent a black hole. When an off-axis galaxy lenses an object far behind it, the image can be distorted into several arcs (Fig. 5.121). More accurately, the phenomenon can be represented via waves passing through a distorting medium (like those in Fig. 5.114*a*) using Huygens's construction (p. 104). As we'll see later when we study diffraction (Fig. 10.7*d*), that approach establishes that there will always be an odd number of miraged images, including a central undiffracted one. Figure 5.122 shows how a cluster of

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



Figure 5.120 An aspherical lens used to simulate the effects of gravitational lensing by a large massive object like a galaxy.

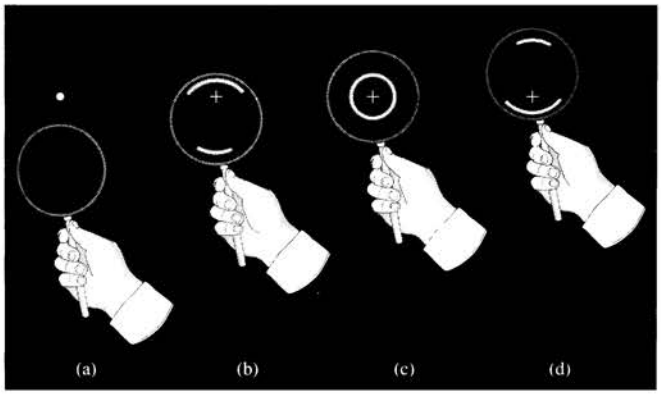
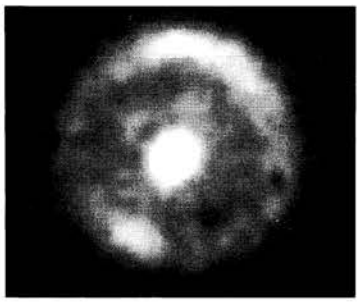


Figure 5.121 An asphere, like that in Fig. 5.120, used to simulate gravitational lensing by a galaxy.



A photo of the first complete Einstein Ring ever observed, taken by the Hubble Space Telescope in 1998. It resulted from the near perfect alignment of the Earth and two galaxies, one behind the other (see Fig. 5.121c).

lensing galaxies images a single distant galaxy as a shower of arcs more or less concentric with the cluster's center-of-mass.

Einstein, who began thinking about gravitational lensing as early as 1912, suggested that in the unlikely event the alignment was near perfect, with all three participants precisely on axis (Fig. 5.121*c*), the image would be smeared out into a ring. In 1998 the Hubble Space Telescope photographed a complete Einstein Ring for the first time (see photo).

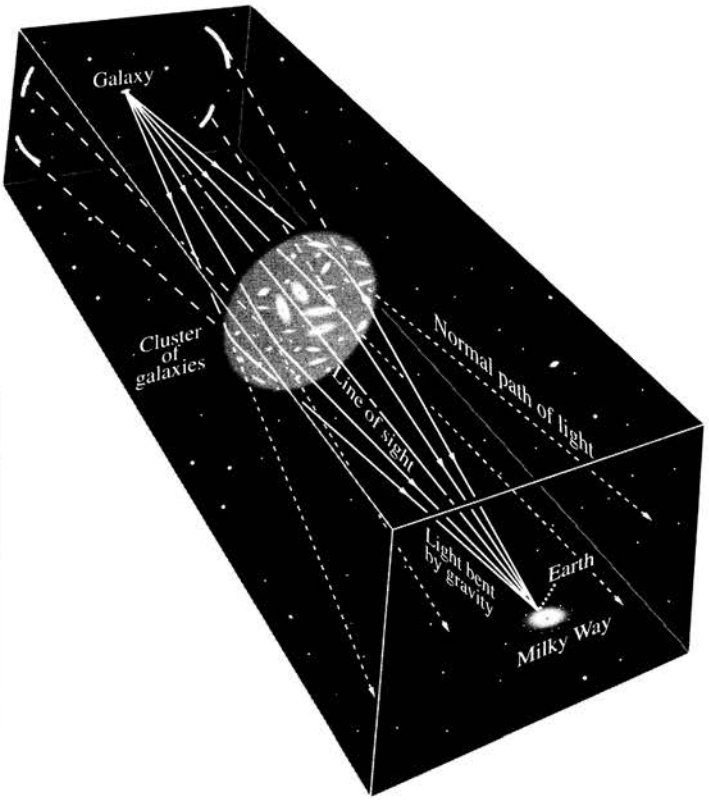
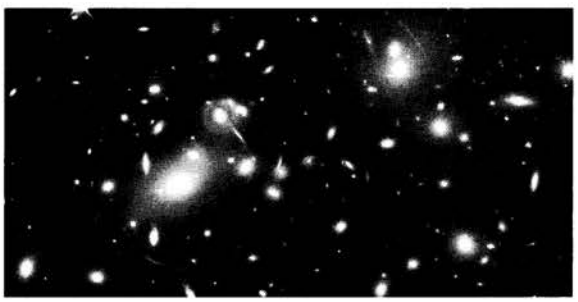


Figure 5.122 Gravitational lensing by a cluster of galaxies.



The galaxy cluster Abell 2218 is so massive and compact that light passing through it is deflected by the enormous gravitation field. The process magnifies, brightens, and distorts the images of galaxies that lie far behind it. The numerous arcs in the picture are the contorted images of galaxies 5 to 10 times farther away than the lensing cluster.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

PROBLEMS

Complete solutions to all problems—except those with an asterisk—can be found in the back of the book.

5.1 The shape of the interface pictured in Fig. P.5.1 is known as a Cartesian oval after René Descartes who studied it in the early 1800s. It's the perfect configuration to carry any ray from S to the interface to P. Prove that the defining equation is

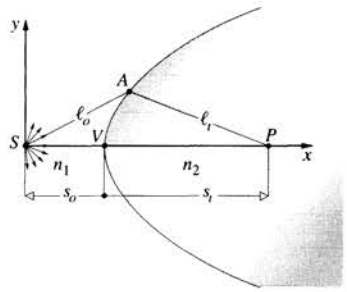
$$\ell_o n_1 + \ell_i n_2 = \text{constant}$$

Show that this is equivalent to

$$n_1(x^2 + y^2)^{1/2} + n_2[y^2 + (s_o + s_i - x^2)]^{1/2} = \text{constant}$$

where x and y are the coordinates of point A.

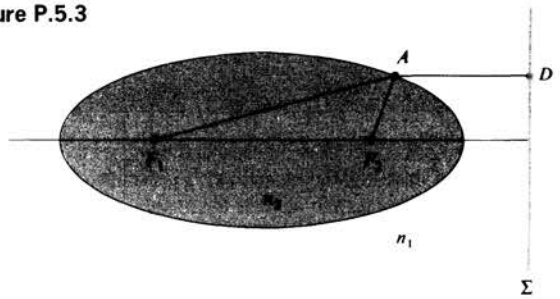
Figure P.5.1



5.2 Construct a Cartesian oval such that the conjugate points will be separated by 11 cm when the object is 5 cm from the vertex. If $n_1 = 1$ and $n_2 = \frac{3}{2}$, draw several points on the required surface.

5.3* Use Fig. P.5.3 to show that if a point source is placed at the focus F_1 of the ellipsoid, plane waves will emerge from the far side. Remember that the defining requirement for an ellipse is that the net distance from one focus to the curve and back to the other focus is constant.

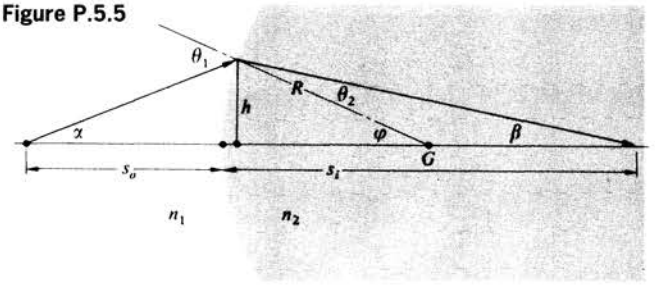
Figure P.5.3



5.4 Diagrammatically construct both a sphero-elliptic positive lens and an ellipto-spheric negative lens, showing rays and wavefronts as they pass through the lens. Do the same for an oval-spheric positive lens.

5.5* Making use of Fig. P.5.5, Snell's Law, and the fact that in the paraxial region $\alpha = h/s_o$, $\varphi \approx h/R$, and $\beta \approx h/s_i$, derive Eq. (5.8).

Figure P.5.5

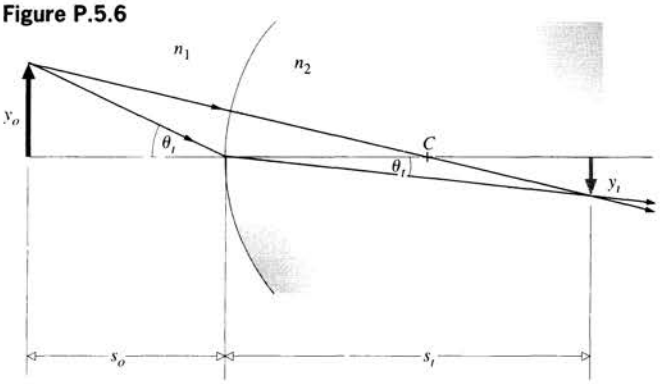


5.6* Show that, in the paraxial domain, the magnification produced by a single spherical interface between two continuous media, as shown in Fig. P.5.6, is given by

$$M_T = -\frac{n_1 s_i}{n_2 s_o}$$

Use the small-angle approximation for Snell's Law and approximate the angles by their tangents.

Figure P.5.6



5.7* Imagine a hemispherical interface, with a radius of curvature of radius 5.00 cm, separating two media: air on the left, water on the right. A 3.00-cm-tall toad is on the central axis, in air, facing the con-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

vex interface and 30.0 cm from its vertex. Where in the water will it be imaged? How big will it appear to a fish in the water? Use the results of the previous problem, even though our frog is pushing the paraxial approximation.

5.8 Locate the image of an object placed 1.2 m from the vertex of a gypsy's crystal ball, which has a 20-cm diameter ($n = 1.5$). Make a sketch (of the rays, not the gypsy).

5.9* Return to Problem 5.7 and suppose we cut off the medium on the right forming a thick water biconvex lens, with each surface having a radius of curvature of 5.00 cm. If the lens is 10.0 cm thick, determine the total magnification and everything you can about the toad's image.

5.10* A biconvex glass ($n_l = 1.5$) thin lens is to have a +10.0-cm focal length. If the radius of curvature of each surface is measured to be the same, what must it be? Show that a spider standing 1.0 cm from the lens will be imaged at -1.1 cm. Describe that image and draw a ray diagram.

5.11* Going back to Section 5.2.3, prove that for a thin lens immersed in a medium of index n_m

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

That done, imagine a double-concave air lens surrounded by water; determine if it's converging or diverging.

5.12* A meniscus concave glass ($n_l = 1.5$) thin lens (see Fig. 5.12) has radii of curvature of +20.0 cm and +10.0 cm. If an object is placed 20.0 cm in front of the lens, show that the image distance will be -13.3 cm. Describe that image and draw a ray diagram.

5.13 A biconcave lens ($n_l = 1.5$) has radii of 20 cm and 10 cm and an axial thickness of 5 cm. Describe the image of an object 1-inch tall placed 8 cm from the first vertex. Use the thin-lens equation to see how far off it is in determining the final-image location.

5.14* A 35-mm camera has a single thin lens having a 50.0-mm focal length. A woman 1.7 m tall stands 10.0 m in front of the camera. (a) Show that the lens-film distance must be 50.3 mm. (b) How tall is her image on the film?

5.15 Prove that the minimum separation between conjugate *real* object and image points for a thin positive lens is $4f$.

5.16 An object 2 cm high is positioned 5 cm to the right of a positive thin lens with a focal length of 10 cm. Describe the resulting image completely, using *both* the Gaussian and Newtonian equations.

5.17 Make a rough graph of the Gaussian Lens Equation; that is, plot s_i versus s_o, using unit intervals of f along each axis. (Get both segments of the curve.)

5.18* A parallel bundle of rays from a very distant point source is incident on a thin negative lens having a focal length of -50.0 cm. The rays make an angle of 6.0° with the optical axis of the lens. Locate the image of the source.

5.19 What must the focal length of a thin negative lens be for it to form a virtual image 50 cm away (measured from the lens) of an ant located 100 cm away (measured from the lens)? Given that the ant is to the right of the lens, locate and describe its image.

5.20* Compute the focal length in air of a thin biconvex lens ($n_l = 1.5$) having radii of 20 and 40 cm. Locate and describe the image of an object 40 cm from the lens.

5.21 Determine the focal length of a planar-concave lens ($n_l = 1.5$) having a radius of curvature of 10 cm. What is its power in diopters?

5.22* Determine the focal length in air of a thin spherical planar-convex lens having a radius of curvature of 50.0 mm and an index of 1.50. What, if anything, would happen to the focal length if the lens were placed in a tank of water?

5.23* We would like to place an object 45 cm in front of a lens and have its image appear on a screen 90 cm behind the lens. What must be the focal length of the appropriate positive lens?

5.24 The horse in Fig. 5.26 is 2.25 m tall, and it stands with its face 15.0 m from the plane of the thin lens whose focal length is 3.00 m.

(a) Determine the location of the image of the equine nose.
(b) Describe the image in detail—type, orientation, and magnification.
(c) How tall is the image?
(d) If the horse's tail is 17.5 m from the lens, how long, nose-to-tail, is the image of the beast?

5.25* A candle that is 6.00 cm tall is standing 10 cm from a thin concave lens whose focal length is -30 cm. Determine the location of the image and describe it in detail. Draw an appropriate ray diagram.

5.26* The image projected by an equiconvex lens ($n = 1.50$) of a frog 5.0 cm tall, who is located 0.60 m from a screen, is to be 25 cm high. Please compute the necessary radii of the lens.

5.27* A thin piece of wire 4.00 mm long is located in a plane perpendicular to the optical axis and 60.0 cm in front of a thin lens. The sharp image of the wire formed on a screen is 2.00 mm long. What is the focal length of the lens? When the screen is moved farther from the lens by 10.0 mm, the image blurs to a width of 0.80 mm. What is

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

the diameter of the lens? [*Hint:* Image a source point on the axis.]

5.28 A thin double-convex glass lens (with an index of 1.56) while surrounded by air has a 10-cm focal length. If it is placed under water (having an index of 1.33) 100 cm beyond a small fish, where will the guppy's image be formed?

5.29 Consider a homemade television projection system that uses a large positive lens to cast the image of the TV screen onto a wall. The projected picture is enlarged three times, and although dim, it's nice and clear. If the lens has a focal length of 60 cm, what should be the distance between the screen and the wall? Why use a large lens? How should we mount the set with respect to the lens?

5.30 Write an expression for the focal length (f_w) of a thin lens immersed in water ($n_w = \frac{4}{3}$) in terms of its focal length when it's in air (f_a).

5.31* Observe the three vectors $\vec{\mathbf{A}}$, $\vec{\mathbf{B}}$, and $\vec{\mathbf{C}}$ in Fig. P.5.31, each of which has a length of $0.10f$ where f is the focal length of the thin positive lens. The plane formed by $\vec{\mathbf{A}}$ and $\vec{\mathbf{B}}$ is at a distance of $1.10f$ from the lens. Describe the image of each vector.

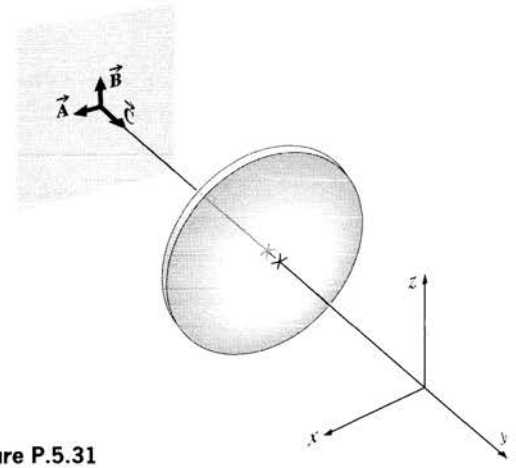


Figure P.5.31

5.32* A convenient way to measure the focal length of a positive lens makes use of the following fact. If a pair of conjugate object and (real) image points (S and P) are separated by a distance $L > 4f$, there will be two locations of the lens, a distance d apart, for which the same pair of conjugates obtain. Show that

$$f = \frac{L^2 - d^2}{4L}$$

Note that this avoids measurements made specifically from the vertex, which are generally not easy to do.

5.33* Two positive lenses with focal lengths of 0.30 m and 0.50 m are separated by a distance of 0.20 m. A small butterfly rests on the central axis 0.50 m in front of the first lens. Locate the resulting image with respect to the second lens.

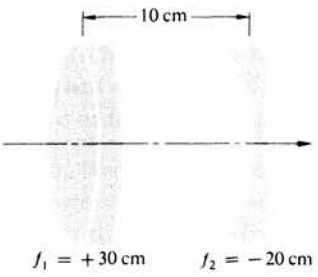
5.34 In the process of constructing a doublet, an equiconvex thin lens L_1 is positioned in intimate contact with a thin negative lens, L_2, such that the combination has a focal length of 50 cm in air. If their indices are 1.50 and 1.55, respectively, and if the focal length of L_2 is -50 cm, determine all the radii of curvature.

5.35 Verify Eq. (5.34), which gives M_T for a combination of two thin lenses.

5.36* A blade of grass standing 10.0 mm tall is 150 mm in front of a thin positive lens having a 100 mm focal length; 250 mm behind that first lens is a thin negative lens with a focal length of -75.0 mm. (a) Show that the first lens forms an image 300 mm behind it. (b) Describe that image. (c) What's its magnification? (d) Prove that the final image formed by both lenses is located 150 mm behind the negative lens. (e) What is the total magnification of the combination?

5.37 Compute the image location and magnification of an object 30 cm from the front doublet of the thin-lens combination in Fig. P.5.37. Do the calculation by finding the effect of each lens separately. Make a sketch of appropriate rays.

Figure P.5.37



5.38* Two thin lenses having focal lengths of $+15.0$ cm and -15.0 cm are positioned 60.0 cm apart. A page of print is held 25.0 cm in front of the positive lens. Describe, in detail, the image of the print (i.e., insofar as it's paraxial).

5.39* Draw a ray diagram for the combination of two positive lenses wherein their separation equals the sum of their respective focal

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Figure P.5.43a

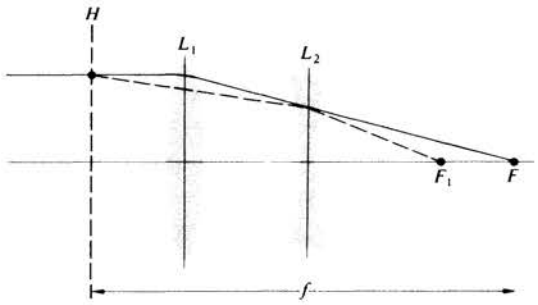
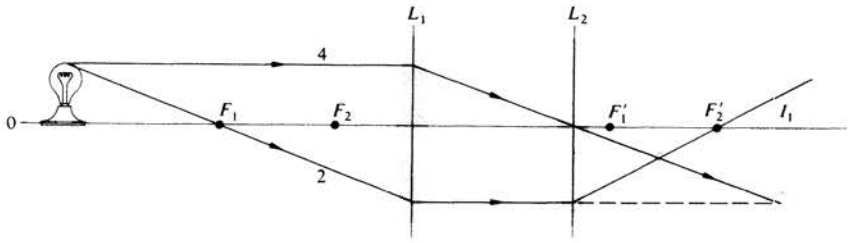


Figure P.5.43b



lengths. Do the same thing for the case in which one of the lenses is negative.

5.40* Two positive lenses are to be used as a laserbeam expander. An axial 1.0-mm diameter beam enters a short focal length positive lens, which is followed by a somewhat longer focal length positive lens from which it emerges with a diameter of 8.0 mm. Given that the first lens has a 50.0 mm focal length, determine the focal length of the second lens and the separation between the lenses. Draw a diagram.

5.41 Redraw the ray diagram for a compound microscope (Fig. 5.99), but this time treat the intermediate image as if it were a real object. This approach should be a bit simpler.

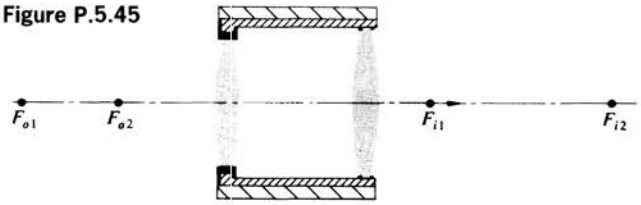
5.42* Consider a thin positive lens L_1, and using a ray diagram, show that if a second lens L_2 is placed at the focal point of L_1, the magnification does not change. That's a good reason to wear eyeglasses, whose lenses are different, at the correct distance from the eye.

5.43* Figures P.5.43*a* and P.5.43*b* are taken from an introductory physics book. What's wrong with them?

5.44 Consider the case of two positive thin lenses, L_1 and L_2, separated by 5 cm. Their diameters are 6 and 4 cm, respectively, and their focal lengths are $f_1 = 9$ cm and $f_2 = 3$ cm. If a diaphragm with a hole 1 cm in diameter is located between them, 2 cm from L_2, find (a) the aperture stop and (b) the locations and sizes of the pupils for an axial point, S, 12 cm in front of (to the left of) L_1.

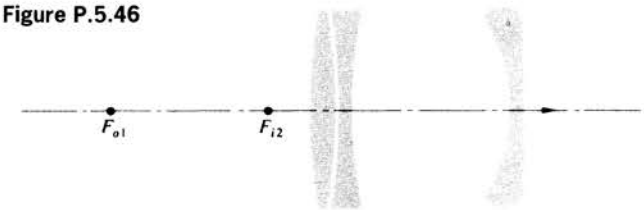
5.45 Make a sketch roughly locating the aperture stop and entrance and exit pupils for the lens in Fig. P.5.45.

Figure P.5.45



5.46 Make a sketch roughly locating the aperture stop and entrance and exit pupils for the lens in Fig. P.5.46, assuming the object point to be beyond (to the left of) F_{o1}.

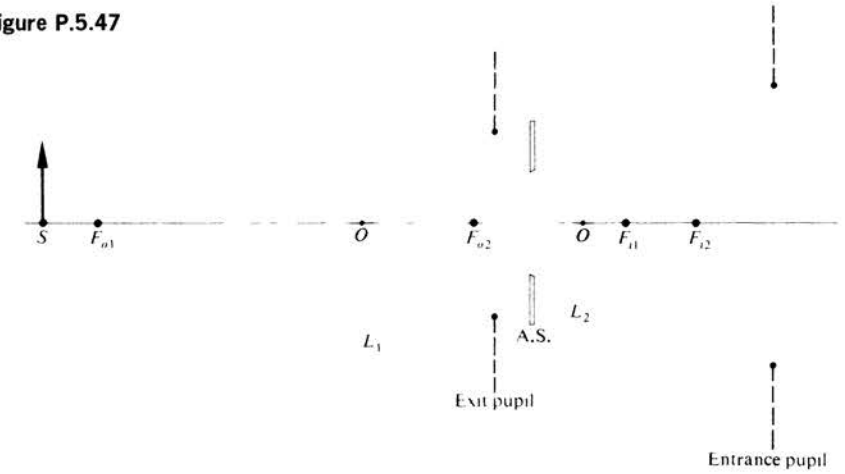
Figure P.5.46



5.47 Figure P.5.47 shows a lens system, an object, and the appropriate pupils. Diagrammatically locate the image.

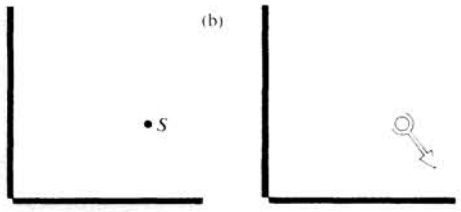
Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Figure P.5.47



5.48 Draw a ray diagram locating the images of a point source as formed by a pair of mirrors at 90° (Fig. P.5.48*a*). Now create a ray diagram locating the images of the arrow shown in Fig. P.5.48*b*.

Figure P.5.48 (a) (b)



5.49 Examine Velasquez's painting of *Venus and Cupid* (Fig. P.5.49). Is Venus looking at herself in the mirror? Explain.

Figure P.5.49 *The Toilet of Venus* by Diego Rodriguez de Silva y Veláquez (Courtesy of the Trustees, The National Gallery, London)



5.50 Manet's painting *The Bar at the Folies Bergères* (Fig. P.5.50) shows a girl standing in front of a large planar mirror. Reflected in it is her back and a man in evening dress with whom she appears to be talking. It would seem that Manet's intent was to give the uncanny feeling that the viewer is standing where that gentleman must be. From the laws of Geometrical Optics, what's wrong?

Figure P.5.50 *The Bar at the Folies Bergere* by Édouard Manet (Courtesy of the Courtauld Institute Galleries, London. Courtauld Collection.)



5.51 Show that Eq. (5.48) for a spherical surface is equally applicable to a plane mirror.

5.52 Locate the image of a paperclip 100 cm away from a convex spherical mirror having a radius of curvature of 80 cm.

5.53* Imagine that you are standing 5 feet from, and looking direct-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

ly toward, a brass ball 1 foot in diameter hanging in front of a pawn shop. Describe the image you would see in the ball.

5.54* A thin lens having a focal length of +50.0 cm is positioned 250 cm in front of (i.e., to the left of) a plane mirror. An ant sits on the central axis 250 cm in front of (i.e., to the left of) the lens. Locate the three images of the ant.

5.55 The image of a red rose is formed by a concave spherical mirror on a screen 100 cm away. If the rose is 25 cm from the mirror, determine its radius of curvature.

5.56 From the image configuration determine the shape of the mirror hanging on the back wall in van Eyck's painting of *John Arnolfini and His Wife* (Fig. P.5.56).

5.57* There are several varieties of retro-reflector that are commercially available; one type is comprised of transparent spheres, the backs of which are silvered. Light is refracted at the front surface, focused onto the rear surface, and there reflected back out in the direction it came. Determine the necessary index of refraction of the spheres. Assume the incident light is collimated.

5.58* Design an eye for a robot using a concave spherical mirror such that the image of an object 1.0 m tall and 10 m away fills its 1.0-cm-square photosensitive detector (which is movable for focusing purposes). Where should this detector be located with respect to the mirror? What should be the focal length of the mirror? Draw a ray diagram.

5.59 Design a little dentist's mirror to be fixed at the end of a shaft for use in the mouth of some happy soul. The requirements are (1) that the image be erect as seen by the dentist and (2) that when held 1.5 cm from a tooth the mirror produces an image twice life-size.

5.60 An object is located at a distance s_o from a spherical mirror of radius R. Show that the resulting image will be magnified by an amount

$$M_T = \frac{R}{2s_o + R}$$

5.61* A device used to measure the radius of curvature of the cornea of the eye is called a keratometer. This is useful information when fitting contact lenses. In effect, an illuminated object is placed a known distance from the eye, and the image reflected off the cornea is observed. The instrument allows the operator to measure the size of that virtual image. If the magnification is found to be 0.037× when the object distance is set at 100 mm, what is the radius of curvature?

5.62* Considering the operation of a spherical mirror, prove that the locations of the object and image are given by

$$s_o = f(M_T - 1)/M_T \quad \text{and} \quad s_i = -f(M_T - 1)$$

5.63 A man whose face is 25 cm away looks into the bowl of a soup-

Figure P.5.56 Detail of *John Arnolfini and His Wife* by Jan van Eyck. (Courtesy of the Trustees, The National Gallery, London)



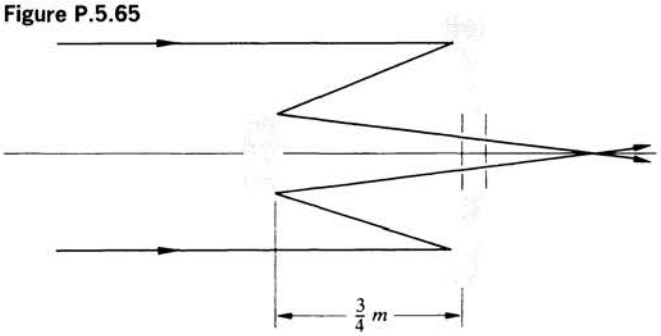
Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

spoon and sees his image reflected with a magnification of -0.064. Determine the radius of curvature of the spoon.

5.64* In an amusement park a large upright convex spherical mirror is facing a plane mirror 10.0 m away. A girl 1.0 m tall standing midway between the two sees herself twice as tall in the plane mirror as in the spherical one. In other words, the angle subtended at the observer by the image in the plane mirror is twice the angle subtended by the image in the spherical mirror. What is the focal length of the latter?

5.65* A homemade telephoto "lens" (Fig. P.5.65) consists of two spherical mirrors. The radius of curvature is 2.0 m for the primary and 60 cm for the secondary. How far from the smaller mirror should the film plane be located if the object is a star? What is the effective focal length of the system?

Figure P.5.65



5.66* A point source S sitting on the central axis of a positive thin lens is located (to the left) between one and two focal lengths from the lens. A concave spherical mirror is to be positioned to the right of the lens so that the final real image also lies at point S. Where should the mirror be placed? Where should a convex spherical mirror be located to accomplish the same feat?

5.67* Suppose you have a concave spherical mirror with a focal length of 10 cm. At what distance must an object be placed if its image is to be erect and one and a half times as large? What is the radius of curvature of the mirror? Check with Table 5.5.

5.68 Describe the image that would result for an object 3 inches tall placed 20 cm from a spherical concave shaving mirror having a radius of curvature of -60 cm.

5.69* Parallel rays along the central axis enter a biconcave lens, both of whose radii of curvature are equal. Some of the light is reflected from the first surface, and the remainder passes through the lens. Show that, if the index of refraction of the lens (which is surrounded by air) is 2.00, the reflected image will fall at the same point as the image formed by the lens.

5.70 Referring to the dove prism in Fig. 5.63, rotate it through 90° about an axis along the ray direction. Sketch the new configuration and determine the angle through which the image is rotated.

5.71 Determine the numerical aperture of a single clad optical fiber, given that the core has an index of 1.62 and the clad 1.52. When immersed in air, what is its maximum acceptance angle? What would happen to a ray incident at, say, 45°?

5.72 Given a fused silica fiber with an attenuation of 0.2 dB/km, how far can a signal travel along it before the power level drops by half?

5.73 The number of modes in a stepped-index fiber is provided by the expression

$$N_m \simeq \tfrac{1}{2}(\pi D \, \mathrm{NA}/\lambda_0)^2$$

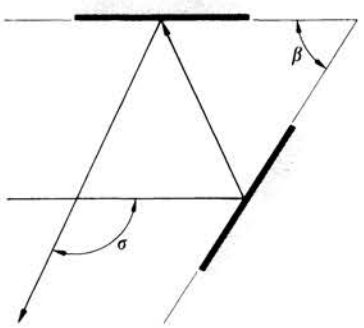
Given a fiber with a core diameter of 50 μm and $n_c = 1.482$ and $n_f = 1.500$, determine N_m when the fiber is illuminated by an LED emitting at a central wavelength of 0.85 μm.

5.74* Determine the intermodal delay (in ns/km) for a stepped-index fiber with a cladding of index 1.485 and a core of index 1.500.

5.75 Using the information on the eye in Section 5.7.1, compute the approximate size (in millimeters) of the image of the Moon as cast on the retina. The Moon has a diameter of 2160 miles and is roughly 230 000 miles from here, although this, of course, varies.

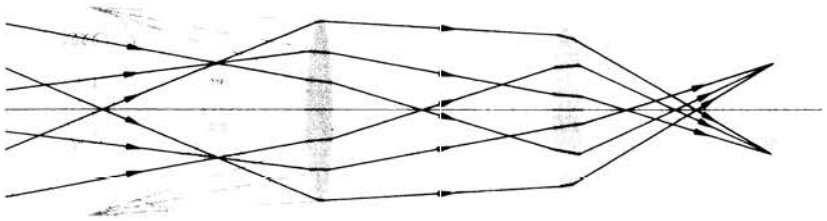
5.76* Figure P.5.76 shows an arrangement in which the beam is deviated through a constant angle σ, equal to twice the angle β between the plane mirrors, regardless of the angle-of-incidence. Prove that this is indeed the case.

Figure P.5.76



Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Figure P.5.78

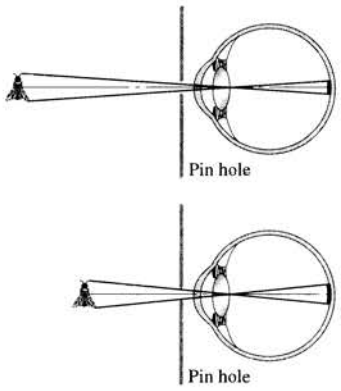


5.77 An object 20 m from the objective (f_o = 4 m) of an astronomical telescope is imaged 30 cm from the eyepiece (f_o = 60 cm). Find the total linear magnification of the scope.

5.78* Figure P.5.78, which purports to show an erecting lens system, is taken from an old, out-of-print optics text. What's wrong with it?

5.79* Figure P.5.79 shows a pin hole in an opaque screen being used for something practical. Explain what's happening and why it works. Try it.

Figure P.5.79



5.80* If a photograph of a moving merry-go-round is perfectly exposed, but blurred, at $\frac{1}{30}$ s and $f/11$, what must the diaphragm setting be if the shutter speed is raised to $\frac{1}{120}$ s in order to "stop" the motion?

5.81 The field of view of a simple two-element astronomical telescope is restricted by the size of the eye-lens. Make a ray sketch showing the vignetting that arises.

5.82 A *field-lens*, as a rule, is a positive lens placed at (or near) the intermediate image plane in order to collect the rays that would otherwise miss the next lens in the system. In effect, it increases the field of view without changing the power of the system. Redraw the ray diagram of the previous problem to include a field-lens. Show that as a consequence the eye relief is reduced somewhat.

5.83* Describe completely the image that results when a bug sits at the vertex of a thin positive lens. How does this relate directly to the manner in which a field-lens works? (See Problem 5.82.)

5.84* It is determined that a patient has a near point at 50 cm. If the eye is approximately 2.0 cm long,

(a) How much power does the refracting system have when focused on an object at infinity? when focused at 50 cm?
(b) How much accommodation is required to see an object at a distance of 50 cm?
(c) What power must the eye have to see clearly an object at the standard near-point distance of 25 cm?
(d) How much power should be added to the patient's vision system by a correcting lens?

5.85* An optometrist finds that a farsighted person has a near point at 125 cm. What power will be required for contact lenses if they are effectively to move that point inward to a more workable distance of 25 cm so that a book can be read comfortably? Use the fact that if the object is imaged at the near point, it can be seen clearly.

5.86 A farsighted person can see very distant mountains with relaxed eyes while wearing +3.2-D contact lenses. Prescribe spectacle lenses that will serve just as well when worn 17 mm in front of the cornea. Locate and compare the far point in both cases.

5.87* A jeweler is examining a diamond 5.0 mm in diameter with a loupe having a focal length of 25.4 mm.

(a) Determine the maximum angular magnification of the loupe.
(b) How big does the stone appear through the magnifier?
(c) What is the angle subtended by the diamond at the unaided eye when held at the near point?
(d) What angle does it subtend at the aided eye?

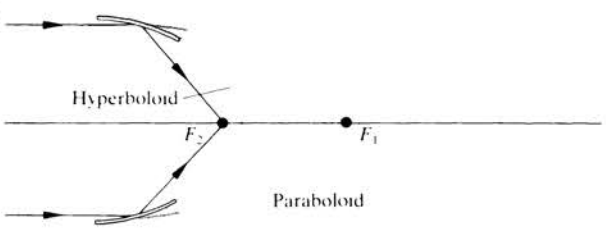
Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

5.88 Suppose we wish to make a microscope (that can be used with a relaxed eye) out of two positive lenses, both with a focal length of 25 mm. Assuming the object is positioned 27 mm from the objective, (a) how far apart should the lenses be, and (b) what magnification can we expect?

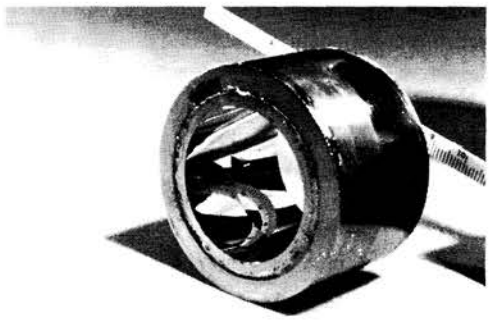
5.89* Figure P.5.89 shows a glancing-incidence X-ray focusing system designed in 1952 by Hans Wolter. How does it work? Microscopes with this type of system have been used to photograph, in X-rays, the implosion of fuel pellet targets in laser fusion research. Similar X-ray optical arrangements have been used in astronomical telescopes (see photos on p. 79).

Figure P.5.89

(a)

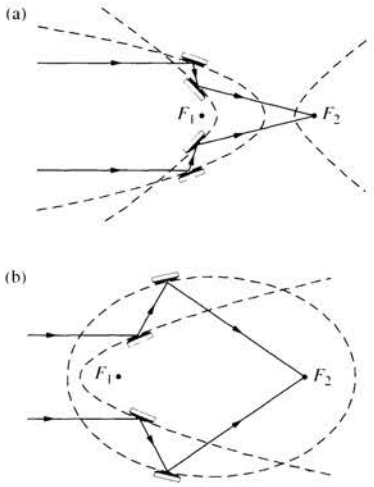


(b)



5.90* The two glancing-incidence aspherical mirror systems depicted in Fig. P.5.90 are designed to focus X-rays. Explain how each works: identify the shapes of the mirrors, discuss the locations of their various foci, and so on.

Figure P.5.90



5.91* The orbiting Hubble Space Telescope has a 2.4-m primary, which we will assume to be diffraction limited. Suppose we wanted to use it to read the print on the side of a distant Russian satellite. Assuming that a resolution of 1.0 cm at the satellite will do, how far away could it be from the HST?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

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 carries 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 little 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 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. 5.16, p. 160) that a ray traversing the

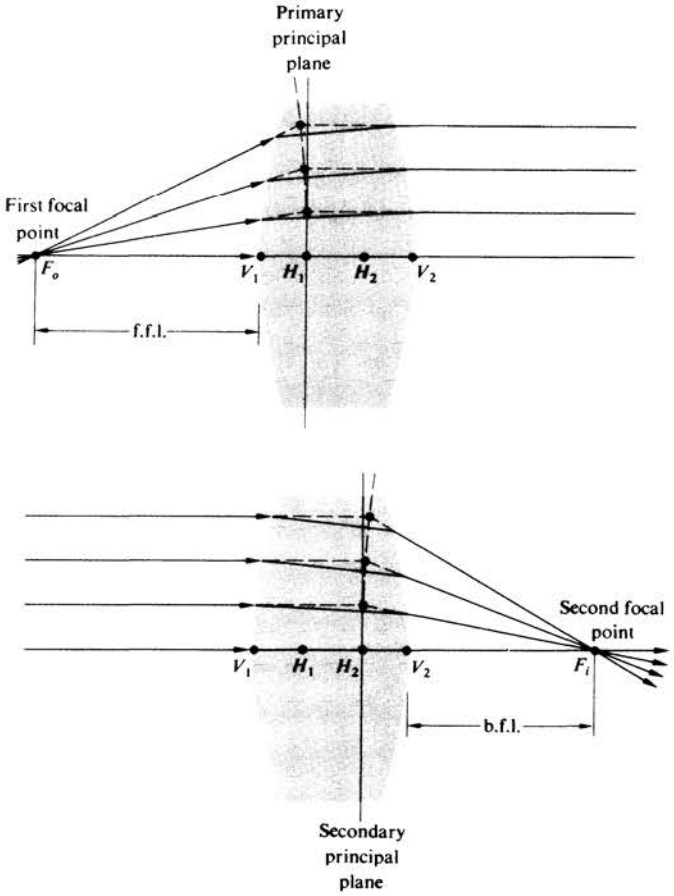


Figure 6.1 A thick lens.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

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, though 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 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_l 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_l 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 Gaussian form,

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

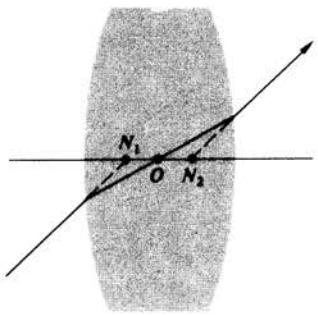


Figure 6.2 Nodal points.

*For the complete derivation, see Morgan, *Introduction to Geometrical and Physical Optics*, p. 57. We will be deriving much of this material using matrices in Section 6.2.1.

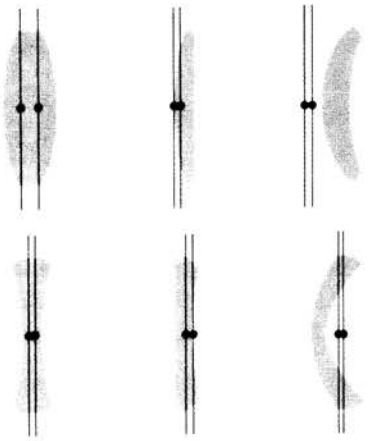


Figure 6.3 Lens bending.

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

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

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 the arrangement of the various quantities. The values of h_1 and h_2 are (Problem 6.18) given by

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

and

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

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

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

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

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

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

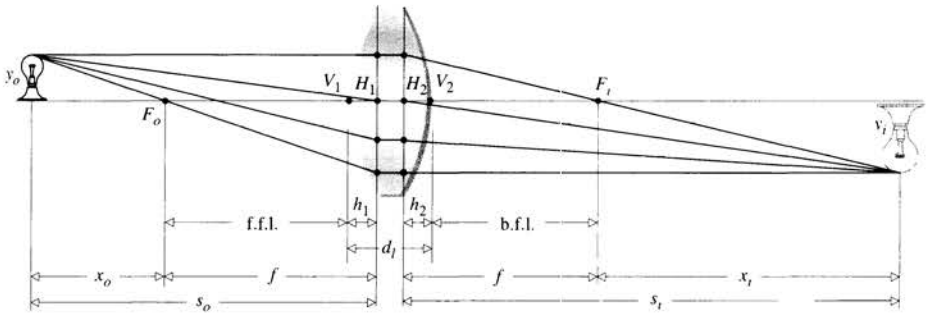


Figure 6.4 Thick-lens geometry.

Obviously, if $d_l \to 0$, Eqs. (6.1), (6.2), and (6.5) are transformed into the thin-lens expressions Eqs. (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

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

and $f = 26.8$ cm. Furthermore,

$$h_1 = -\frac{26.8(0.50)1}{-40(1.5)} = +0.22 \text{ cm}$$

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, $s_o = 30 + 0.22$, whereupon

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

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

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*. 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 com-

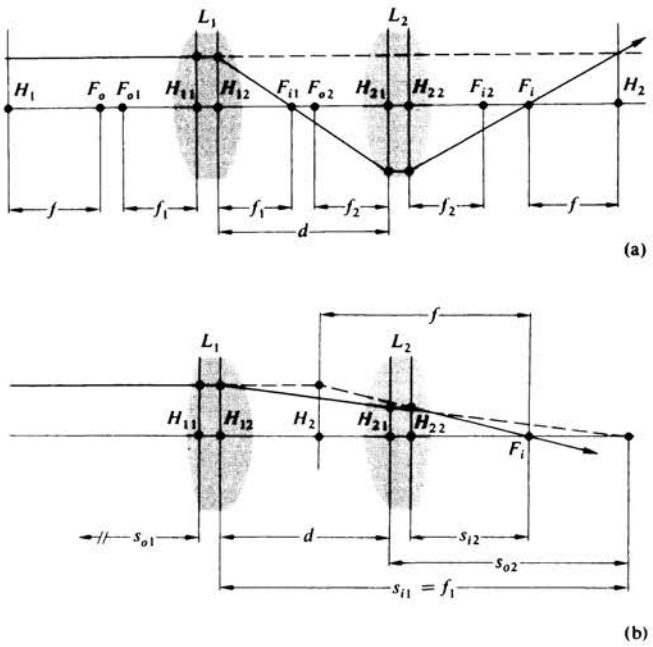


Figure 6.5 A compound thick lens.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

bination 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

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

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 $f_1 = -30$ cm, $f_2 = 20$ cm, and $d = 10$ cm, as in Fig. 6.6,

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

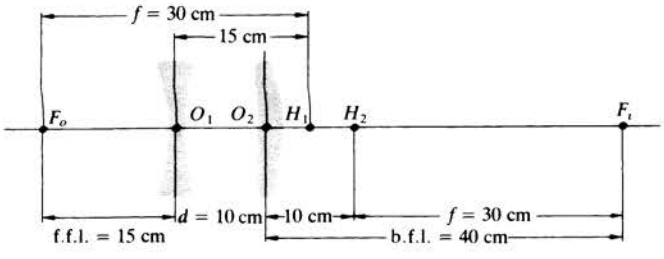


Figure 6.6 A compound lens.

so $f = 30$ cm. We found earlier (p. 168) 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

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

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.

