13

INTERFERENCE OF TWO BEAMS OF LIGHT

It was stated at the beginning of the last chapter that two beams of light can be made to cross each other without either one producing any modification of the other after it passes beyond the region of crossing. In this sense the two beams do not interfere with each other. However, in the region of crossing, where both beams are acting at once, we are led to expect from the considerations of the preceding chapter that the resultant amplitude and intensity may be very different from the sum of those contributed by the two beams acting separately. This modification of intensity obtained by the superposition of two or more beams of light we call *interference.* If the resultant intensity is zero or in general less than we expect from the separate intensities, we have *destructive* interference, while if it is greater, we have *constructive* interference. The phenomenon in its simpler aspects is rather difficult to observe, because of the very short wavelength of light, and therefore was not recognized as such before 1800, when the corpuscular theory of light was predominant. The first man successfully to demonstrate the interference of light, and thus establish its wave character, was Thomas Young. In order to understand his crucial experiment performed in 1801, we must first consider the application to light of an important principle which holds for any type of wave motion.