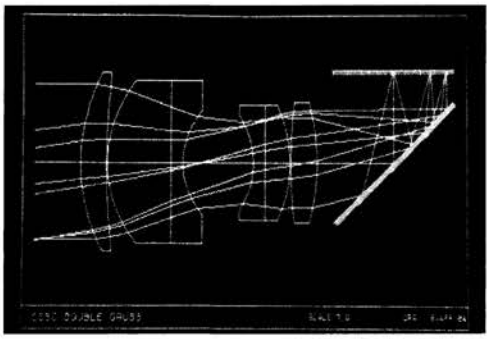
6.2 Analytical Ray Tracing

Ray tracing is unquestionably one of the designer's chief tools. Having formulated an optical system on paper, one can mathematically shine virtual 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

$$n_i(\hat{\mathbf{k}}_i \times \hat{\mathbf{u}}_n) = n_t(\hat{\mathbf{k}}_t \times \hat{\mathbf{u}}_n) \tag{4.6}$$

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 computer, which simply takes a trifle longer to make the

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



Computer ray tracing. (Photo courtesy of Optical Research Associates, Pasadena, California.)

trace. Thus, whereas it would probably take 10 or 15 minutes for a skilled person with a calculator to evaluate the trajectory of a single skew ray through a single surface, a computer would 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. 6.7 at point P_1 yields

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

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

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 yields

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

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

$$\mathscr{D}_1 = \frac{(n_{t1} - n_{i1})}{R_1}$$

Hence
$$n_{t1}\alpha_{t1} = n_{i1}\alpha_{i1} - \mathscr{D}_1 y_1 \tag{6.12}$$

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 $d_{21} \approx \overline{V_2V_1}$ and y_2 is easily computed. Equations (6.12) and (6.13) 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.

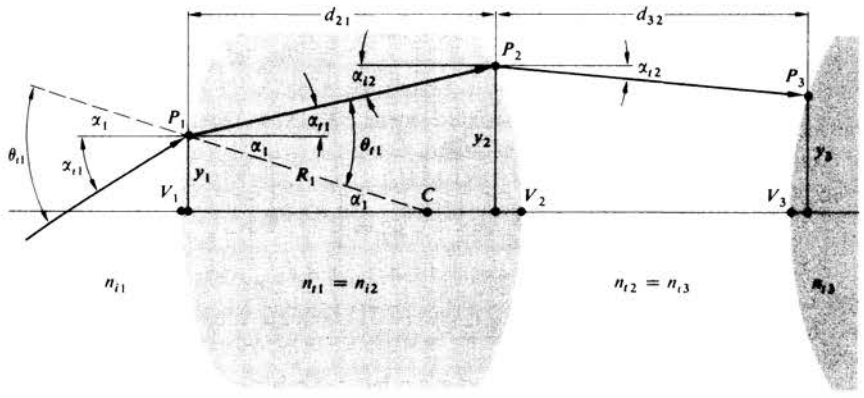


Figure 6.7 Ray geometry.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

6.2.1 Matrix Methods

In the beginning of the 1930s, T. Smith formulated an 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 30 years. However, the early 1960s saw a rebirth of interest in this approach.* We shall only outline some of the salient features of the method, leaving a more detailed study to the references.

Matrix Analysis of Lenses

Let's begin by writing the formulas

$$n_{t1}\alpha_{t1} = n_{i1}\alpha_{i1} - \mathscr{D}_1 y_{i1} \tag{6.14}$$

and

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

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 $\mathscr{r}_{t1}$ and $\mathscr{r}_{i1}$ for the two rays, we can write

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

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

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

and Eq. (6.16) can be stated concisely as

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

which just says that $\mathscr{R}_1$ transforms the rays $\mathscr{r}_{i1}$ into the ray $\mathscr{r}_{t1}$ during refraction at the first interface. Notice that the way we arranged the terms in Eqs. (6.14) and (6.15) determined the form of the refraction matrix. Accordingly, several equivalent variations of the matrix can be found in the literature.

From Fig. 6.7 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

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

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., $\mathscr{r}_{t1}$) and transforms it into the incident ray at P_2:

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

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

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

*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*.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

If use is made of Eq. (6.20), this becomes

$$\mathbf{r}_{i2} = \mathcal{T}_{21}\mathcal{R}_1\mathbf{r}_{i1} \tag{6.26}$$

The 2×2 matrix formed by the product of the transfer and refraction matrices $\mathcal{T}_{21}\mathcal{R}_1$ will carry the ray incident at P_1 into the ray incident at P_2. Notice that the determinant of $\mathcal{T}_{21}$, denoted by $|\mathcal{T}_{21}|$, equals 1; that is, $(1)(1) - (0)(d_{21}/n_{t1}) = 1$. Similarly $|\mathcal{R}_1| = 1$, and since the determinant of a matrix product equals the product of the individual determinants, $|\mathcal{T}_{21}\mathcal{R}_1| = 1$. This provides a quick check on the computations. Carrying the procedure through the second interface (Fig. 6.7) of the lens, which has a refraction matrix $\mathcal{R}_2$, it follows that

$$\mathbf{r}_{t2} = \mathcal{R}_2\mathbf{r}_{i2} \tag{6.27}$$

where

$$\mathcal{R}_2 \equiv \begin{bmatrix} 1 & -\mathcal{D}_2 \\ 0 & 1 \end{bmatrix}$$

and the power of the second surface is

$$\mathcal{D}_2 = \frac{(n_{t2} - n_{i2})}{R_2}$$

From Eq. (6.26)

$$\mathbf{r}_{t2} = \mathcal{R}_2\mathcal{T}_{21}\mathcal{R}_1\,\mathbf{r}_{i1} \tag{6.28}$$

The **system matrix** $\mathcal{A}$ is then defined as

$$\mathcal{A} \equiv \mathcal{R}_2\mathcal{T}_{21}\mathcal{R}_1 \tag{6.29}$$

and has the form

$$\mathcal{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{6.30}$$

Inasmuch as

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

or

$$\mathcal{A} = \begin{bmatrix} 1 & -\mathcal{D}_2 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -\mathcal{D}_1 \\ \dfrac{d_{21}}{n_{t1}} & 1 - \dfrac{\mathcal{D}_1 d_{21}}{n_{t1}} \end{bmatrix}$$

it follows that

$$\mathcal{A} = \begin{bmatrix} 1 - \dfrac{\mathcal{D}_2 d_{21}}{n_{t1}} & -\mathcal{D}_1 - \mathcal{D}_2 + \dfrac{\mathcal{D}_2\mathcal{D}_1 d_{21}}{n_{t1}} \\ \dfrac{d_{21}}{n_{t1}} & 1 - \dfrac{\mathcal{D}_1 d_{21}}{n_{t1}} \end{bmatrix}$$

and again $|\mathcal{A}| = 1$ (Problem 6.17). Because we are working with only one lens, let's simplify the notation a little letting $d_{21} = d_l$ and $n_{t1} = n_l$. Consequently,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 1 - \dfrac{\mathcal{D}_2 d_l}{n_l} & -\mathcal{D}_1 - \mathcal{D}_2 + \dfrac{\mathcal{D}_1\mathcal{D}_2 d_l}{n_l} \\ \dfrac{d_l}{n_l} & 1 - \dfrac{\mathcal{D}_1 d_l}{n_l} \end{bmatrix} \tag{6.31}$$

The value of each element in $\mathcal{A}$ is expressed in terms of the physical lens parameters, such as thickness, index, and radii (via $\mathcal{D}$). Thus the cardinal points, which are properties of the lens determined solely by its makeup, should be deducible from $\mathcal{A}$. The system matrix in this case, Eq. (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 $\mathcal{A}_{21}$.

The concept of image formation enters rather directly (Fig. 6.8) after introduction of appropriate object and image planes. Consequently, the first operator $\mathcal{T}_{1O}$ transfers the reference point from the object (i.e., P_O to P_1). The next operator $\mathcal{A}_{21}$ then carries the ray through the lens, and a final transfer $\mathcal{T}_{I2}$ brings it to the image plane (i.e., P_I). Thus the ray at the image point ($\mathbf{r}_I$) is given by

$$\mathbf{r}_I = \mathcal{T}_{I2}\mathcal{A}_{21}\mathcal{T}_{1O}\mathbf{r}_O \tag{6.32}$$

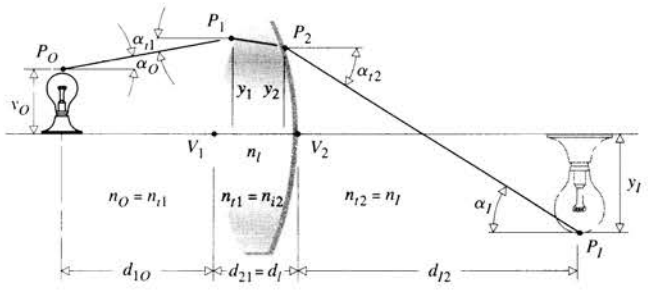


Figure 6.8 Image geometry.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

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

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

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

$$-a_{12} = \mathscr{D}_1 + \mathscr{D}_2 - \mathscr{D}_1\mathscr{D}_2 d_I/n_I$$

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

$$\mathscr{D}_1 = \frac{n_I - 1}{R_1} \quad \text{and} \quad \mathscr{D}_2 = \frac{n_I - 1}{-R_2}$$

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

$$-a_{12} = (n_I - 1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n_I - 1)d_I}{R_1R_2n_I}\right]$$

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

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

Thus the power of the lens as a whole is given by

$$-a_{12} = \mathscr{D}_I = \mathscr{D}_1 + \mathscr{D}_2 - \frac{\mathscr{D}_1\mathscr{D}_2 d_I}{n_I}$$

If the embedding media were different on each side of the lens (Fig. 6.9), this would become

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

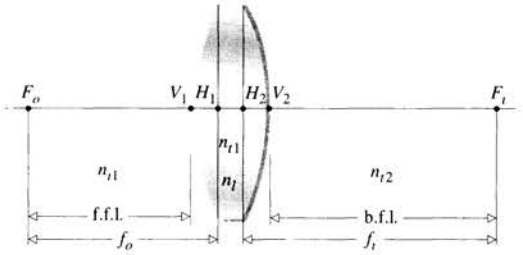


Figure 6.9 Principal planes and focal lengths.

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.

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. 6.10. 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

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

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

*This particular example was chosen primarily because Nussbaum's book *Geometric Optics* contains a simple Fortran computer program written specifically for this lens. It would be almost silly to evaluate the system matrix by hand.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

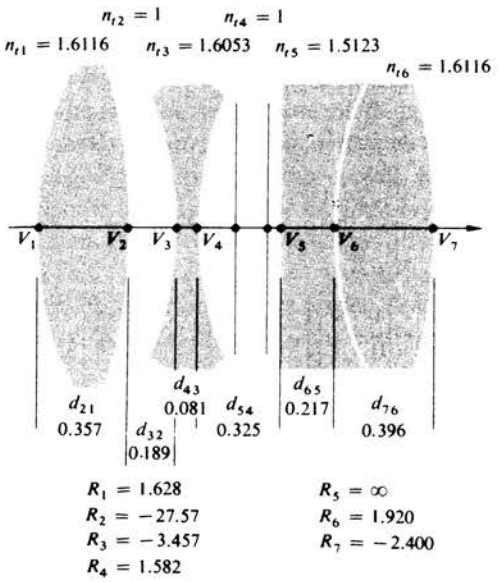


Figure 6.10 A Tessar.

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, though 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 $\overline{V_7H_2} = -0.67$.

Thin Lenses

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 $d_l \to 0$, it corresponds to a thin lens. This is equivalent to making $\mathscr{T}_{21}$ a unit matrix; thus

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

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

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

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

$$\mathscr{A} = \begin{bmatrix} 1 - d/f_2 & -1/f_1 + d/f_1f_2 - 1/f_2 \\ d & -d/f_1 + 1 \end{bmatrix}$$

Clearly then,

$$-a_{12} = \frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1f_2}$$

and from Eqs. (6.36) and (6.37)

$$\overline{O_1H_1} = fd/f_2 \qquad \overline{O_2H_2} = -fd/f_1$$

all of which should be quite familiar by now. 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.

Matrix Analysis of Mirrors

To derive the appropriate matrix for reflection, consult Fig. 6.11, which depicts a concave spherical mirror, and write down two equations that describe the incident and reflected rays. Again, the final form of the matrix depends on how we arrange these two equations and the signs we assign to the various quantities. What's needed is an expression relating the ray angles and another relating their heights at the point of interaction with the mirror.

First let's consider the ray angles. The Law of Reflection is $\theta_i = \theta_r$; therefore from the geometry $\tan(\alpha_i - \theta_i) = y_i/R$, and

$$(\alpha_i - \theta_i) \approx y_i/R \tag{6.38}$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

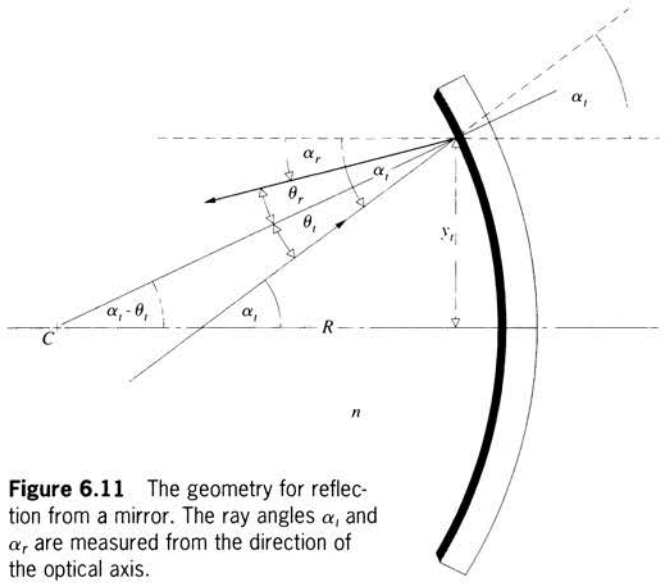


Figure 6.11 The geometry for reflection from a mirror. The ray angles α_i and α_r are measured from the direction of the optical axis.

Taking these angles to be positive, y is positive, but R isn't, and this equation will be in error as soon as we enter a negative value for the radius. Therefore rewrite it as $(\alpha_i - \theta_i) = -y_i/R$. Now to get α_r into the analysis, note that $\alpha_i = \alpha_r + 2\theta_i$ and $\theta_i = (\alpha_i - \alpha_r)/2$. Substituting this into Eq. (6.38) yields $\alpha_r = -\alpha_i - 2y_i/R$, and multiplying by n, the index of the surrounding medium (where usually $n = 1$), leads to

$$n\alpha_r = -n\alpha_i - 2ny_i/R$$

The second necessary equation is simply $y_r = y_i$ and so

$$\begin{bmatrix} n\alpha_r \\ y_r \end{bmatrix} = \begin{bmatrix} -1 & -2n/R \\ 0 & 1 \end{bmatrix} \begin{bmatrix} n\alpha_i \\ y_i \end{bmatrix}$$

Thus the mirror matrix $\mathcal{M}$ for a spherical configuration is given by

$$\mathcal{M}_o = \begin{bmatrix} -1 & -2n/R \\ 0 & 1 \end{bmatrix} \tag{6.39}$$

remembering from Eq. (5.49) that $f = -2/R$.

Flat Mirrors and the Planar Optical Cavity

For a flat mirror ($R \to \infty$) in air ($n = 1$), the matrix is

$$\mathcal{M}_| = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

where the minus sign in the first position reverses the ray upon reflection. Figure 6.12 shows two planar mirrors facing each other forming an **optical cavity** (p. 591). Light leaves point O, traverses the gap in the positive direction, is reflected by mirror-1, retraces the gap in the negative direction, and is reflected by mirror-2. The system matrix is

$$\mathcal{A} = \mathcal{M}_{|2}\mathcal{T}_{21}\mathcal{M}_{|1}\mathcal{T}_{12}$$

$$\mathcal{A} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -d & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ d & 1 \end{bmatrix}$$

and

$$\mathcal{A} = \begin{bmatrix} 1 & 0 \\ 2d & 1 \end{bmatrix}$$

where again the determinant of the system matrix is one: $|\mathcal{A}| = 1$. Presumably, if the initial ray was axial ($\alpha = 0$), the system matrix should bring it back to its starting point so that the final ray $\mathbf{r}_f$ is identical to the initial ray $\mathbf{r}_i$. That is,

$$\mathcal{A}\mathbf{r}_i = \mathbf{r}_f = \mathbf{r}_i$$

This is a special kind of mathematical relationship known as an ***eigenvalue equation*** where, a bit more generally,

$$\mathcal{A}\mathbf{r}_i = a\mathbf{r}_i$$

and a is a constant. In other words,

$$\begin{bmatrix} 1 & 0 \\ 2d & 1 \end{bmatrix} \begin{bmatrix} \alpha_i \\ y_i \end{bmatrix} = a \begin{bmatrix} \alpha_i \\ y_i \end{bmatrix}$$

If $\alpha_i = 0$ and the initial ray is launched axially, then $y_i = ay_i$

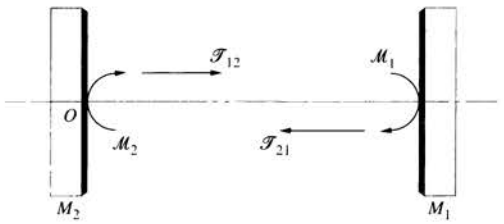


Figure 6.12 A schematic representation of a planar cavity formed by mirrors M_1 and M_2.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

and it follows that $a = 1$. The system matrix functions like a unit matrix that carries $\blacktriangleleft_i$ into $\blacktriangleleft_i$ after two reflections. Axial rays of light travel back and forth across the so-called ***resonant cavity*** without escaping.

Cavities can be constructed in a number of different ways using a variety of mirrors (Fig. 13.12, p. 594). If after traversing a cavity some number of times the light ray returns to its original location and orientation, the beam will be trapped and the cavity is said to be *stable*; that's why the eigenvalue discussion is important. To analyze the *confocal cavity* composed of two concave spherical mirrors facing each other, see Problem 6.24.

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 quasimonochromatic, and they in turn fall into two subgroupings. There are monochromatic aberrations such as *spherical aberration, coma*, and *astigmatism* that deteriorate the image, making it unclear. 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 farther 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.

6.3.1 Monochromatic Aberrations

The paraxial treatment was based on the assumption that $\sin\varphi$, as in Fig. 5.6, could be represented satisfactorily by φ alone; that is, the system was restricted to operating in an extremely narrow region about the optical axis. Obviously, if rays from the periphery of a lens are to be included in the formation of an image, the statement $\sin\varphi \approx \varphi$ is somewhat unsatisfactory. Recall that we also occasionally wrote Snell's Law simply as $n_i\theta_i = n_t\theta_t$, which again would be inappropriate. In any event, if the first two terms in the expansion

$$\sin\varphi = \varphi - \frac{\varphi^3}{3!} + \frac{\varphi^5}{5!} - \frac{\varphi^7}{7!} + \cdots \qquad [5.7]$$

are retained as an improved approximation, we have the so-called *third-order theory*. Departures from first-order theory that then result are embodied in the five *primary aberrations* (spherical aberration, coma, astigmatism, field curvature, and distortion). These were first studied in detail by Ludwig von Seidel (1821–1896) in the 1850s. Accordingly, they are frequently spoken of as the ***Seidel aberrations***. In addition to the first two contributions, the series obviously contains many other terms, smaller to be sure, but still to be reckoned with. Thus there are most certainly *higher-order aberrations*. The difference between the results of exact ray tracing and the computed primary aberrations can therefore be thought of as the sum of all contributing higher-order aberrations. We shall restrict this discussion exclusively to the primary aberrations.

Spherical Aberration

Let's return for a moment to Section 5.2.2 (p. 153), where we computed the conjugate points for a single refracting spherical interface. We found that for the paraxial region,

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

If the approximations for ℓ_o and ℓ_i are improved a bit (Prob-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

lem 6.26), we get the third-order expression:

$$\frac{n_1}{s_o}+\frac{n_2}{s_i}=\frac{n_2-n_1}{R}+h^2\left[\frac{n_1}{2s_o}\left(\frac{1}{s_o}+\frac{1}{R}\right)^2+\frac{n_2}{2s_i}\left(\frac{1}{R}-\frac{1}{s_i}\right)^2\right] \tag{6.40}$$

The additional term, which varies approximately as h^2, is clearly a measure of the deviation from first-order theory. As shown in Fig. 6.13, rays striking the surface at greater distances above the axis (h) are focused nearer the vertex. In brief, spherical aberration, or SA, corresponds to a dependence of focal length on aperture for nonparaxial rays. Similarly, for a converging lens, as in Fig. 6.14, the marginal rays will, in effect, be bent too much, being focused in front of the paraxial rays. Keep in mind that spherical aberration pertains only to object points that are on the optical axis. The distance between the axial intersection of a ray and the paraxial focus, F_i, is known as the **longitudinal spherical aberration**, or L · SA, of that ray. In this case, the SA is *positive*. In contrast, the marginal rays for a diverging lens will generally intersect the axis behind the paraxial focus, and its spherical aberration is therefore *negative*.

To better appreciate aberrations in terms of their effects on the wavefronts, consider the light from a point source traversing an optical system. Ideally, if the transmitted wavefront at the exit pupil is a sphere centered on the Gaussian image point (P), then the image is perfect; if not, it's aberrated (Fig. 6.15). ***Wave (or wavefront) aberrations*** are the deviations in optical path length between the actual and ideal wavefronts, often specified by the maximum values given in microns, nanometers, or wavelengths. Thus, the peak-to-peak deviation of the wavefront in Fig. 6.15 from the ideal spherical surface converging to P is some fraction of a wavelength, λ/N. With this in mind J.W. Strutt, better known as Lord Rayleigh, suggested a practical criterion of optical quality: an optical instrument will produce a noticeably degraded image when the wavefront aberration at 550 nm (yellow-green) exceeds $\lambda/4$.

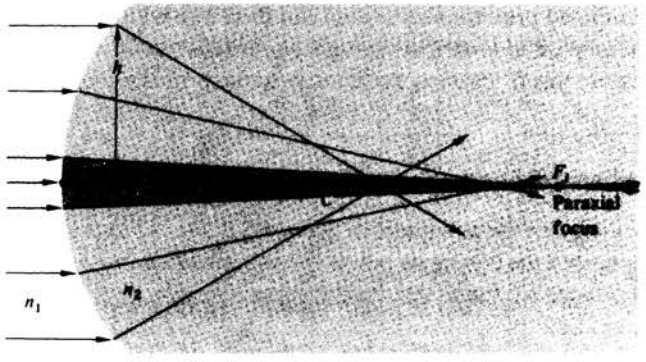


Figure 6.13 Spherical aberration resulting from refraction at a single interface.

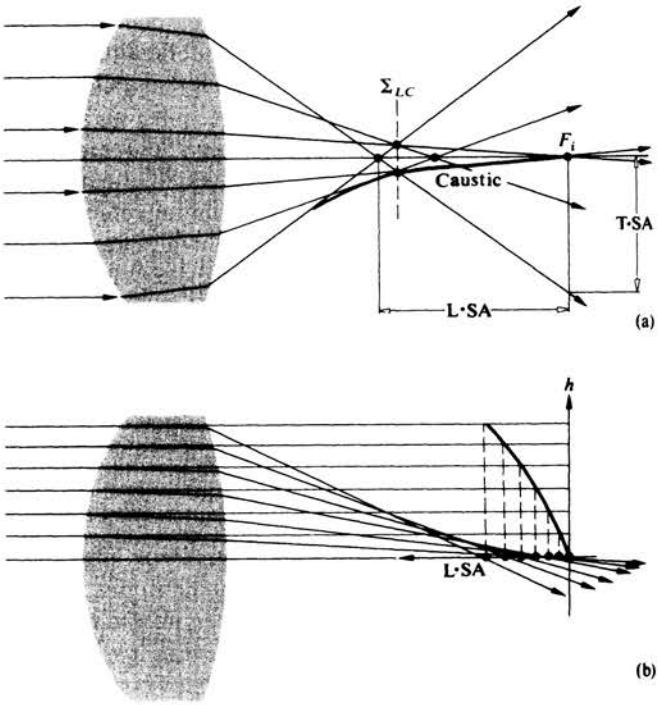


Figure 6.14 Spherical aberration for a lens. The envelope of the refracted rays is called a caustic. The intersection of the marginal rays and the caustic locates Σ_{LC}.

The idea of an optical system forming a point image is, of course, physically unrealistic (if for no other reason than the irradiance would be infinite and Nature abhors infinities). Under the very best conditions, a lens will form an image of a point source (e.g., a star) which is a tiny bright circular disk surrounded by rings that are so faint they're hardly noticeable (see the photo on p. 470 and Fig. 10.23); that's the Airy pattern. In Fig. 6.15 it's represented at P by a tall central irradiance peak surrounded by tiny maxima corresponding to a cross section of the rings.

Spherical aberration essentially shifts light out of the central disk into the surrounding rings, which become far more

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

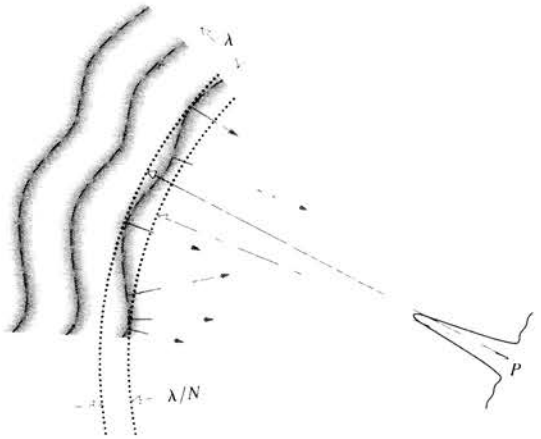


Figure 6.15 Since this wavefront deviates from a portion of a sphere (converging to the Gaussian image point), it is said to be aberrated. The extent of that deviation measured peak-to-peak is an indication of how far from perfection the image will be.

prominent. For example, Rayleigh established that one quarter-wave of spherical aberration diminishes the irradiance of the image disk by about 20%. You can see how that happens in general in Fig. 6.15 where rays (perpendicular to the distorted wavefront) head out away from the central spot toward the rings. Notice that even if the overall wavefront deviation is $\lambda/4$, when the wavefront has tight wiggles in it, lots of light will go out to the rings creating a hazy image. That's what you can expect if the surfaces are not smooth.

Returning to Fig. 6.14, if a screen is placed at F_i the image of a star will appear as a bright central spot on the axis surrounded by a symmetrical halo delineated by the cone of marginal rays. For an extended image, SA would reduce the contrast and degrade the details.

The height above the axis where a given ray strikes this screen is called the **traverse** (or **lateral**) **spherical aberration**, or T·SA for short. Evidently, SA can be reduced by stopping down the aperture—but that reduces the amount of light entering the system as well. Notice that if the screen is moved to the position labeled Σ_{LC}, the image blur will have its smallest diameter. This is known as the ***circle of least confusion***, and Σ_{LC} is generally the best place to observe the image. If a lens exhibits appreciable SA, it will have to be refocused after it is stopped down because the position of Σ_{LC} will approach F_i as the aperture decreases.

The amount of spherical aberration, when the aperture and focal length are fixed, varies with both the object distance and the lens shape. For a converging lens, the nonparaxial rays are too strongly bent. Yet if we imagine the lens as roughly resembling two prisms joined at their bases, it is evident that *the incident ray will undergo a minimum deviation when it makes, more or less, the same angle as does the emerging ray* (Section 5.5.1). A striking example is illustrated in Fig. 6.16, where simply turning the lens around markedly reduces the SA. When the object is at "infinity," a simple concave or convex lens that has an almost, but not quite, flat rear side will suffer a minimum amount of spherical aberration. In the same way, if the object and image distances are to be equal ($s_o = s_i = 2f$), the lens should be equiconvex to minimize SA. A combination of a converging and a diverging lens (as in an achromatic doublet) can also be utilized to diminish spherical aberration.

Recall that the aspherical lenses of Section 5.2.1 were completely free of spherical aberration for a specific pair of conjugate points. Moreover, Huygens seems to have been the first to discover that two such axial points exist for spherical surfaces as well. These are shown in Fig. 6.17*a*, which depicts rays issuing from P and leaving the surface as if they came from P'. It is left as a problem to show that the appropriate locations of P and P' are those indicated in the figure. Just as with the aspherical lenses, spherical lenses can be formed that have this same zero SA for the pair of points P and P'. One simply

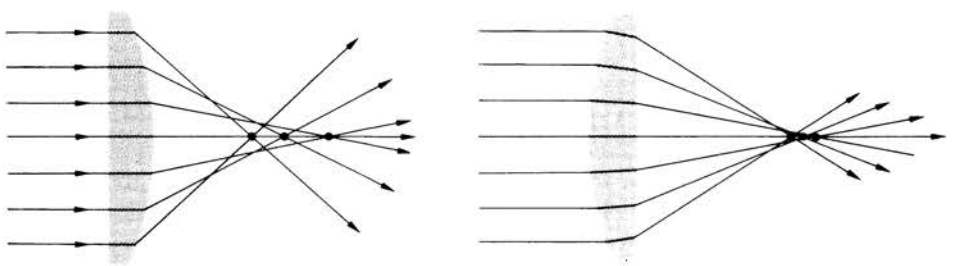


Figure 6.16 SA for a planar-convex lens.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

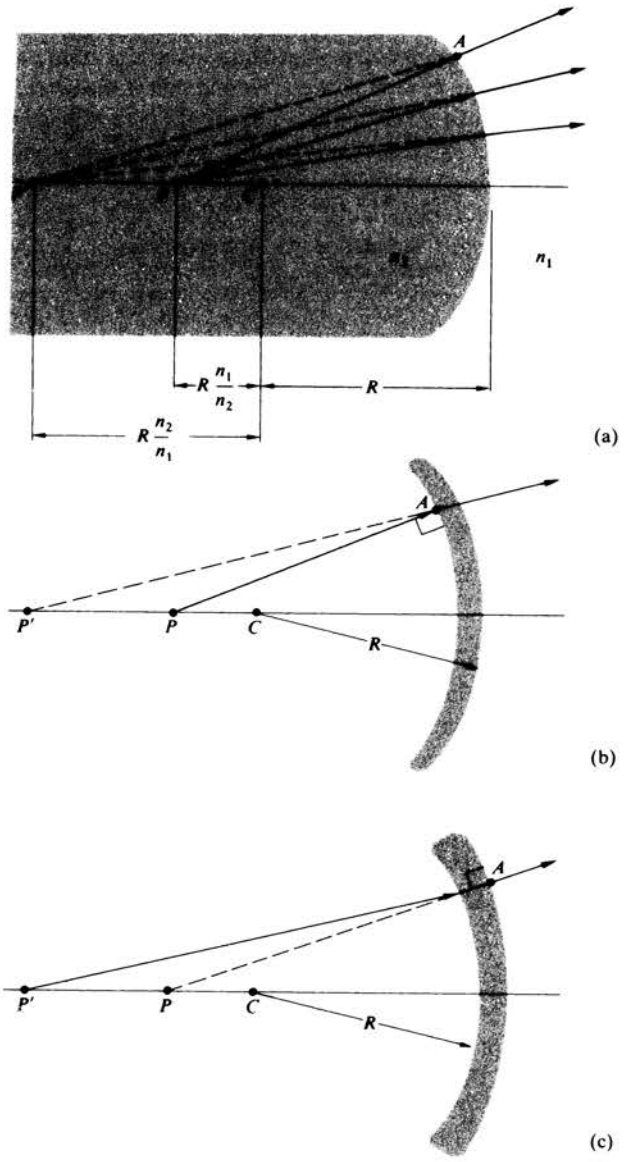


Figure 6.17 Corresponding axial points for which SA is zero.

grinds another surface of radius $\overline{PA}$ centered on P to form either a positive- or negative-meniscus lens. The oil-immersion microscope objective uses this principle to great advantage. The object under study is positioned at P and surrounded by oil of index n_2, as in Fig. 6.18. P and P' are the proper conjugate points for zero SA for the first element, and P' and P'' are those for the meniscus lens.

Soon after the Hubble Space Telescope (HST) was placed in orbit in April 1990, it became obvious that there was something terribly wrong. The pictures it was returning remained blurred, despite all attempts to adjust the orientation and location of the secondary mirror (p. 224). For a distant star, which was essentially a point source, the size of the image disk was close to the expected diffraction-limited value (about 0.1 arcsecond in diameter), but only about 12% of the radiant energy was there, instead of the expected 70% (roughly 84% is the ideal limit). The disk was surrounded by a halo extending out in diameter to about 1.5 arc-seconds containing some 70% of the light. The remaining radiant energy was unavoidably distributed beyond the halo in a radial tendril pattern as a result of a combination of mirror micro-roughness and diffraction from the struts holding the secondary (Fig. 6.19*b*). The situation was a classic example of spherical aberration.

As scientists later determined, the primary mirror (p.180) had been polished incorrectly; it was too flat at its periphery by about half a wavelength. Rays from its central region were focusing on the optical axis in front of those from the edges. The people at Perkin-Elmer, the company that fashioned the 2.4-m hyperboloid, had polished it superbly well, but to the wrong figure, or curvature. A series of blunders, starting with a 1.3-mm error in the position of a component in the shape-testing device, ultimately led to the flaw. The $1.6 billion telescope ended up with a debilitating longitudinal spherical aberration of 38 mm (Fig. 6.19*a*).

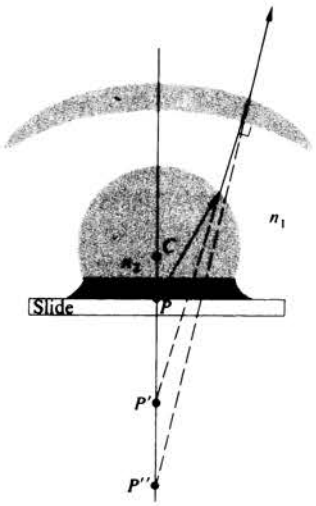


Figure 6.18 An oil-immersion microscope objective.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

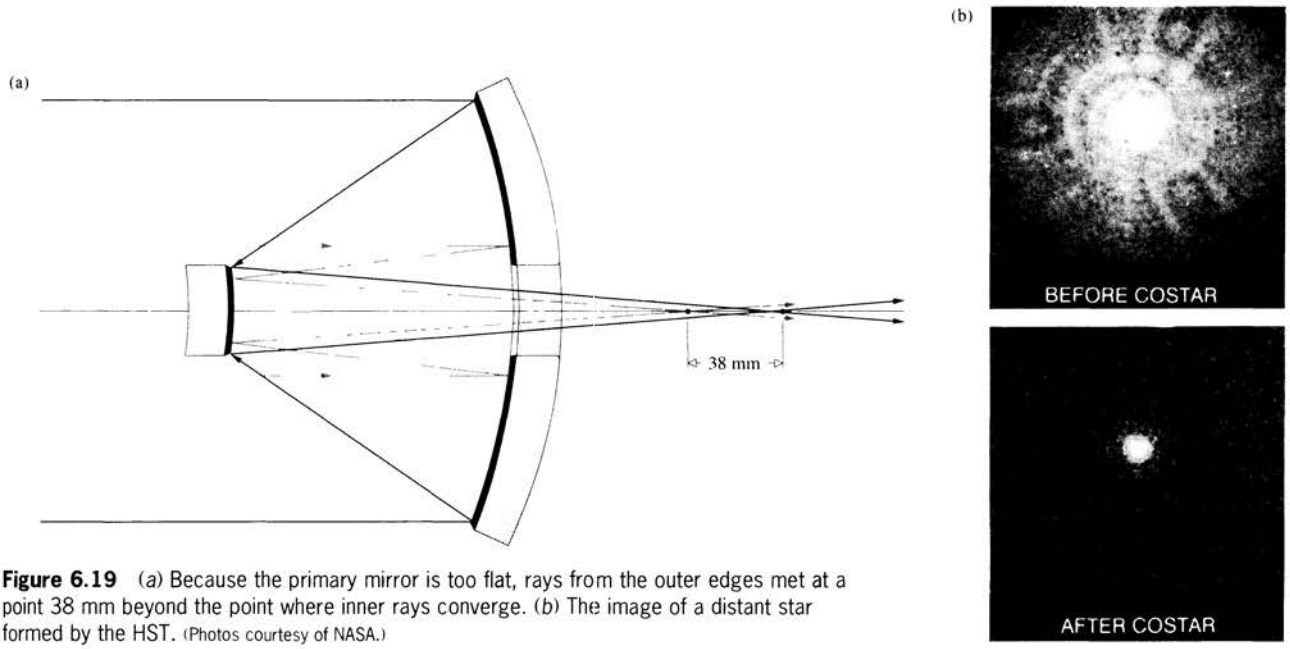
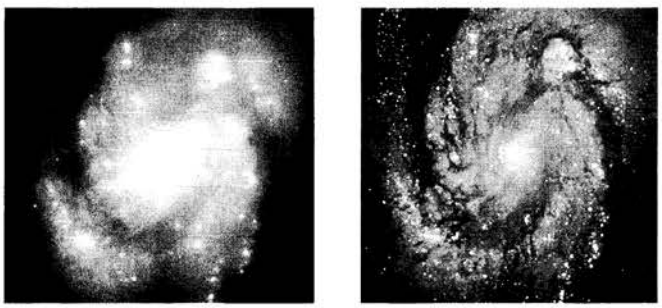


Figure 6.19 (*a*) Because the primary mirror is too flat, rays from the outer edges met at a point 38 mm beyond the point where inner rays converge. (*b*) The image of a distant star formed by the HST. (Photos courtesy of NASA.)

In 1993 astronauts from the *Endeavor* Space Shuttle successfully executed a dramatic repair mission. They installed a new Wide-Field Planetary Camera (with its own corrective optics that adds about half a wavelength to the edges) and the Corrective Optics Space Telescope Axial Replacement (COSTAR) module. The job of COSTAR is to reshape the aberrated wavefronts entering the three remaining scientific instruments. It inserts a pair of small mirrors (10 mm and 30 mm) into the beam heading toward each instrument aperture. One of these mirrors simply redirects the light to the other, which is a complex asymmetrical aspheric. That off-axis correcting mirror is configured with the inverse of the spherical aberration of the primary, so that upon reflection the wavefront is reshaped into a perfect wave directed toward the intended aperture. Now better than 70% of the light energy resides in the central image disk, and celestial objects are about 6.5 times brighter than before. People at NASA like to point out that with its vision clearer than ever (see photo), and its light-gathering ability improved, the HST could now spot a firefly over a distance equivalent to roughly halfway around the world. (Of course, the bug would have to stay stuck at maximum emission for about 90 minutes.) Moreover, the HST could distinguish two such persistent fireflies provided they were at least 3 m apart.

The Arecibo Observatory in Puerto Rico is home to the largest single radiotelescope in the world. Its objective is a 1000-ft diameter stationary spherical dish antenna operating at wavelengths from 3 cm to 6 m. By comparison, steerable radiotelescopes (p.179) are usually parabolic because that configuration can focus the radiation from a source, located in the forward direction, to a small axial image spot. But the 1000-ft dish had to be immobile, and so its designers opted for a compromise. The primary mirror was made spherical and therefore could collect radiation from a wide range of directions, in each



HST images of the M-100 galaxy with (before repair) and without (after repair) spherical aberration. (Photos courtesy of NASA.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

case focusing it at a "point" along the axis connecting the dish and the source. High above the mirror, they suspended a movable radio receiver whose position determined which part of the sky the telescope was looking at. Nonetheless, although a spherical mirror is omnidirectional, it's also equally imperfect in all directions. It suffers from spherical aberration just like the convex lens (Fig. 6.14). Instead of a single focal point there is an axial focal line. This was dealt with as best as possible by detecting the signals at several axial points and combining them via so-called line feeds, but the scheme was inefficient and the instrument rarely operated at its full potential.

In 1997 the Arecibo telescope underwent a major upgrade with the installation of a set of off-axis aspherical mirrors (Fig. 6.20) that compensate for spherical aberration in much the same way as do the corrective mirrors added to the Hubble Space Telescope. Named after James Gregory, the man who introduced a reflecting telescope with a concave secondary in 1661 (p. 223), the 90-ton Gregorian receiver dome is suspended 450 ft above the main reflector. Within an aluminum housing it contains a 72-ft diameter secondary mirror that receives the upwardly reflected EM-radiation from the primary. It, in turn, reflects this radiation down onto a 26-ft diameter tertiary mirror that focuses the beam upward to a spot at the receiver. The surfaces are so configured that the optical path length traversed by each ray is identical and all arrive (within a one-eight-inch circle) at the focus in-phase.

The device can also operate in reverse as a 1-megawatt radar transmitter, which is used for planetary studies. By sending out and receiving back reflected radar signals, the telescope can resolve features about a half mile across on the surface of Venus. It could detect a conductor the size of a golf ball on the Moon.

Coma

Coma, or *comatic aberration*, is an image-degrading, monochromatic, primary aberration associated with an object point even a short distance from the axis. Its origins lie in the fact that the principal "planes" can actually be treated as planes only in the paraxial region. They are, in fact, principal curved surfaces (Fig. 6.1). In the absence of SA, a parallel bundle of

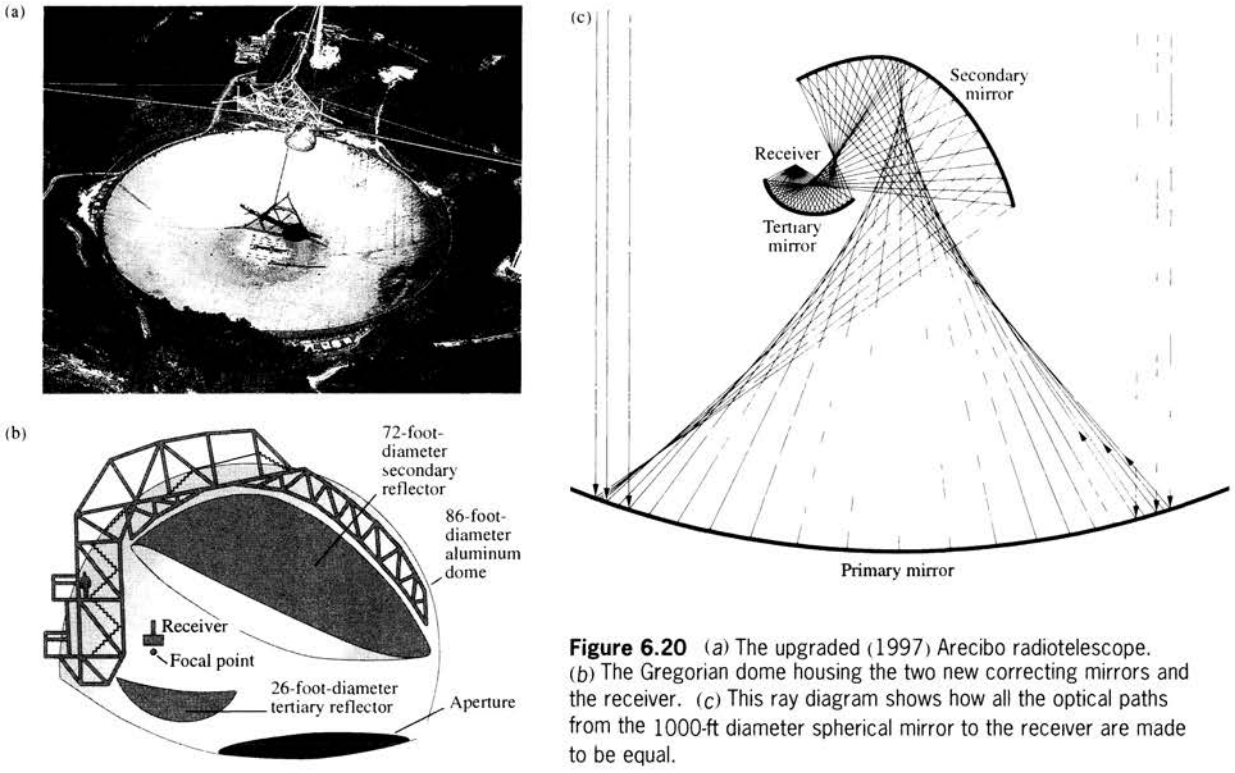


Figure 6.20 (*a*) The upgraded (1997) Arecibo radiotelescope. (*b*) The Gregorian dome housing the two new correcting mirrors and the receiver. (*c*) This ray diagram shows how all the optical paths from the 1000-ft diameter spherical mirror to the receiver are made to be equal.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

rays will focus at the axial point F_i, a distance b.f.l. from the rear vertex. Yet the effective focal lengths, and therefore the transverse magnifications, will differ for rays traversing off-axis regions of the lens. When the image point is on the optical axis, this situation is of little consequence, but when the ray bundle is oblique and the image point is off-axis, coma will be evident.

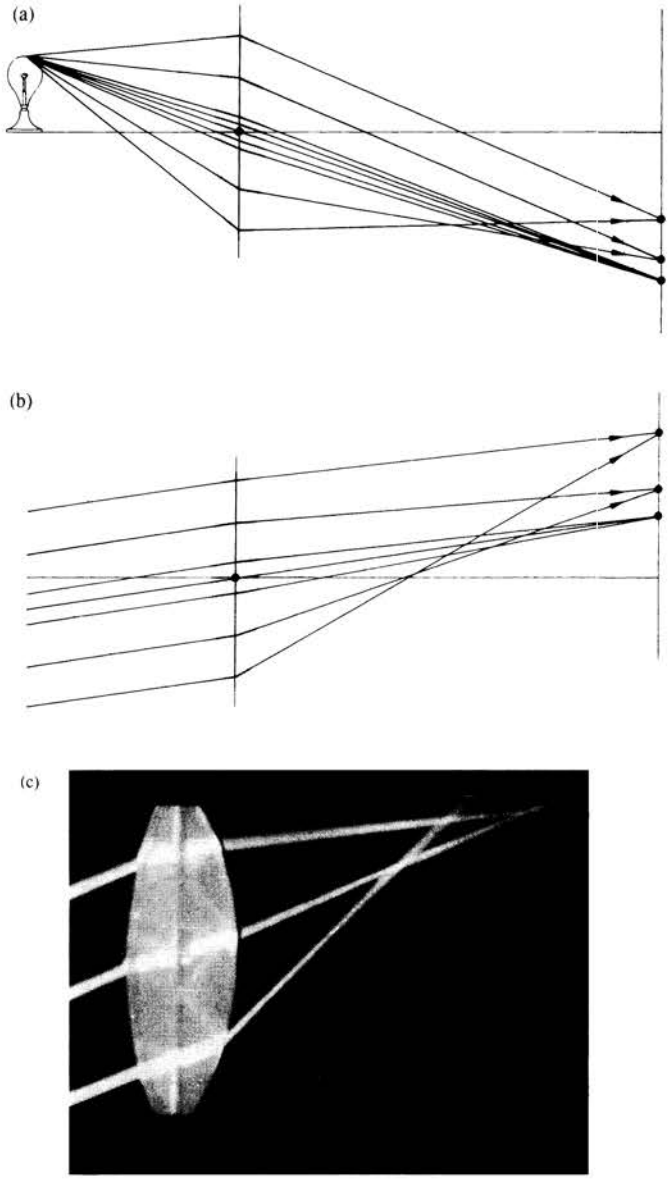


Figure 6.21 (*a*) Negative coma. (*b*) and (*c*) Positive coma. (Photo by E.H.)

The dependence of M_T on h, the ray height at the lens, is shown in Fig. 6.21*a*. Here meridional rays traversing the extremities of the lens arrive at the image plane closer to the axis than do the rays in the vicinity of the *principal ray* (i.e., the ray that passes through the principal points). In this instance, the least magnification is associated with the marginal rays that would form the smallest image—the coma is negative. By comparison, the coma in Figs. 6.21*b* and *c* is positive because the marginal rays focus farther from the axis.

Several skew rays are drawn from an extra-axial object point S in Fig. 6.22 to illustrate the formation of the geometrical comatic image of a point. Observe that each circular cone of rays whose endpoints (1-2-3-4-1-2-3-4) form a ring on the lens is imaged in what H. Dennis Taylor called a *comatic circle* on Σ_i. This case corresponds to positive coma, so the larger the ring on the lens, the more distant its comatic circle from the axis. When the outer ring is the intersection of marginal rays, the distance from 0 to 1 in the image is the ***tangential coma***, and the length from 0 to 3 on Σ_i is termed the ***sagittal coma***. A little more than half of the energy in the image appears in the roughly triangular region between 0 and 3. The coma flare, which owes its name to its cometlike tail, is often thought to be the worst of all aberrations, primarily because of its asymmetric configuration.

It's not the purview of Geometrical Optics to be concerned with interference, but when light reaches the screen in Fig. 6.22, it's certainly to be expected. The coma cone, just like the Gaussian image point, is an oversimplification. The image point is really an image disk-ring system, and the coma cone is actually a complicated asymmetrical diffraction pattern. The more coma there is, the more the cone departs from the Airy pattern into an elongated structure of blotches and arcs that only vaguely suggests the disk-ring structure from which it evolved (Fig. 6.23).

Like SA, coma is dependent on the shape of the lens. Thus a strongly concave positive-meniscus lens) with the object at infinity will have a large negative coma. Bending the lens so that it becomes planar-convex), then equiconvex), convex-planar (, and finally convex-meniscus (will change the coma from negative, to zero, to positive. The fact that it can be made exactly zero for a single lens with a given object distance is quite significant. The particular shape it then has ($s_o = \infty$) is almost convex-planar and nearly the configuration for minimum SA.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

It is important to realize that *a lens that is well corrected for the case in which one conjugate point is at infinity* ($s_o = \infty$) *may not perform satisfactorily when the object is nearby*. One would therefore do well, when using off-the-shelf lenses in a system operating at finite conjugates, to combine two infinite conjugate corrected lenses, as in Fig. 6.24. In other words, since it is unlikely that a lens with the desired focal length, which is also corrected for the particular set of finite conjugates, can be obtained ready-made, this back-to-back lens approach is an appealing alternative.

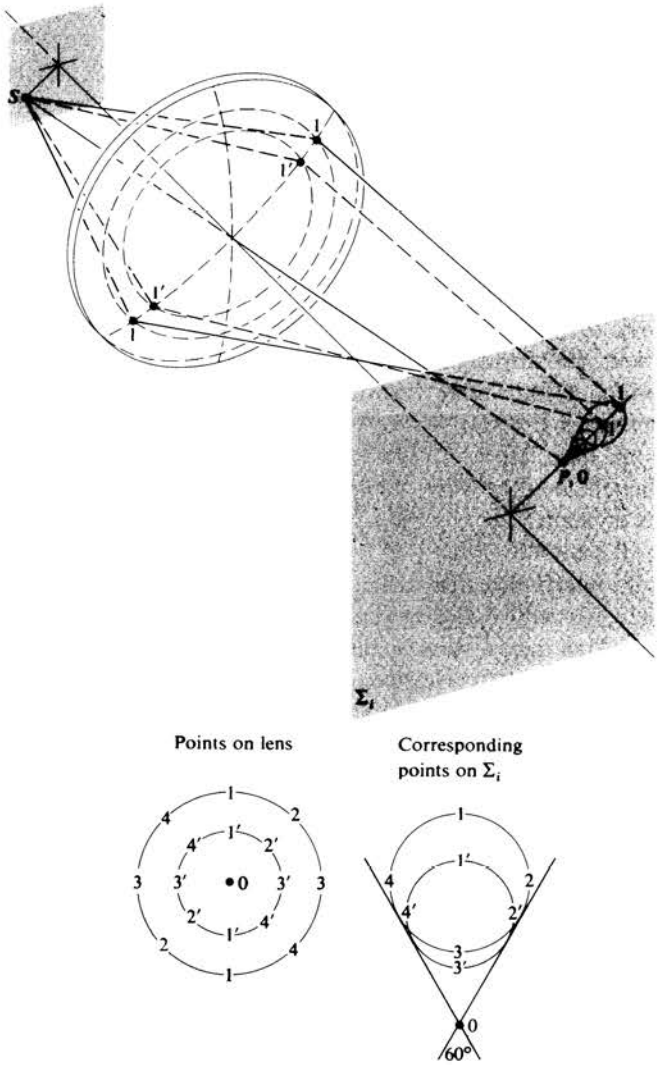


Figure 6.22 The geometrical coma image of a monochromatic point source. The central region of the lens forms a point image at the vertex of the cone.

Coma can also be negated by using a stop at the proper location, as William Hyde Wollaston (1766–1828) discovered in 1812. The order of the list of primary aberrations (SA, coma, astigmatism, Petzval field curvature, and distortion) is significant, because any one of them, except SA and Petzval curvature, will be affected by the position of a stop, but only if one of the preceding aberrations is also present in the system. Thus, while SA is independent of the location along the axis of a stop, coma will not be, as long as SA is present. This can be appreciated by examining the representation in Fig. 6.25. With the stop at Σ_1, ray 3 is the chief ray and there is SA but no coma; that is, the ray pairs meet on 3. If the stop is moved to Σ_2, the symmetry is upset, ray 4 becomes the chief ray, and the rays on either side of it, such as 3 and 5, meet above, not on it—there is positive coma. With the stop at Σ_3, rays 1 and 3 intersect below the chief ray, 2, and there is negative coma. In this way, controlled amounts of the aberration can be introduced into a compound lens in order to cancel coma in the system as a whole.

The **optical sine theorem** is an important relationship that must be introduced here even if space precludes its formal proof. It was discovered independently in 1873 by Abbe and Helmholtz, although a different form of it was given 10 years earlier by R. Clausius (of thermodynamics fame). In any event, it states that

$$n_o y_o \sin \alpha_o = n_i y_i \sin \alpha_i \tag{6.41}$$

where n_o, y_o, α_o and n_i, y_i, α_i are the index, height, and slope angle of a ray in object and image space, respectively, at any aperture size* (Fig. 6.9). If coma is to be zero,

$$M_T = \frac{y_i}{y_o} \tag{5.24}$$

must be constant for all rays. Suppose then that we send a marginal and a paraxial ray through the system. The former will comply with Eq. (6.41), the latter with its paraxial version (in

*To be precise, the sine theorem is valid for all values of α_o only in the sagittal plane (from the Latin *sagitta*, meaning arrow), which is discussed in the next section.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

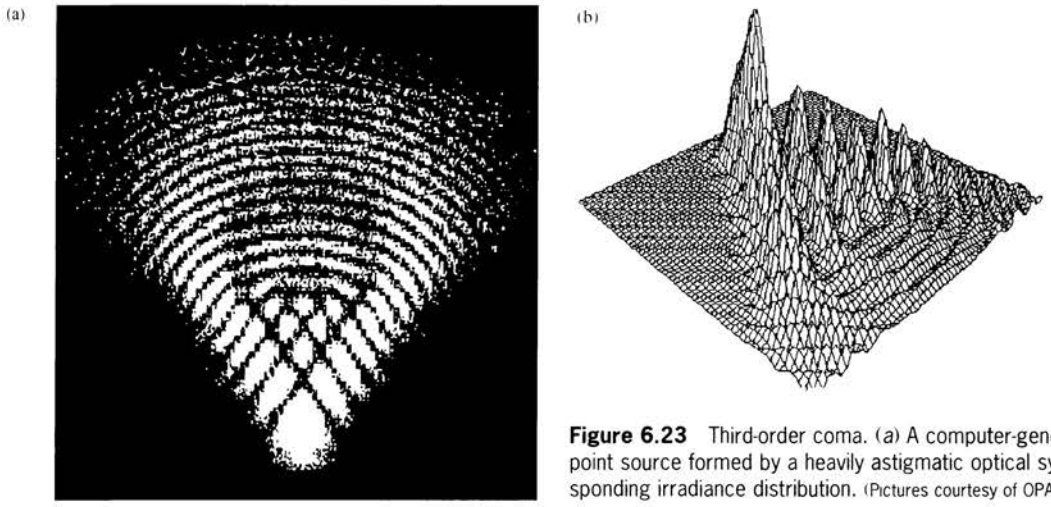


Figure 6.23 Third-order coma. (*a*) A computer-generated diagram of the image of a point source formed by a heavily astigmatic optical system. (*b*) A plot of the corresponding irradiance distribution. (Pictures courtesy of OPAL Group, St. Petersburg, Russia.)

which $\sin \alpha_o = \alpha_{op}$, $\sin \alpha_i = \alpha_{ip}$). Since M_T is to be constant over the entire lens, we equate the magnification for both marginal and paraxial rays to get

$$\frac{\sin \alpha_o}{\sin \alpha_i} = \frac{\alpha_{op}}{\alpha_{ip}} = \text{constant} \tag{6.42}$$

which is known as the **Sine Condition**. A necessary criterion for the absence of coma is that the system meet the Sine Condition. If there is no SA, compliancy with the Sine Condition will be both necessary and sufficient for zero coma.

It's an easy matter to observe coma. In fact, anyone who has focused sunlight with a simple positive lens has no doubt seen the effects of this aberration. A slight tilt of the lens, so that the nearly collimated rays from the Sun make an angle with the optical axis, will cause the focused spot to flare out into the characteristic comet shape.

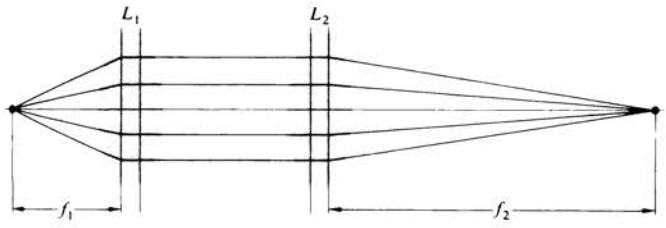


Figure 6.24 A combination of two infinite conjugate lenses yielding a system operating at finite conjugates.

Astigmatism

When an object point lies an appreciable distance from the optical axis, the incident cone of rays will strike the lens asymmetrically, giving rise to a third primary aberration known as **astigmatism**. The word derives from the Greek *a-*, meaning not, and *stigma*, meaning spot or point. To facilitate its description, envision the meridional plane (also called the *tangential plane*) containing both the chief ray (i.e., the one passing through the center of the aperture) and the optical axis. The *sagittal plane* is then defined as the plane containing the chief ray, which, in addition, is perpendicular to the meridion-

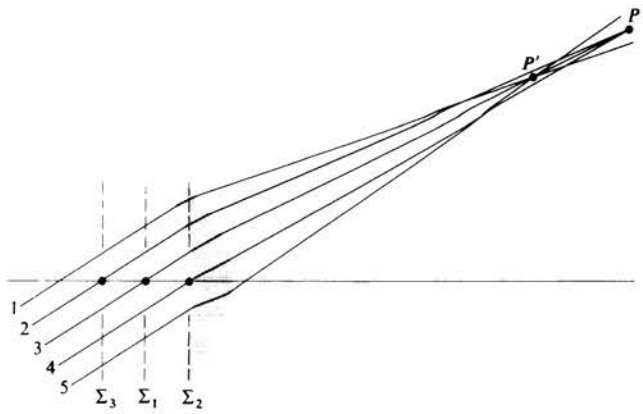


Figure 6.25 The effect of stop location on coma.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

al plane (Fig. 6.26). Unlike the latter, which is unbroken from one end of a complicated lens system to the other, the sagittal plane generally changes slope as the chief ray is deviated at the various elements. Hence to be accurate we should say that there are actually several sagittal planes, one attendant with each region within the system. Nevertheless, all skew rays from the object point lying in a sagittal plane are termed ***sagittal rays***.

In the case of an axial object point, the cone of rays is symmetrical with respect to the spherical surfaces of a lens. There is no need to make a distinction between meridional and sagittal planes. The ray configurations in all planes containing the optical axis are identical. In the absence of spherical aberration, all the focal lengths are the same, and consequently all rays arrive at a single focus. In contrast, the configuration of an oblique, parallel ray bundle will be different in the meridional and sagittal planes. As a result, the focal lengths in these planes will be different as well. In effect, here the meridional rays are tilted more with respect to the lens than are the sagittal rays, and they have a shorter focal length. It can be shown,* using Fermat's Principle, that the *focal length difference* depends effectively on the power of the lens (as opposed to the shape or index) and the angle at which the rays are inclined. This *astigmatic difference*, as it is often called, increases rapidly as the rays become more oblique, that is, as the object point moves farther off the axis, and is, of course, zero on axis.

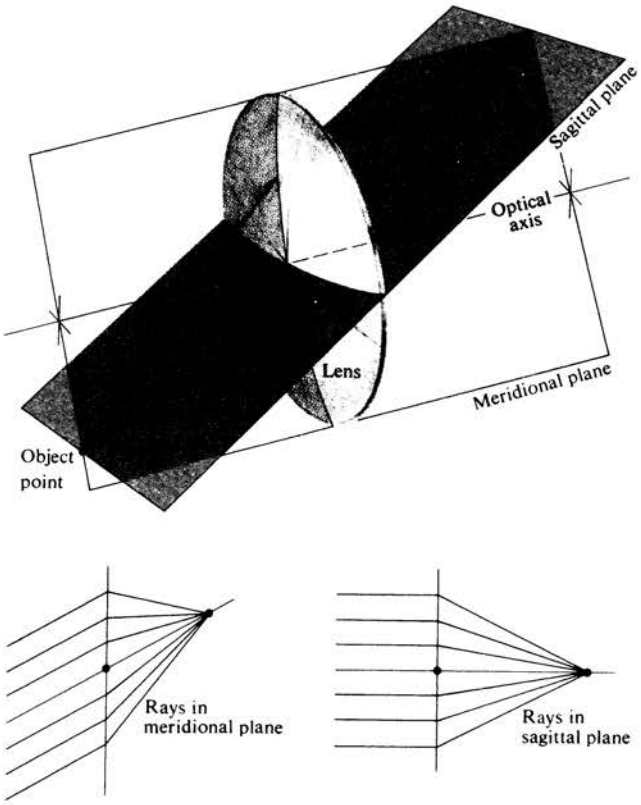


Figure 6.26 The sagittal and meridional planes.

Having two distinct focal lengths, the incident conical bundle of rays takes on a considerably altered form after refraction (Fig. 6.27). The cross section of the beam as it leaves the lens is initially circular, but it gradually becomes elliptical with the major axis in the sagittal plane, until at the *tangential* or *meridional focus* F_T, the ellipse degenerates into a "*line*" (at least in third-order theory). Actually, it's a complicated elongated diffraction pattern that looks more linelike the more astigmatism is present. All rays from the object point traverse this "line," which is known as the *primary image*. Beyond this point, the beam's cross section rapidly opens out until it is again circular. At that location, the image is a circular blur known as the *circle of least confusion*. Moving farther from the lens, the beam's cross section again deforms into a "line," called the *secondary image*. This time it's in the meridional plane at the *sagittal focus*, F_S.

The image of a point source formed by a slightly astigmatic optical system ($\lesssim 0.2\lambda$), in the vicinity of the circle of least confusion, looks very much like the Airy disk-ring pattern, but it's somewhat asymmetrical. As the amount of astigmatism increases (upwards of roughly 0.5λ), the biaxial asymmetry becomes more apparent. The image transforms into a complex distribution of bright and dark regions (resembling the Fresnel diffraction patterns for rectangular openings, p. 499) and only very subtly retains a curved structure arising from the circular aperture. Remember that in all of this we are assuming the absence of SA and coma.

Since the circle of least confusion increases in diameter as the astigmatic difference increases (i.e., as the object moves farther off-axis), the image will deteriorate, losing definition around its edges. Observe that the secondary "line" image will change in orientation with changes in the object position, but it will always point toward the optical axis; that is, it will be radial. Similarly, the primary "line" image will vary in orien-

*See A. W. Barton, *A Text Book on Light*, p. 124.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

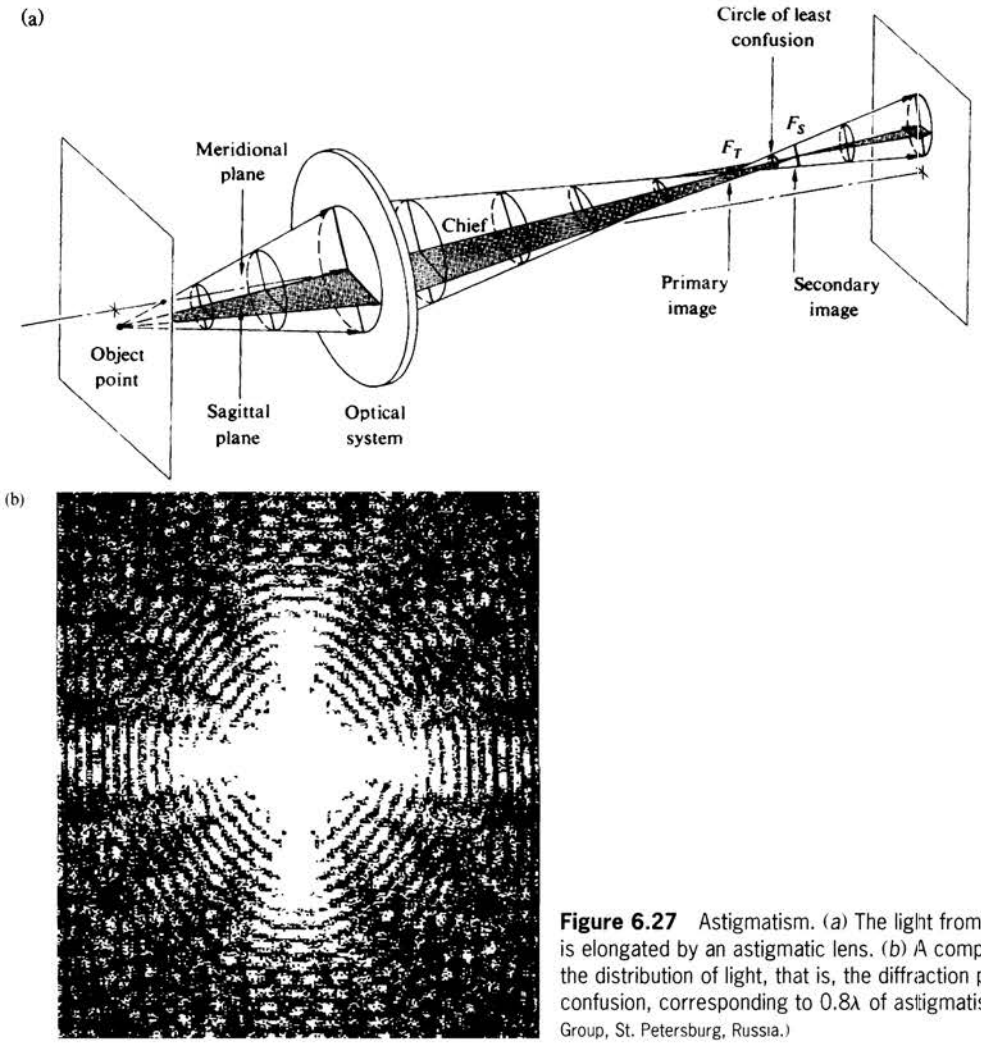


Figure 6.27 Astigmatism. (*a*) The light from a monochromatic point source is elongated by an astigmatic lens. (*b*) A computer-generated diagram showing the distribution of light, that is, the diffraction pattern, near the circle of least confusion, corresponding to 0.8λ of astigmatism. (Computer picture courtesy OPAL Group, St. Petersburg, Russia.)

tation, but it will remain normal to the secondary image. This arrangement causes the interesting effect shown in Fig. 6.28 when the object is made up of radial and tangential elements. The primary and secondary images are, in effect, formed of transverse and radial dashes, which increase in size with distance from the axis. In the latter case, the dashes point like arrows toward the center of the image—ergo, the name *sagitta*.

The existence of the sagittal and tangential foci can be verified directly with a fairly simple arrangement. Place a positive lens with a short focal length (about 10 or 20 mm) in the beam of a He–Ne laser. Position another positive test lens with a somewhat longer focal length far enough away so that the now diverging beam fills that lens. A convenient object, to be located between the two lenses, is a piece of ordinary wire screening (or a transparency). Align it so the wires are horizontal (x) and vertical (y). If the test lens is rotated roughly 45° about the vertical (with the x-, y-, and z-axes fixed in the lens), astigmatism should be observable. The meridional is the xz-plane (z being the lens axis, now at about 45° to the laser axis), and the sagittal plane corresponds to the plane of y and the laser axis. As the wire mesh is moved toward the test lens, a point will be reached where the horizontal wires are in focus on a screen beyond the lens, whereas the vertical wires are not. This is the

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

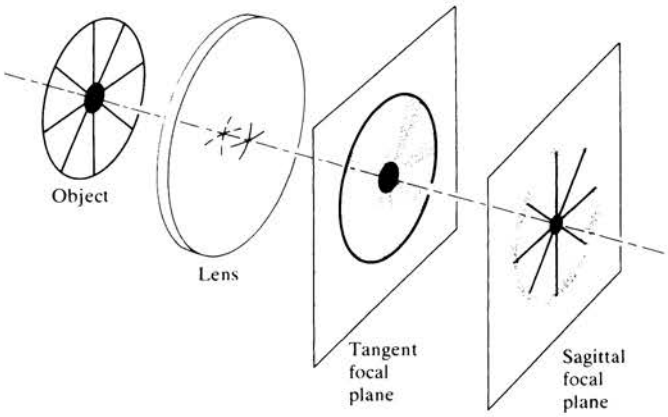


Figure 6.28 Images in the tangent and sagittal focal planes.

location of the sagittal focus. Each point on the object is imaged as a short line in the meridional (horizontal) plane, which accounts for the fact that only the horizontal wires are in focus. Moving the mesh slightly closer to the lens will bring the vertical lines into clarity while the horizontal ones are blurred. This is the tangential focus. Try rotating the mesh about the central laser axis while at either focus.

Note that unlike visual astigmatism (p. 209), which arose from an actual asymmetry in the surfaces of the optical system, the third-order aberration by that same name applies to spherically symmetrical lenses.

Mirrors, with the singular exception of the plane mirror, suffer much of the same monochromatic aberrations as do lenses. Thus, although a paraboloidal mirror is free of SA for an infinitely distant axial object point, its off-axis imagery is quite poor due to astigmatism and coma. This strongly restricts its use to narrow field devices, such as searchlights and astronomical telescopes. A concave spherical mirror shows SA, coma, and astigmatism. Indeed, one could draw a diagram just like Fig. 6.27 with the lens replaced by an obliquely illuminated spherical mirror. Incidentally, such a mirror displays appreciably less SA than would a simple convex lens of the same focal length.

Field Curvature

Suppose we had an optical system that was free of all the aberrations thus far considered. There would then be a one-to-one correspondence between points on the object and image surfaces (i.e., stigmatic imagery). We mentioned earlier (Section 5.2.3) that a planar object normal to the axis will be imaged approximately as a plane only in the paraxial region. At finite apertures the resulting curved stigmatic image surface is a manifestation of the primary aberration known as **Petzval field curvature**, after the Hungarian mathematician Josef Max Petzval (1807–1891). The effect can readily be appreciated by examining Figs. 5.19 (p. 161) and 6.29. A spherical object segment σ_o is imaged by the lens as a spherical segment σ_i, both centered at O. Flattening out σ_o into the plane σ_o' will cause each image point to move toward the lens along the concomitant chief ray, thus forming a paraboloidal *Petzval surface* Σ_P. Whereas the Petzval surface for a positive lens curves *inward* toward the object plane, for a negative lens it curves *outward* away from that plane. Evidently, a suitable combination of positive and negative lenses will negate field curvature. Indeed, the displacement Δx of an image point at height y_i on

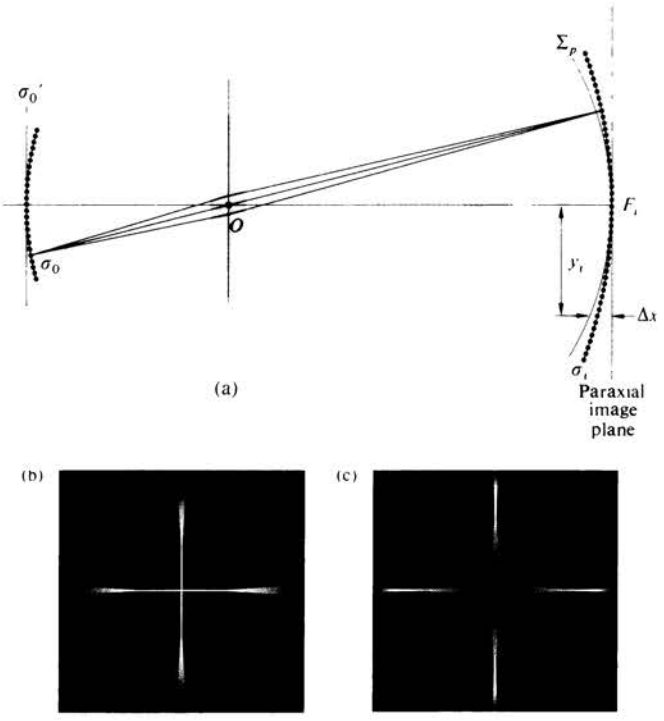


Figure 6.29 Field curvature. (*a*) When the object corresponds to σ_o', the image will correspond to surface Σ_P. (*b*) The image formed on a flat screen near the paraxial image plane will be in focus only at its center. (*c*) Moving the screen closer to the lens will bring the edges into focus.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

the Petzval surface from the paraxial image plane is given by

$$\Delta x = \frac{y_i^2}{2} \sum_{j=1}^{m} \frac{1}{n_j f_j} \tag{6.43}$$

where n_j and f_j are the indices and focal lengths of the m thin lenses forming the system. This implies that the Petzval surface will be unaltered by changes in the positions or shapes of the lenses or in the location of the stop, as long as the values of n_j and f_j are fixed. Notice that for the simple case of two thin lenses ($m = 2$) having any spacing, Δx *can be made zero* provided that

$$\frac{1}{n_1 f_1} + \frac{1}{n_2 f_2} = 0$$

or, equivalently,

$$n_1 f_1 + n_2 f_2 = 0 \tag{6.44}$$

This is the so-called **Petzval condition**. As an example of its use, suppose we combine two thin lenses, one positive, the other negative, such that $f_1 = -f_2$ and $n_1 = n_2$. Since

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

$$f = \frac{f_1^2}{d}$$



(a) Petzval lens with field flattener



(b) 16 mm projection lens

Figure 6.30 The field flattener.

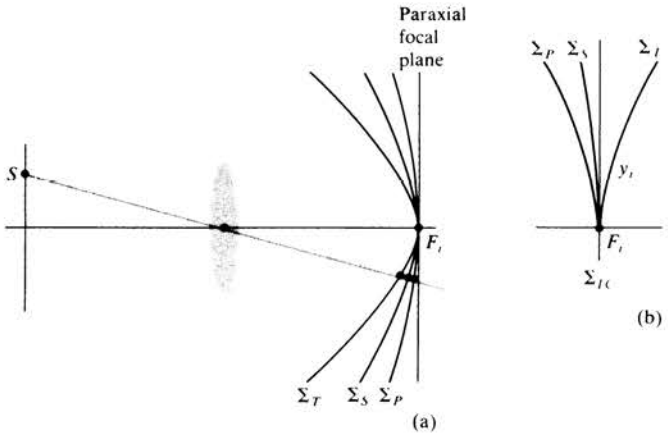


Figure 6.31 The tangential, sagittal, and Petzval image surfaces.

the system can satisfy the Petzval condition, have a flat field, and still have a finite positive focal length.

In visual instruments a certain amount of curvature can be tolerated, because the eye can accommodate for it. Clearly, in photographic lenses field curvature is most undesirable, since it has the effect of rapidly blurring the off-axis image when the film plane is at F_i. An effective means of nullifying the inward curvature of a positive lens is to place a negative *field flattener* lens near the focal plane. This is often done in projection and photographic objectives when it is not otherwise practicable to meet the Petzval condition (Fig. 6.30). In this position the flattener will have little effect on other aberrations.

Astigmatism is intimately related to field curvature. In the presence of the former aberration, there will be *two* paraboloidal image surfaces, the tangential, Σ_T, and the sagittal, Σ_S (as in Fig. 6.31). These are the loci of all the primary and secondary images, respectively, as the object point roams over the object plane. At a given height (y_i), a point on Σ_T always lies three times as far from Σ_P as does the corresponding point on Σ_S, and both are on the same side of the Petzval surface (Fig. 6.31). When there is no astigmatism, Σ_S and Σ_T coalesce on Σ_P. It is possible to alter the shapes of Σ_S and Σ_T by bending or relocating the lenses or by moving the stop. The configuration of Fig. 6.31*b* is known as an *artificially flattened* field. A stop in front of an inexpensive meniscus box camera lens is usually arranged to produce just this effect. The surface of least confusion, Σ_{LC}, is planar, and the image there is tolerable, losing definition at the margins because of the astigmatism. That is to say, although their loci form Σ_{LC}, the circles of

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

least confusion increase in diameter with distance off the axis. Modern good-quality photographic objectives are generally ***anastigmats***; that is, they are designed so that Σ_S and Σ_T cross each other, yielding an additional off-axis angle of zero astigmatism. The Cooke Triplet, Tessar, Orthometer, and Biotar (Fig. 5.104) are all anastigmats, as is the relatively fast Zeiss Sonnar, whose residual astigmatism is illustrated graphically in Fig. 6.32. Note the relatively flat field and small amount of astigmatism over most of the film plane.

Let's return briefly to the Schmidt camera shown in Fig. 5.113 (p. 226), since we are now in a better position to appreciate how it functions. With a stop at the center of curvature of the spherical mirror, all chief rays, which by definition pass through C, are incident normally on the mirror. Moreover, each pencil of rays from a distant object point is symmetrical about its chief ray. In effect, each chief ray serves as an optical axis, so there are no off-axis points and, in principle, no coma or astigmatism. Instead of attempting to flatten the image surface, the designer has coped with curvature by simply shaping the film plate to conform with it.

Distortion

The last of the five primary, monochromatic aberrations is **distortion**. Its origin lies in the fact that the transverse magnification, M_T, may be a function of the off-axis image distance, y_i. Thus, that distance may differ from the one predicted by paraxial theory in which M_T is constant. In other words, distortion arises because different areas of the lens have different

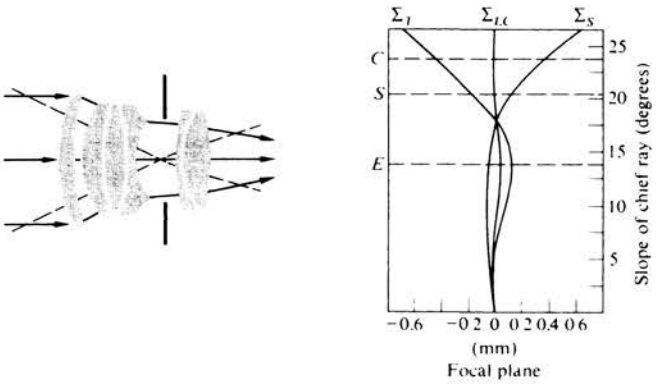


Figure 6.32 A typical Sonnar. The markings C, S, and E denote the limits of the 35-mm film format (field stop), that is, corners, sides, and edges. The Sonnar family lies between the double Gauss and the triplet.

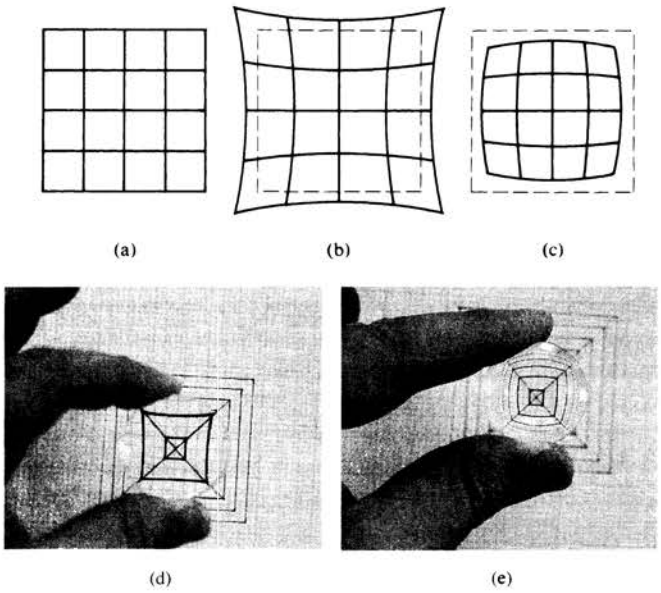


Figure 6.33 (*a*) Undistorted object. (*b*) When the magnification on the optical axis is less than the off-axis magnification, pincushion distortion results. (*c*) When it is greater on axis than off, barrel distortion results. (*d*) Pincushion distortion in a single thin lens. (*e*) Barrel distortion in a single thin lens. (Photos by E.H.)

focal lengths and different magnifications. In the absence of any of the other aberrations, distortion is manifest in a misshaping of the image as a whole, even though each point is sharply focused. Consequently, when processed by an optical system suffering ***positive*** or ***pincushion distortion***, a square array deforms, as in Fig. 6.33*b*. In that instance, each image point is displaced radially outward from the center, with the most distant points moving the greatest amount (i.e., M_T increases with y_i). Similarly, ***negative*** or ***barrel distortion*** corresponds to the situation in which M_T decreases with the axial distance, and in effect, each point on the image moves radially inward toward the center (Fig. 6.33*c*).

Distortion can easily be seen by just looking through an aberrant lens at a piece of lined or graph paper. Fairly thin lenses will show essentially no distortion, whereas ordinary positive or negative, thick, simple lenses will generally suffer positive or negative distortion, respectively. The introduction of a stop into a system of thin lenses is invariably accompanied by distortion, as indicated in Fig. 6.34. One exception is the case in which the aperture stop is at the lens, so that the chief ray is, in effect, the principal ray (i.e., it passes through

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

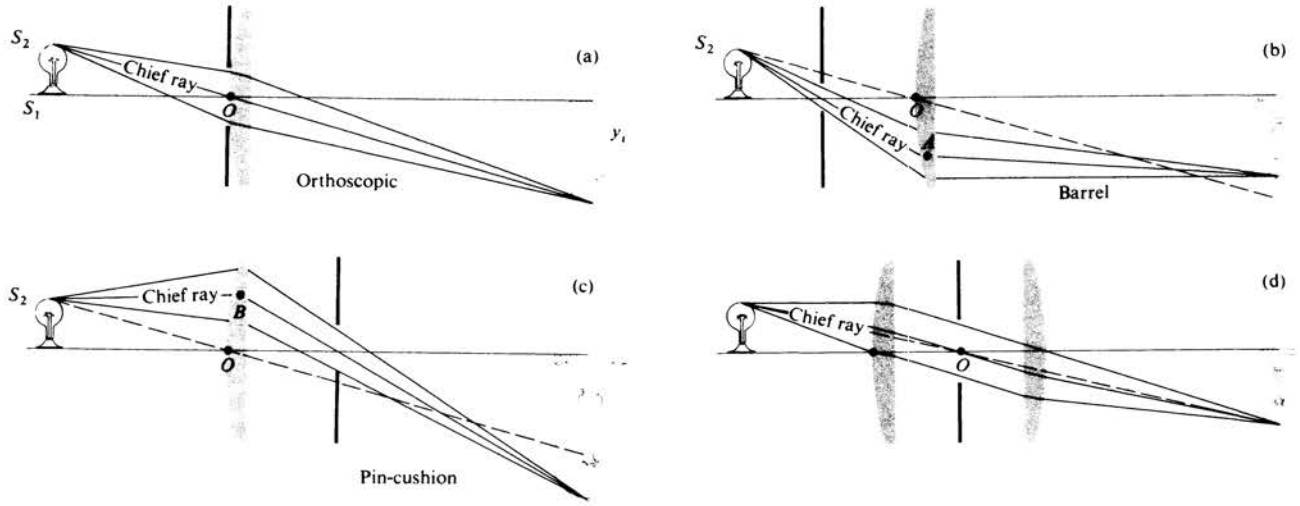


Figure 6.34 The effect of stop location on distortion.

the principal points, here coalesced at O). If the stop is in front of a positive lens, as in Fig. 6.34*b*, the object distance measured along the chief ray will be greater than it was with the stop at the lens ($S_2A > S_2O$). Thus x_o will be greater and [Eq. (5.26)] M_T will be smaller—ergo, barrel distortion. In other words, M_T for an off-axis point will be less with a front stop in position than it would be without it. The difference is a measure of the aberration, which, by the way, exists regardless of the size of the aperture. In the same way, a rear stop (Fig. 6.34*c*) decreases x_o along the chief ray (i.e., $S_2O > S_2B$), thereby increasing M_T and introducing pincushion distortion. *Interchanging the object and image thus has the effect of changing the sign of the distortion* for a given lens and stop. The aforementioned stop positions will produce the opposite effect when the lens is negative.

All of this suggests the use of a stop midway between identical lens elements. The distortion from the first lens will precisely cancel the contribution from the second. This approach has been used to advantage in the design of a number of photographic lenses (Fig. 5.104). To be sure, if the lens is perfectly symmetrical and operating as in Fig. 6.34*d*, the object and image distances will be equal, hence $M_T = 1$. (Incidentally, coma and lateral color will then be identically zero as well.) This applies to (finite conjugate) copy lenses used, for example, to record data. Nonetheless, even when M_T is not 1, making the system approximately symmetrical about a stop is a very common practice, since it markedly reduces these several aberrations.

Distortion can arise in compound lens systems, as for example in the telephoto arrangement shown in Fig. 6.35. For a distant object point, the margin of the positive achromat serves as the aperture stop. In effect, the arrangement is like a negative lens with a front stop, so it displays positive or pincushion distortion.

Suppose a chief ray enters and emerges from an optical system in the same direction as, for example, in Fig. 6.34*d*. The point at which the ray crosses the axis is the optical center of the system, but since this is a chief ray, it is also the center of the aperture stop. This is the situation approached in Fig. 6.34*a*, with the stop up against the thin lens. In both instances

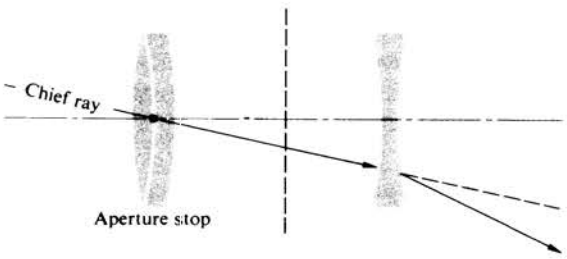


Figure 6.35 Distortion in a compound lens.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

the incoming and outgoing segments of the chief ray are parallel, and there is zero distortion; that is, the system is *orthoscopic*. This also implies that the entrance and exit pupils will correspond to the principal planes (if the system is immersed in a single medium—see Fig. 6.2). Bear in mind that the chief ray is now a principal ray. *A thin-lens system will have zero distortion if its optical center is coincident with the center of the aperture stop.* By the way, in a pinhole camera, the rays connecting conjugate object and image points are straight and pass through the center of the aperture stop. The entering and emerging rays are obviously parallel (being one and the same), and there is no distortion.

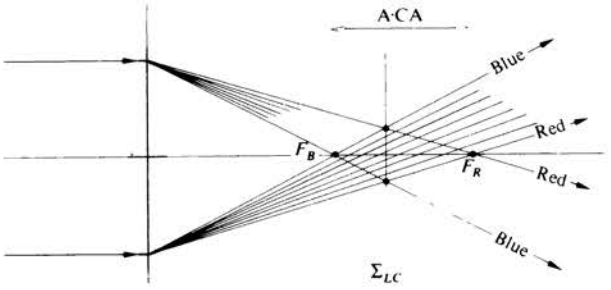


Figure 6.36 Axial chromatic aberration.

6.3.2 Chromatic Aberrations

The five primary or Seidel aberrations have been considered in terms of monochromatic light. To be sure, if the source has a broad spectral bandwidth, these aberrations are influenced accordingly; but the effects are inconsequential, unless the system is quite well corrected. There are, however, **chromatic aberrations** that arise specifically in polychromatic light, which are far more significant. The ray-tracing equation [Eq. (6.12)] is a function of the indices of refraction, which in turn vary with wavelength. Different "colored" rays will traverse a system along different paths, and this is the quintessential feature of chromatic aberration.

Since the thin-lens equation

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

is wavelength-dependent via $n_l(\lambda)$, the focal length must also vary with λ. In general (Fig. 3.40, p. 72), $n_l(\lambda)$ decreases with wavelength over the visible region, and thus $f(\lambda)$ increases with λ. The result is illustrated in Fig. 6.36, where the constituent colors in a collimated beam of white light are focused at different points on the axis. The axial distance between two such focal points spanning a given frequency range (e.g., blue to red) is termed the ***axial*** (or *longitudinal*) ***chromatic aberration***, A · CA for short.

It's an easy matter to observe chromatic aberrations, or CA, with a thick, simple converging lens. When illuminated by a polychromatic point source (a candle flame will do), the lens will cast a real image surrounded by a halo. If the plane of observation is then moved nearer the lens, the periphery of the blurred image will become tinged in orange-red. Moving it back away from the lens, beyond the best image, will cause the outlines to become tinted in blue-violet. The location of the circle of least confusion (i.e., the plane Σ_{LC}) corresponds to the position where the best image will appear. Try looking directly through the lens at a source—the coloration will be far more striking.

The image of an off-axis point will be formed of the constituent frequency components, each arriving at a different height above the axis (Fig. 6.37). In essence, the frequency dependence of f causes a frequency dependence of the transverse magnification as well. The vertical distance between two such image points (most often taken to be blue and red) is a measure of the ***lateral chromatic aberration***, L · CA, or ***lateral color***. Consequently, a chromatically aberrant lens illuminated by white light will fill a volume of space with a continuum of more or less overlapping images, varying in size and color. Because the eye is most sensitive to the yellow-green portion of the spectrum, the tendency is to focus the lens for that region. With such a configuration one would see all the other colored images superimposed and slightly out of focus, producing a whitish blur or hazed overlay.

When the blue focus, F_B, is to the left of the red focus, F_R, the A · CA is said to be positive, as it is in Fig. 6.36. Conversely, a negative lens would generate negative A · CA, with the

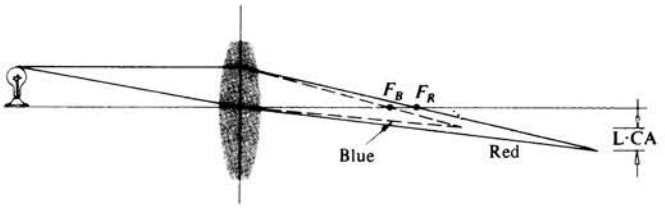


Figure 6.37 Lateral chromatic aberration.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

more strongly deviated blue rays appearing to originate at the right of the red focus. Physically, what is happening is that the lens, whether convex or concave, is prismatic in shape; that is, it becomes either thinner or thicker as the radial distance from the axis increases. As you well know, rays are therefore deviated either toward or away from the axis, respectively. In both cases the rays are bent toward the thicker "base" of the prismatic cross section. But the angular deviation is an increasing function of n, and therefore it decreases with λ. Hence blue light is deviated the most and is focused nearest the lens. In other words, for a convex lens the red focus is farthest and to the right; for a concave lens it is farthest and to the left.

The human eye has a substantial amount of chromatic aberration which is compensated for by several psychophysical mechanisms. Still, it's possible to see the effect with a small purple dot: held close to the eye, it will appear blue at the center surrounded by red; farther away it will appear red surrounded by blue.

Thin Achromatic Doublets

All of this suggests that a combination of two thin lenses, one positive and one negative, could conceivably result in the precise overlapping of F_R and F_B (Fig. 6.38). Such an arrangement is said to be *achromatized* for those two specific wavelengths. Notice that what we would like to do is effectively eliminate the total dispersion (i.e., the fact that each color is deviated by a different amount) and not the total deviation itself. With the two lenses separated by a distance d,

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2} \qquad [6.8]$$

Rather than writing out the second term in the Thin-Lens Equation [Eq. (5.16), p. 158] let's abbreviate the notation and use $1/f_1 = (n_1 - 1)\rho_1$ and $1/f_2 = (n_2 - 1)\rho_2$ for the two elements. Then

$$\frac{1}{f} = (n_1 - 1)\rho_1 + (n_2 - 1)\rho_2 - d(n_1 - 1)\rho_1(n_2 - 1)\rho_2 \qquad (6.45)$$

This expression will yield the focal length of the doublet for red (f_R) and blue (f_B) light when the appropriate indices are introduced, namely, n_{1R}, n_{2R}, n_{1B}, and n_{2B}. But if f_R is to equal f_B, then

$$\frac{1}{f_R} = \frac{1}{f_B}$$

and, using Eq. (6.45),

$$(n_{1R} - 1)\rho_1 + (n_{2R} - 1)\rho_2 - d(n_{1R} - 1)\rho_1(n_{2R} - 1)\rho_2 = (n_{1B} - 1)\rho_1 + (n_{2B} - 1)\rho_2 - d(n_{1B} - 1)\rho_1(n_{2B} - 1)\rho_2 \qquad (6.46)$$

One case of particular importance corresponds to $d = 0$; that is, the two lenses are in contact. Expanding out Eq. (6.46) with $d = 0$ then leads to

$$\frac{\rho_1}{\rho_2} = -\frac{n_{2B} - n_{2R}}{n_{1B} - n_{1R}} \qquad (6.47)$$

The focal length of the compound lens (f_Y) can conveniently be specified as that associated with yellow light, roughly midway between the blue and red extremes. For the component lenses in yellow light, $1/f_{1Y} = (n_{1Y} - 1)\rho_1$ and $1/f_{2Y} = (n_{2Y} - 1)\rho_2$. Hence

$$\frac{\rho_1}{\rho_2} = \frac{(n_{2Y} - 1)}{(n_{1Y} - 1)}\frac{f_{2Y}}{f_{1Y}} \qquad (6.48)$$

Equating Eqs. (6.47) and (6.48) leads to

$$\frac{f_{2Y}}{f_{1Y}} = -\frac{(n_{2B} - n_{2R})/(n_{2Y} - 1)}{(n_{1B} - n_{1R})/(n_{1Y} - 1)} \qquad (6.49)$$

The quantities

$$\frac{n_{2B} - n_{2R}}{n_{2Y} - 1} \quad \text{and} \quad \frac{n_{1B} - n_{1R}}{n_{1Y} - 1}$$

are known as the **dispersive powers** of the two materials forming the lenses. Their reciprocals, V_2 and V_1, are variously known as the *dispersive indices*, *V-numbers*, or **Abbe numbers**. The lower the Abbe numbers, the greater the dispersive

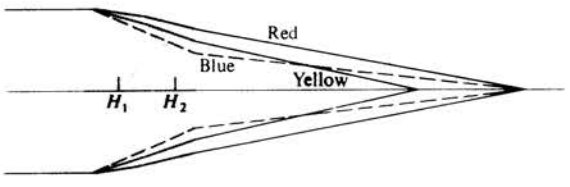


Figure 6.38 An achromatic doublet. The paths of the rays are much exaggerated.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

power. Thus

$$\frac{f_{2Y}}{f_{1Y}} = -\frac{V_1}{V_2}$$

or

$$f_{1Y}V_1 + f_{2Y}V_2 = 0 \tag{6.50}$$

Since the dispersive powers are positive, so too are the V-numbers. This implies, as we anticipated, that one of the two component lenses must be negative, and the other positive, if Eq. (6.50) is to obtain; that is, if f_R is to equal f_B.

At this point we could presumably design an *achromatic doublet*, and indeed we presently shall, but a few additional points must be made first. The designation of wavelengths as red, yellow, and blue is far too imprecise for practical application. Instead it is customary to refer to specific spectral lines whose wavelengths are known with great precision. The **Fraunhofer lines**, as they are called, serve as the needed reference markers across the spectrum. Several of these for the visible region are listed in Table 6.1. The lines F, C, and d (i.e., D_3) are most often used (for blue, red, and yellow), and one generally traces paraxial rays in d-light. Glass manufacturers will usually list their wares in terms of the Abbe number, as in Fig. 6.39, which is a plot of the refractive index versus

$$V_d = \frac{n_d - 1}{n_F - n_C} \tag{6.51}$$

TABLE 6.1 Several Strong Fraunhofer Lines

Designation	Wavelength (Å)*	Source
C	6562.816 Red	H
D_1	5895.923 Yellow	Na
D	Center of doublet 5892.9	Na
D_2	5889.953 Yellow	Na
D_3 or d	5875.618 Yellow	He
b_1	5183.618 Green	Mg
b_2	5172.699 Green	Mg
c	4957.609 Green	Fe
F	4861.327 Blue	H
f	4340.465 Violet	H
g	4226.728 Violet	Ca
K	3933.666 Violet	Ca

*1 Å = 0.1 nm.

TABLE 6.2 Optical Glass

Type Number	Name	n_D	V_D
511:635	Borosilicate crown—BSC-1	1.5110	63.5
517:645	Borosilicate crown—BSC-2	1.5170	64.5
513:605	Crown—C	1.5125	60.5
518:596	Crown	1.5180	59.6
523:586	Crown—C-1	1.5230	58.6
529:516	Crown flint—CF-1	1.5286	51.6
541:599	Light barium crown—LBC-1	1.5411	59.9
573:574	Barium crown—LBC-2	1.5725	57.4
574:577	Barium crown	1.5744	57.7
611:588	Dense barium crown—DBC-1	1.6110	58.8
617:550	Dense barium crown—DBC-2	1.6170	55.0
611:572	Dense barium crown—DBC-3	1.6109	57.2
562:510	Light barium flint—LBF-2	1.5616	51.0
588:534	Light barium flint—LBF-1	1.5880	53.4
584:460	Barium flint—BF-1	1.5838	46.0
605:436	Barium flint—BF-2	1.6053	43.6
559:452	Extra light flint—ELF-1	1.5585	45.2
573:425	Light flint—LF-1	1.5725	42.5
580:410	Light flint—LF-2	1.5795	41.0
605:380	Dense flint—DF-1	1.6050	38.0
617:366	Dense flint—DF-2	1.6170	36.6
621:362	Dense flint—DF-3	1.6210	36.2
649:338	Extra dense flint—EDF-1	1.6490	33.8
666:324	Extra dense flint—EDF-5	1.6660	32.4
673:322	Extra dense flint—EDF-2	1.6725	32.2
689:309	Extra dense flint—EDF	1.6890	30.9
720:293	Extra dense flint—EDF-3	1.7200	29.3

From T. Calvert, "Optical Components," *Electromechanical Design*, May 1971. Type number is given by $(n_D - 1)$:$(10\ V_D)$, where n_D is rounded off to three decimal places. For more data, see Smith, *Modern Optical Engineering*, Fig. 7.5.

(Take a look at Table 6.2 as well.) Thus Eq. (6.50) might better be written as

$$f_{1d}V_{1d} + f_{2d}V_{2d} = 0 \tag{6.52}$$

where the numerical subscripts pertain to the two glasses used in the doublet, and the letter relates to the d-line.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

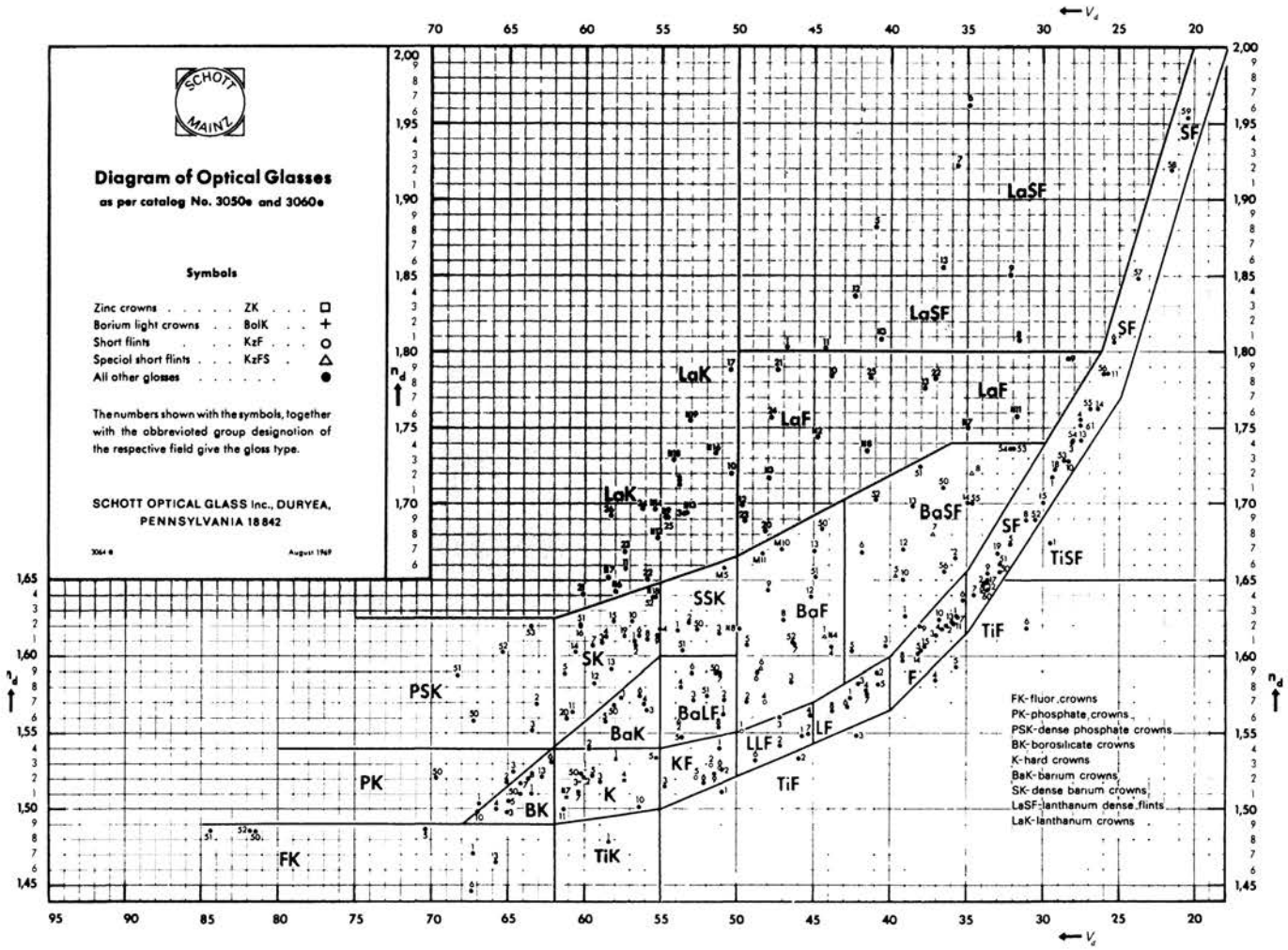


Figure 6.39 Refractive index versus Abbe number for various glasses. The specimens in the upper shaded area are the rare-earth glasses, which have high indices of refraction and low dispersions.

Incidentally, Newton erroneously concluded, on the basis of experiments with the very limited range of materials available at the time, that the dispersive power was constant for all glasses. This is tantamount to saying [Eq. (6.52)] that $f_{1d} = -f_{2d}$, in which case the doublet would have zero power. Newton, accordingly, shifted his efforts from the refracting to the reflecting telescope, and this fortunately turned out to be a good move in the long run. The achromat was invented around 1733 by Chester Moor Hall, Esq., but it lay in limbo until it was seemingly reinvented and patented in 1758 by the London optician John Dollond.

Several forms of the achromatic doublet are shown in Fig. 6.40. Their configurations depend on the glass types selected, as well as on the choice of the other aberrations to be controlled. By the way, when purchasing off-the-shelf doublets of unknown origin, be careful not to buy a lens that has been deliberately designed to include certain aberrations in order to compensate for errors in the original system from

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

which it came. Perhaps the most commonly encountered doublet is the cemented Fraunhofer achromat. It's formed of a crown* double-convex lens in contact with a concave-planar (or nearly planar) flint lens. The use of a crown front element is quite popular because of its resistance to wear. Since the overall shape is roughly convex-planar, by selecting the proper glasses, both spherical aberration and coma can be corrected as well. Suppose that we wish to design a Fraunhofer achromat of focal length 50 cm. We can get some idea of how to select glasses by solving Eq. (6.52) simultaneously with the compound-lens equation

$$\frac{1}{f_{1d}} + \frac{1}{f_{2d}} = \frac{1}{f_d}$$

to get

$$\frac{1}{f_{1d}} = \frac{V_{1d}}{f_d(V_{1d} - V_{2d})} \tag{6.53}$$

and

$$\frac{1}{f_{2d}} = \frac{V_{2d}}{f_d(V_{2d} - V_{1d})} \tag{6.54}$$

Thus, in order to avoid small values of f_{1d} and f_{2d}, which would necessitate strongly curved surfaces on the component lenses, the difference $V_{1d} - V_{2d}$ should be made large (roughly 20 or more is convenient). From Fig. 6.39 (or its equivalent) we select, say, BK1 and F2. These have catalogued indices of $n_C = 1.507\,63$, $n_d = 1.510\,09$, $n_F = 1.515\,66$ and $n_C = 1.615\,03$, $n_d = 1.620\,04$, $n_F = 1.632\,08$, respectively. Likewise, their V-numbers are generally given rather accurately, and we needn't compute them. In this instance they are $V_{1d} = 63.46$ and $V_{2d} = 36.37$, respectively. The focal lengths, or if you will, the powers of the two lenses, are given by Eqs. (6.53) and (6.54):

$$\mathcal{D}_{1d} = \frac{1}{f_{1d}} = \frac{63.46}{0.50(27.09)}$$

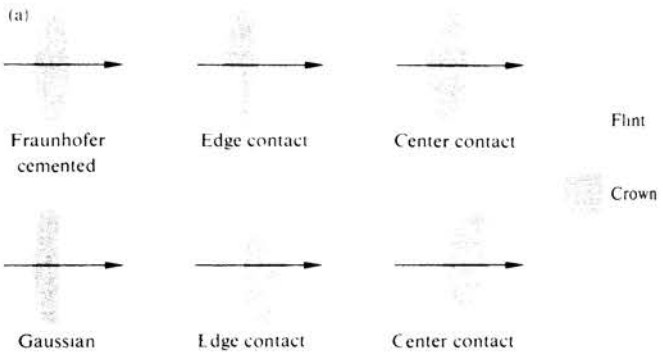
and

$$\mathcal{D}_{2d} = \frac{1}{f_{2d}} = \frac{36.37}{0.50(-27.09)}$$

Hence $\mathcal{D}_{1d} = 4.685$ D and $\mathcal{D}_{2d} = -2.685$ D, the sum being 2 D, which is $1/0.5$, as it should be. For ease of fabrication let the first or positive lens be equiconvex. Consequently, its radii R_{11} and R_{12} are equal in magnitude. Hence

$$\rho_1 = \frac{1}{R_{11}} - \frac{1}{R_{12}} = \frac{2}{R_{11}}$$

*Traditionally, the glasses roughly in the range $n_d > 1.60$, $V_d > 50$, as well as $n_d < 1.60$, $V_d > 55$ are known as *crowns*, and the others are *flints*. Note the letter designations in Fig. 6.39.



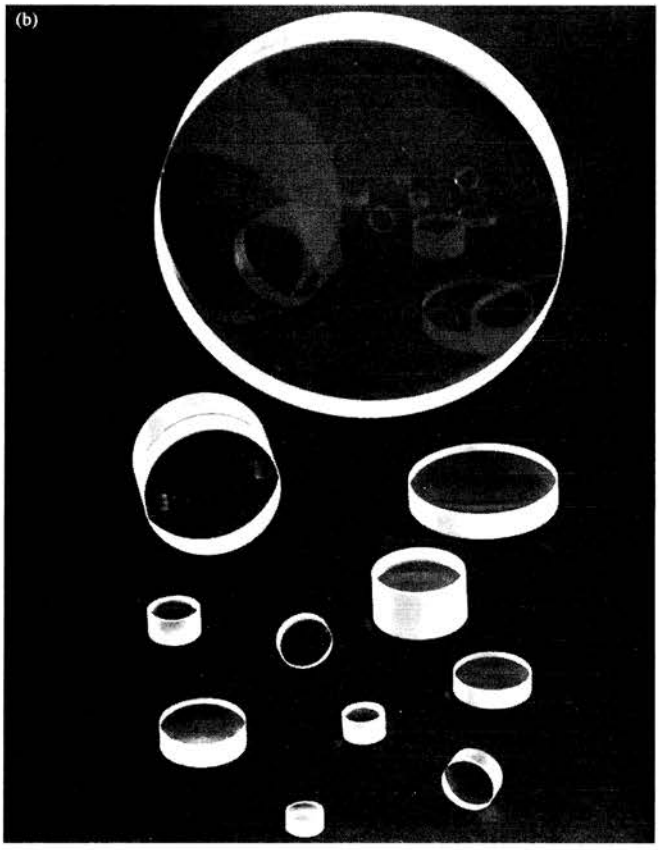


Figure 6.40 (*a*) Achromatic doublets. (*b*) Doublets and triplets. (Photo courtesy Melles Griot.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

or, equivalently,

$$\frac{2}{R_{11}} = \frac{\mathscr{D}_{1d}}{n_{1d} - 1} = \frac{4.685}{0.51009} = 9.185$$

Thus $R_{11} = -R_{12} = 0.2177$ m. Furthermore, having specified that the lenses be in intimate contact, we have $R_{12} = R_{21}$; that is, the second surface of the first lens matches the first surface of the second lens. For the second lens

$$\rho_2 = \frac{1}{R_{21}} - \frac{1}{R_{22}} = \frac{\mathscr{D}_{2d}}{n_{2d} - 1}$$

or

$$\frac{1}{-0.2177} - \frac{1}{R_{22}} = \frac{-2.685}{0.62004}$$

and $R_{22} = -3.819$ m. In summary, the radii of the crown element are $R_{11} = 21.8$ cm and $R_{12} = -21.8$ cm while the flint has radii of $R_{21} = -21.8$ cm and $R_{22} = -381.9$ cm.

Note that for a thin-lens combination the principal planes coalesce, so that achromatizing the focal length corrects both A·CA and L·CA. In a thick doublet, however, even though the focal lengths for red and blue are made identical, the different wavelengths may have different principal planes. Consequently, although the magnification is the same for all wavelengths, the focal points may not coincide; in other words, correction is made for L·CA but not for A·CA.

In the above analysis, only the *C*- and *F*-rays were brought to a common focus, and the *d*-line was introduced to establish a focal length for the doublet as a whole. It is not possible for *all* wavelengths traversing a doublet achromat to meet at a common focus. The resulting residual chromatism is known as *secondary spectrum*. The elimination of secondary spectrum is particularly troublesome when the design is limited to the glasses currently available. Nevertheless, a fluorite (CaF_2) element combined with an appropriate glass element can form a doublet achromatized at three wavelengths and having very little secondary spectrum. More often triplets are used for color correction at three or even four wavelengths. The secondary spectrum of a binocular can easily be observed by looking at a distant white object. Its borders will be slightly haloed in magenta and green—try shifting the focus forward and backward.

Separated Achromatic Doublets

It is also possible to achromatize the focal length of a doublet composed of two widely separated elements of the same glass. Return to Eq. (6.46) and set $n_{1R} = n_{2R} = n_R$ and $n_{1B} = n_{2B} = n_B$. After a bit of straightforward algebraic manipulation, it becomes

$$(n_R - n_B)[(\rho_1 + \rho_2) - \rho_1\rho_2 d(n_B + n_R - 2)] = 0$$

or

$$d = \frac{1}{(n_B + n_R - 2)}\left(\frac{1}{\rho_1} + \frac{1}{\rho_2}\right)$$

Again introducing the yellow reference frequency, as we did before, namely, $1/f_{1Y} = (n_{1Y} - 1)\rho_1$ and $1/f_{2Y} = (n_{2Y} - 1)\rho_2$, we can replace ρ_1 and ρ_2. Hence

$$d = \frac{(f_{1Y} + f_{2Y})(n_Y - 1)}{n_B + n_R - 2}$$

where $n_{1Y} = n_{2Y} = n_Y$. Assuming $n_Y = (n_B + n_R)/2$, we have

$$d = \frac{f_{1Y} + f_{2Y}}{2}$$

or in *d*-light

$$d = \frac{f_{1d} + f_{2d}}{2} \tag{6.55}$$

This is precisely the form taken by the Huygens ocular (Section 5.7.4). Since the red and blue focal lengths are the same, but the corresponding principal planes for the doublet need not be, the two rays will generally not meet at the same focal point. Thus the ocular's lateral chromatic aberration is well corrected, but axial chromatic aberration is not.

In order for a system to be free of both chromatic aberrations, the red and blue rays must emerge parallel to each other (no L·CA) and must intersect the axis at the same point (no A·CA), which means they must overlap. Since this is effectively the case with a thin achromat, it implies that multi-element systems, as a rule, should consist of achromatic components in order to keep the red and blue rays from separating (Fig. 6.41). As with all such invocations there are exceptions. The Taylor triplet (Section 5.7.7) is one. The two colored rays for which it is achromatized separate within the lens but are recombined and emerge together.

6.4 GRIN Systems

An ordinary homogeneous lens has two physical features that contribute to the manner in which it reconfigures a wavefront: the difference between its index of refraction and that of the

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

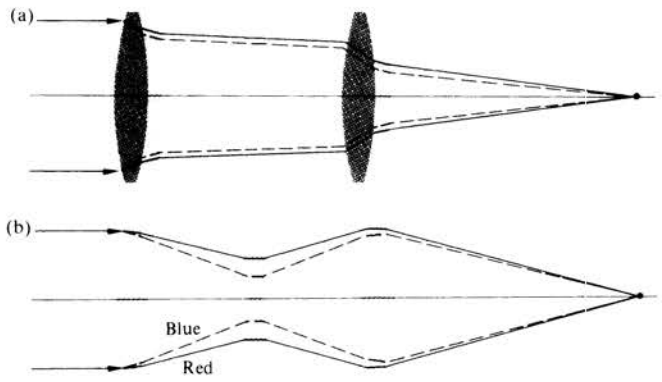


Figure 6.41 Achromatized lenses.

surrounding medium, and the curvature of its interfaces. But as we have already seen, when light propagates through an inhomogeneous medium, wavefronts essentially slow down in optically dense regions and speed up in less dense regions, and bending again occurs. In principle, then, it should be possible to make a lens from some inhomogeneous material, one where there is a GRadient in the INdex of refraction; such a device is known as a **GRIN lens**. A powerful incentive for developing such systems is that they provide the optical designer with an additional set of new parameters with which to control aberrations.

To get a rough sense of how a GRIN lens might work, consider the device pictured in Fig. 6.42 where, for simplicity, we assume $f > r$. This is a flat disk of glass that has been treated so that it has an index $n(r)$ that drops off radially in some as yet undetermined fashion from a maximum value of n_{max} on the optical axis. Accordingly, it's called a **radial-GRIN** device. A ray that traverses the disk on the optical axis passes along an optical path length of $(OPL)_o = n_{max}d$, whereas for a ray traversing at a height r, overlooking the slight bending of its path, $(OPL)_r \approx n(r)d$. Since a planar wavefront must bend into a spherical wavefront, the OPLs from one to the other, along any route must be equal (p. 150):

$$(OPL)_r + \overline{AB} = (OPL)_o$$

and

$$n(r)d + \overline{AB} = n_{max}d$$

But $\overline{AF} \approx \sqrt{r^2 + f^2}$; moreover, $\overline{AB} = \overline{AF} - f$ and so

$$n(r) = n_{max} - \frac{\sqrt{r^2 + f^2} - f}{d}$$

Rewriting the square root via the Binomial Theorem, $n(r)$ becomes

$$n(r) = n_{max} - \frac{r^2}{2fd}$$

This tells us that if the index of refraction drops off parabolically from its high along the central axis, the GRIN plate will focus a collimated beam at F and serve as a positive lens. Although this is a rather simplistic treatment, it does make the point: a parabolic refractive index profile will focus parallel light.

Today a variety of radial-GRIN lenses are commercially available, and tens of millions of them are already in service in laser printers, photocopiers, and fax machines. The most common device is a GRIN cylinder a few millimeters in diameter, similar in kind to the optical fiber shown in Fig. 5.72*b*. Monochromatically, they provide nearly diffraction-limited performance on axis. Polychromatically, they offer substantial benefits over aspherics.

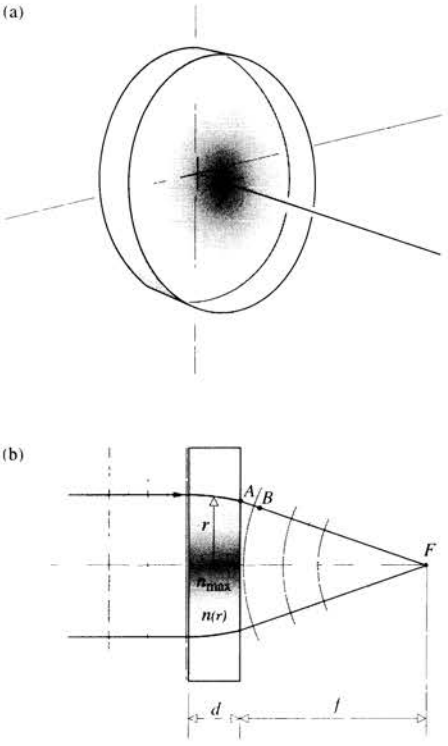


Figure 6.42 A disk of transparent glass whose index of refraction decreases radially out from the central axis. (*b*) The geometry corresponding to the focusing of parallel rays by a GRIN lens.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

These small-diameter GRIN rods are usually fabricated via ionic diffusion. A homogeneous base glass is immersed in a molten salt bath for many hours during which ion diffusion/exchange slowly occurs. One type of ion migrates out of the glass, and another from the bath takes its place, changing the index of refraction. The process works its way inward radially toward the optical axis, and the time it takes is roughly proportional to the rod's diameter squared. For a parabolic profile, that sets the practical limit on the aperture size. The focal length is determined by the index change, Δn, and the faster the lens the larger must be Δn. Even so, Δn is usually constrained to be less than about 0.10 for production reasons. Most GRIN cylinders have a parabolic index profile typically expressed as

$$n(r) = n_{\max}(1 - ar^2/2)$$

Figure 6.43 shows one such radial-GRIN rod of length L, under monochromatic illumination. Meridional rays travel sinusoidal paths within the plane-of-incidence, which is perpendicular to the circumference. These sinusoids have a period in space of $2\pi/\sqrt{a}$, where the **gradient constant** $\sqrt{a}$, is a function of λ and depends on the specific GRIN material. The cross-sectional view in Fig. 6.43*a* shows how a radial-GRIN lens can form an erect, real, magnified image. By changing the object distance or the length of the lens L, a wide range of images can be produced. It's even possible to have the object and image planes on the face of the rod (Figs. 6.43*b* and *c*).

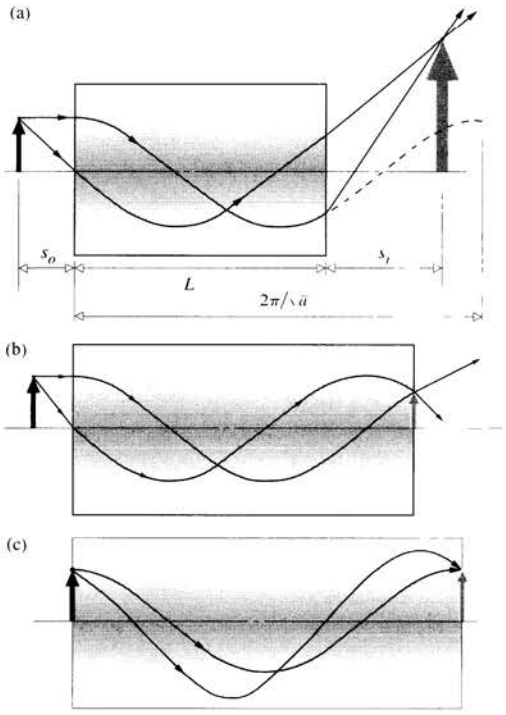


Figure 6.43 (*a*) A radial-GRIN rod producing a real, magnified, erect image. (*b*) Here the image is formed on the face of the rod. (*c*) This is a convenient setup for use in a copy machine.

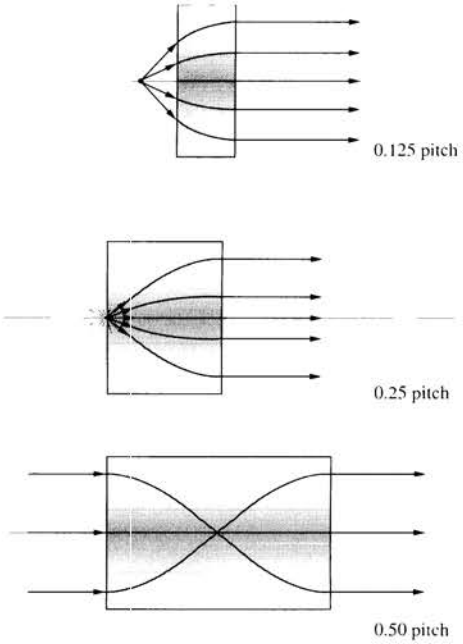


Figure 6.44 Radial GRIN lenses with several pitches used in a few typical ways.

Radial-GRIN lenses are often specified in terms of their length or, equivalently, their **pitch** (Fig. 6.44). A radial-GRIN rod with a pitch of 1.0 is one sine-wave long: $L = 2\pi/\sqrt{a}$. A rod with a pitch of 0.25 has a length of a quarter of a sine-wave $(\pi/2\sqrt{a})$.

An alternative approach to the flat-faced radial-GRIN rod is the axial-GRIN lens which is generally polished with spherical surfaces. As such, it's similar to a bi-aspheric, but without the difficulty of creating the complex surfaces. Usually, a stack of glass plates with appropriate indices are fused together. At high temperatures the glasses meld, diffusing into each other, creating a block of glass with a continuous index profile that can be made linear, quadratic, or even cubic (Fig. 6.45*a*). When such a block is ground into a lens, the process cuts back

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

on the glass and exposes a range of indices. Every annulus (concentric with the optical axis) on the lens's face has a gradually changing index. Rays impinging at different heights above the optical axis encounter glass with different indices and bend appropriately. The spherical aberration evident in Fig. 6.45*c* arises because the edges of the spherical lens refract too strongly. Gradually lowering the index of refraction out toward the edges allows the axial-GRIN lens to correct for spherical aberration.

Generally, introducing a GRIN element into the design of a compound lens greatly simplifies the system, reducing the number of elements by as much as a third while maintaining overall performance.

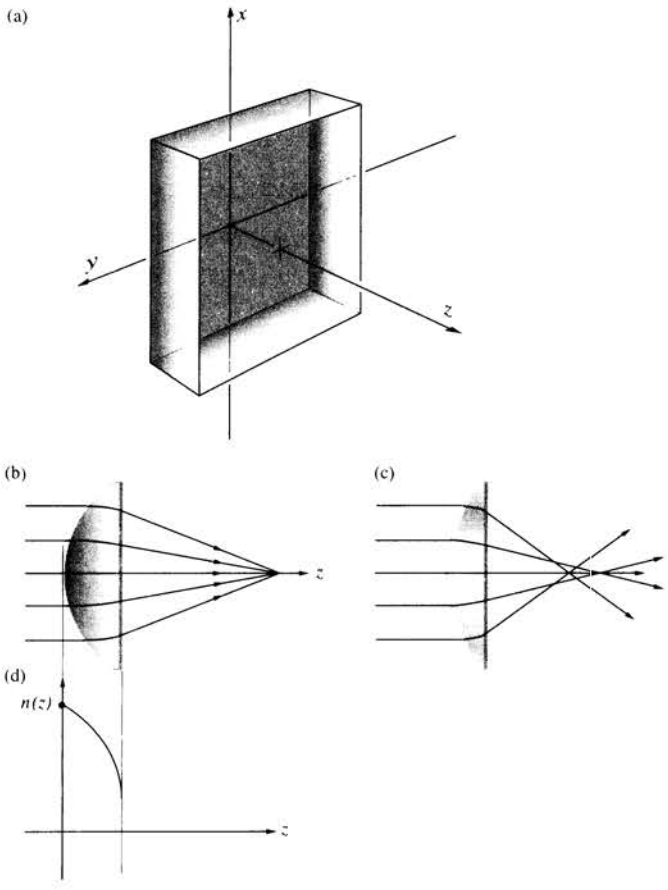


Figure 6.45 (*a*) A slab of axial-GRIN material for which the index of refraction is $n(z)$. (*b*) An axial-GRIN lens for which there is no spherical aberration. (*c*) An ordinary lens having SA. (*d*) The index profile.

6.5 Concluding Remarks

For the practical reason of manufacturing ease, the vast majority of optical systems are limited to lenses having spherical surfaces. There are, to be sure, toric and cylindrical lenses as well as many other aspherics. Indeed, very fine, and as a rule very expensive devices, such as high-altitude reconnaisance cameras and tracking systems, may have several aspherical elements. Even so, spherical lenses are here to stay and with them are their inherent aberrations which must satisfactorily be dealt with. As we have seen, the designer (and his faithful electronic companion) must manipulate the system variables (indices, shapes, spacings, stops, etc.) in order to balance out offensive aberrations. This is done to whatever degree and in whatever order is appropriate for the specific optical system. Thus one might tolerate far more distortion and curvature in an ordinary telescope than in a good photographic objective. Similarly, there is little need to worry about chromatic aberration if you want to work exclusively with laser light of almost a single frequency.

In any event, this chapter has only touched on the problems (more to appreciate than solve them). That they are most certainly amenable to solution is evidenced, for example, by the accompanying aerial photographs, which speak rather eloquently for themselves, especially when you consider that a good spy satellite will do upwards of ten times better than this.

(*a*) New Orleans and the Mississippi River photographed from 12,500 m (41,000 ft) with Itek's Metritek-21 camera (f = 21 cm). Ground resolution, 1 m; scale, 1:59,492. (*b*) Photo scale, 1:10,000. (*c*) Photo scale, 1:2500. (Photos courtesy of Litton/Itek Optical Systems.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



(a)



(b)



(c)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

PROBLEMS

Complete solutions to all problems—except those with an asterisk—can be found in the back of the book.

6.1* Work out the details leading to Eq. (6.8).

6.2 According to the military handbook MIL-HDBK-141 (23.3.5.3), the Ramsden eyepiece (Fig. 5.94) is made up of two planar-convex lenses of equal focal length f' separated by a distance $2f'/3$. Determine the overall focal length f of the thin-lens combination and locate the principal planes and position of the field stop.

6.3 Write an expression for the thickness d_l of a double-convex lens such that its focal length is infinite.

6.4 Suppose we have a positive meniscus lens of radii 6 and 10 and a thickness of 3 (any units, as long as you're consistent), with an index of 1.5. Determine its focal length and the locations of its principal points (compare with Fig. 6.3).

6.5* Prove that if the principal points of a biconvex lens of thickness d_l overlap midway between the vertices, the lens is a sphere. Assume the lens is in air.

6.6 Using Eq. (6.2), derive an expression for the focal length of a homogeneous transparent sphere of radius R. Locate its principal points.

6.7* A spherical glass bottle 20 cm in diameter with walls that are negligibly thin is filled with water. The bottle is sitting on the back seat of a car on a nice sunny day. What's its focal length?

6.8* With the previous two problems in mind, compute the magnification that results when the image of a flower 4.0 m from the center of a solid, clear-plastic sphere with a 0.20-m diameter (and a refractive index of 1.4) is cast on a nearby wall. Describe the image in detail.

6.9* A thick glass lens of index 1.50 has radii of +23 cm and +20 cm, so that both vertices are to the left of the corresponding centers of curvature. Given that the thickness is 9.0 cm, find the focal length of the lens. Show that in general $R_1 - R_2 = d/3$ for such afocal zero-power lenses. Draw a diagram showing what happens to an axial incident parallel bundle of rays as it passes through the system.

6.10 It is found that sunlight is focused to a spot 29.6 cm from the back face of a thick lens, which has its principal points H_1 at +0.2 cm and H_2 at −0.4 cm. Determine the location of the image of a candle that is placed 49.8 cm in front of the lens.

6.11* Please establish that the separation between principal planes for a thick glass lens is roughly one-third its thickness. The simplest geometry occurs with a planar-convex lens tracing a ray from the object focus. What can you say about the relationship between the focal length and the thickness for this lens type?

6.12 A crown glass double-convex lens, 4.0 cm thick and operating at a wavelength of 900 nm, has an index of refraction of 3/2. Given that its radii are 4.0 cm and 15 cm, locate its principal points and compute its focal length. If a television screen is placed 1.0 m from the front of the lens, where will the real image of the picture appear?

6.13* Imagine two identical double-convex thick lenses separated by a distance of 20 cm between their adjacent vertices. Given that all the radii of curvature are 50, the refractive indices are 1.5, and the thickness of each lens is 5.0 cm, calculate the combined focal length.

6.14* A compound lens is composed of two thin lenses separated by 10 cm. The first of these has a focal length of +20 cm, and the second a focal length of −20 cm. Determine the focal length of the combination and locate the corresponding principal points. Draw a diagram of the system.

6.15* A convex-planar lens of index 3/2 has a thickness of 1.2 cm and a radius of curvature of 2.5 cm. Determine the system matrix when light is incident on the curved surface.

6.16* A positive meniscus lens with an index of refraction of 2.4 is immersed in a medium of index 1.9. The lens has an axial thickness of 9.6 mm and radii of curvature of 50.0 mm and 100 mm. Compute the system matrix when light is incident on the convex face and show that its determinant is equal to 1.

6.17* Prove that the determinant of the system matrix in Eq. (6.31) is equal to 1.

6.18 Establish that Eqs. (6.36) and (6.37) are equivalent to Eqs. (6.3) and (6.4), respectively.

6.19 Show that the planar surface of a concave-planar or convex-planar lens doesn't contribute to the system matrix.

6.20 Compute the system matrix for a thick biconvex lens of index 1.5 having radii of 0.5 and 0.25 and a thickness of 0.3 (in any units you like). Check that $|\mathscr{A}| = 1$.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

6.21* The system matrix for a thick biconvex lens in air is given by

$$\begin{bmatrix} 0.6 & -2.6 \\ 0.2 & 0.8 \end{bmatrix}$$

Knowing that the first radius is 0.5 cm, that the thickness is 0.3 cm, and that the index of the lens is 1.5, find the other radius.

6.22* A concave-planar glass ($n = 1.50$) lens in air has a radius of 10.0 cm and a thickness of 1.00 cm. Determine the system matrix and check that its determinant is 1. At what positive angle (in radians measured above the axis) should a ray strike the lens at a height of 2.0 cm, if it is to emerge from the lens at the same height but parallel to the optical axis?

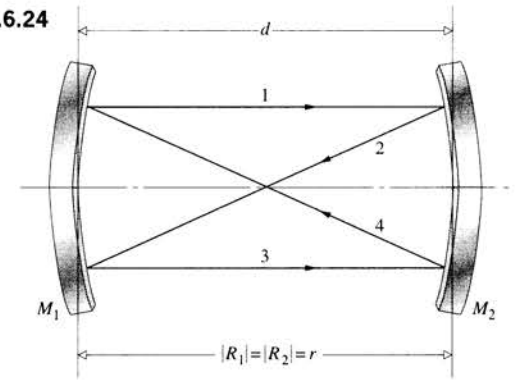
6.23* Considering the lens in Problem 6.20, determine its focal length and the location of the focal points with respect to its vertices V_1 and V_2.

6.24* Figure P.6.24 shows two identical concave spherical mirrors forming a so-called confocal cavity. Show, without first specifying the value of d, that after traversing the cavity two times the system matrix is

$$\begin{bmatrix} \left(\frac{2d}{r} - 1\right)^2 - \frac{2d}{r} & \frac{4}{r}\left(\frac{d}{r} - 1\right) \\ 2d\left(1 - \frac{d}{r}\right) & 1 - 2\frac{d}{r} \end{bmatrix}$$

Then for the specific case of $d = r$ show that after four reflections the system is back where it started and the light will retrace its original path.

Figure P.6.24

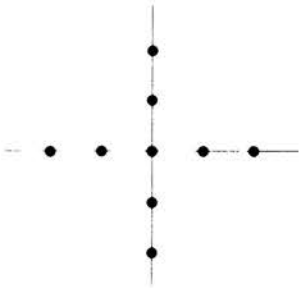


6.25 Referring back to Fig. 6.17*a*, show that when $\overline{P'C} = Rn_2/n_1$ and $\overline{PC} = Rn_1/n_2$ all rays originating at P appear to come from P'.

6.26 Starting with the exact expression given by Eq. (5.5), show that Eq. (6.40) results, rather than Eq. (5.8), when the approximations for ℓ_o and ℓ_i are improved a bit.

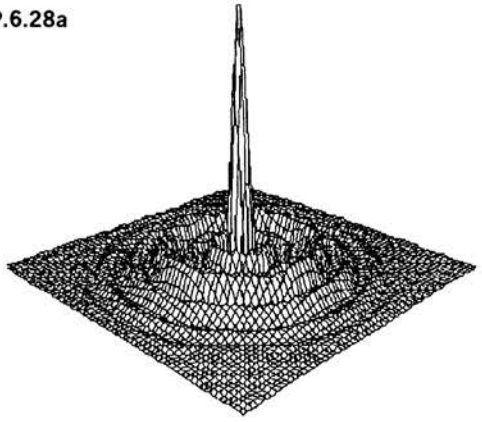
6.27 Supposing that Fig. P.6.27 is to be imaged by a lens system suffering spherical aberration only, make a sketch of the image.

Figure P.6.27



6.28* Figure P.6.28 shows the image irradiance distributions arising when a monochromatic point source illuminates three different optical systems, each having only one type of aberration. From the graphs identify that aberration in each case and justify your answer.

Figure P.6.28a



Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Figure P.6.28b

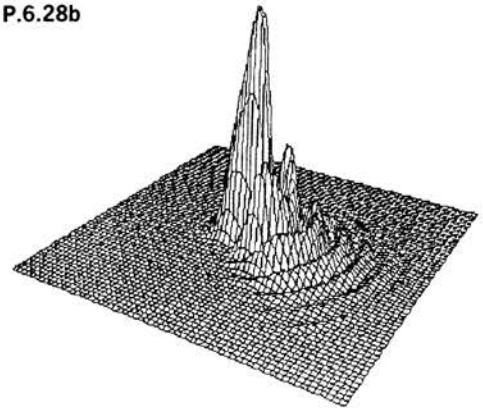
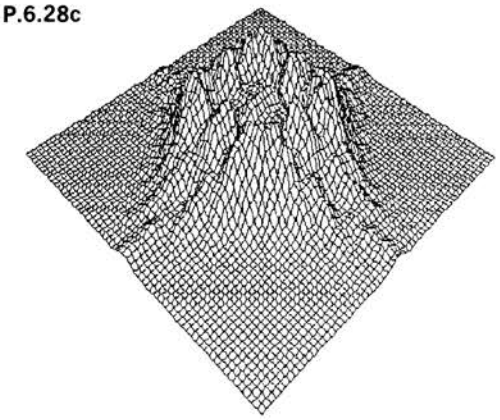
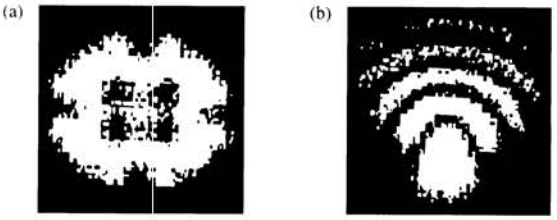


Figure P.6.28c



6.29* Figure P.6.29 shows the distribution of light corresponding to the image arising when a monochromatic point source illuminates two different optical systems each having only one type of aberration. Identify the aberration in each case and justify your answer.

Figure P.6.29



Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

7 The Superposition of Waves

In succeeding chapters we shall study the phenomena of polarization, interference, and diffraction. These all share a common conceptual basis in that they deal, for the most part, with various aspects of the same process. Stating this in the simplest terms, we are really concerned with what happens when two or more lightwaves overlap in some region of space. The precise circumstances governing this superposition determine the final optical disturbance. Among other things we are interested in learning how the specific properties of each constituent wave (amplitude, phase, frequency, etc.) influence the ultimate form of the composite disturbance.

Recall that each field component of an electromagnetic wave (E_x, E_y, E_z, B_x, B_y, and B_z) satisfies the scalar three-dimensional differential wave equation,

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2} \qquad [2.60]$$

A significant feature of this expression is that it is *linear*; $\psi(\vec{r}, t)$ and its derivatives appear only to the first power. Consequently, if $\psi_1(\vec{r}, t), \psi_2(\vec{r}, t), \ldots, \psi_n(\vec{r}, t)$ are individual solutions of Eq. (2.60), *any linear combination* of them will, in turn, be a solution. Thus

$$\psi(\vec{r}, t) = \sum_{i=1}^{n} C_i\psi_i(\vec{r}, t) \qquad (7.1)$$

satisfies the wave equation, where the coefficients C_i are simply arbitrary constants. Known as the ***Principle of Superposition***, this property suggests that the resultant disturbance at any point in a medium is the algebraic sum of the separate constituent waves (Fig. 7.1). At this time we are interested only in linear systems where the superposition principle is applicable. Do keep in mind, however, that large-amplitude waves, whether sound waves or waves on a string, can generate a nonlinear response. The focused beam of a high-intensty laser (where the electric field might be as high as 10^{10} V/cm) is easily capable of eliciting nonlinear effects (see Chapter 13).

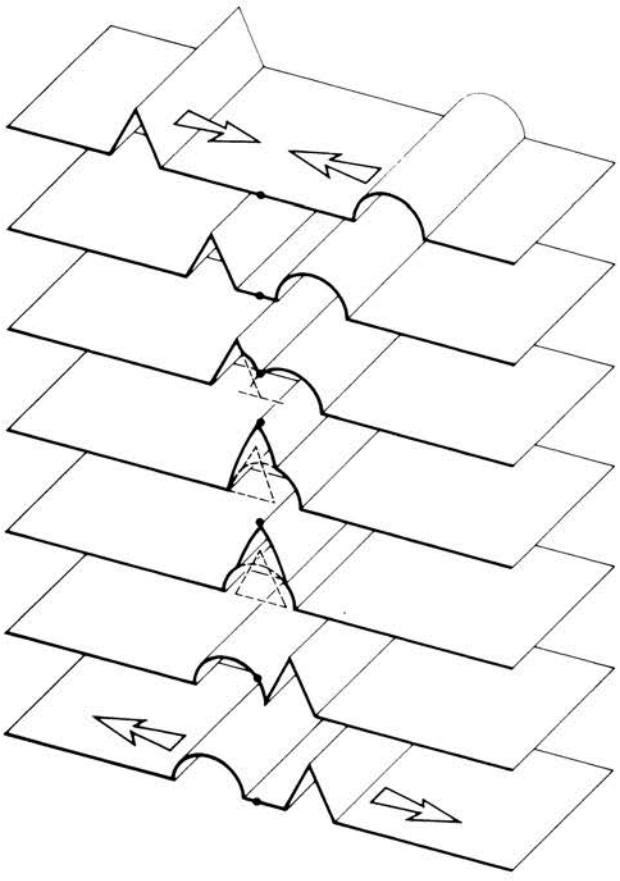


Figure 7.1 The superposition of two disturbances.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

By comparison, the electric field associated with sunlight here on Earth has an amplitude of only about 10 V/cm.

In many instances we need not be concerned with the vector nature of light, and for the present we will restrict ourselves to such cases. For example, if the lightwaves all propagate along the same line and share a common constant plane of vibration, they can each be described in terms of one electric-field component. These would all be either parallel or antiparallel at any instant and could thus be treated as scalars. A good deal more will be said about this point as we progress; for now, let's represent the optical disturbance as a scalar function $E(\vec{r}, t)$, which is a solution of the differential wave equation. This approach leads to a simple scalar theory that is highly useful as long as we are careful about applying it.

7.1 The Addition of Waves of the Same Frequency

There are several equivalent ways of mathematically adding two or more overlapping waves that have the same frequency and wavelength. Let's examine these different approaches so that, in any particular situation, we can use the one most suitable.

7.1.1 The Algebraic Method

A solution of the differential wave equation can be written in the form

$$E(x, t) = E_0 \sin [\omega t - (kx + \varepsilon)] \tag{7.2}$$

in which E_0 is the amplitude of the harmonic disturbance propagating along the positive x-axis. To separate the space and time parts of the phase, let

$$\alpha(x, \varepsilon) = -(kx + \varepsilon) \tag{7.3}$$

so that

$$E(x, t) = E_0 \sin [\omega t + \alpha(x, \varepsilon)] \tag{7.4}$$

Suppose then that there are two such waves

$$E_1 = E_{01} \sin (\omega t + \alpha_1) \tag{7.5a}$$

and

$$E_2 = E_{02} \sin (\omega t + \alpha_2) \tag{7.5b}$$

each with the same frequency and speed, coexisting in space. The resultant disturbance is the linear superposition of these waves:

$$E = E_1 + E_2$$

Here it helps to know beforehand what we're looking for. The sum should resemble Eq. (7.4); you can't add two signals of the same frequency and get a resultant with a different frequency. That makes a lot of sense when you remember that the frequency of a photon corresponds to its energy and that doesn't change. In any event, we can anticipate a sinusoidal function of frequency ω with an amplitude (E_0) and phase (α) that have to be determined.

Forming the sum and expanding Eqs. (7.5a) and (7.5b) leads to

$$E = E_{01} (\sin \omega t \cos \alpha_1 + \cos \omega t \sin \alpha_1) + E_{02} (\sin \omega t \cos \alpha_2 + \cos \omega t \sin \alpha_2)$$

When we separate out the time-dependent terms, this becomes

$$E = (E_{01} \cos \alpha_1 + E_{02} \cos \alpha_2) \sin \omega t + (E_{01} \sin \alpha_1 + E_{02} \sin \alpha_2) \cos \omega t \tag{7.6}$$

Since the parenthetical quantities are constant in time, let

$$E_0 \cos \alpha = E_{01} \cos \alpha_1 + E_{02} \cos \alpha_2 \tag{7.7}$$

and

$$E_0 \sin \alpha = E_{01} \sin \alpha_1 + E_{02} \sin \alpha_2 \tag{7.8}$$

This is not an obvious substitution, but it will be legitimate as long as we can solve for E_0 and α. To that end, and remembering that $\cos^2\alpha + \sin^2\alpha = 1$, square and add Eqs. (7.7) and (7.8) to get

$$E_0^2 = E_{01}^2 + E_{02}^2 + 2E_{01}E_{02} \cos (\alpha_2 - \alpha_1) \tag{7.9}$$

That's the sought-after expression for the amplitude (E_0) of the resultant wave. Now to get the phase, divide Eq. (7.8) by (7.7):

$$\tan \alpha = \frac{E_{01} \sin \alpha_1 + E_{02} \sin \alpha_2}{E_{01} \cos \alpha_1 + E_{02} \cos \alpha_2} \tag{7.10}$$

Provided these last two expressions are satisfied for E_0 and α, the situation of Eqs. (7.7) and (7.8) is valid. The total distur-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

bance [Eq. (7.6)] then becomes

$$E = E_0 \cos \alpha \sin \omega t + E_0 \sin \alpha \cos \omega t$$

or

$$E = E_0 \sin (\omega t + \alpha) \tag{7.11}$$

where you can use Eq. (7.9) to determine E_0 and Eq. (7.10) to compute α. A single disturbance results from the superposition of the sinusoidal waves E_1 and E_2. *The composite wave [Eq. (7.11)] is harmonic and of the same frequency as the constituents, although its amplitude and phase are different.*

Note that when $E_{01} >> E_{02}$ in Eq. (7.10), $\alpha \approx \alpha_1$ and when $E_{02} >> E_{01}$, $\alpha \approx \alpha_2$; the resultant is in-phase with the dominant component wave (take another look at Fig. 4.11). The flux density of a lightwave is proportional to its amplitude squared, by way of Eq. (3.44). It follows from Eq. (7.9) that the resultant flux density is not simply the sum of the component flux densities; there is an additional contribution $2E_{01}E_{02} \cos (\alpha_2 - \alpha_1)$, known as the **interference term**. The crucial factor is the difference in phase between the two *interfering* waves E_1 and E_2, $\delta \equiv (\alpha_2 - \alpha_1)$. When $\delta = 0, \pm 2\pi, \pm 4\pi, \ldots$ the resultant amplitude is a maximum, whereas $\delta = \pm\pi, \pm 3\pi, \ldots$ yields a minimum at any point in space (Problem 7.3). In the former case, the waves are said to be in-phase; crest overlaps crest. In the latter instance, the waves are 180° out-of-phase and trough overlaps crest, as shown in Fig. 7.2. Realize that the *phase difference* may arise from a difference in path length traversed by the two waves, as well as a difference in the initial phase angle; that is,

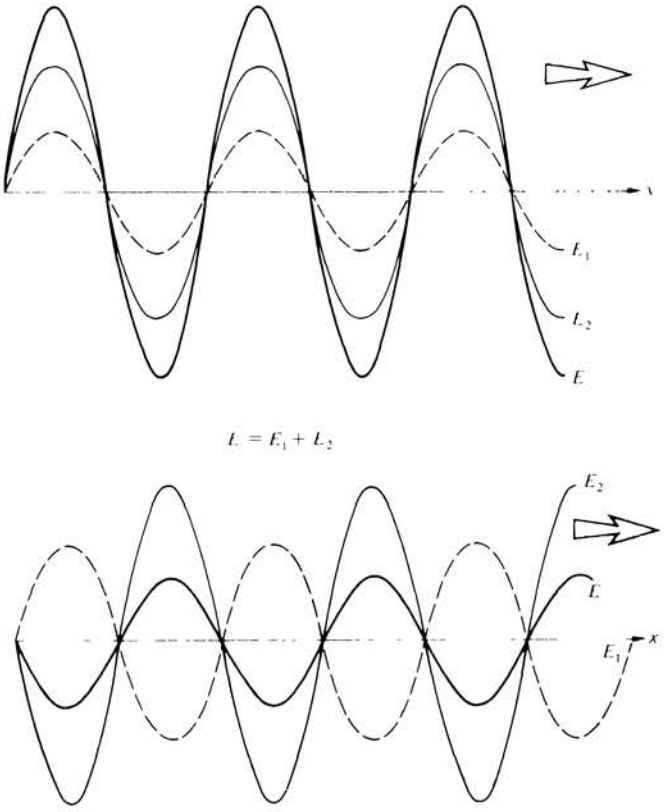


Figure 7.2 The superposition of two harmonic waves in-phase and out-of-phase.

$$\delta = (kx_1 + \varepsilon_1) - (kx_2 + \varepsilon_2) \tag{7.12}$$

or

$$\delta = \frac{2\pi}{\lambda}(x_1 - x_2) + (\varepsilon_1 - \varepsilon_2) \tag{7.13}$$

Here x_1 and x_2 are the distances from the sources of the two waves to the point of observation, and λ is the wavelength in the pervading medium. If the waves are initially in-phase at their respective emitters, then $\varepsilon_1 = \varepsilon_2$, and

$$\delta = \frac{2\pi}{\lambda}(x_1 - x_2) \tag{7.14}$$

This would also apply to the case in which two disturbances from the same source traveled different routes before arriving at the point of observation. Since $n = c/v = \lambda_0/\lambda$,

$$\delta = \frac{2\pi}{\lambda_0} n(x_1 - x_2) \tag{7.15}$$

The quantity $n(x_1 - x_2)$ is known as the **optical path difference** and will be represented by the abbreviation OPD or by the symbol Λ. It's the difference in the two optical path lengths [Eq. (4.9)]. It is possible, in more complicated situations, for each wave to travel through a number of different thicknesses of different media (Problem 7.6). Notice too that $\Lambda/\lambda_0 = (x_1 - x_2)/\lambda$ is the number of waves in the medium corresponding to the path difference; one route is that many wavelengths longer than the other. Since each wavelength is associated with a 2π radian phase change, $\delta = 2\pi(x_1 - x_2)/\lambda$, or

$$\delta = k_0 \Lambda \tag{7.16}$$

k_0 being the propagation number in vacuum; that is, $2\pi/\lambda_0$. One route is essentially δ radians longer than the other.

Waves for which $\varepsilon_1 - \varepsilon_2$ is *constant*, regardless of its value, are said to be **coherent**, a situation we shall assume obtains throughout most of this discussion.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

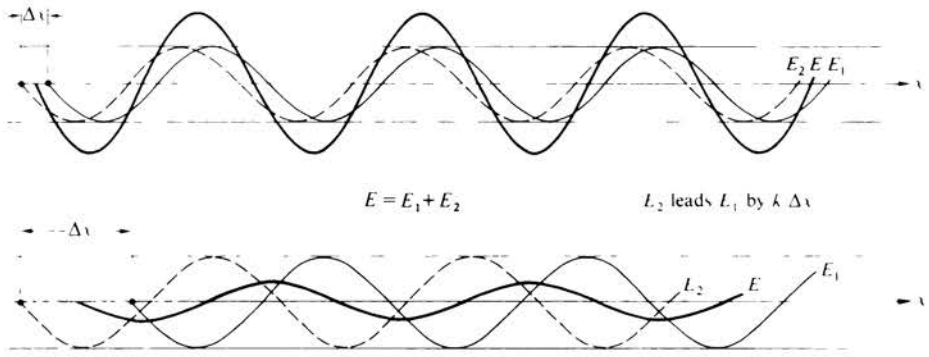


Figure 7.3 Waves out-of-phase by $k\Delta x$ radians.

One special case of some interest is the superposition of two waves that travel slightly different distances (Δx) in the same direction:

$$E_1 = E_{01} \sin [\omega t - k(x + \Delta x)]$$

and

$$E_2 = E_{02} \sin (\omega t - kx)$$

where in particular $E_{01} = E_{02}$ and $\alpha_2 - \alpha_1 = k\,\Delta x$. It is left to Problem 7.7 to show that in this case Eqs. (7.9), (7.10), and (7.11) lead to a resultant wave of

$$E = 2E_{01} \cos\left(\frac{k\,\Delta x}{2}\right) \sin\left[\omega t - k\left(x + \frac{\Delta x}{2}\right)\right] \quad (7.17)$$

This brings out rather clearly the dominant role played by the path length difference, Δx, especially when the waves are emitted in-phase ($\varepsilon_1 = \varepsilon_2$). There are many practical instances in which one arranges just these conditions, as will be seen later. If $\Delta x << \lambda$, the resultant has an amplitude that is nearly $2E_{01}$, whereas if $\Delta x = \lambda/2$, since $k = 2\pi/\lambda$, the cosine term is zero and $E = 0$. Recall that the former situation (p. 20) is referred to as **constructive interference**, and the latter as **destructive interference** (see Fig. 7.3).

To underscore the potential for practical application of these ideas, consider Fig. 7.4. It shows a jet fighter that has been illuminated by microwaves from a hostile ground-based radar transmitter. To the considerable consternation of the pilot (and unlike the F-117 Stealth fighter, p. 100) the plane reflects a substantial amount of radiant energy down toward the radar antenna. But all is not lost; having detected the beam, the plane matches its frequency and amplitude, and transmits a $\lambda/2$ phase-shifted radar wave of its own. Propagating back to the source in nearly the same direction, the two waves (reflected and emitted) interfere destructively [via Eq. (7.17)], thereby eliminating the radar echo in the specific direction of the enemy detector. Of course, if there happened to be several ground receivers, the pilot could be in big trouble.

The Superposition of Many Waves

By repeated applications of the procedure used to arrive at Eq. (7.11), we can show that the ***superposition of any number of coherent harmonic waves having a given frequency and traveling in the same direction leads to a harmonic wave of that same frequency*** (Fig. 7.5). We happen to have chosen to represent the two waves above in terms of sine functions, but the same results would prevail if cosine functions were used. In general, then, the sum of N such waves,

$$E = \sum_{i=1}^{n} E_{0i} \cos (\alpha_i \pm \omega t)$$

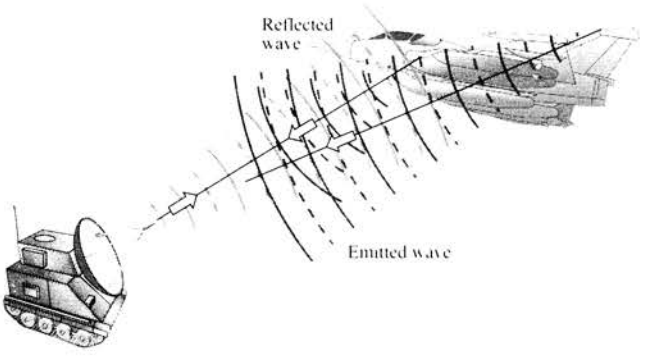


Figure 7.4 The French fighter Rafale uses active cancellation to confound radar detection. It sends out a signal that's one half a wavelength out-of-phase with the radar wave it reflects. The reflected and emitted waves cancel in the direction of the enemy receiver.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

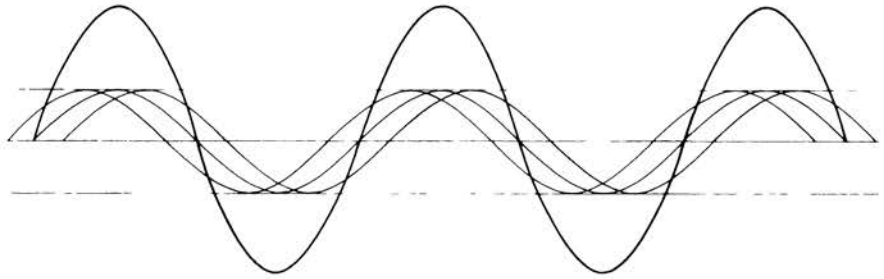


Figure 7.5 The superposition of three harmonic waves yields a harmonic wave of the same frequency.

is given by $$E = E_0 \cos(\alpha \pm \omega t) \qquad (7.18)$$

where

$$E_0^2 = \sum_{i=1}^{N} E_{0i}^2 + 2\sum_{j>i}^{N}\sum_{i=1}^{N} E_{0i}E_{0j}\cos(\alpha_i - \alpha_j) \qquad (7.19)$$

and $$\tan\alpha = \frac{\sum_{i=1}^{N} E_{0i}\sin\alpha_i}{\sum_{i=1}^{N} E_{0i}\cos\alpha_i} \qquad (7.20)$$

Pause for a moment and satisfy yourself that these relations are indeed true.

Consider a number (N) of atomic emitters constituting an ordinary source (an incandescent bulb, candle flame, or discharge lamp). A flood of light is emitted that presumably corresponds to a torrent of photons, which manifest themselves *en masse* as an electromagnetic wave. To keep things in a wave perspective, it's useful to imagine the photon as somehow associated with a short-duration oscillatory wave pulse. Each atom is effectively an independent source of photon wavetrains (Section 3.4.4), and these, in turn, extend in time for roughly 1 to 10 ns. In other words, the atoms can be thought of as emitting wavetrains that have a sustained phase for only up to about 10 ns. After that a new wavetrain may be emitted with a totally random phase, and it too will be sustained for less than approximately 10 ns, and so forth. On the whole each atom emits a disturbance (composed of a stream of photons) that varies in its phase rapidly and randomly.

In any event, the phase of the light from one atom, $\alpha_i(t)$, will remain constant with respect to the phase from another atom $\alpha_j(t)$, for only a time of at most 10 ns before it changes randomly: the atoms are coherent for up to about 10^{-8} s. Since flux density is proportional to the time average of E_0^2, generally taken over a comparatively long interval of time, it follows that the second summation in Eq. (7.19) will contribute terms proportional to $\langle\cos[\alpha_i(t) - \alpha_j(t)]\rangle$, each of which will average out to zero because of the random rapid nature of the phase changes. Only the first summation remains in the time average, and its terms are constants. If each atom is emitting wavetrains of the same amplitude E_{01}, then

$$E_0^2 = NE_{01}^2 \qquad (7.21)$$

The resultant flux density arising from N sources having random, rapidly varying phases is given by N times the flux density of any one source. In other words, *it is determined by the sum of the individual flux densities.*

A flashlight bulb, whose atoms are all emitting a random tumult, puts out light which (as the superposition of these essentially "incoherent" wavetrains) is itself rapidly and randomly varying in phase. Thus two or more such bulbs will emit light that is essentially incoherent (i.e., for durations longer than about 10 ns), light whose total combined irradiance will simply equal the sum of the irradiances contributed by each individual bulb. This is also true for candle flames, flashbulbs, and all thermal (as distinct from laser) sources. ***We cannot expect to see interference when the lightwaves from two reading lamps overlap.***

At the other extreme, if the sources are coherent and in-phase at the point of observation (i.e., $\alpha_i = \alpha_j$), Eq. (7.19) will become

$$E_0^2 = \sum_{i=1}^{N} E_{0i}^2 + 2\sum_{j>i}^{N}\sum_{i=1}^{N} E_{0i}E_{0j}$$

or, equivalently,

$$E_0^2 = \left(\sum_{i=1}^{N} E_{0i}\right)^2 \qquad (7.22)$$

Again, supposing that each amplitude is E_{01}, we get

$$E_0^2 = (NE_{01})^2 = N^2E_{01}^2 \qquad (7.23)$$

In this case of in-phase coherent sources, we have a situation

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

in which the amplitudes are added first and then squared to determine the resulting flux density. The superposition of coherent waves generally has the effect of altering the spatial distribution of the energy but not the total amount present. If there are regions where the flux density is greater than the sum of the individual flux densities, there will be regions where it is less than that sum.

7.1.2 The Complex Method

It's often mathematically convenient to make use of the complex representation when dealing with the superposition of harmonic disturbances. The wave

$$E_1 = E_{01} \cos (kx \pm \omega t + \varepsilon_1)$$

or

$$E_1 = E_{01} \cos (\alpha_1 \mp \omega t)$$

can then be written as

$$\tilde{E}_1 = E_{01} e^{i(\alpha_1 \mp \omega t)} \tag{7.24}$$

if we remember that we are interested only in the real part (see Section 2.5). Suppose that there are N such overlapping waves having the same frequency and traveling in the *positive x-direction*. The resultant wave is given by

$$\tilde{E} = E_0 e^{i(\alpha + \omega t)}$$

which is equivalent to Eq. (7.18) or, upon summation of the component waves,

$$\tilde{E} = \left[\sum_{j=1}^{N} E_{0j} e^{i\alpha_j} \right] e^{+i\omega t} \tag{7.25}$$

The quantity

$$E_0 e^{i\alpha} = \sum_{j=1}^{N} E_{0j} e^{i\alpha_j} \tag{7.26}$$

is known as the *complex amplitude* of the composite wave and is simply the sum of the complex amplitudes of the constituents. Since

$$E_0^2 = (E_0 e^{i\alpha})(E_0 e^{i\alpha})^* \tag{7.27}$$

we can always compute the resultant irradiance from Eqs. (7.26) and (7.27). For example, if $N = 2$,

$$E_0^2 = (E_{01} e^{i\alpha_1} + E_{02} e^{i\alpha_2})(E_{01} e^{-i\alpha_1} + E_{02} e^{-i\alpha_2})$$

$$E_0^2 = E_{01}^2 + E_{02}^2 + E_{01} E_{02} [e^{i(\alpha_1 - \alpha_2)} + e^{-i(\alpha_1 - \alpha_2)}]$$

or

$$E_0^2 = E_{01}^2 + E_{02}^2 + 2E_{01} E_{02} \cos (\alpha_1 - \alpha_2)$$

which is identical to Eq. (7.9).

7.1.3 Phasor Addition

The summation described in Eq. (7.26) can be represented graphically as an addition of vectors in the complex plane (recall the discussion on p. 21). In the parlance of electrical engineering, the complex amplitude is known as a **phasor**, and it's specified by its magnitude and phase, often written simply as $E_0 \angle \alpha$. Imagine, then, that we have a disturbance described by

$$E_1 = E_{01} \sin (\omega t + \alpha_1)$$

In Fig. 7.6*a* the wave is represented by a vector of length E_{01} rotating counterclockwise at a rate ω such that its projection on the vertical axis is $E_{01} \sin (\omega t + \alpha_1)$. If we were concerned with cosine waves, we would take the projection on the horizontal axis. Incidentally, the rotating vector is, of course, a phasor $E_{01} \angle \alpha_1$, and the R and I designations signify the real and imaginary axes. Similarly, a second wave

$$E_2 = E_{02} \sin (\omega t + \alpha_2)$$

is depicted along with E_1 in Fig. 7.6*b*. Their algebraic sum, $E = E_1 + E_2$, is the projection on the I-axis of the resultant phasor determined by the vector addition of the component phasors, as in Fig. 7.6*c*. The law of cosines applied to the triangle of sides E_{01}, E_{02}, and E_0 yields

$$E_0^2 = E_{01}^2 + E_{02}^2 + 2E_{01} E_{02} \cos (\alpha_2 - \alpha_1)$$

where use was made of the fact that $\cos [\pi - (\alpha_2 - \alpha_1)] = -\cos (\alpha_2 - \alpha_1)$. This is identical to Eq. (7.9), as it must be. Using the same diagram, observe that $\tan \alpha$ is given by Eq. (7.10) as well. We are usually concerned with finding E_0 rather than $E(t)$, and since E_0 is unaffected by the constant

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

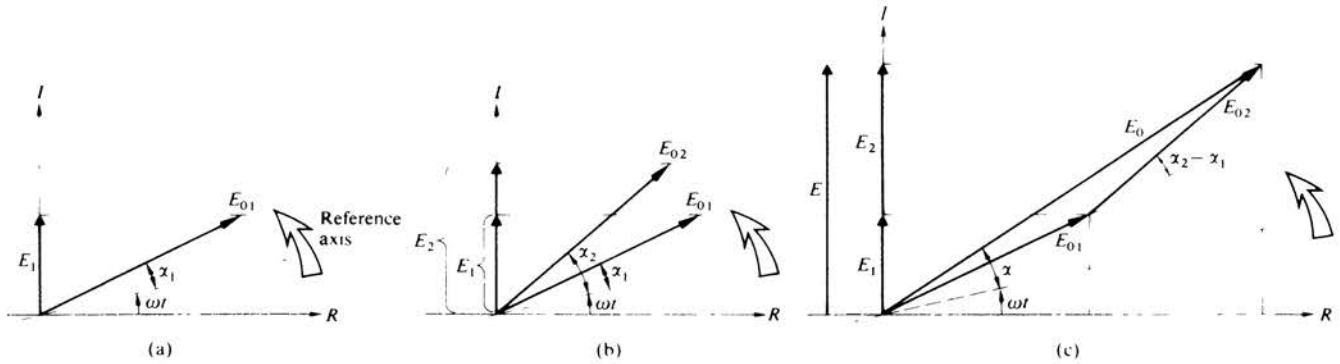


Figure 7.6 Phasor addition.

revolving of all the phasors, it will often be convenient to set $t = 0$ and eliminate that rotation.

Some rather elegant schemes, such as the *vibration curve* and the *Cornu spiral* (Chapter 10), will be predicated on the technique of phasor addition. As a further example, let's briefly examine the wave resulting from the addition of

$$E_1 = 5 \sin \omega t$$

$$E_2 = 10 \sin (\omega t + 45°)$$

$$E_3 = \sin (\omega t - 15°)$$

$$E_4 = 10 \sin (\omega t + 120°)$$

and

$$E_5 = 8 \sin (\omega t + 180°)$$

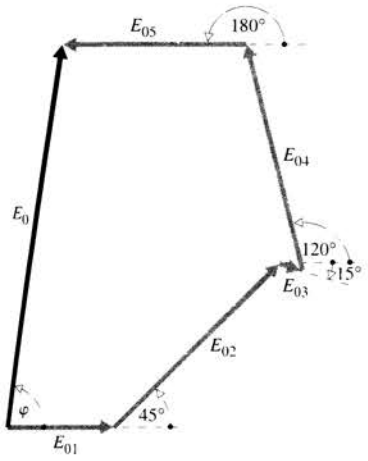


Figure 7.7 The phasor sum of E_1, E_2, E_3, E_4, and E_5.

where ω is in degrees per second. The appropriate phasors $5\angle 0°$, $10\angle 45°$, $1\angle -15°$, $10\angle 120°$, and $8\angle 180°$ are plotted in Fig. 7.7. Notice that each phase angle, whether positive or negative, is referenced to the horizontal. One need only read off $E_0\angle\varphi$ with a scale and protractor to get $E = E_0 \sin(\omega t + \alpha)$. It is evident that this technique offers a tremendous advantage in speed and simplicity, if not in accuracy.

To exemplify the method, Fig. 7.8 depicts the superposition of two different-amplitude waves of the same frequency that are out-of-phase by an amount α. Notice that the amplitude of

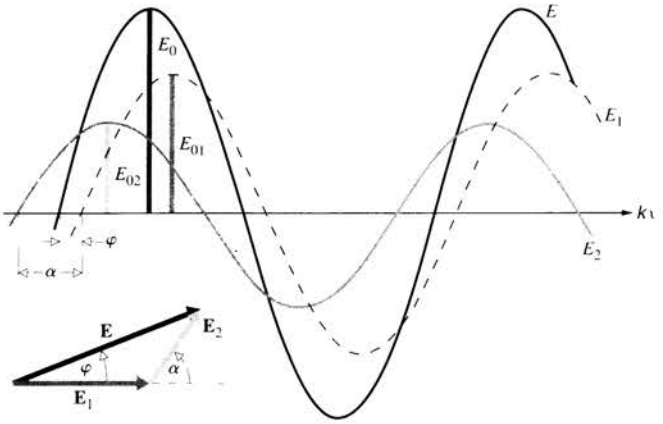


Figure 7.8 The summation of two sinusoidal functions of the same frequency using phasor addition. Here $\mathbf{E}_1$ is taken as the reference phasor, and since E_2 leads E_1 (i.e., its peak occurs at an earlier location) the angle α is positive. Thus φ is positive and the resultant E also leads E_1.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

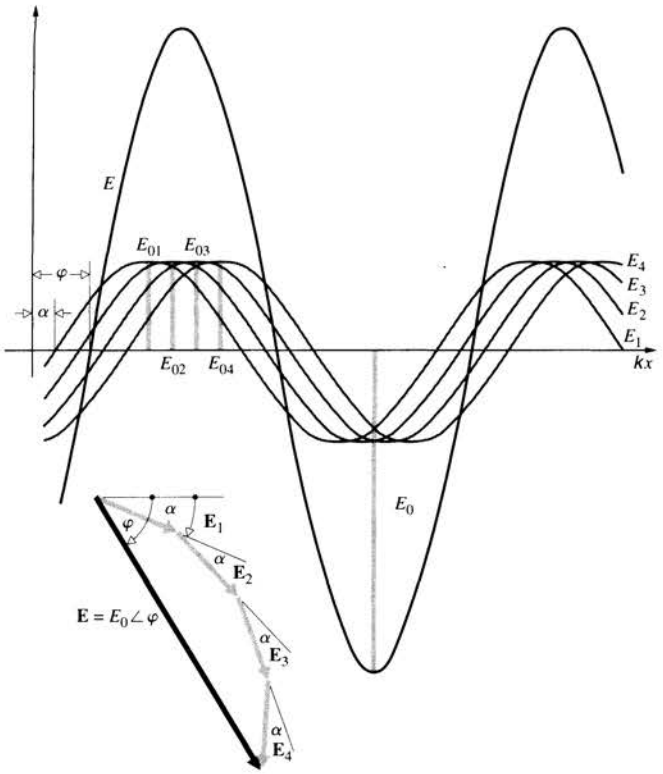


Figure 7.9 The summation of four sinusoidal waves of the same frequency. To further explore the phasor method, this time we'll take the origin as the zero of phase and reference everything with respect to it. The wave E_1 lags the origin by α; that is, the magnitude of the wave is zero at a larger value of kx than 0. Moreover, each wave lags the previous one by that same angle α. Accordingly, we draw the phasor $\mathbf{E}_1$ below the horizontal reference level and so lagging by α. All the other phasors successively lag one another by that same amount. Note that the length of the resultant phasor equals the amplitude of the resultant wave.

each wave (E_{01} or E_{02}) is the amplitude of the corresponding phasor. The length of the resultant phasor (E_0) equals the amplitude of the resultant wave, and its phase angle is a bit less than α. In a similar situation to that of Fig. 7.5, picture four equal-amplitude waves of the same frequency (Fig. 7.9), each shifted from the previous one by the same small amount α. The resultant phasor is $\mathbf{E} = E_0 \angle \varphi$, and it has the amplitude and phase of the resultant wave. That amplitude is substantial, but interestingly, if there were more constituent waves their phasors, tipped-to-tailed, would spiral around and E_0 would start to decrease. That's obvious in the phasor diagram, though it's not nearly so apparent in the wave representation.

7.1.4 Standing Waves

We saw earlier (p. 20) that the sum of solutions to the differential wave equation is itself a solution. Thus, in general,

$$\psi(x, t) = C_1 f(x - vt) + C_2 g(x + vt)$$

satisfies the differential wave equation. In particular let's examine *two harmonic waves of the same frequency propagating in opposite directions*. A situation of practical concern arises when the incident wave is reflected backward off some sort of mirror; a rigid wall will do for sound waves or a conducting sheet for electromagnetic waves. Imagine that an incoming wave traveling to the left,

$$E_I = E_{0I} \sin (kx + \omega t + \varepsilon_I) \tag{7.28}$$

strikes a mirror at $x = 0$ and is reflected to the right in the form

$$E_R = E_{0R} \sin (kx - \omega t + \varepsilon_R) \tag{7.29}$$

The composite wave in the region to the right of the mirror is $E = E_I + E_R$. In other words, the two waves (one traveling to the right, the other to the left) exist simultaneously in the region between the source and the mirror.

We could perform the indicated summation and arrive at a general solution* much like that of Section 7.1. However, some valuable physical insights can be gained by taking a slightly more restricted approach.

The initial phase ε_I may be set to zero by merely starting our clock at a time when $E_I = E_{0I} \sin kx$. Certain qualifications determined by the physical setup must be met by the mathematical solution, and these are known formally as ***boundary conditions***. For example, if we were talking about a rope with one end tied to a wall at $x = 0$, that point must always have a zero displacement. The two overlapping waves, one incident and the other reflected, would have to add in such a way as to yield a zero resultant wave at $x = 0$. Similarly, at the boundary of a perfectly conducting sheet, the resultant electromagnetic wave must have a zero electric-field component parallel to the surface. Assuming $E_{0I} = E_{0R}$, the boundary conditions require that at $x = 0$, $E = 0$, for all values of t and since $\varepsilon_I = 0$, it follows from Eqs. (7.28) and (7.29) that $\varepsilon_R = 0$. In other words, at $x = 0$, $E_I = E_{0I} \sin (+\omega t)$ and $E_R = E_{0I} \sin (-\omega t)$; the two are 180° out-of-phase, $E_I = -E_R$, and they

*See, for example, J. M. Pearson, *A Theory of Waves*.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

cancel at any time t. The composite disturbance is then

$$E = E_{0l} [\sin (kx + \omega t) + \sin (kx - \omega t)]$$

Applying the identity

$$\sin \alpha + \sin \beta = 2 \sin \tfrac{1}{2}(\alpha + \beta) \cos \tfrac{1}{2}(\alpha - \beta)$$

yields

$$E(x, t) = 2E_{0l} \sin kx \cos \omega t \qquad (7.30)$$

This is the equation for a **standing** or **stationary wave**, as opposed to a traveling wave (Fig. 7.10). Its profile does not move through space; it is clearly not of the form $f(x \pm vt)$. At any point $x = x'$, the amplitude is a constant equal to $2E_{0l} \sin kx'$, and $E(x', t)$ varies harmonically as $\cos \omega t$. At certain points, namely, $x = 0, \lambda/2, \lambda, 3\lambda/2, \ldots$, the disturbance will be zero at all times. These are known as **nodes** or *nodal points* (Fig. 7.11). Halfway between each adjacent node, that is, at $x = \lambda/4, 3\lambda/4, 5\lambda/4, \ldots$, the amplitude has a maximum value of $\pm 2E_{0l}$, and these points are known as the **antinodes**. The disturbance $E(x, t)$ will be zero at all values of x whenever $\cos \omega t = 0$, that is, when $t = (2m + 1)\tau/4$, where $m = 0, 1, 2, 3, \ldots$ and τ is the period of the component waves.

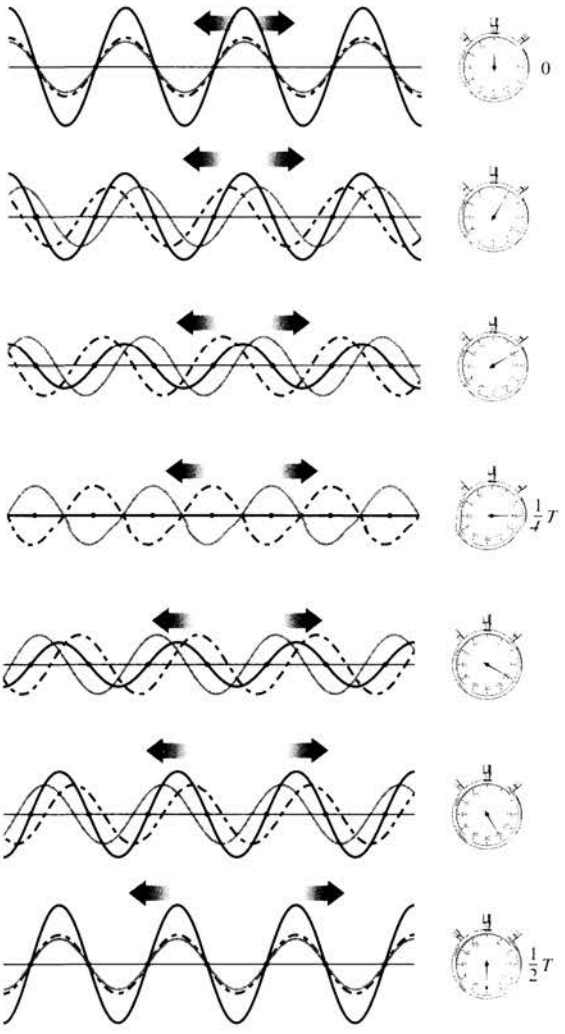


Figure 7.10 The creation of standing waves. Two waves of the same amplitude and wavelength traveling in opposite directions form a stationary disturbance that oscillates in place.

Figure 7.12 illustrates how the standing-wave pattern is generated from a phasor perspective. There are two harmonic waves, so begin with two phasors $\mathbf{E}_1$ and $\mathbf{E}_2$. The waves are 180° out-of-phase at the boundary $x = 0$, and hence the two phasors must have initial values of $E_{01}\angle 0$ and $E_{02}\angle \pi$. Earlier (Section 2.6) we saw that a phasor rotating counterclockwise at a rate ω is equivalent to a wave traveling to the left (decreasing x), and similarly, one rotating clockwise corresponds to a wave traveling to the right (increasing x). Let the phasors $\mathbf{E}_1$ represent a wave traveling to the left, and $\mathbf{E}_2$ a wave to the right. The resultant phasor is $\mathbf{E}_1 + \mathbf{E}_2 = \mathbf{E} = E_0\angle\varphi$, where E_0 is the amplitude of the curve (i.e., the resultant disturbance) at any chosen moment. It's gotten by tip-to-tailing $\mathbf{E}_1$ and $\mathbf{E}_2$. Wanting to reproduce Fig. 7.11, let the two waves have the same amplitude, $E_{01} = E_{02}$. Keeping the two phasors tip-to-tail and having $\mathbf{E}_1$ rotate counterclockwise as $\mathbf{E}_2$ rotates (at the same rate) clockwise generates $\mathbf{E}$ as a function of t. Notice that the triangles formed by the three phasors are always isosceles, with $\mathbf{E}$ being vertical. It doesn't rotate at all, and the resultant wave it represents doesn't progress through space—it's a standing wave.

Going back to Fig. 7.10, if the reflection is not perfect, as is often the case, the composite disturbance will not have zero amplitude at the nodes (Fig. 7.13). That's most easily seen with the phasors $\mathbf{E}_1$ and $\mathbf{E}_2$ where this time $E_{01} > E_{02}$. Now $\mathbf{E}$ rotates in the same direction (counterclockwise) as the larger of the two component phasors, namely, $\mathbf{E}_1$. The composite wave contains a traveling component along with the stationary wave and that component moves off to the left with an amplitude equal to $E_{01} - E_{02}$. Under such conditions there will be a net transfer of energy, whereas for the pure standing wave there is none. It's also possible to write an expression for the resultant partial standing wave in the form $E = E_0(x) \cos [\omega t - \varphi(x)]$ *where the amplitude varies from point to point,* while at each value of x the wave oscillates in time cosinu-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

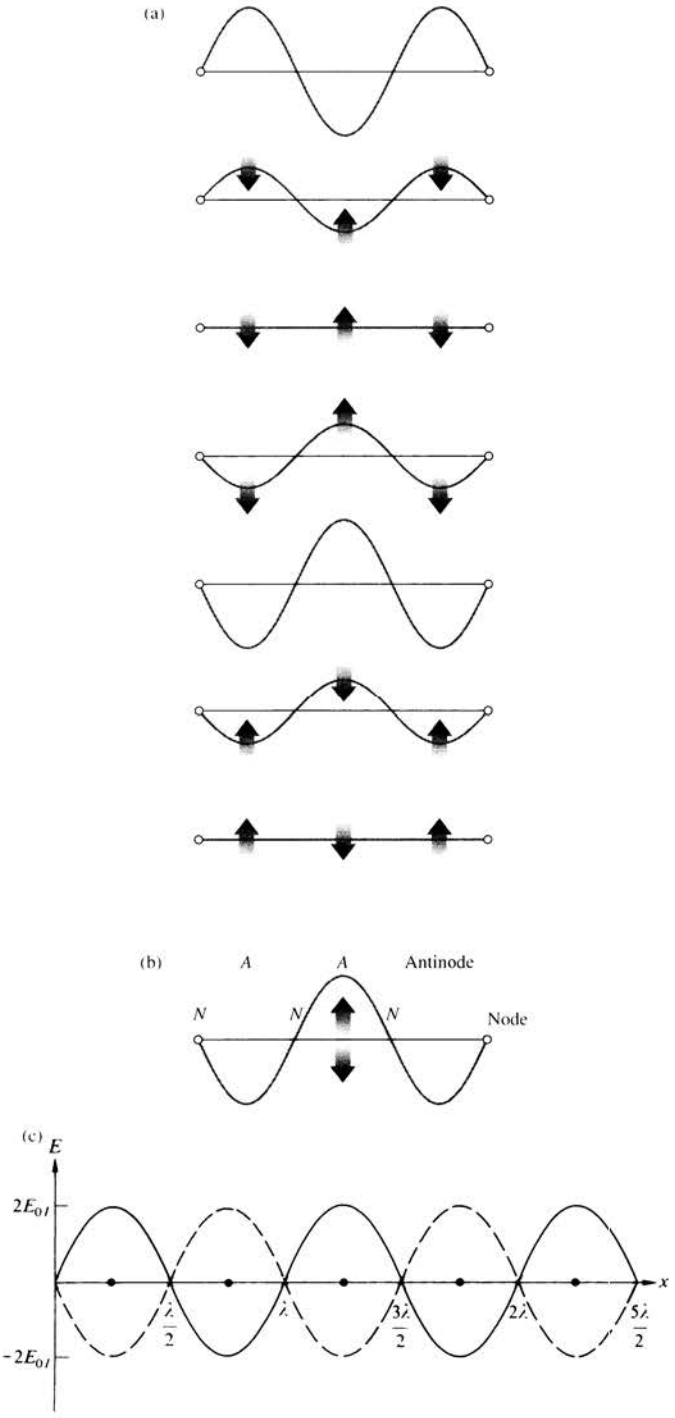
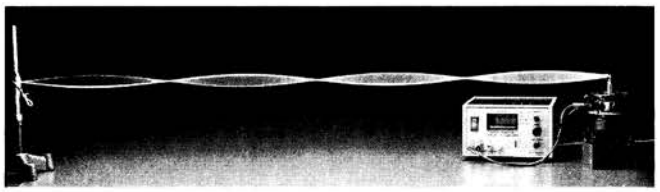


Figure 7.11 A standing wave at various times.



Standing waves on a vibrating string. (Photo courtesy PASCO.)

soidally. You can see from the phasor diagram (Fig. 7.13*b*), applying the Law of Cosines, that the position-dependent amplitude is $E_0(x) = (E_{01}^2 + E_{02}^2 + 2E_{01}E_{02}\cos 2kx)^{1/2}$.

Although the analysis carried out above is essentially one-dimensional, standing waves exist in two and three dimensions as well. The phenomenon is extremely commonplace: standing waves occur in one dimension on guitar strings and diving boards, in two dimensions on the surface of a drum or in a jiggled pail of water (see photo), and in three dimensions when you sing in a shower stall. In fact, standing waves are created within the cavities inside your head whenever you sing, no matter where you are.

If a standing-wave system is driven by an oscillating source, it will efficiently absorb energy provided that the vibrations match one of its standing-wave modes. That process is known as resonance, and it happens every time your house buzzes when an airplane flies low overhead or when a heavy truck passes by. If the source continues to supply ener-



A pail used to wash a floor contained a suspension of fine dirt particles in water. When placed in a curved sink, the pail gently rocked along a fixed axis, setting up standing waves and distributing the particles in ridges as they settled. (Photo by E.H.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

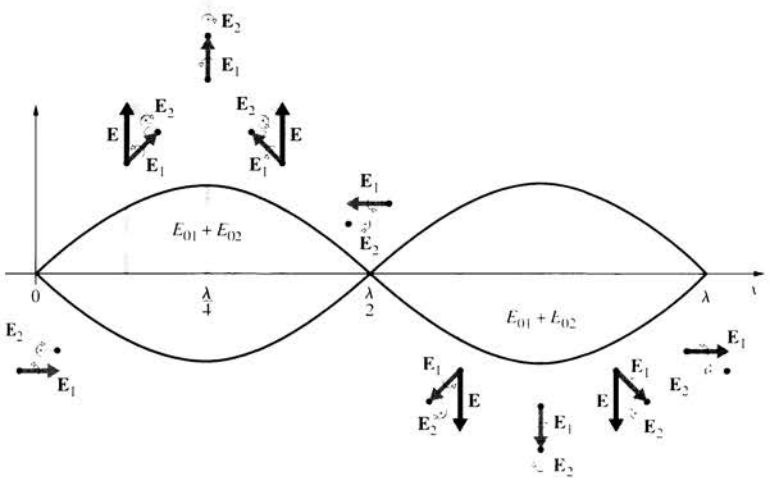


Figure 7.12 The generation of a standing wave from the perspective of phasor addition. The two phasors rotate at the same rate but in opposite directions. Here both waves have the same amplitudes, and that produces complete cancelation at the nodes.

gy, the wave will continue to build until the system's inherent losses equal the energy input and equilibrium is reached. This ability to sustain and simplify an input is an extremely important feature of standing-wave systems. The ear's auditory canal is just such a resonant cavity. It amplifies (by about 100%) sounds in the range from ≈3 kHz to ≈4 kHz. Similarly, the laser builds its powerful emission within a standing-



A standing-wave pattern on the side of a car due to vibrations caused by its running engine. The scale is in microns, where 1 μm = 10 $^{-6}$ m. The photo was made using a holographic technique. (Photo courtesy HOLO3/FTPO.)

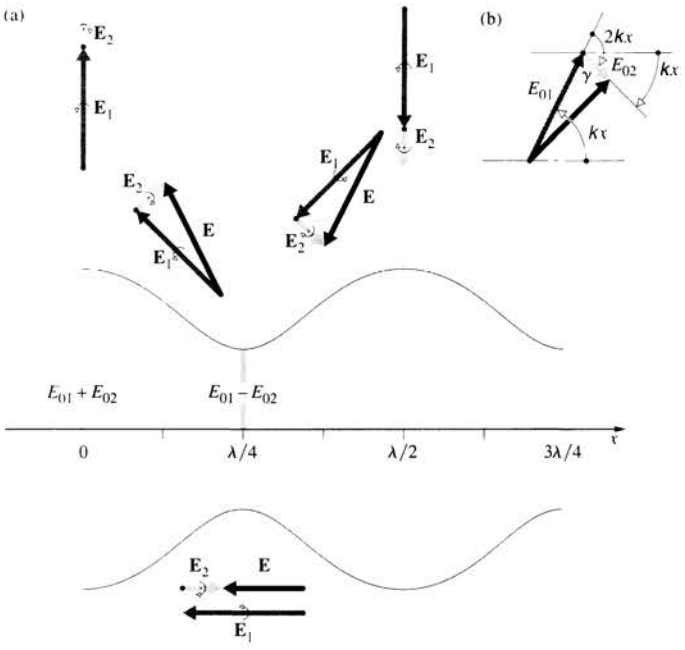


Figure 7.13 (*a*) The generation of a partial standing wave from the perspective of phasor addition. This time the two waves have different amplitudes, and that produces nonzero nodes. Accordingly, the disturbance has a traveling wave component that propagates in the direction of the larger constituent wave. (*b*) The wave can be written in the form $E = E_0(x) \cos [\omega t - \varphi(x)]$. Applying the Law of Cosines, the amplitude is given by $E_0(x) = (E_{01}^2 + E_{02}^2 + 2E_{01}E_{02} \cos 2kx)^{1/2}$.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

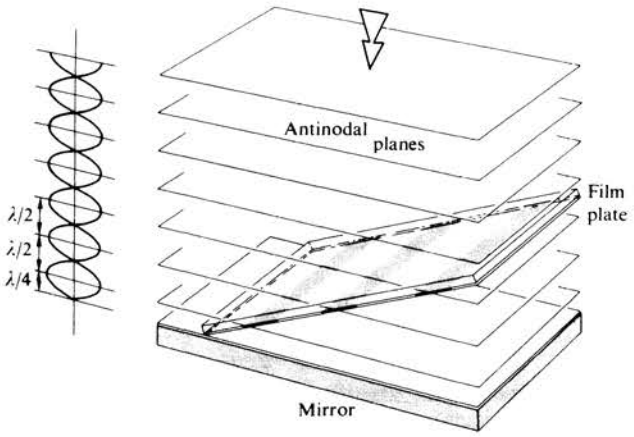
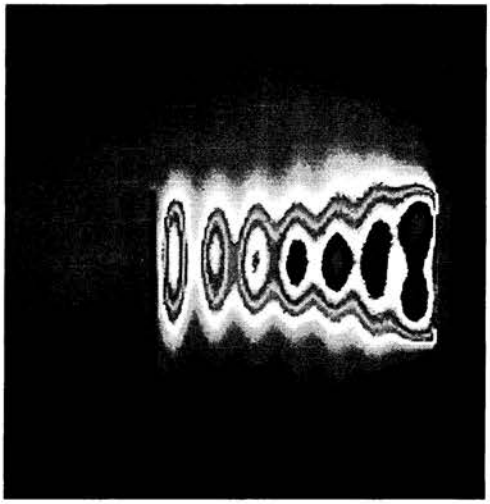


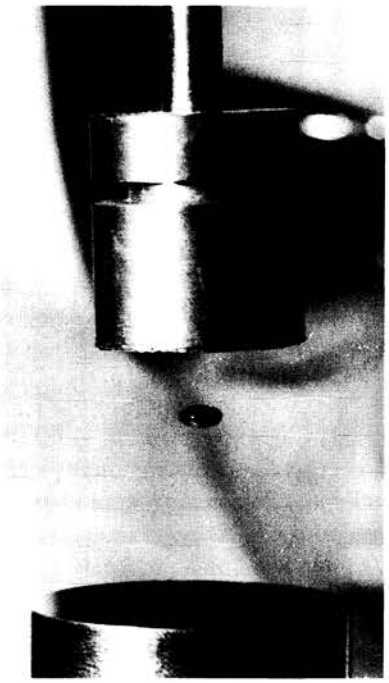
Figure 7.14 Wiener's experiment.

wave cavity (p. 591).

It was by measuring the distances between the nodes of standing waves that Hertz was able to determine the wavelength of the radiation in his historic experiments (see Section 3.6.1). A few years later, in 1890, Otto Wiener first demonstrated the existence of standing lightwaves. The arrangement he used is depicted in Fig. 7.14. It shows a normally incident parallel beam of quasimonochromatic light reflecting off a front-silvered mirror. A transparent photographic film, less than $\lambda/20$ thick, deposited on a glass plate, was inclined to the mirror at an angle of about 10^{-3} radians. In that way the film plate cut across the pattern of standing plane waves. After developing the emulsion, it was found to be blackened along a series of equidistant parallel bands. These corresponded to the regions where the photographic layer had intersected the antinodal planes. Significantly, there was no blackening of the emulsion at the mirror's surface. It can be shown that the nodes and antinodes of the magnetic field component of an electromagnetic standing wave alternate with those of the electric field (Problem 7.11). We might suspect as much from the fact that at $t = (2m + 1)\tau/4$, $E = 0$ for all values of x, so to conserve energy it follows that $B \neq 0$. In agreement with theory, Hertz had previously (1888) determined the existence of a nodal point of the electric field at the surface of his reflector. Accordingly, Wiener could conclude that the blackened

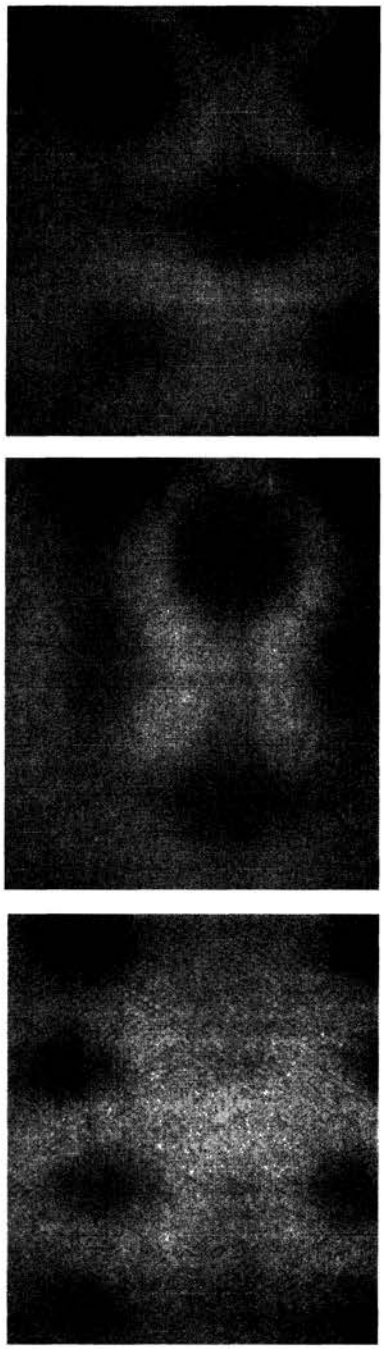


A two-dimensional standing-wave pattern formed between a source and a reflector. EM-waves from a 3.9-GHz antenna enter from the right. They reflect off a metal rod and travel back to the antenna. The pattern is made visible by absorbing the microwave radiation and recording the resulting temperature distribution with an IR camera. (Photo courtesy H. H. Pohle, Phillips Laboratory, Kirtland Air Force Base.)



Ultrasonic levitation. Ultrasonic waves, one traveling up, the other down form a standing-wave pattern. Here a droplet of water is suspended in a nodal region. (Photo courtesy NASA.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



Slices of the three-dimensional electromagnetic standing-wave pattern at different heights inside a microwave oven. (Photos courtesy Alistair Steyn-Ross, University of Waikato.)

regions were associated with antinodes of the $\vec{\mathbf{E}}$-field. ***It is the electric field that triggers the photochemical process.***

In a similar way Drude and Nernst showed that the $\vec{\mathbf{E}}$-field is responsible for fluorescence. These observations are all quite understandable, since the force exerted on an electron by the $\vec{\mathbf{B}}$-field component of an electromagnetic wave is generally negligible in comparison to that of the $\vec{\mathbf{E}}$-field. It is for these reasons that the electric field is referred to as the *optic disturbance* or *light field.*

Standing waves generated by two oppositely propagating disturbances represent a special case of the broader subject of double-beam interference (p. 389). Consider the two point sources sending out waves in Fig. 7.15. When point P, the point of observation somewhere near the middle is far from the sources, angle ϕ is small, the two waves superimpose, and there results a complicated interference pattern (that will be treated in detail in Chapter 9). Suffice it to say here that the space surrounding the sources will be filled with a system of bright and dark bands where the interference is alternately constructive and destructive. As P comes closer and ϕ gets larger, the fringes become finer, that is, narrower, until P is on the line joining the sources and $\phi = 180°$. Then standing waves are set up, and the "fringes" are the finest they'll get, namely, half a wavelength peak-to-peak.

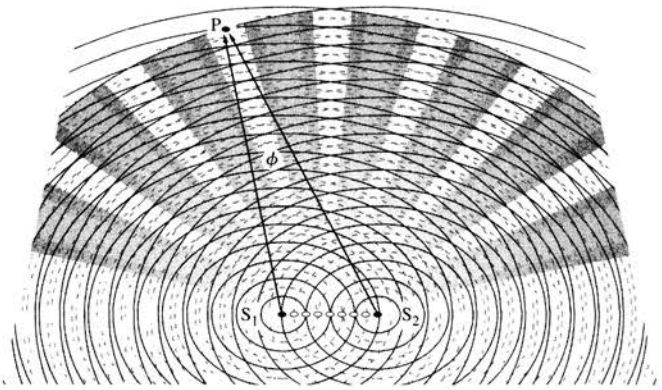


Figure 7.15 Two monochromatic point sources. At any point P the resultant wave is maximum where peak (—) overlaps peak (—) or trough (– –) overlaps trough (– –). It's minimum where peak overlaps trough. The maxima that form along the $\overline{S_1S_2}$ line correspond to standing waves.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

7.2 The Addition of Waves of Different Frequency

Thus far the analysis has been restricted to the superposition of waves all having the same frequency. Yet one never actually has disturbances, of any kind, that are strictly monochromatic. It will be far more realistic, as we shall see, to speak of quasi-monochromatic light, which is composed of a narrow range of frequencies. The study of such light will lead us to the important concepts of bandwidth and coherence time.

The ability to modulate light effectively (Section 8.11.3) makes it possible to couple electronic and optical systems in a way that has had and will certainly continue to have far-reaching effects on the entire technology. Moreover, with the advent of electro-optical techniques, light has taken on a significant role as a carrier of information. This section is devoted to developing some of the mathematical ideas needed to appreciate this new emphasis.

7.2.1 Beats

We'll start with the especially simple case of two waves of different frequency traveling in the same direction. Consider the composite disturbance arising from a combination of the waves

$$E_1 = E_{01} \cos (k_1 x - \omega_1 t)$$

and

$$E_2 = E_{01} \cos (k_2 x - \omega_2 t)$$

where $k_1 > k_2$ and $\omega_1 > \omega_2$. These waves have equal amplitudes and zero initial phase angles. The net composite wave

$$E = E_{01}[\cos (k_1 x - \omega_1 t) + \cos (k_2 x - \omega_2 t)]$$

can be reformulated as

$$E = 2E_{01} \cos \tfrac{1}{2}[(k_1 + k_2)x - (\omega_1 + \omega_2)t] \times \cos \tfrac{1}{2}[(k_1 - k_2)x - (\omega_1 - \omega_2)t]$$

using the identity

$$\cos \alpha + \cos \beta = 2 \cos \tfrac{1}{2}(\alpha + \beta) \cos \tfrac{1}{2}(\alpha - \beta)$$

Now define the quantities $\bar{\omega}$ and $\bar{k}$, which are the ***average angular frequency*** and ***average propagation number***, respectively. Similarly, the quantities ω_m and k_m are designated the ***modulation frequency*** and ***modulation propagation number***, respectively. Let

$$\bar{\omega} \equiv \tfrac{1}{2}(\omega_1 + \omega_2) \qquad \omega_m \equiv \tfrac{1}{2}(\omega_1 - \omega_2) \tag{7.31}$$

and

$$\bar{k} \equiv \tfrac{1}{2}(k_1 + k_2) \qquad k_m \equiv \tfrac{1}{2}(k_1 - k_2) \tag{7.32}$$

thus

$$E = 2E_{01} \cos (k_m x - \omega_m t) \cos (\bar{k}x - \bar{\omega}t) \tag{7.33}$$

The total disturbance may be regarded as a traveling wave of frequency $\bar{\omega}$ having a time-varying or modulated amplitude $E_0(x, t)$ such that

$$E(x, t) = E_0(x, t) \cos (\bar{k}x - \bar{\omega}t) \tag{7.34}$$

where

$$E_0(x, t) = 2E_{01} \cos (k_m x - \omega_m t) \tag{7.35}$$

In applications of interest here, ω_1 and ω_2 will always be rather large. In addition, if they are comparable to each other, $\omega_1 \approx \omega_2$, then $\bar{\omega} >> \omega_m$ and $E_0(x, t)$ will change slowly, whereas $E(x, t)$ will vary quite rapidly (Fig. 7.16). The irradiance is proportional to

$$E_0^2(x, t) = 4E_{01}^2 \cos^2 (k_m x - \omega_m t)$$

or

$$E_0^2(x, t) = 2E_{01}^2[1 + \cos (2k_m x - 2\omega_m t)]$$

Notice that $E_0^2(x, t)$ oscillates about a value of $2E_{01}^2$ with an angular frequency of $2\omega_m$ or simply $(\omega_1 - \omega_2)$, which is known as the **beat frequency**. That is, E_0 varies at the modulation frequency, whereas E_0^2 varies at twice that, namely, the beat frequency.

When the two overlapping harmonic waves have different amplitudes, they still produce beats, but the cancelation is incomplete—there's less contrast. Figure 7.17 depicts such a pattern and indicates how the two phasors $\mathbf{E}_1$ and $\mathbf{E}_2$ give rise to it. Remember that the resultant phasor $\mathbf{E} = E_0(x, t)\angle\varphi$ provides the **amplitude** and **relative phase** of the composite disturbance. The slowly oscillating envelope is a plot of $E_0(x, t)$ as it changes in time. The resultant phasor does not give us the instantaneous magnitude of the oscillating carrier.

The two waves travel in the same direction, and consequently their phasors rotate the same way—one at ω_1 and the other at ω_2. Rather than letting both phasors revolve around at their different frequencies, we can simplify things a little. Sup-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

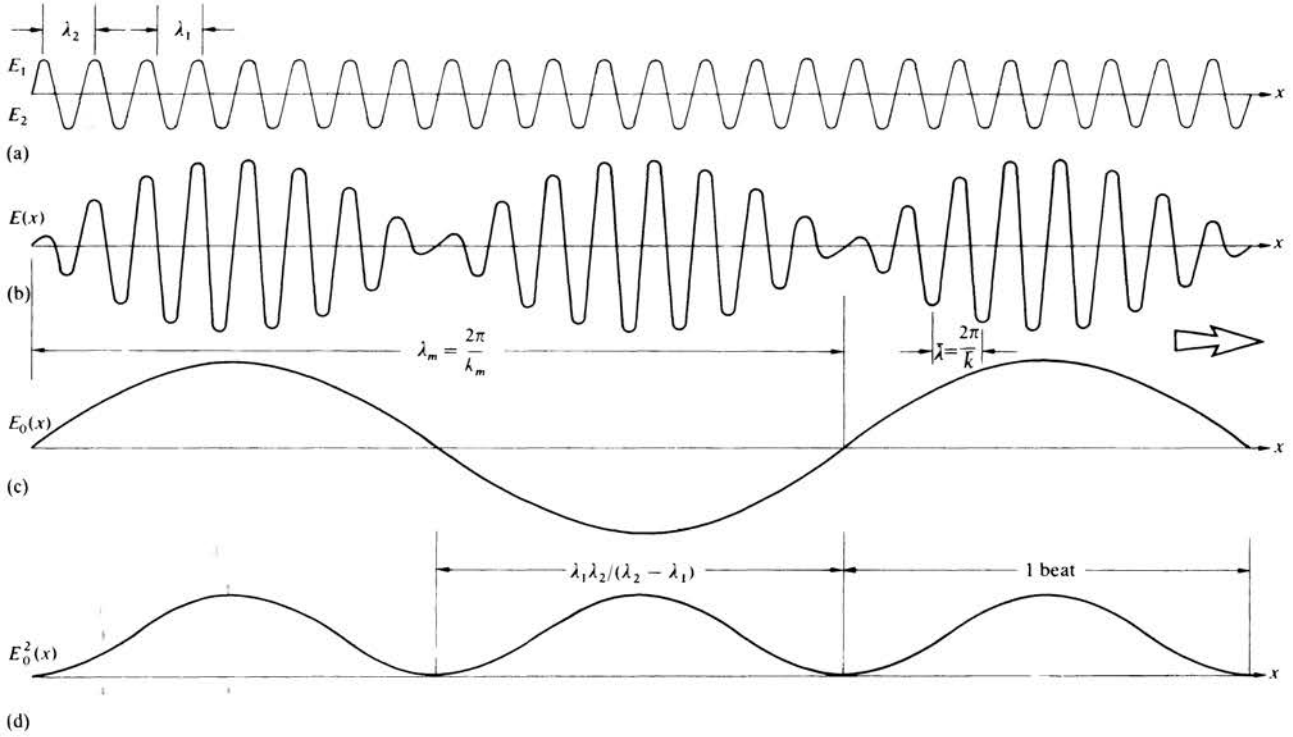


Figure 7.16 The superposition of two equal-amplitude harmonic waves of different frequency producing a beat pattern.

pose that $\omega_1 > \omega_2$ and place the higher-frequency phasor $\mathbf{E}_1$ at the tip of the lower-frequency one $\mathbf{E}_2$ (Fig. 7.17*b*). Redraw $\mathbf{E}_2$ fixed along the horizontal zero-phase reference line. The angle α that $\mathbf{E}_1$ makes (Fig. 7.17*c*) with the horizontal (i.e., with $\mathbf{E}_2$) at any instant is its phase with respect to $\mathbf{E}_2$, and so $\mathbf{E}_1$ rotates at a rate $(\omega_1 - \omega_2)$ and $\alpha = (\omega_1 - \omega_2)t$. The amplitude $E_0(x, t)$ of the resultant (the envelope of the carrier) oscillates between values of $E_{02} + E_{01}$ and $E_{02} - E_{01}$. The angle that $\mathbf{E}$ makes with the horizontal (φ) is the phase of the resultant wave with respect to $\mathbf{E}_2$, and it gradually oscillates as $\mathbf{E}_1$ rotates around in a circle.

Note that in the case of Fig. 7.16, where $E_{01} = E_{02}$, E_0 oscillates between 0 and $2E_{01}$. Moreover, $2\varphi = \alpha$, and the resultant phasor $\mathbf{E}$—which corresponds to the amplitude of the disturbance—rotates at a rate $\omega_m = \frac{1}{2}(\omega_1 - \omega_2)$, all of which agrees with Eq. (7.33).

Beats are commonplace in sound: piano tuners have always done their work beating the notes of vibrating strings against tuning forks. But the effect was only first observed with light in 1955 by Forrester, Gudmundsen, and Johnson.* To obtain two waves of slightly different frequency they used the Zeeman Effect. When the atoms of a discharge lamp, in this case mercury, are subjected to a magnetic field, their energy levels split. As a result, the emitted light contains two frequency components, ν_1 and ν_2, which differ in proportion to the magnitude of the applied field. When these components are recombined at the surface of a photoelectric mixing tube, the beat frequency, $\nu_1 - \nu_2$, is generated. Specifically, the field was adjusted so that $\nu_1 - \nu_2 = 10^{10}$ Hz, which conveniently corresponds to a 3-cm microwave signal. The recorded photoelectric current had the same form as the $E_0^2(x)$ curve in Fig. 7.16*d*.

The advent of the laser has since made the observation of beats using light considerably easier. Even a beat frequency of a few Hz out of 10^{14} Hz can be seen as a variation in phototube

*A. T. Forrester, R. A. Gudmundsen, and P. O. Johnson, "Photo-electric Mixing of Incoherent Light." *Phys. Rev.* **99**, 1691 (1955).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

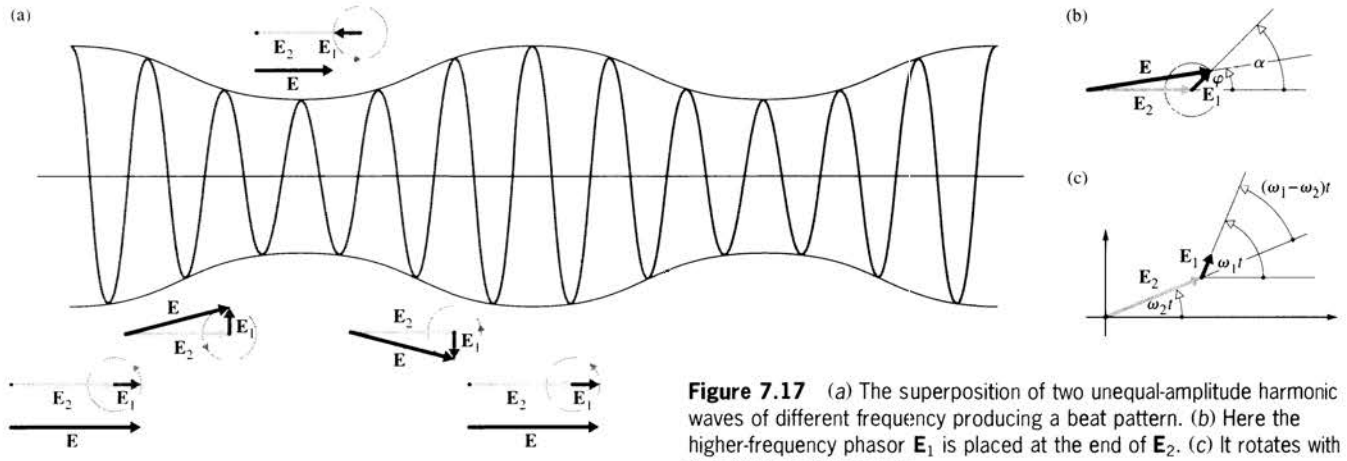


Figure 7.17 (*a*) The superposition of two unequal-amplitude harmonic waves of different frequency producing a beat pattern. (*b*) Here the higher-frequency phasor $\mathbf{E}_1$ is placed at the end of $\mathbf{E}_2$. (*c*) It rotates with the difference frequency.

current. The observation of beats now represents a particularly sensitive and fairly simple means of detecting small frequency differences. The ring laser (Section 9.8.3), functioning as a gyroscope, utilizes beats to measure frequency differences induced as a result of the rotation of the system. The Doppler Effect, which accounts for the frequency shift when light is reflected off a moving surface, provides another series of applications of beats. By scattering light off a target, whether solid, liquid, or even gaseous, and then beating the original and reflected waves, we get a precise measure of the target speed. In much the same way on an atomic scale, laser light will shift in phase upon interacting with sound waves moving in a material. (This phenomenon is called Brillouin Scattering.) Thus $2\omega_m$ becomes a measure of the speed of sound in the medium.

7.2.2 Group Velocity

The specific relationship between ω and k determines v, the phase velocity of a wave. In a nondispersive medium, and vacuum is the only truly nondispersive environment, $v = \omega/k$ [Eq. (2.33)], and a plot of ω versus k is a straight line. The frequency and wavelength change so as to keep v constant. All waves of a particular type (e.g., all EM-waves) travel with the same phase speed in a nondispersive medium. By contrast, in a dispersive medium (anything other than vacuum) every wave propagates at a speed that depends on its frequency.

When a number of different-frequency harmonic waves superimpose to form a composite disturbance, the resulting modulation envelope will travel at a speed different from that of the constituent waves. This raises the important notion of the ***group velocity*** and its relationship to the phase velocity. The concept was first put forward (1839) by the great Irish physicist and mathematician Sir William Rowan Hamilton, though it got little attention until Stokes reintroduced it in 1876 in the context of hydrodynamics. Assuming we can recognize some constant feature in the shape of a pulse, like its leading edge, we'll take the rate at which that feature moves to be the velocity of the group of waves as a whole.

The disturbance examined in the previous section,

$$E(x, t) = E_0(x, t) \cos(\bar{k}x - \bar{\omega}t) \qquad [7.34]$$

consists of a high-frequency ($\bar{\omega}$) ***carrier wave***, *amplitude-modulated* by a cosine function. Suppose, for a moment, that the wave in Fig. 7.16*b* was not modulated; that is, E_0 = constant. Each small peak in the carrier would travel to the right with the usual phase velocity. In other words,

$$v = -\frac{(\partial\varphi/\partial t)_x}{(\partial\varphi/\partial x)_t} \qquad [2.32]$$

From Eq. (7.34) the phase is given by $\varphi = (\bar{k}x - \bar{\omega}t)$, hence

$$v = \bar{\omega}/\bar{k} \qquad (7.36)$$

This is the phase velocity of the carrier, whether it's modulated or not. In the former case, the peaks simply change amplitude periodically as they stream along.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Evidently, there is another motion to be concerned with, and that's the propagation of the modulation envelope. Return to Fig. 7.16*a* and suppose that the constituent waves, $E_1(x, t)$ and $E_2(x, t)$, advance with the same speed, $v_1 = v_2$. Imagine, if you will, the two harmonic functions having different wavelengths and frequencies drawn on separate sheets of clear plastic. When these are overlaid in some way [as in Fig. 7.16*a*], the resultant is a stationary beat pattern. If the sheets are both moved to the right at the same speed so as to resemble traveling waves, the beats will obviously move with that same speed. *The rate at which the modulation envelope advances* is known as the **group velocity**, or v_g. In this instance, the group velocity equals the phase velocity of the carrier (the average speed, $\bar{\omega}/\bar{k}$). In other words, $v_g = v = v_1 = v_2$. This applies specifically to nondispersive media in which the phase velocity is independent of wavelength so that the two waves could have the same speed.

For a more generally applicable solution, examine the expression for the modulation envelope:

$$E_0(x, t) = 2E_{01} \cos (k_m x - \omega_m t) \qquad [7.35]$$

The speed with which that wave moves is again given by Eq. (2.32), but now we can forget the carrier wave. The modulation advances at a rate dependent on the phase of the envelope $(k_m x - \omega_m t)$, and

$$v_g = \frac{\omega_m}{k_m}$$

or

$$v_g = \frac{\omega_1 - \omega_2}{k_1 - k_2} = \frac{\Delta\omega}{\Delta k}$$

Recall that in ordinary media ω is dependent on λ, or equivalently on k. The particular function $\omega = \omega(k)$ is called a *dispersion relation*. When the frequency range $\Delta\omega$, centered about $\bar{\omega}$, is small, $\Delta\omega/\Delta k$ is approximately equal to the derivative of the dispersion relation evaluated at $\bar{\omega}$; that is,

$$v_g = \left(\frac{d\omega}{dk}\right)_{\bar{\omega}} \qquad (7.37)$$

The modulation or signal propagates at a speed v_g that may be greater than, equal to, or less than v, the phase velocity of the carrier.

Strictly speaking, any real wave is finite in spatial extent: it's turned on (or received) at some specific time and, presumably, shut off at some later time. A real wave is therefore actually a pulse, though it could be a rather long one. As we're about to learn (p. 308), any such pulse is identical to a superposition of numerous different-frequency sine waves (often called Fourier components), each with a specific amplitude and phase. Accordingly, envision not just two constituent waves as in Fig (7.16), but upwards of a thousand, all with different frequencies. If, as is certainly possible, the sinusoids cancel each other everywhere except over a region where they are in-phase, or nearly so, the resulting disturbance will resemble a localized pulse, often called a *wave packet* (Fig. 7.18) to remind us that it's just that. Here again it's natural to think about the group velocity. Equation (7.37) will be true, more or less, for any collection of overlapping sinusoidal waves, as long as Δk, their range of values of k, is narrow. As we'll learn, a narrow range of k (or equivalently of λ) means that we

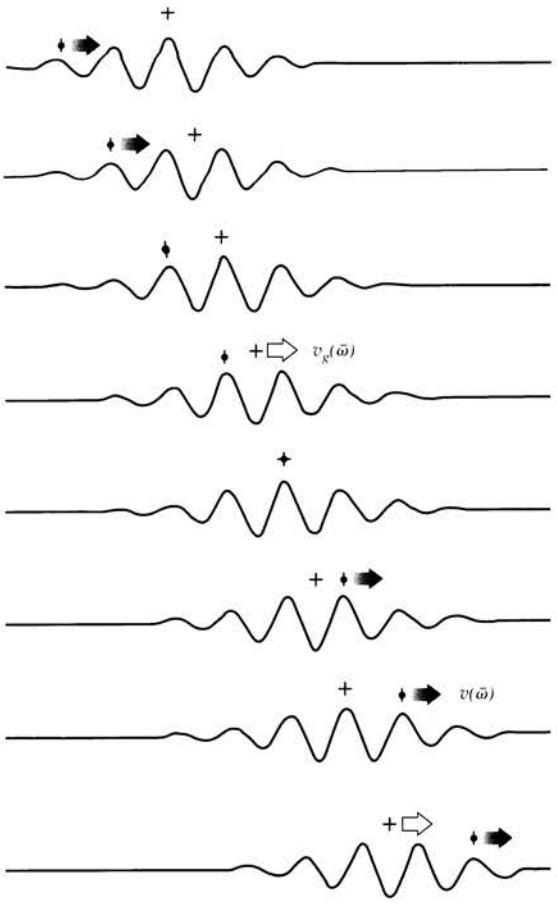


Figure 7.18 A wave pulse in a dispersive medium.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

have a wide packet in space. By contrast, if the pulse is narrow in space, there will be a large number of sinusoidal components present that have a correspondingly wide range of k values. Since each wavelength component travels at a different phase velocity in a dispersive medium, such a pulse would change shape as it moved along, making v_g a less than precise concept to deal with experimentally.

Recall that a typical medium in the vicinity of a resonance (ν_0) has an $n(\nu)$-versus-ν curve resembling Fig. 7.19. Radiant energy corresponding in frequency to the central region, where the slope of the curve is negative, is very strongly absorbed, and so this is called the *absorption band*. On either side of it, $n(\nu)$ increases with increasing ν, and this is the domain of *normal dispersion*. Inside the absorption band, $n(\nu)$ decreases with increasing ν, and this is the domain of *anomalous dispersion*.

A plot of the dispersion relation (Fig. 7.20) produces a curve passing through the origin that is convex upward for normal dispersion and concave downward for anomalous dispersion (p. 67). In either case, the slope of a line drawn from the origin to any point (ω, k) on the curve is the phase velocity at that frequency. Similarly, the slope of the curve at the point $(\bar{\omega}, \bar{k})$ is $(d\omega/dk)_{\bar{\omega}}$, and that's the group velocity for the set of component waves centered at $\bar{\omega}$. In normal dispersion, sinusoidal waves of high frequency (e.g., blue) have larger indices and travel slower than low-frequency waves (e.g., red).

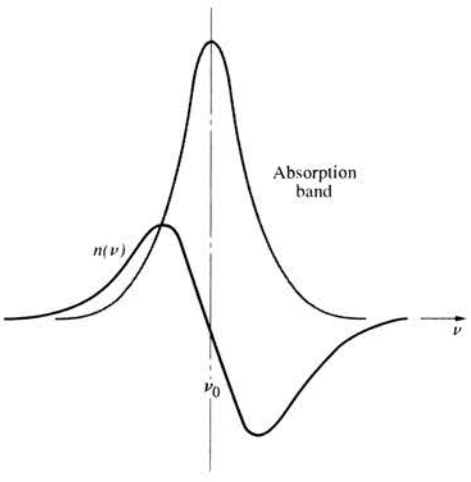


Figure 7.19 A typical representation of the frequency dependence of the index of refraction in the vicinity of an atomic resonance. Also shown is the absorption curve centered on the resonant frequency.

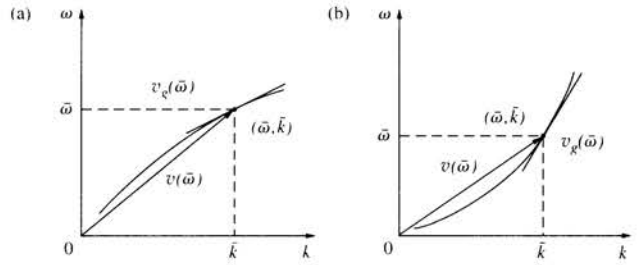


Figure 7.20 A plot of the dispersion relation. (*a*) In normal dispersion $v(\bar{\omega}) > v_g(\bar{\omega})$ whereas in (*b*) anomalous dispersion $v_g(\bar{\omega}) > v(\bar{\omega})$. The phase velocity v, of a wave of any frequency ω, is the slope of the line drawn from the origin to that point. The group velocity is the slope of the tangent to the curve at $(\bar{\omega}, \bar{k})$ where $\bar{\omega}$ is the mean frequency of the waves composing the group.

Moreover, the slope of the dispersion curve (v_g) is always less than the slope of the line (v); that is, $v_g < v$, whereas in anomalous dispersion $v_g > v$.

Since $\omega = kv$, Eq. (7.37) yields

$$v_g = v + k\frac{dv}{dk} \tag{7.38}$$

As a consequence, in an idealized nondispersive medium in which v is independent of λ, $dv/dk = 0$ and $v_g = v$. Specifically, in vacuum $\omega = kc$, $v = c$, and $v_g = c$.

Real substantial media are all more or less dispersive ($v_1 \neq v_2$, as is the case of Fig. 7.21). Given that $n(k)$ is known, $\omega = kc/n$, and it's then useful to reformulate v_g as

$$v_g = \frac{c}{n} - \frac{kc}{n^2}\frac{dn}{dk}$$

or

$$v_g = v\left(1 - \frac{k}{n}\frac{dn}{dk}\right) \tag{7.39}$$

For optical media, in regions of normal dispersion, the refractive index increases with frequency ($dn/dk > 0$), and as a result $v_g < v$. Clearly, one should also define a ***group index of refraction***

$$n_g \equiv c/v_g \tag{7.40}$$

which must be carefully distinguished from n. In 1885 A. A. Michelson measured n_g in carbon disulfide using pulses of white light and obtained 1.758 in comparison to $n = 1.635$.

Return to Fig. 7.18 where the medium is normally disper-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

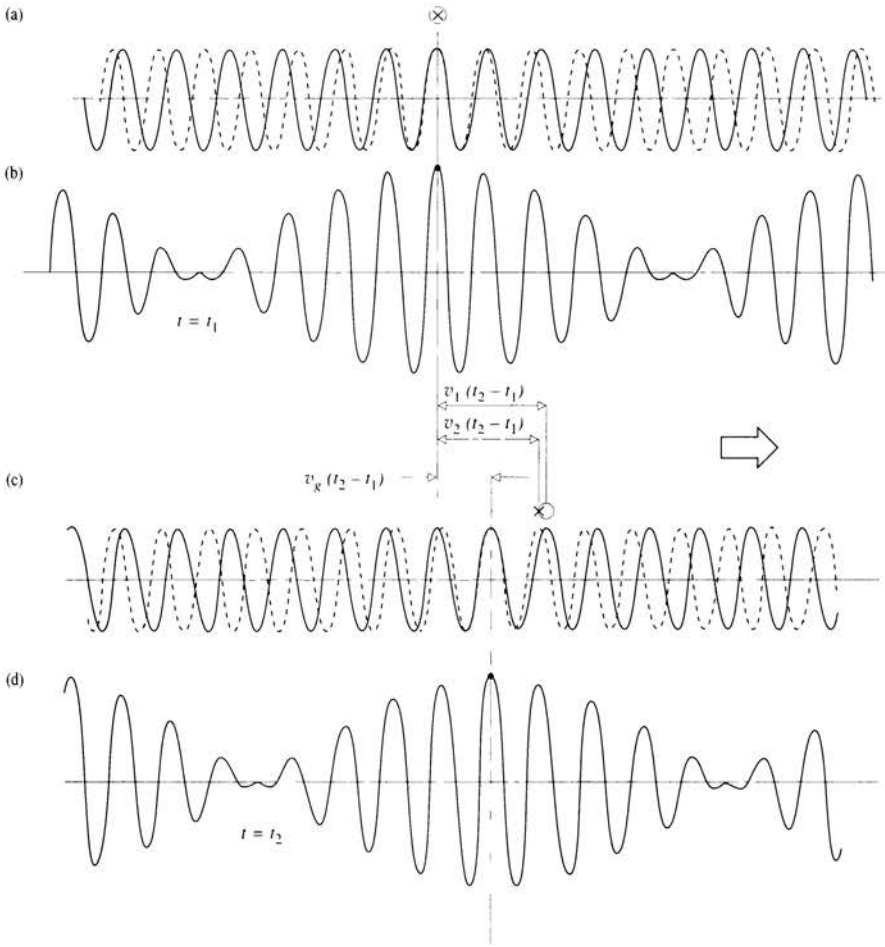


Figure 7.21 Group and phase velocities. In (*a*) the two waves coincide at the point indicated by an ⊗. And in (*b*) the peak of the modulated wave occurs at that point. But the waves travel at different speeds in (*c*) and the two original peaks (marked by x and ○) separate. A different pair now coincide in (*d*) to form the high point of the modulated wave, which therefore travels at yet a different speed. Here $v_1 > v_2 > v_g$, and since $\lambda_1 > \lambda_2$ this is a case of normal dispersion.

sive and the phase velocity will be taken as the velocity of the carrier, that is, of the roughly sinusoidal wave of frequency $\bar{\omega}$. Because the peaks of the carrier travel faster than does the pulse as a whole, they appear to enter it at the left, sweep through it, and vanish off at the right. Although each peak of the carrier changes height as it progresses across the pulse, $v(\bar{\omega})$ is the speed of any such peak and it's therefore properly the speed of the condition of constant phase. By contrast, the modulation envelope travels at a speed $v_g(\bar{\omega}) = (d\omega/dk)_{\bar{\omega}}$, which in this particular instance equals one quarter of $v(\bar{\omega})$. Any point on the envelope (e.g., the maximum at the center of the pulse) moves at a speed $v_g(\bar{\omega})$, which is the speed of the condition of constant magnitude.

Superluminal Light

The title of this section announces that it will treat "faster than light" light, which certainly seems strange, but the phrase makes for great headlines in the news media and in recent years it's become part of the popular scientific discourse.

The Special Theory of Relativity maintains that there are no circumstances under which a signal (i.e., a communicative instrumentality, which perforce carries energy) can propagate at a speed greater than c. Yet we have already seen that under certain circumstances (Section 3.5.1) the phase velocity can do just that. Indeed, as early as 1904 R. W. Wood showed experimentally that white light passing through a chamber contain-

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

ing sodium vapor could have phase velocities exceeding c. He studied the region of anomalous dispersion in the vicinity of the two closely spaced yellow sodium D resonances (having wavelengths of 589.0 nm and 589.6 nm).

At frequencies far from the resonant frequencies of the vapor, the index of refraction was slightly greater than one, as expected. Moreover, little or no light was transmitted in the frequency range of the absorption band. But for light with a frequency close to the D lines, the index $n(\nu)$ began to show signs of anomalous dispersion. As the frequency approached the resonances from the high-frequency low-wavelength side, n rapidly decreased, becoming much less than one ($v > c$). So superluminal phase velocities have been well known for some time.

The contradiction of Relativity is only an apparent one, arising from the fact that although a monochromatic wave can have a speed in excess of c, it cannot convey information. In contrast, a signal in the form of any modulated wave will propagate at the group velocity, which is always less than c in normally dispersive media.*

Starting in the 1980s and continuing to the present, a number of experimenters† have worked to establish that the group velocity could also exceed c. A light pulse of frequency ν will have a group index of refraction given by

$$n_g = n(\nu) + \nu \frac{dn(\nu)}{d\nu}$$

(It's left for Problem 7.25 to prove that that's the case.) This suggests that the place to go to create superluminal pulses is a region of anomalous dispersion where $n(\nu)$ changes rapidly with ν. We want $n_g < 1$, so we need a negative value of $dn(\nu)/d\nu$; that's just what obtains inside an absorption band—the slope of the $n(\nu)$-versus-ν curve is negative.

The problem with that approach is that it's also a place of considerable absorption and the pulses would either be severely distorted or attenuated, making the results ambiguous. That difficulty could be addressed using a medium that had *gain*, one that amplified light. This was recently accomplished in a small cell containing cesium gas. The desired index profile was produced by pumping the cesium atoms using two different-frequency laserbeams. A region of lossless anomalous dispersion was thereby created between the two resulting gain lines (Fig. 7.22).

A diode laser then fired a very long 3.7-μs— nearly Gaussian, or bell-shaped (p. 12)—pulse toward the cell. Amazingly, an essentially identical pulse appeared at the far side of the cell even before the peak of the incoming pulse reached the entrance. The measured lead time was 62 ns, the equivalent of the exiting pulse getting about 20 m ahead of the entering pulse. That's $\approx$310 times farther than if the pulse had traveled the 6-cm length of the cell in vacuum (which would take a mere 0.2 ns).

When $dn(\nu)/d\nu$ is very large and negative, even though it's counterintuitive, it is possible for n_g to be negative. Indeed, in this experiment $n_g = -310$. To appreciate what that means, consider that it takes a pulse a time $L/v_g = n_g L/c$ to traverse a medium of length L, as compared to the time it would take (L/c) to cover the same distance in vacuum. The difference between these two intervals, $\Delta t = L/v_g - L/c = (n_g - 1)L/c$, is the delay the pulse experiences in crossing the medium as opposed to vacuum. However, when $n_g < 1$, $\Delta t < 0$ and there is no such "delay," the pulse arrives early; it appears on the far

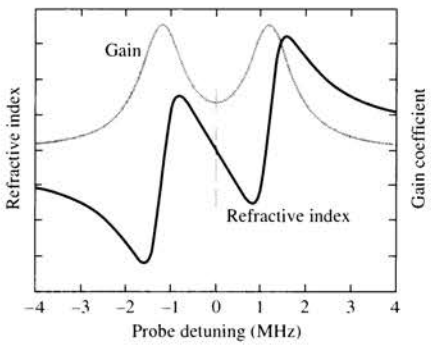


Figure 7.22 Gain-assisted linear anomalous dispersion used to demonstrate superluminal group velocity. The index of refraction and gain coefficient for a cesium gas with two closely spaced gain lines.

*In regions of anomalous dispersion (Section 3.5.1) where $dn/dk < 0$, v_g may be greater than c. Here, however, the signal propagates at yet a different speed, known as the *signal velocity*, v_s. Thus $v_s = v_g$ except in a resonance absorption band. In all cases v_s corresponds to the velocity of energy transfer and never exceeds c.

†S. Chu and S. Wong, "Linear Pulse Propagation in an Absorbing Medium," *Phys. Rev. Lett.* **48**, 738 (1982). L. J. Wang, A. Kuzmich, and A. Dogariu, "Gain-assisted superluminal light propagation," *Nature* **406**, 277 (2000). D. Mognai, A. Ranfagni, and R. Ruggeri, "Observation of Superluminal Behavior in Wave Propagation," *Phys. Rev. Lett.* **84**, 4830 (2000).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

side sooner than if it had traveled the distance L in vacuum.

To begin to understand how that could happened, imagine a Gaussian wave packet whose amplitude falls toward zero ahead and behind the central region. Physically, it's entirely equivalent to a large group of overlapping sine waves that are all in-phase at the one point where the peak of the pulse happens to be at any time. Because their wavelengths are different, moving out from the center of the peak these Fourier component waves individually fall in- and out-of-phase with distance. The jumble of sine waves on either side of the peak increasingly cancel each other, forming the long tapered "wings" of the pulse.

The central insight is that regardless of the amplitude of a wing at any location, it still contains exactly the same sine-wave distribution as does the main peak. It's just that in the outskirts of the pulse, the component waves overlap in such a way as to produce a highly diminished net result. When the leading wing of the wave packet traverses the cell, the cesium atoms take up and re-emit the constituent sine waves, shifting their relative phases (in a frequency-dependent way). That has the effect of reconstituting a clone of the original wave packet. This composite pulse appears at the far end of the cell as if it had traveled at a rate far in excess of c, while the incident pulse vanishes within the gas.

Subluminal Light

At the same time that researchers were producing superluminal wave packets, others were making equally dramatic breakthroughs in slowing down and even stopping light pulses.*

In one experiment sodium atoms were chilled to nanokelvin temperatures via laser cooling (p. 65), followed by evaporative cooling. When the gas fell below 435 nK, it transitioned into a Bose-Einstein condensate (BEC)—a dense cloud of atoms all in the same quantum state. Increasing the density (here at maximum about 5×10^{12} atoms/cm^3) is desirable because it increases the steepness of the $n(\nu)$-versus-ν curve.

Ordinarily, a dense gas would manifest a great deal of dissipative absorption in the vicinity of any one of its resonances (i.e., spectral lines), and that's just where we want to send our laser pulses (each centered at ν_p). In other words (Fig. 7.23*a*), in a dense gas the transition from the ground state $|1\rangle$, to the first excited state $|3\rangle$ will result in dissipative absorption of the light at that frequency (ν_0). An atom absorbs a photon and gets excited, but before it can reradiate, it collides with a neighboring atom and loses that energy. As a consequence, the medium is opaque to pulses centered on ν_0.

This difficulty can be surmounted with a technique called **electromagnetically induced transparency** (EIT). Using magnetic filtering, all of the atoms are first put into the $|1\rangle$ state. The gas is then illuminated by a second so-called coupling laserbeam (ν_c). It's tuned to the transition between a close-by unpopulated hyperfine ground-state level $|2\rangle$ and the same excited state $|3\rangle$. What results is a coupling of the two ground states (a quantum interference effect) that does not allow light in a narrow band around ν_0 to be absorbed; it closes out the $|1\rangle \rightarrow |3\rangle$ transition. In other words, with the coupling laser turned on and all the atoms in $|1\rangle$, the system is in a "dark state" wherein its atoms cannot absorb light of frequency ν_0. When the incident probe light (ν_p) falls within the transmission band around ν_0, it "sees" an essentially transparent medium. There is the usual dispersion, but no absorption and no subsequent dissipation of energy out of the pulse. Furthermore, to avoid distortion of the signal pulse, it was made adequately long in time so that its frequency spectrum would be narrow enough to fit within the transparency band.

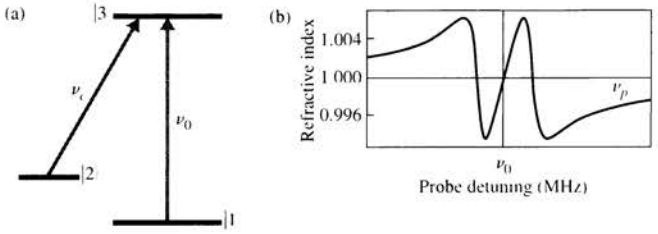


Figure 7.23 (*a*) The energy-level configuration involved in producing electromagnetically induced transparency. (*b*) The index of refraction versus frequency curve for sodium showing the region of high-slope normal dispersion around the resonant frequency.

*Lene Vestergaard Hau, S. E. Harris, Z. Dutton, and C. H. Behroozi, "Light speed reduction to 17 metres per second in an ultracold atomic gas," *Nature* **397**, 594 (1999). Chien Liu, Z. Dutton, and C. H. Behroozi, and Lene Vestergaard Hau, "Observation of coherent optical information storage in an atomic medium using halted light pulses," *Nature* **409**, 490 (2001). D. F. Phillips, A Fleischhauer, A. Mair, R. L. Walsworth, and M. D. Lukin, "Storage of Light in Atomic Vapor," *Phys. Rev. Lett.* **86**, 783 (2001). Also take a look at Kirk T. McDonald, "Slow light," *Am. J. Phys.* **68**, 293 (2000). For a short review, see Barbara Gross Levi, "Researchers Stop, Store, and Retrieve Photons—or at Least the Information They Carry," *Phys. Today* **54**, 17 (2001).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

The frequency dependence of the index of refraction is sketched in Fig. 7.23*b*. At ν_0 the index is one, and so the second term in

$$n_g = n(\nu) + \nu \frac{dn(\nu)}{d\nu}$$

dominates. The steep portion of the curve, where $dn(\nu)/d\nu$ is positive and large, corresponds to a region of normal dispersion with a tremendous group index n_g. Pulses centered at $\nu_p = \nu_0$ propagate through the gas with group velocities as low as 17 m/s. Not long after these results were obtained, by coupling the sodium D_2 line to the sodium D_1 probe line, researchers were able to take the group velocity down to 0.44 m/s (a mere 1 mph).

In January 2001 two independent teams at Harvard (one using cold sodium and the other warm rubidium) brought pulses of light to a crawl, and then, by shutting off the coupling laser—turning the medium opaque again—they stopped the light altogether. Of course, every time you blink you stop light in its tracks, but this was very different. Here the light was first coupled to a system of atoms, and the information characterizing the component sinusoids of the signal pulse (frequency, amplitude, and angular momentum) was imprinted on the gas as a coherent ordering of its atomic spins. This information was later transferred back to the light field and the signal pulse reappeared. The following briefly describes the way this was accomplished.

As the signal pulse (which had a free-space length of 3.4 km) entered the dense dark-state gas, it was thereupon compressed by a factor of c/v_g. (You can imagine this happening as the leading edge of the pulse enters the medium and slows; the fast-moving remainder of the pulse compacts in on itself. The situation can be simulated by a line of runners several strides apart, one behind the other, on a dry road. Suppose the leader suddenly enters a large puddle of knee-deep water with everyone following behind. By the time the last guy reaches the water, the "pulse" of runners will be far more compact and traveling much slower.)

Everything was prearranged so that the compressed signal pulse, of about 27×10^3 photons, just fit inside the ultracold sodium cloud (339 μm). And it was traveling very slowly; at that moment, much of the probe-pulse energy had been transferred to the coupling light field (via stimulated emission, p. 586) and had left the cell. The atoms within the active region of the pulse were in a superposition state determined by the amplitudes and phases of the two laser fields.

Just as the signal pulse disappeared into the cloud, and before it could emerge, the coupling beam was abruptly shut off. The very small amount of energy still associated with the pulse went into a collective spin excitation of the gas cloud. The imprinted atoms retained the information about the physical characteristics of the constituent sine-wave components for up to about 1 ms. When the coupling beam was promptly turned back on, a duplicate of the original pulse re-emerged from the gas. In other words, operating as a coherent quantum-mechanical system, the activated atoms of the gas stored a template of the pulse. When the dark state was switched back on, and electromagnetic energy was thereby made available (via the coupling beam), the atoms reconstituted the signal pulse.

Everything we've talked about in this section relates to pulses of light and their group velocities, whether they're greater than or less than c. In either case, photons only exist at c and they either exist or they don't. Photons never speed up and never slow down, and they certainly never stop and wait around, motionless.

7.3 Anharmonic Periodic Waves

We've already introduced the fascinating notion that any real wave can be synthesized out of appropriately selected sine waves. Now let's see how that's actually done.

Figure 7.24 depicts a disturbance that arises from the superposition of two harmonic functions having different amplitudes and wavelengths. Notice that something rather obvious but interesting has taken place—the composite disturbance is **anharmonic**; it's not sinusoidal.

7.3.1 Fourier Series

Figure 7.25 begins to suggest that by using a number of sinusoidal functions whose amplitudes, wavelengths, and relative phases have been judiciously selected, it would be possible to synthesize some interesting wave profiles. The wavelengths of the sinusoids in that illustration, λ_1 and λ_2, are different, and after one cycle they're out-of-phase. But after N_1 cycles of one and N_2 cycles of the other, such that $\lambda_1 N_1 = \lambda_2 N_2$, they'll be back in-phase and the resultant will repeat itself over and over again; the synthesized function is periodic. The harmonic components that constitute a periodic function need

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

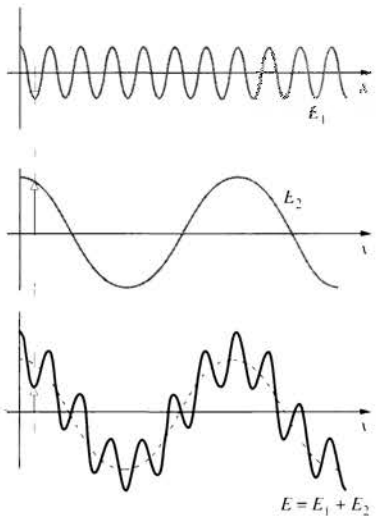


Figure 7.24 The superposition of two harmonic waves of different frequency produces a periodic anharmonic wave.

not all be in-phase at the origin (Fig. 7.26*a*). Take note of the amplitudes of each constituent sinusoid in Fig. 7.26*b*; when we display them at their corresponding frequencies, as in Fig. 7.26*c*, the result is called the **frequency spectrum**.

An exceptionally beautiful mathematical technique for analyzing periodic functions was devised by the French physicist Jean Baptiste Joseph, Baron de Fourier (1768–1830). That theory is predicated on what has come to be known as ***Fourier's Theorem***, which states that *a function f(x), having a spatial period λ, can be synthesized by a sum of harmonic functions whose wavelengths are integral submultiples of λ (that is, λ, λ/2, λ/3, etc.)*. This Fourier-series representation has the mathematical form

$$f(x) = C_0 + C_1 \cos\left(\frac{2\pi}{\lambda}x + \varepsilon_1\right) + C_2 \cos\left(\frac{2\pi}{\lambda/2}x + \varepsilon_2\right) + \ldots \quad (7.41)$$

where the C-values are constants, and of course the profile $f(x)$ may correspond to a traveling wave $f(x - vt)$. To get some sense of how this scheme works, observe that although C_0 by itself is obviously a poor substitute for the original function, it will be appropriate at those few points where it crosses the $f(x)$ curve. In the same way, adding on the next term improves

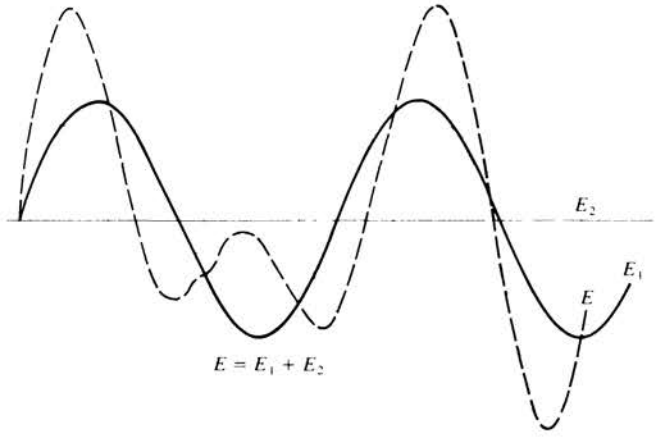


Figure 7.25 The superposition of two harmonic waves of different frequency. The resultant wave is periodic but anharmonic.

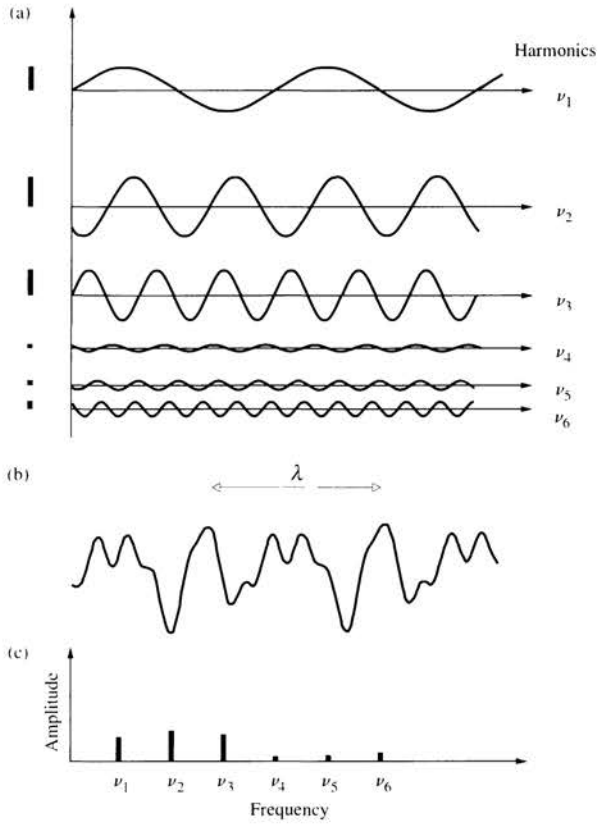


Figure 7.26 (*a*) The superposition of six harmonic waves with different amplitudes and frequencies. (*b*) The resultant periodic function. (*c*) The frequency spectrum.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

things a bit, since the function

$$[C_0 + C_1 \cos (2\pi x/\lambda + \varepsilon_1)]$$

will be chosen so as to cross the $f(x)$ curve even more frequently. If the synthesized function [the right-hand side of Eq. (7.41)] comprises an infinite number of terms, selected to intersect the anharmonic function at an infinite number of points, the series will presumably be identical to $f(x)$.

It is usually more convenient to reformulate Eq. (7.41) by making use of the trigonometric identity

$$C_m \cos (mkx + \varepsilon_m) = A_m \cos mkx + B_m \sin mkx$$

where $k = 2\pi/\lambda$, λ being the wavelength of $f(x)$, $A_m = C_m \cos \varepsilon_m$, and $B_m = -C_m \sin \varepsilon_m$. Thus

$$f(x) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos mkx + \sum_{m=1}^{\infty} B_m \sin mkx \quad (7.42)$$

The first term is written as $A_0/2$ because of the mathematical simplification it will lead to later on. The process of determining the coefficients A_0, A_m, and B_m for a specific periodic function $f(x)$ is referred to as **Fourier analysis**. We'll spend a moment now deriving a set of equations for these coefficients that can be used henceforth. To that end, integrate both sides of Eq. (7.42) over any spatial interval equal to λ, for example, from 0 to λ or from $-\lambda/2$ to $+\lambda/2$ or, more generally, from x' to $x' + \lambda$. Since over any such interval

$$\int_0^\lambda \sin mkx\, dx = \int_0^\lambda \cos mkx\, dx = 0$$

there is only one nonzero term to be evaluated, namely,

$$\int_0^\lambda f(x)\, dx = \int_0^\lambda \frac{A_0}{2}\, dx = A_0 \frac{\lambda}{2}$$

and thus

$$A_0 = \frac{2}{\lambda} \int_0^\lambda f(x)\, dx \quad (7.43)$$

To find A_m and B_m we will make use of the *orthogonality of sinusoidal functions* (Problem 7.32), that is, the fact that

$$\int_0^\lambda \sin akx \cos bkx\, dx = 0 \quad (7.44)$$

$$\int_0^\lambda \cos akx \cos bkx\, dx = \frac{\lambda}{2} \delta_{ab} \quad (7.45)$$

$$\int_0^\lambda \sin akx \sin bkx\, dx = \frac{\lambda}{2} \delta_{ab} \quad (7.46)$$

where a and b are nonzero positive integers and δ_{ab}, known as the *Kronecker delta*, is a shorthand notation equal to zero when $a \neq b$ and equal to 1 when $a = b$. To find A_m we now multiply both sides of Eq. (7.42) by $\cos \ell kx$, ℓ being a positive integer, and then integrate over a spatial period. Only one term is nonvanishing, and that is the single contribution in the first sum, which corresponds to $\ell = m$, in which case

$$\int_0^\lambda f(x) \cos mkx\, dx = \int_0^\lambda A_m \cos^2 mkx\, dx = \frac{\lambda}{2} A_m$$

Thus

$$A_m = \frac{2}{\lambda} \int_0^\lambda f(x) \cos mkx\, dx \quad (7.47)$$

This expression can be used to evaluate A_m *for all values of* m, *including* $m = 0$, as is evident from a comparison of Eqs. (7.43) and (7.47). Similarly, multiplying Eq. (7.42) by $\sin \ell kx$ and integrating, leads to

$$B_m = \frac{2}{\lambda} \int_0^\lambda f(x) \sin mkx\, dx \quad (7.48)$$

In summary, a periodic function $f(x)$ can be represented as a Fourier series

$$f(x) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos mkx + \sum_{m=1}^{\infty} B_m \sin mkx \quad [7.42]$$

where, knowing $f(x)$, the coefficients are computed using

$$A_m = \frac{2}{\lambda} \int_0^\lambda f(x) \cos mkx\, dx \quad [7.47]$$

and

$$B_m = \frac{2}{\lambda} \int_0^\lambda f(x) \sin mkx\, dx \quad [7.48]$$

Be aware that there are some mathematical subtleties related to the convergence of the series and the number of singularities in $f(x)$, but we need not be concerned with these matters here.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Certain symmetry conditions are well worth recognizing, because they lead to some computational shortcuts. Thus if a function $f(x)$ is *even*, that is, if $f(-x) = f(x)$, or equivalently, if it is symmetric about $x = 0$, its Fourier series will contain only cosine terms ($B_m = 0$ for all m) that are themselves even functions. Likewise, *odd* functions that are antisymmetric about $x = 0$, that is, $f(-x) = -f(x)$, will have series expansions containing only sine functions ($A_m = 0$ for all m). In either case, one need not bother to calculate both sets of coefficients. This is particularly helpful when the location of the origin ($x = 0$) is arbitrary, and we can choose it so as to make life as simple as possible. Nonetheless, keep in mind that many common functions are neither odd nor even (e.g., e^x).

The "saw-tooth" function (Fig. 7.27) is odd; whatever its value is a certain distance to the right of the origin, its value is the negative of that, at the same distance to the left of the origin. Thus it can be synthesized out of sinusoids alone—the component harmonic functions are all in-phase and zero at the origin. Here there are only six component sinusoids, so our saw tooth is rather wiggly; add in another few dozen terms and it will sharpen up nicely.

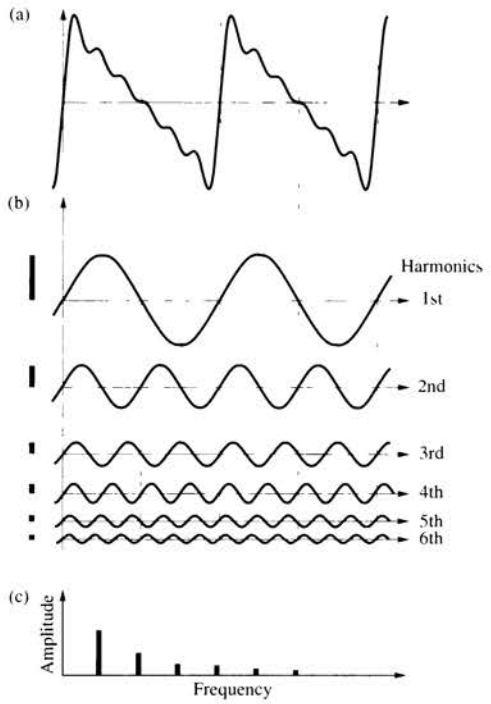


Figure 7.27 (*a*) An approximation of a saw-tooth wave. Ordinarily such a wave would be composed of thousands of sine components and have straight edges that meet at sharp points. (*b*) The six harmonic waves, with different amplitudes and fequencies, that constitue this wiggly saw tooth. (*c*) The frequency spectrum.

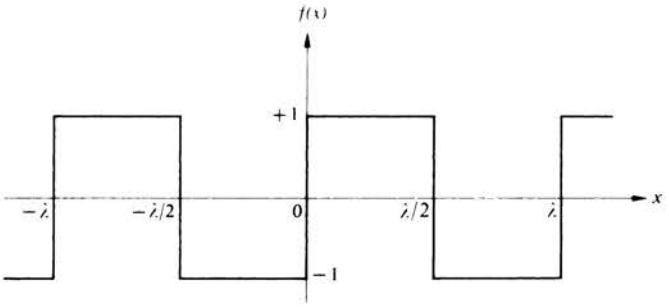


Figure 7.28 The profile of a periodic square wave.

As an example of the technique, let's compute the Fourier series that corresponds to a square wave. Select the location of the origin as shown in Fig. 7.28, whereupon

$$f(x) = \begin{cases} +1 & \text{when } 0 < x < \lambda/2 \\ -1 & \text{when } \lambda/2 < x < \lambda \end{cases}$$

Since $f(x)$ is odd, $A_m = 0$, and

$$B_m = \frac{2}{\lambda}\int_0^{\lambda/2} (+1)\sin mkx\,dx + \frac{2}{\lambda}\int_{\lambda/2}^{\lambda} (-1)\sin mkx\,dx$$

thus

$$B_m = \frac{1}{m\pi}[-\cos mkx]_0^{\lambda/2} + \frac{1}{m\pi}[\cos mkx]_{\lambda/2}^{\lambda}$$

Remembering that $k = 2\pi/\lambda$, we obtain

$$B_m = \frac{2}{m\pi}(1 - \cos m\pi)$$

The Fourier coefficients are therefore

$$B_1 = \frac{4}{\pi}, \qquad B_2 = 0, \qquad B_3 = \frac{4}{3\pi},$$

$$B_4 = 0, \qquad B_5 = \frac{4}{5\pi}, \ldots,$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

and the required series is simply

$$f(x) = \frac{4}{\pi}(\sin kx + \tfrac{1}{3}\sin 3kx + \tfrac{1}{5}\sin 5kx + \cdots) \quad (7.49)$$

Figure 7.29 is a plot of a few partial sums of the series as the number of terms increases. We could pass over to the time domain to find $f(t)$ by just changing kx to ωt. Suppose that we have three ordinary electronic oscillators whose output voltages vary sinusoidally and are controllable in both frequency and amplitude. If these are connected in series with their frequencies set at ω, 3ω, and 5ω and the total signal is examined on an oscilloscope, we can synthesize any of these curves. Similarly, we might simultaneously strike three keys on an appropriately tuned piano with just the correct force on each to create a chord, or composite sound wave, having the curve in Fig. 7.29*c* as its profile. Curiously enough, the human ear–brain audio system is capable of Fourier analysis of a simple composite wave into its harmonic constituents. Presumably there are people who could even name each note in the chord.

Earlier we postponed any detailed consideration of anharmonic periodic functions and restricted the analysis to purely sinusoidal waves. We now have a cogent rationale for having done so. From here on we can envision this kind of disturbance as a superposition of harmonic constituents of different frequencies whose individual behavior can be studied separately. Accordingly, we can write

$$f(x \pm vt) = \frac{A_0}{2} + \sum_{m=1}^{\infty} A_m \cos mk(x \pm vt) + \sum_{m=1}^{\infty} B_m \sin mk(x \pm vt) \quad (7.50)$$

or equivalently

$$f(x \pm vt) = \sum_{m=0}^{\infty} C_m \cos [mk(x \pm vt) + \varepsilon_m] \quad (7.51)$$

for any such *anharmonic periodic wave*.

As a last example, let's analyze the square wave of Fig. 7.30 into its Fourier components. Notice that with the origin chosen as shown, the function is even, and all the B_m terms are zero. The appropriate Fourier coefficients (Problem 7.33) are then

$$A_0 = \frac{4}{a} \quad \text{and} \quad A_m = \frac{4}{a}\left(\frac{\sin m2\pi/a}{m2\pi/a}\right) \quad (7.52)$$

Unlike the previous function, this one has a nonzero value of A_0. You might have already noticed that $A_0/2$ is actually the *mean value* of $f(x)$, and since the curve lies completely above the axis, it will clearly not be zero.

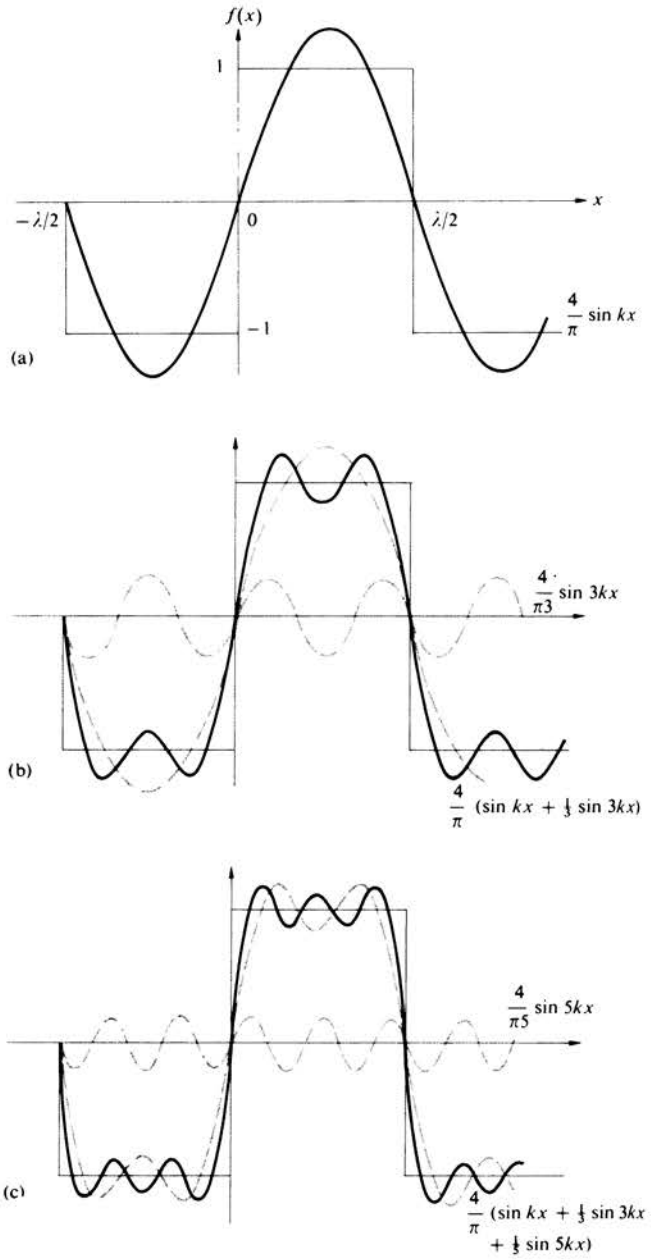


Figure 7.29 Continued.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

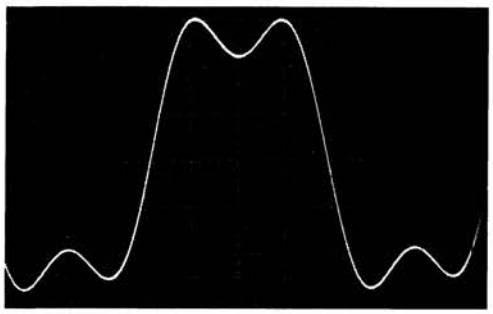


Figure 7.29 Synthesis of the profile of a periodic square wave. Notice that all of the constituent waves are in-phase and zero wherever the square wave is zero. Since all the sine waves are in-phase at $x = 0$, all the B_m coefficients are positive. (Photo by E. H.)

The expression $(\sin u)/u$, which we studied earlier (p. 48), was given the name **sinc** u, and its values are listed in Table 1 in the appendix. Since the limit of sinc u as u goes to zero is 1, A_m can represent all the coefficients, if we let $m = 0, 1, 2, \ldots$. Notice, too, that because the sinc function has negative values, some of the A_m coefficients will be negative. That means that some of the higher order cosines will be 180° out-of-phase with the $m = 1$ cosine term.

Three things distinguish the functions in Figs. 7.28 and 7.30, which otherwise have the same shape: the location of the $x = 0$ axis, the location of the $f(x) = 0$ axis, and the height of the steps. Consequently, beyond the constant A_0, the constituent harmonic terms must have the same relation to either $f(x)$ when they are plotted. In other words, moving the $x = 0$ axis from where it is in Fig. 7.28 to where it is in Fig. 7.30 will change sines into cosines in the analysis, but otherwise leave the constituent harmonic functions in Fig. 7.29 unaltered. The sinusoids that make up the square pulse in Fig. 7.29 will be the cosinusoids that make it up in Fig. 7.30. With the vertical axis in the middle of the square peak, it's clear from Fig. 7.29 that alternate cosines will have to be negative at $x = 0$.

The width of the square peak, $2(\lambda/a)$, can be any fraction of the total wavelength, depending on a. The Fourier series is then

$$f(x) = \frac{2}{a} + \sum_{m=1}^{\infty} \frac{4}{a} \operatorname{sinc} m2\pi/a \cos mkx \qquad (7.53)$$

If we were synthesizing the corresponding function of time, $f(t)$, having a square peak of width $2(\tau/a)$, the same expression, Eq. (7.53), would apply where kx was simply replaced by ωt. Here ω is the *angular temporal frequency* of the periodic function $f(t)$ and is known as the **fundamental**. It is the lowest frequency of the cosine term and arises when $m = 1$. Frequencies of 2ω, 3ω, $4\omega, \ldots$, are known as ***harmonics*** of the fundamental and are associated with $m = 2, 3, 4, \ldots$. In much the same way, since λ is the *spatial period*, $\kappa \equiv 1/\lambda$ is the **spatial frequency**, and $k = 2\pi\kappa$ might be called the **angular spatial frequency**. Again one speaks of the harmonics, of frequency $2k$, $3k$, $4k, \ldots$, where these are spatial alternations. Evidently, the dimensions of κ are cycles per unit length (e.g., cycles per mm or possibly just cm^{-1}), and those of k are radians per unit length. Thus a fence could have a spatial frequency of 1 picket per foot, while a comb might have a k of 10 teeth per inch.

Let's clarify a few points so as to avoid a common confusion concerning the use of the terms *spatial frequency* and *spatial period* (or wavelength). Figure 7.30 shows a one-dimensional periodic square-wave function spread out in space along the x-axis. This might be a pattern seen on the face of an oscilloscope or the profile of a rather extraordinary disturbance moving along a taut rope. In either case, it repeats itself in space over a distance known as the wavelength, and one over that is the spatial frequency.

Now suppose instead that the pattern corresponds to an irradiance distribution, a series of bright and dark stripes, for instance, the kind of thing you might see looking through a narrow horizontal slit against a picket fence or, even better, while scanning on a line across a group of alternately clear and opaque bands (Fig. 13.25) illuminated by monochromatic light. Again the pattern will have some spatial period and frequency determined by the rate at which it repeats in space. As ever the light field will also have a spatial frequency (κ) and period (λ), as well as a temporal frequency and period, quite apart from the other. The pattern might have a wavelength (λ) of 20 cm, and the light generating it a wavelength (λ) of 500 nm. Herein lies

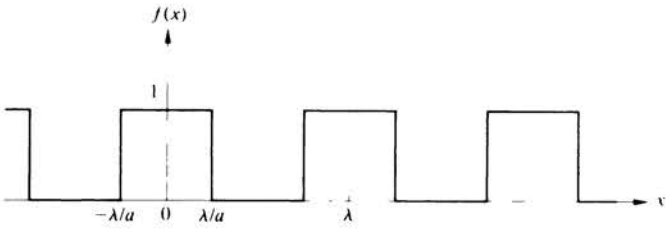


Figure 7.30 A periodic anharmonic function.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

the area of potential confusion. ***Henceforth, we will reserve the symbols k and λ for the lightwave itself and use k* and λ to describe spatial optical patterns.**

Now return to the square function of Fig. 7.30 where now $a = 4$, or in other words the peak has a width of $\lambda/2$. In that instance

$$f(x) = \frac{1}{2} + \frac{2}{\pi}(\cos kx - \tfrac{1}{3}\cos 3kx + \tfrac{1}{5}\cos 5kx - \cdots) \tag{7.54}$$

As a matter of fact, *if the graph of the function $f(x)$ is such that a horizontal line could divide it into equally shaped segments, above and below that line, the Fourier series will consist of only odd harmonics.*

Were we to plot the curve representing the partial sum of the terms through $m = 9$, it would closely resemble the square wave. In contrast, if the width of the peak is reduced, the number of terms in the series needed to produce the same general resemblance to $f(x)$ will increase. This can be appreciated by examining the ratio

$$\frac{A_m}{A_1} = \frac{\sin m2\pi/a}{m \sin 2\pi/a} \tag{7.55}$$

Observe that for $a = 4$, the ninth term (i.e., $m = 9$) is fairly small, $A_9 \approx 10\%\, A_1$. In comparison, for a peak 100 times narrower (that is, $a = 400$), $A_9 \approx 99\%\, A_1$. Similarly, whereas it takes terms through $m = 4$ to duplicate the curve of Fig. 7.29*b* when $a = 4$, it will take up to $m = 8$ to produce roughly the equivalent profile when $a = 8$. Making the peak narrower has the effect of introducing higher-order harmonics, which in turn have smaller wavelengths. We might guess, then, that it is not the total number of terms in the series that is of prime importance but rather the relative dimensions of the smallest features being reproduced and the corresponding wavelengths available.* If there are fine details in the profile, the series must contain comparatively short-wavelength (or in the time domain, short-period) contributions.

The negative values of A_m in Eq. (7.53) should simply be thought of as the amplitudes of those harmonic contributions that are to be added into the synthesis with their phases shifted by 180°, as compared with the positive terms. The equivalence between a negative amplitude and a π-rad phase shift is clear from the fact that $A_m \cos(kx + \pi) = -A_m \cos kx$.

*Evidently, one is not going to be able to build a castle of blocks unless the blocks are a good deal smaller than the castle.

7.4 Nonperiodic Waves

All real waves are pulses, albeit sometimes rather long ones, and so it's important to learn how to analyze nonperiodic functions. Such functions are of great practical interest in physics, particularly in Optics and Quantum Mechanics.

We saw earlier (Fig. 7.16) how adding two sinusoids produces beats; the sines fall out-of-phase creating a minimum in the envelope and then come back in-phase to produce a maximum. We might guess that packing in more frequency components would necessitate a greater distance in space before they all could come back in-phase to form the next maximum in the envelope (Fig. 7.31). In other words, the presence of more frequency components might well have the effect of separating the maxima. Remember too that the carrier in the beat pattern was at the average frequency (call it k_p because it's going to turn out to be the peak frequency present) so if we add in sinusoids symmetrically around k_p the carrier oscillation shouldn't change frequency; we can see as much in Fig. 7.31*b* and via Problem 7.18. The agenda that lies before us if we are to generate a single solitary pulse out of harmonic components (Fig. 7.31*e*) is to determine exactly which frequencies need to be added in and how much of each should be included.

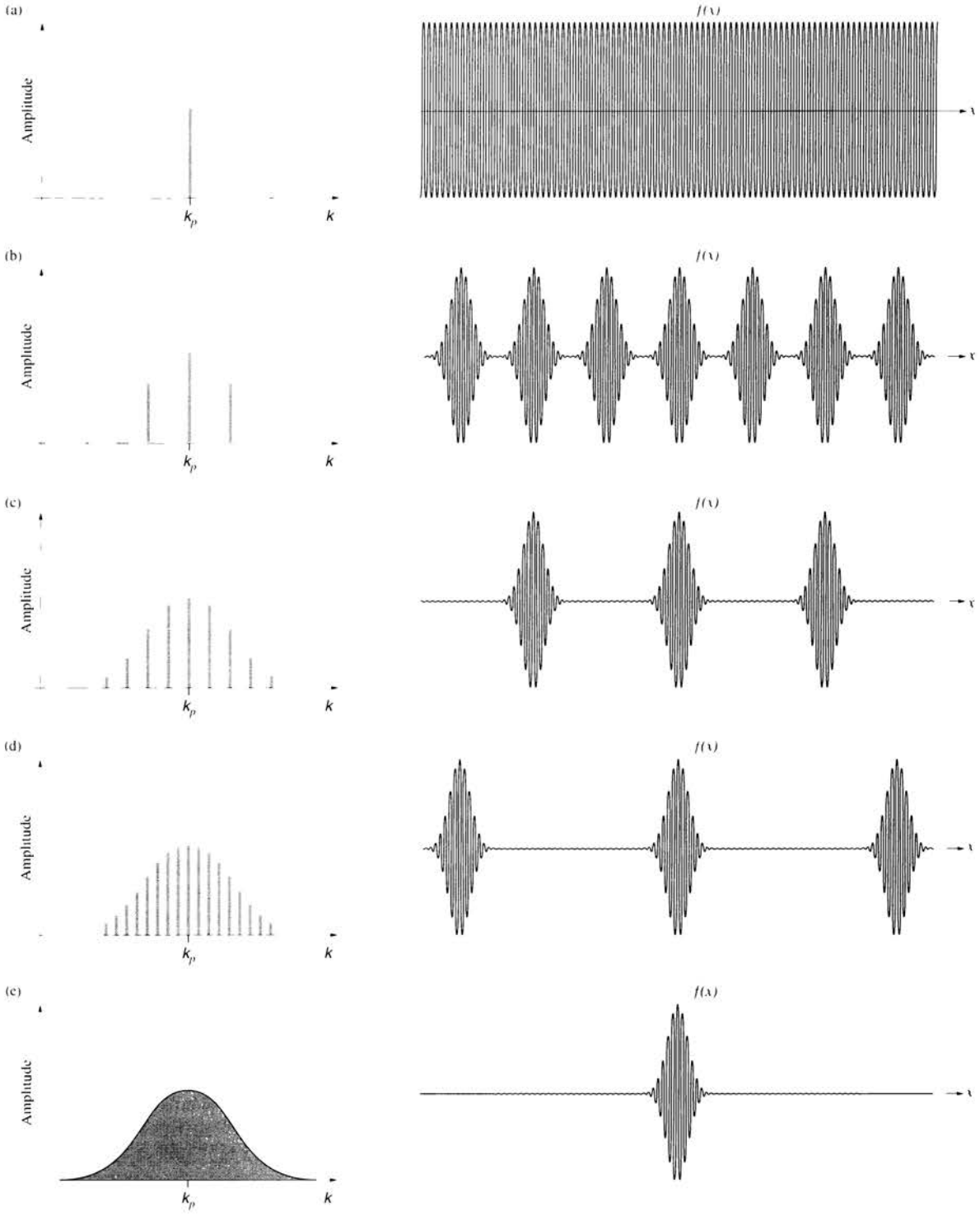
Return to Fig. 7.30 and suppose that we keep the width of the square peak constant while λ is made to increase without limit. As λ approaches infinity, the resulting function will no longer appear periodic. We then have one single square pulse, the adjacent peaks having moved off to infinity. This suggests a possible way of generalizing the method of Fourier series to include nonperiodic functions.

7.4.1 Fourier Integrals

To essentially stretch out the function, let's initially set $a = 4$ and choose some value of λ; anything will do, say, $\lambda = 1$ cm. The peak then has a width of $\frac{1}{2}$ cm, that is, $2(\lambda/a)$, centered at

Figure 7.31 Here we start with a single sinusoid at a spatial frequency (k_p) that's called the carrier (or peak) frequency. Adding in two more frequency components symmetrical about k_p leaves the carrier (or average) frequency unchanged but creates beats. Adding in still more pairs of sinusoids further separates the pulses without changing their shapes. As we'll see in Fig. 7.37, if the amplitudes of the constituents form an envelope that is Gaussian, the envelopes of the pulses will also be Gaussian.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027



Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

$x = 0$, as illustrated in Fig. 7.32*a*. The importance of each particular frequency, mk, can be appreciated by examining the value of the corresponding Fourier coefficient, in this case A_m. The coefficients may be thought of as weighting factors that appropriately emphasize the various harmonics. Figure 7.32*a* contains a plot of a number of values of A_m (where $m = 0, 1, 2,\ldots$) versus mk for the foregoing square wave. Recall that such a curve is known as the ***spatial frequency spectrum***.

We can regard A_m as a function, $A(mk)$, of mk, which may be nonzero only at values of $m = 0, 1, 2,\ldots$. If the quantity a is now made equal to 8 while λ is increased to 2 cm, the peak width will be completely unaffected. The only alteration is a doubling of the space between peaks. Yet a very interesting change in the spatial frequency spectrum is evident in Fig. 7.32*b*. Note that the density of components along the mk-axis has increased markedly. Nonetheless, $A(mk)$ is still zero when $mk = 4\pi, 8\pi, 12\pi,\ldots$, but since k is now π rather than 2π, there will be more terms between these zero points. Finally, let $a = 16$ and increase λ to 4 cm. Again the individual peaks are unaltered in shape, but the terms in the frequency spectrum are now even more densely packed. In effect, the pulse, as compared with λ, is getting smaller and smaller, thereby requiring higher frequencies to synthesize it.

Observe that the envelope of the curve, which was barely discernible in Fig. 7.32*a*, is quite evident in Fig. 7.32*c*. In fact, the envelope is identical in each case, except for a scale factor. It is determined only by the shape of the original signal and will be quite different for other configurations. We conclude that as λ increases and the function takes on the appearance of a single square pulse, the space between each of the $A(mk)$ contributions in the spectrum will decrease. The discrete spectral lines, while decreasing in amplitude, will gradually merge, becoming individually unresolvable. In the limit as λ approaches ∞, the spectral lines will become infinitely close to each other. As k becomes extremely small, m must consequently become exceedingly large, if mk is to be at all appreciable. Changing notation, replace mk, the angular frequency of the harmonics, by k_m. Although it comprises discrete terms, in the limit k_m will be transformed into k (i.e., a continuous frequency distribution). The function $A(k_m)$ in the limit will become the envelope shown in Fig. 7.32. It is obviously no longer meaningful to talk about the fundamental frequency and its harmonics. The pulse being synthesized, $f(x)$, has no apparent fundamental frequency.

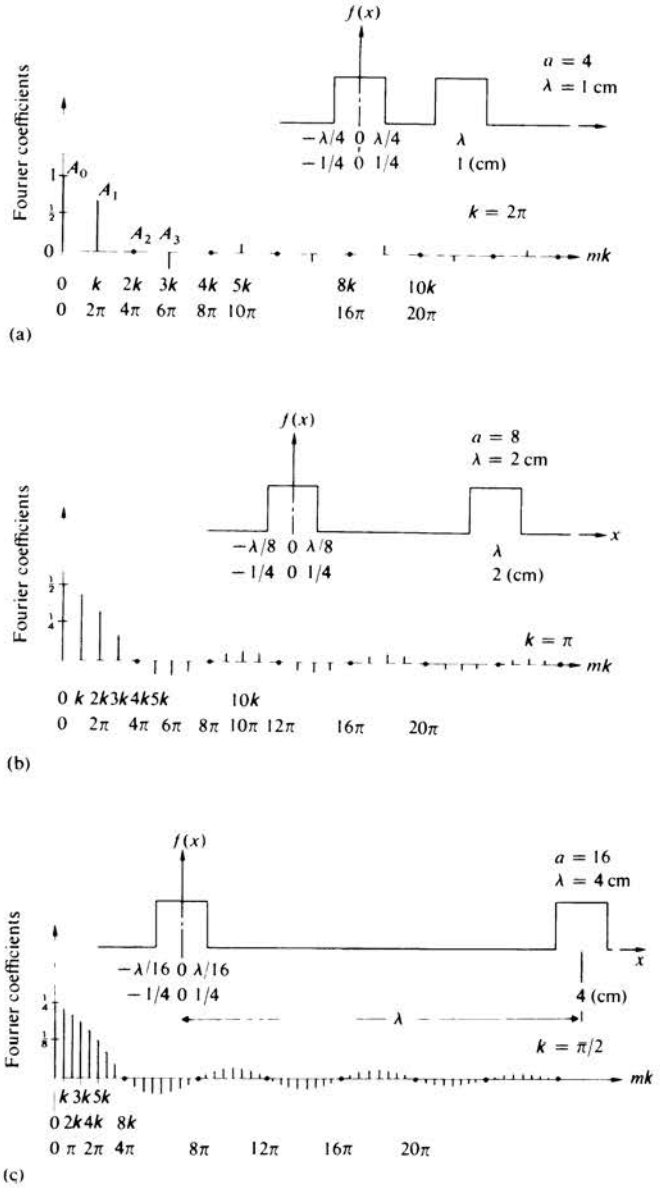


Figure 7.32 The square pulse as a limiting case. The negative coefficients correspond to a phase shift of π radians. As more and more frequency terms are added to the synthesis, the peaks on either side of the one at the origin move out toward ±infinity, respectively. Ultimately when there is a continuous range of component frequencies present, they will combine to produce a single square pulse at the origin.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Recall that an integral is actually the limit of a sum as the number of elements goes to infinity and their size approaches zero. Thus it should not be surprising that the *Fourier series* must be replaced by the so-called **Fourier integral** as λ goes to infinity. That integral, which is stated here without proof, is

$$f(x) = \frac{1}{\pi}\left[\int_0^\infty A(k)\cos kx\,dk + \int_0^\infty B(k)\sin kx\,dk\right] \quad (7.56)$$

provided that

$$A(k) = \int_{-\infty}^{\infty} f(x)\cos kx\,dx$$

and

$$B(k) = \int_{-\infty}^{\infty} f(x)\sin kx\,dx \quad (7.57)$$

The similarity with the series representation should be obvious. The quantities $A(k)$ and $B(k)$ are interpreted as the amplitudes of the sine and cosine contributions in the range of angular spatial frequency between k and $k + dk$. They are the **Fourier cosine** and **sine transforms**, respectively. In the foregoing example of a square pulse, it is the cosine transform, $A(k)$, that will be found to correspond to the envelope in Fig. 7.32.

A careful examination of Fig. 7.32 and Eq. (7.53) reveals that except for the zero-frequency term, the amplitudes of the contributions to the synthesis vary as $(4/a)$ sinc $m2\pi/a$: the envelope of the curve is a sinc function. Remember that the first term in the series is $\frac{1}{2}A_0$, which suggests another way to represent the frequency spectrum. Inasmuch as $\cos(mkx) = \cos(-mkx)$, we can divide the amplitude of every contribution beyond $m = 0$ in half and plot it twice, once with a positive value of k and again with a negative one (Fig. 7.33). This mathematical contrivance provides a nice symmetrical curve; it's introduced here because it is common practice to represent frequency spectra in that fashion.

As we will see in Chapter 11, the most powerful Fourier transform methods involve a complex representation that automatically gives rise to a symmetrical distribution of positive and negative spatial frequency terms. Certain optical phenomena (such as diffraction) also occur symmetrically in space, and a marvelous relationship can be constructed with the spatial frequency spectrum, provided that it encompasses positive and negative frequencies. Thus the negative frequency is a useful mathematical device that allows us to describe physical systems that are symmetrical (going off in opposite directions from a central point).

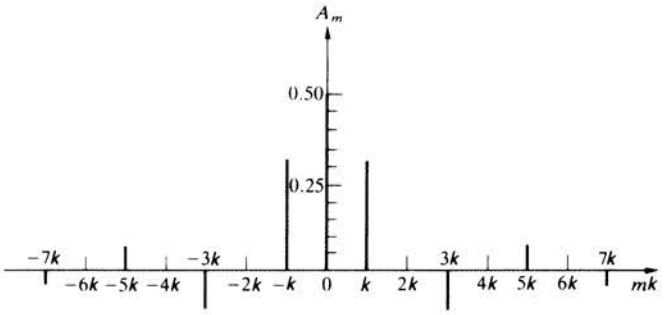


Figure 7.33 A symmetrical frequency spectrum for the waveform in Figure 7.21*a*. Note that the zeroth term is actually $A_0/2$, which is indeed the amplitude of the $m = 0$ contribution to the series.

7.4.2 Pulses and Wave Packets

Let's now determine the Fourier-integral representation of the square pulse in Fig. 7.34, which is described by the function

$$f(x) = \begin{cases} E_0 & \text{when } |x| < L/2 \\ 0 & \text{when } |x| > L/2 \end{cases}$$

For the moment we'll limit the analysis to positive values of k. Since $f(x)$ is an even function, the sine transform, $B(k)$, will be found to be zero. Pressing on,

$$A(k) = \int_{-\infty}^{\infty} f(x)\cos kx\,dx = \int_{-L/2}^{+L/2} E_0\cos kx\,dx$$

Hence

$$A(k) = \frac{E_0}{k}\sin kx\Big|_{-L/2}^{+L/2} = \frac{2E_0}{k}\sin kL/2$$

Multiplying numerator and denominator by L and rearranging terms, we have

$$A(k) = E_0 L\,\frac{\sin kL/2}{kL/2}$$

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

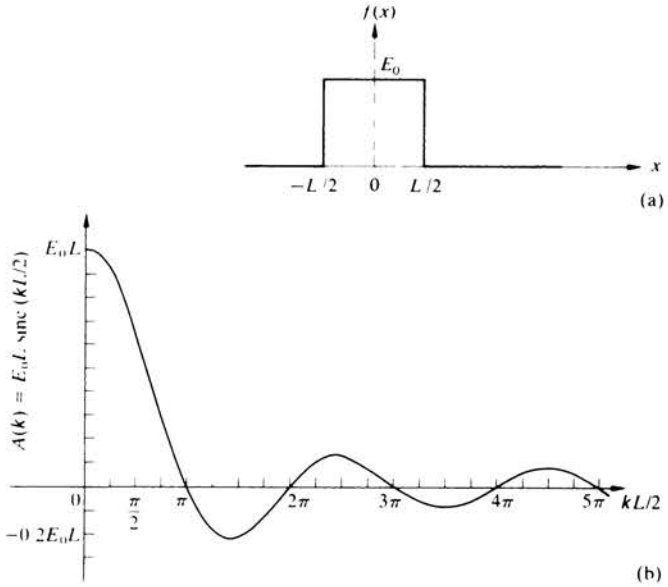


Figure 7.34 The square pulse and its transform.

or equivalently

$$A(k) = E_0 L \text{ sinc } (kL/2) \tag{7.58}$$

The Fourier transform of the square pulse is plotted in Fig. 7.34*b* and should be compared with the envelope in Fig. 7.32. As L increases, the spacing between successive zeros of $A(k)$ decreases and vice versa. Moreover, when $k = 0$, it follows from Eq. (7.58) that $A(0) = E_0 L$.

It is a simple matter to write out the integral representation of $f(x)$ using Eq. (7.56):

$$f(x) = \frac{1}{\pi} \int_0^\infty E_0 L \text{ sinc } (kL/2) \cos kx \, dx \tag{7.59}$$

An evaluation of this integral is left for Problem 7.37.

The Cosine Wavetrain

Earlier, when we talked about monochromatic waves, we pointed out that they were in fact fictitious, at least physically. There will always have been some point in time when the generator, however perfect, was turned on. Figure 7.35 depicts a somewhat idealized harmonic pulse with a carrier frequency k_p corresponding to the function

$$E(x) = \begin{cases} E_0 \cos k_p x & \text{when } -L \le x \le L \\ 0 & \text{when } |x| > L \end{cases}$$

We chose to work in the space domain but could certainly have envisioned the disturbance as a function of time. We are effectively examining the spatial profile of the wave $E(x - vt)$ at $t = 0$ rather than the temporal profile at $x = 0$. The spatial frequency k_p is that of the harmonic region of the pulse itself. Proceeding with the analysis, note that $E(x)$ is an even function; consequently, $B(k) = 0$ and

$$A(k) = \int_{-L}^{+L} E_0 \cos k_p x \cos kx \, dx$$

This is identical to

$$A(k) = \int_{-L}^{+L} E_0 \tfrac{1}{2} [\cos (k_p + k)x + \cos (k_p - k)x] \, dx$$

which integrates to

$$A(k) = E_0 L \left[\frac{\sin (k_p + k)L}{(k_p + k)L} + \frac{\sin (k_p - k)L}{(k_p - k)L} \right]$$

or, if you like,

$$A(k) = E_0 L [\text{sinc } (k_p + k)L + \text{sinc } (k_p - k)L] \tag{7.60}$$

When there are many waves in the train ($\lambda_p << L$), $k_p L >> 2\pi$. Thus $(k_p + k)L >> 2\pi$, and therefore sinc $(k_p + k)L$ is down to fairly small values. In contrast, when $k_p = k$, the second sinc function in the brackets has a maximum value of 1. In other words, the function given by Eq. (7.60) can be thought of as having a peak at $k = -k_p$ as shown in part (*b*) of the drawing. If we limit the treatment to only positive values of k, only the tail of that left-side peak that crosses into the positive k region will contribute. As we have just seen, such contributions will be negligible far from $k = -k_p$, especially when $L >> \lambda_p$ and the peaks are both narrow and widely spaced. The positive tail of the left-side peak then falls off rapidly beyond $k = -k_p$. Consequently, we can neglect the first sinc in this particular case and write the transform as

$$A(k) = E_0 L \text{ sinc } (k_p - k)L \tag{7.61}$$

(Fig. 7.35*c*). Even though the wavetrain is very long, since it is not infinitely long it must be synthesized from a continuous

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

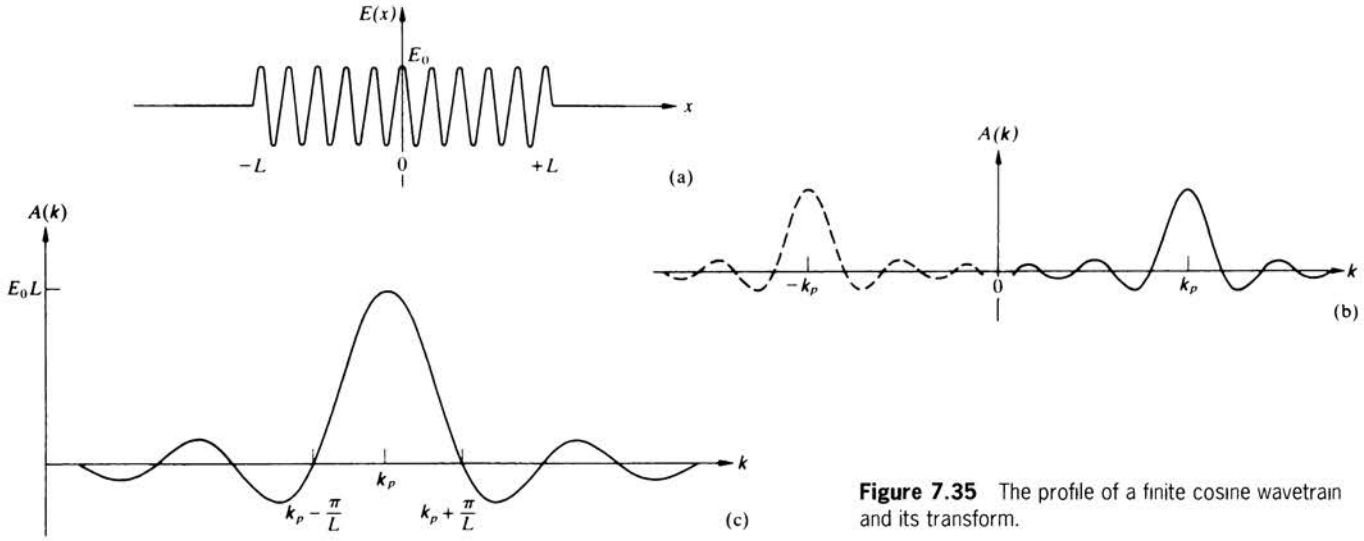


Figure 7.35 The profile of a finite cosine wavetrain and its transform.

range of spatial frequencies. Thus it can be thought of as the composite of an infinite ensemble of harmonic waves. One speaks of such pulses as ***wave packets*** or ***wave groups***. As we might have expected, the dominant contribution is associated with $k = k_p$. Had the analysis been carried out in the time domain, the same results would have obtained where the transform was centered about the temporal angular frequency ω_p. Clearly, as the wavetrain becomes infinitely long (i.e., $L \to \infty$), its frequency spectrum shrinks, and the curve of Fig. 7.35*c* closes down to a single tall spike at k_p (or ω_p). This is the limiting case of the idealized monochromatic wave.

Since we can think of $A(k)$ as the amplitude of the contributions to $E(x)$ in the range k to $k + dk$, $A^2(k)$ must be related to the energy of the wave in that range (Problem 7.38). We'll come back to this point in Chapter 11 when we consider the *power spectrum*. For the moment, merely observe (Fig. 7.35*c*) that most of the energy is carried in the spatial frequency range from $k_p - \pi/L$ to $k_p + \pi/L$, extending between the minima on either side of the central peak. An increase in the length of the wavetrain causes the energy of the wave to become concentrated in an ever narrowing range of k about k_p.

The wave packet in the time domain, that is,

$$E(t) = \begin{cases} E_0 \cos \omega_p t & \text{when } -T \le t \le T \\ 0 & \text{when } |t| > T \end{cases}$$

has the transform

$$A(\omega) = E_0 T \operatorname{sinc} (\omega_p - \omega)T \tag{7.62}$$

where ω and k are related by the phase velocity. The frequency spectrum, except for the notational change from k to ω and L to T, is identical to that of Fig. 7.35*c*.

Frequency Bandwidth

For the particular wave packet being studied, the range of angular frequencies (ω or k) that the transform comprises is certainly not finite. Yet if we were to speak of the *width* of the transform ($\Delta\omega$ or Δk), Fig. 7.35*c* suggests that we use $\Delta k = 2\pi/L$ or $\Delta\omega = 2\pi/T$. In contrast, the spatial or temporal extent of the pulse is unambiguous at $\Delta x = 2L$ or $\Delta t = 2T$, respectively. The product of the width of the packet in what might be called *k-space* and its width in *x-space* is $\Delta k\, \Delta x = 4\pi$ or, analogously, $\Delta\omega\, \Delta t = 4\pi$. The quantities Δk and $\Delta\omega$ are the **frequency bandwidths**. Had we used a differently shaped pulse, the product of the bandwidth and the pulse length might certainly have been somewhat different. The ambiguity arises because we have not yet chosen one of the alternative possibilities for specifying $\Delta\omega$ and Δk. For example, rather than using the first minima of $A(k)$ (there are transforms that have no such minima, such as the Gaussian function of Section 11.2), we could have let Δk be the width of $A^2(k)$ at a point where the curve had dropped to $\frac{1}{2}$ or possibly

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

$1/e$ of its maximum value. In any event, it will suffice for the time being to observe that since $\Delta\omega = 2\pi\Delta\nu$,

$$\Delta\nu \approx 1/\Delta t \tag{7.63}$$

that is, the frequency bandwidth is the same order of magnitude as the reciprocal of the temporal extent of the pulse (Problem 7.39). If the wave packet has a narrow bandwidth, it will extend over a large region of space and time. Accordingly, a radio tuned to receive a bandwidth of $\Delta\nu$ will be capable of detecting pulses of duration no shorter than $\Delta t \approx 1/\Delta\nu$.

These considerations are of profound importance in Quantum Mechanics where wave packets describe particles, and Eq. (7.63) is akin to the Heisenberg Uncertainty Principle.

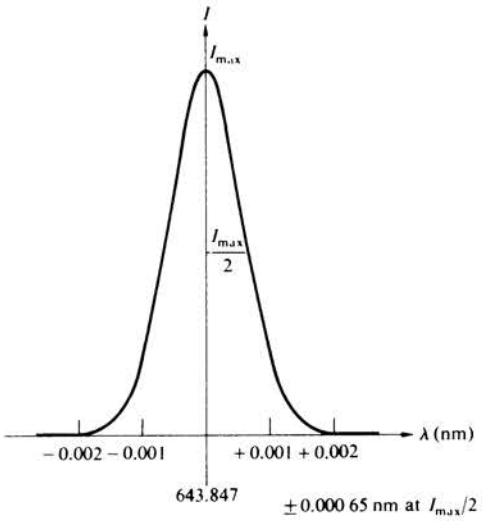


Figure 7.36 The cadmium red ($\bar{\lambda}$ = 643.847 nm) spectral line from low-pressure lamp.

7.4.3 Coherence Length

Let's now consider the light emitted by what is loosely termed a monochromatic source, for example, a sodium discharge lamp. When the beam is passed through some sort of spectrum analyzer, all its various frequency components are observed. Typically, we find that a number of fairly narrow frequency ranges contain most of the energy and that these are separated by much larger regions of darkness. Each such brightly colored band is known as a **spectral line**. There are devices in which the light enters by way of a slit, and each line is actually a colored image of that slit. Other analyzers represent the frequency distribution on the screen of an oscilloscope. In any event, the individual spectral lines are never infinitely sharp. They always consist of a band of frequencies, however small (Fig. 7.36).

The electron transitions responsible for the generation of light have a duration on the order of 10^{-8} s to 10^{-9} s. Because the emitted wavetrains are finite, there will be a spread in the frequencies present, known as the **natural linewidth** (see Section 11.3.4). Moreover, since the atoms are in random thermal motion, the frequency spectrum will be altered by the Doppler Effect. In addition, the atoms suffer collisions that interrupt the wavetrains and again tend to broaden the frequency distribution. The total effect of all these mechanisms is that each spectral line has a bandwidth $\Delta\nu$ rather than one single frequency. The time that satisfies Eq. (7.63) is referred to as the **coherence time** (henceforth to be written Δt_c), and the length Δl_c given by

$$\Delta l_c = c\,\Delta t_c \tag{7.64}$$

is the **coherence length**. As will become evident presently, the coherence length is the extent in space over which the wave is nicely sinusoidal so that its phase can be predicted reliably. The corresponding temporal duration is the coherence time. These concepts are extremely important in considering the interaction of waves, and we will come back to them later in the discussion of interference.

Although the concept of the photon wavetrain is already familiar, we are now in a position, armed with a little Fourier analysis, to deduce something about its configuration. This can be done by essentially working backward from the experimental observation that the frequency distribution of a spectral line from a quasimonochromatic (nonlaser) source can be represented by a bell-shaped Gaussian function (Section 2.1). That is, the irradiance versus frequency is found to be Gaussian. But irradiance is proportional to the electric-field amplitude squared, and since the square of a Gaussian function is a Gaussian function, it follows that the net amplitude of the light field is also bell-shaped.

Now suppose a single photon wavetrain, one of N identical such packets making up the beam, resembles Fig. 7.37*a* in that it is a harmonic function modulated by a Gaussian envelope. Its Fourier transform, $A(\omega)$, is also Gaussian. Imagine that we look at only one and the same harmonic frequency component that goes into making up each photon wavetrain, for example, the one corresponding to ω'. Remember that this component is

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

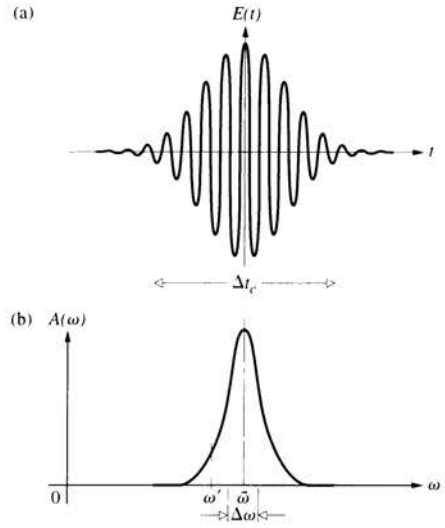


Figure 7.37 A cosinusoidal wavetrain modulated by a Gaussian envelope along with its transform, which is also Gaussian.

an infinitely long, constant-amplitude sinusoid. If every packet is indeed identical, the amplitude of the Fourier component associated with ω' will be the same in each. At any point in a stream of photons these ω'-component monochromatic waves, one from each wavetrain, will have a random relative phase distribution that rapidly changes in time with the arrival of each photon. Thus all such contributions taken together [Eq. (7.21)] will correspond on average to a harmonic wave of frequency ω' having an amplitude proportional to $N^{1/2}$, and this is the ω' part of the net observed field. The same will be true for every other frequency constituting the packets. This means that the same amount of energy is present at each frequency in the net light field of the beam as there is in the totality of the separate constituent wavetrains. Moreover, we know all about this energy-frequency distribution; it's Gaussian, so the transform of the photon wavetrain must be Gaussian too. In other words, the observed spectral line corresponds to the power spectrum of the beam, but it also corresponds to the power spectrum of an individual photon packet. If the irradiance is Gaussian, the photon wavetrain is Gaussian.

As a result of the randomness of the wavetrains, the individual harmonic components of the resultant wave will not have the same relative phases as they did in each packet. Thus the profile of the resultant will differ from that of the separate wave packets, even though the amplitude of each frequency component present in the resultant is simply $N^{1/2}$ times its amplitude in any one packet. The observed spectral line corresponds to the power spectrum of the resultant beam, to be sure, but it also corresponds to the power spectrum of an individual packet. Ordinarily there will be a tremendous number of arbitrarily overlapping wave groups, so that the envelope of the resultant will rarely, if ever, be zero. If the source is quasimonochromatic (i.e., if the bandwidth is small compared with the mean frequency $\bar{\nu}$), we can envision the resultant as being "almost" sinusoidal.

In summary, the composite lightwave can be pictured as in Fig. 7.38. We might imagine the frequency and amplitude to be randomly varying, the former over a range $\Delta\nu$ centered at $\bar{\nu}$. Accordingly, the **frequency stability**, defined as $\Delta\nu/\bar{\nu}$, is a useful measure of spectral purity. Even a coherence time as short as 10^{-9} s corresponds to roughly a few million wavelengths of the rapidly oscillating carrier ($\bar{\nu}$), so that any amplitude or frequency variations will occur quite slowly in comparison. Equivalently, we can introduce a time-varying phase factor such that the disturbance can be written as

$$E(t) = E_0(t) \cos [\varepsilon(t) - 2\pi\bar{\nu}t] \tag{7.65}$$

where the separation between wave crests changes in time.

The average duration of a wave packet is Δt_c, so two points on the wave in Fig. 7.38 separated by more than Δt_c must lie on different contributing wavetrains. These points would thus

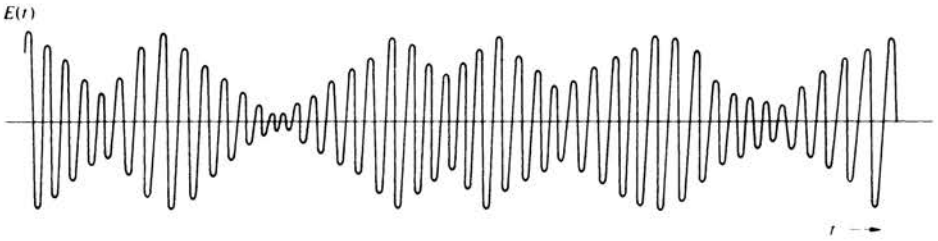


Figure 7.38 A quasimonochromatic lightwave.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

be completely uncorrelated in phase. In other words, if we determined the electric field of the composite wave as it passed by an idealized detector, we could predict its phase fairly accurately for times much less than Δt_c later, but not at all for times greater than Δt_c. In Chapter 12 we will consider the *degree of coherence* that applies over the region between these extremes as well.

White light has a frequency range from 0.4×10^{15} Hz to about 0.7×10^{15} Hz, that is, a bandwidth of about 0.3×10^{15} Hz. The coherence time is then roughly 3×10^{-15} s, which corresponds [Eq. (7.64)] to wavetrains having a spatial extent only a few wavelengths long (Table 7.1). Accordingly, *white light may be envisaged as a random succession of very short pulses*. Were we to synthesize white light, we would have to superimpose a broad, continuous range of harmonic constituents in order to produce the very short wave packets. Inversely, we can pass white light through a Fourier analyzer, such as a diffraction grating or a prism, and in so doing actually generate those components.

The available bandwidth in the visible spectrum ($\approx$300 THz) is so broad that it represents something of a wonderland for the communications engineer. For example, a typical television channel occupies a range of about 4 MHz in the electromagnetic spectrum ($\Delta\nu$ is determined by the duration of the pulses needed to control the scanning electron beam). Thus the visible region could carry roughly 75 million television channels. Needless to say, this is an area of active research (see Section 8.11).

TABLE 7.1 Approximate Coherence Lengths of Several Sources

Source	Mean Wavelength $\bar{\lambda}$(nm)	Linewidth* $\Delta\lambda$ (nm)	Coherence Length Δl_c
Thermal IR (8000–12 000 nm)	10 000	$\approx$4000	$\approx$25 000 nm = $2.5\bar{\lambda}$
Mid-IR (3000–5000 nm)	4000	$\approx$2000	$\approx$8000 nm = $2\bar{\lambda}$
White light	550	$\approx$300	$\approx$900 nm = $1.6\bar{\lambda}$
Mercury arc	546.1	$\approx$1.0	$\lesssim$ 0.03 cm
Kr^{86} discharge lamp	605.6	1.2×10^{-3}	0.3 m
Stabilized He-Ne laser	632.8	$\approx 10^{-6}$	$\lesssim$ 400 m
Special He-Ne laser	1153	8.9×10^{-11}	15×10^{6} m

*To find the corresponding frequency bandwidth use, $\Delta\nu\ \Delta\lambda_0 = \nu\ \bar{\lambda}_0$

Ordinary discharge lamps have relatively large bandwidths leading to coherence lengths only on the order of several millimeters. In contrast, the spectral lines emitted by low-pressure isotope lamps such as Hg^{198} (λ_{air} = 546.078 nm) or the international standard Kr^{86} (λ_{air} = 605.616 nm) have bandwidths of roughly 1000 MHz. The corresponding coherence lengths are approximately 0.3 m, and the coherence times are about 1 ns. The frequency stability is about one part per million—these sources are certainly quasimonochromatic.

The most spectacular of all present-day sources is the laser. Under optimum conditions, with temperature variations and vibrations meticulously suppressed, a laser was actually operated at quite close to its theoretical limit of frequency constancy. A short-term frequency stability of about 8 parts per 10^{14} was attained* with a He–Ne continuous gas laser at λ_0 = 1153 nm. That corresponds to a remarkably narrow bandwidth of about 20 Hz. More common and not very difficult to obtain are frequency stabilities of several parts per 10^9. There are commercially available CO_2 lasers that provide a short-term ($\approx 10^{-1}$ s) $\Delta\nu/\bar{\nu}$ ratio of 10^{-9} and a long-term ($\approx 10^3$ s) value of 10^{-8}.

7.4.4 The Discrete Fourier Transform

A function that describes some physical process can be Fourier analyzed, and its transform can be determined analytically. We've already been introduced to the basics of how that's done, and we'll return to elaborate the effort in Chapter 11. But before leaving the subject, it's important to extend the ideas of Fourier analysis to situations where there are no functional representations of the data. Often one has a collection of data points or perhaps a curve created on a plotter or computer screen. In any event, the information can be digitized; that is, numbers can be associated with points on the curve at convenient intervals. To determine the frequency content of such a limited collection of data, a numerical technique known as the **discrete Fourier transform** is used. Since the treatment is computer based, it will suffice for our purposes just to understand the general scheme and be able to appreciate the results.

*T. S. Jaseja, A. Javan, and C. H. Townes, "Frequency Stability of Helium–Neon Lasers and Measurements of Length," *Phys. Rev. Letters* **10**, 165 (1963).

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

Until now we dealt with functions such as $f(x)$—representing something interesting like an electric field—that provided values for all x. Instead, suppose we have a finite number of points, N, located at 0, x_1, x_2,..., x_{N-1} and the corresponding specific values of whatever quantity is being studied: f_0, f_{x_1}, f_{x_2}, and so on. When the sample points are equally spaced by an interval x_0, they can be represented by the sequence f_0, f_{x_0}, f_{2x_0}, and so forth. In essence, each Fourier integral transform [Eq. (7.57)] is approximated by a summation that is carried out successively, point by point, over the range of the available data: $f_0, f_{x_0}, f_{2x_0}, \ldots$. Figure 7.39 depicts a hand-drawn pulse and the corresponding computer-calculated discrete Fourier transform (displayed with positive and negative frequency values, as in Fig. 7.33).

It's a straightforward business (Section 11.2.2) to extend Fourier analysis to two-dimensional functions, $f(x, y)$. For example, whereas Fig. 7.40*b* is the transform of the one-dimensional unit-square pulse in terms of the angular spatial frequency k, Fig. 7.40*d* is the transform of the two-dimensional unit-square pulse in terms of the angular spatial frequencies k_x and k_y.

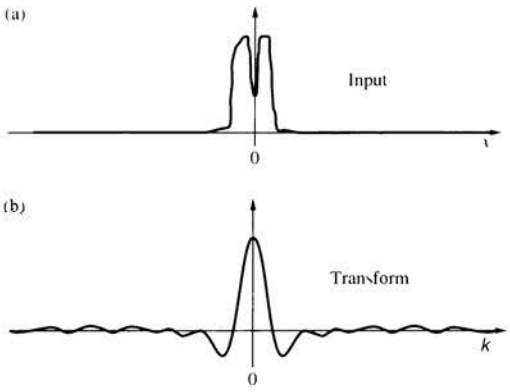
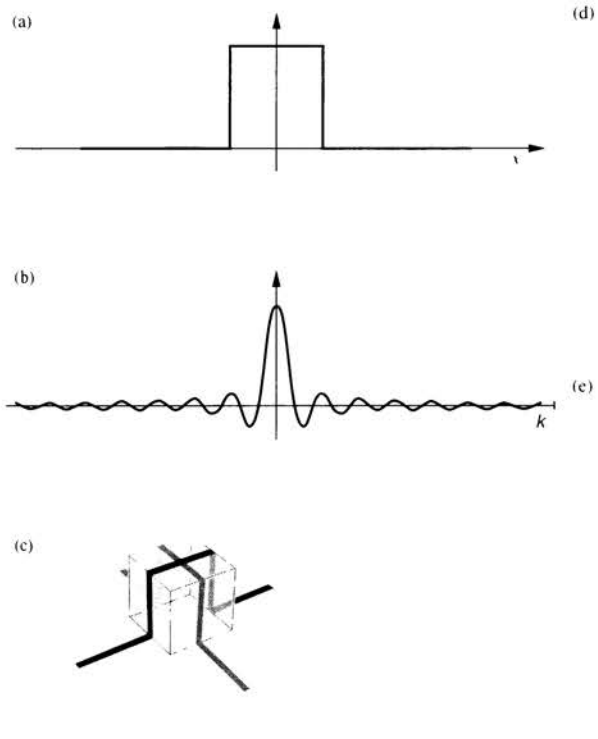


Figure 7.39 An input signal and its discrete Fourier transform.

It's natural for physicists to think about processes in relation to energy, especially if any measurements are to be made. The energy associated with a harmonic wave is proportional to the amplitude squared, and since the transform tells us the amplitudes of all the constituent sinusoids that make up the



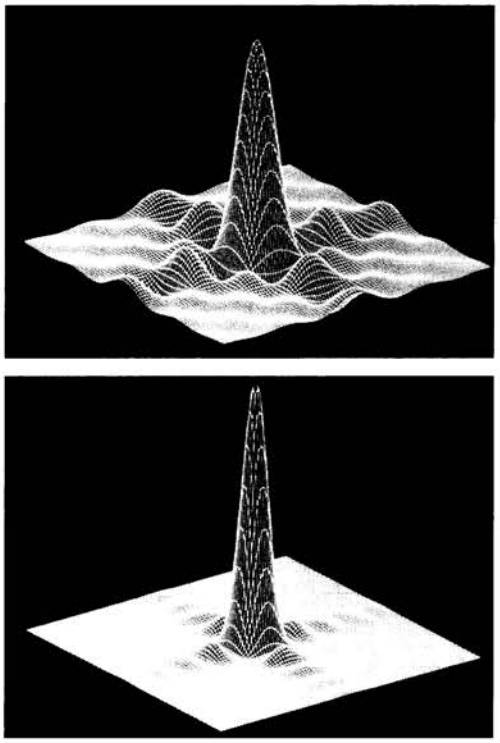


Figure 7.40
(*a*) A one-dimensional square pulse and (*b*) its transform. (*c*) A two-dimensional square pulse and (*d*) its transform. (*e*) The power spectrum of the transform in (*d*) plotted in two-dimensional *k*-space. (Photos courtesy R. G. Wilson, Illinois Wesleyan University.)

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

input signal, the square of the transform provides a measure of the distribution of energy, or power, at each and every component frequency. Consequently, the square of the transform is a function of spatial frequency called the **power spectrum**. Since the transform will most often be written as a complex quantity, the power spectrum can be defined as *the product of the transform and its complex conjugate*, given in units of W/m^{-2} or $W \cdot m^2$.

Figure 7.40*e* is a plot (in *k*-space) of the power spectrum for the two-dimensional square pulse. Notice that it is everywhere positive, which is not the case with the transform. It's clear from the power spectrum that most of the energy in the signal is associated with relatively low frequencies—the frequency increases radially out from the center of the pattern. Because the power spectrum is always positive, it's useful to plot it as a kind of spot diagram in a two-dimensional format; each point then corresponds to the contribution at a particular frequency. Later (p. 539) we'll write the transform in terms of the coordinates (Y, Z) on a distant observing screen and establish that the transform squared is identical to the irradiance distribution in the diffraction pattern on that screen. Expressed in this way, the transform squared (in units of W/m^2) can be called the *irradiance spectrum*. Although there is a mathematical distinction between the power and irradiance spectra, if you were shown an unlabeled representation of each (the former plotted in *k*-space and the latter in ordinary coordinate space), you'd be hard pressed to tell the difference.

When analytic functions are not available, similar results can be accomplished with the discrete Fourier transform. A two-dimensional field of data (e.g., the picture of the Mona Lisa in Fig. 7.41*a*) can be scanned, digitized, and the discrete transform computed. The graph of the transform of so complicated a signal is itself rather complicated, and so the power spectrum (Fig. 7.41*b*) is pictured instead. Because of the way negative frequencies were introduced, the pattern is symmetrical along any diagonal. The bright narrow central cross arises from the sharp boundary edges of the picture. (As we'll see later, the horizontal edge produces the vertical line and the vertical edge produces the horizontal line—take a look at Fig. 13.34. If the higher spatial frequency terms that carry the fine details (the ones far from center) are filtered out and the picture is reconstructed from what remains, a soft blur results (Fig. 7.41*c*). On the other hand, if the low spatial frequency terms are removed by blocking out the center of the transform, the high frequencies that remain will result in a sharp-edged reconstruction (Fig. 7.41*d*).

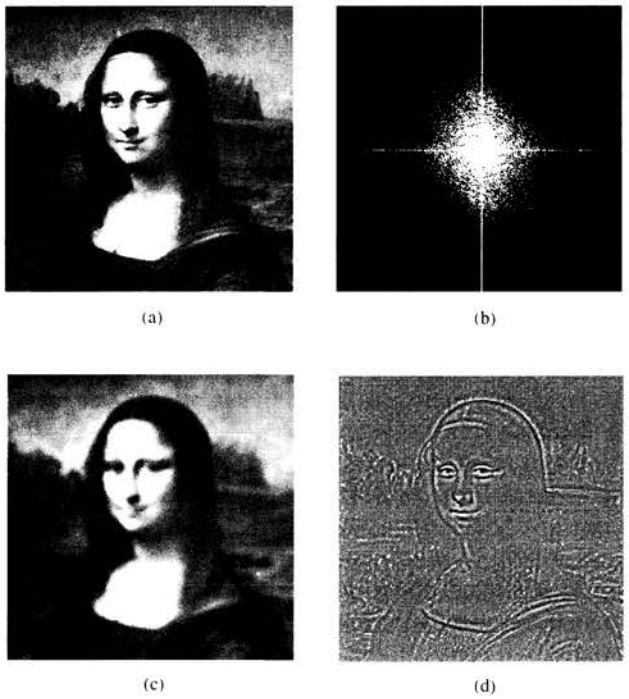


Figure 7.41 (*a*) The Mona Lisa and (*b*) the central portion of its power spectrum. (*c*) Mona, with her high spatial frequencies removed. (*d*) Mona with her low spatial frequencies removed. (Photos courtesy Synoptics Image Processing Systems, Cambridge, UK.)

The form of the elements within a given image determine its transform and therefore its power spectrum. The pictures in Fig. 7.42 were computer-created, with a vertical sinusoidal pattern superimposed in order to illustrate the point. The idea was to successively isolate several subregions of the picture, to study their transforms, and to filter them. The vertical periodic modulation forms a sinusoidal grid or **grating** that has a single fixed spatial frequency (κ_0). Its presence shows up in the computed power spectrum of any portion of the picture, essentially as two bright spots on the horizontal axis at $\pm\kappa_0$. ***Ideally, the power spectrum of a signal in the form of a sinusoidal grating is remarkably simple. It consists of just two spikes, one at plus and the other at minus the grating frequency.***

The filter (represented by the white square with two black spots) was used to create the inserted small images within each photo. It removed frequencies $+\kappa_0$ and $-\kappa_0$ from each power spectrum (the filtered versions of which are shown in the

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

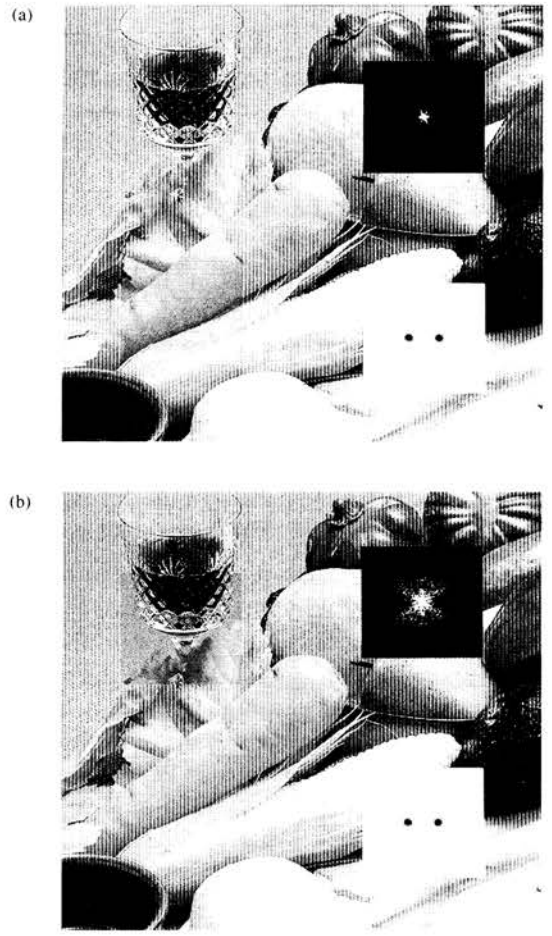


Figure 7.42 Two computer-processed images. The small inserts on the left were created by filtering out the sinusoidal modulation. The white insert represents the filter, and the black one is the filtered power spectrum in each case. (Photos courtesy MountainGate, Reno, Nevada.)

upper right). The subregion images were then reconstructed using the filtered power spectra. Each of the "cleaned up" images, sans sinusoid, was then returned to its place in the original. Notice how different those two spectra are—the facets on the cut-glass cup dominate the spectrum in Fig. 7.42*b*. Clearly, the frequency content of a picture, as spread out before us in the form of either the Fourier transform or the power spectrum, provides a wonderful new way to think about the image.

Fourier Analysis and Diffraction

A discussion of computer image analysis, which is a kind of virtual Optics, can be fascinating in its own right, but it also presages a far more fundamental aspect of diffraction, which can only be touched on in this chapter. The photographic transparency (let it be a slide of the Mona Lisa) shown in Fig. 7.43*a* is a two-dimensional record of the distribution of light

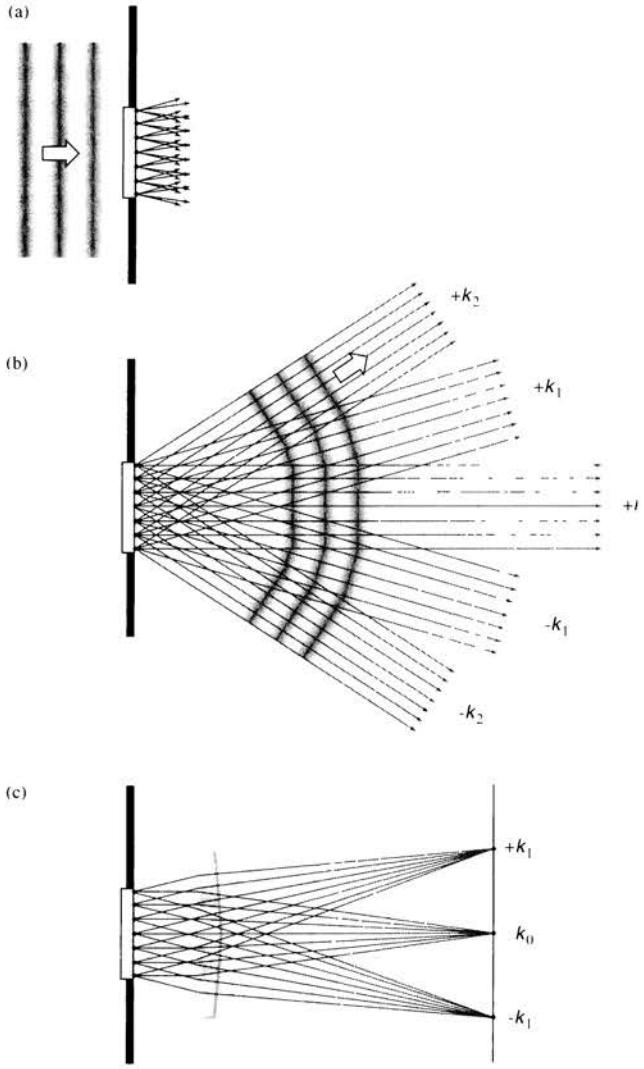


Figure 7.43 An illuminated transparency. (*a*) Incident monochromatic plane waves. (*b*) Scattered parallel bundles of rays (plane waves). (*c*) The projection of the power spectrum onto an observing screen.

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027

that once was an image of the painting. The information so stored can be read out as a signal by illuminating the slide, and that's done here with monochromatic plane waves. Every point on the surface of the slide is a scatterer, and rays emerge from it in a wide range of directions (Fig. 7.43*b*). For every plane wave going off at some angle above the axis, there is one streaming away at the same angle below the axis. Each plane wave (or parallel ray bundle) traveling in a particular κ_t-direction is a Fourier spatial frequency component. A directory of all of these component plane waves constitutes the transform of the transmitted optical field at the transparency. ***The Fourier transform of the electric field at the slide is a weighting function that gives the relative strength of each frequency component composing that field, and therefore each plane-wave stream leaving the aperture***. The sum total of all the plane waves *is* all the transmitted light and must be equivalent to the complicated Mona Lisa wavefront leaving the slide, which is also all the transmitted light.

Another nice way to envision what's happening is to suppose that every picture element with a spatial frequency along any direction in the photo plane acts like a sinusoidal grating. And every such grating essentially diffracts light into two symmetrical streams of plane waves traveling at angles proportional to the grating frequency (p. 476).

The region to the right of the slide is filled with waves that increasingly overlap as the distance from the slide increases. Nearby, the light arriving on a viewing screen would show the Mona Lisa fairly clearly, but as the screen was moved away the image would blur and change until it soon became totally unrecognizable. The region beyond the slide contains an intricate distribution of light, the diffraction pattern of the transparency. Mathematically, there are two regimes: *Fresnel diffraction*, which appears close to the aperture (i.e., the slide) and extends out to the region of *Fraunhoffer diffraction*, which comes into being very far from the aperture and goes on from there (p. 447).

If a lens is placed one focal length from the slide, as in Fig. 7.43*c*, it will cause the parallel ray bundles (which produce Fraunhoffer diffraction beyond a distance so great that it's effectively infinite) to conveniently focus on a nearby screen. There, each point of light in the resulting diagonally symmetrical, two-dimensional irradiance distribution corresponds to a specific value of spatial frequency. ***The amplitude of the electric field everywhere in the Fraunhoffer diffraction pattern corresponds to the Fourier transform of the input signal, that is, the electric-field distribution over the aperture***, although neither is measurable directly.

The observable phenomenon is the two-dimensional irradiance distribution, which is identical to the square of the Fourier transform of the input field (p. 539). It's also a map of the spatial frequency content of the Mona Lisa, and it "matches" the power spectrum pictured in Fig. 7.41*b*. As we'll see (p. 611), it's possible to spatially filter the optical transform, thereby altering the reconstructed image, just as was done via computer to produce Figs. 7.41*c* and *d*.

PROBLEMS

Complete solutions to all problems—except those with an asterisk—can be found in the back of the book.

7.1 Determine the resultant of the superposition of the parallel waves $E_1 = E_{01} \sin(\omega t + \varepsilon_1)$ and $E_2 = E_{02} \sin(\omega t + \varepsilon_2)$ when $\omega = 120\pi$, $E_{01} = 6$, $E_{02} = 8$, $\varepsilon_1 = 0$, and $\varepsilon_2 = \pi$ 2. Plot each function and the resultant.

7.2* Considering Section 7.1, suppose we began the analysis to find $E = E_1 + E_2$ with two cosine functions $E_1 = E_{01} \cos(\omega t + \alpha_1)$ and $E_2 = E_{02} \cos(\omega t + \alpha_2)$. To make things a little less complicated, let $E_{01} = E_{02}$ and $\alpha_1 = 0$. Add the two waves algebraically and make use of the familiar trigonometric identity $\cos\theta + \cos\Phi = 2\cos\frac{1}{2}(\theta + \Phi)\cos\frac{1}{2}(\theta - \Phi)$ in order to show that $E = E_0 \cos(\omega t + \alpha)$, where $E_0 = 2\,E_{01}\cos\alpha_2/2$ and $\alpha = \alpha_2/2$. Now show that these same results follow from Eqs. (7.9) and (7.10).

7.3* Show that when the two waves of Eq. (7.5) are in-phase, the resulting amplitude squared is a maximum equal to $(E_{01} + E_{02})^2$, and when they are out-of-phase it is a minimum equal to $(E_{01} - E_{02})^2$.

7.4* Show that the *optical path length*, defined as the sum of the products of the various indices times the thicknesses of media traversed by a beam, that is, $\Sigma_i n_i x_i$, is equivalent to the length of the path in vacuum that would take the same time for that beam to negotiate.

7.5 Answer the following:

(a) How many wavelengths of $\lambda_0 = 500$ nm light will span a 1-m gap in vacuum?

Giesecke+Devrient GmbH,
Giesecke+Devrient Currency Technology GmbH,
and Papierfabrik Louisenthal GmbH. - Ex. 1027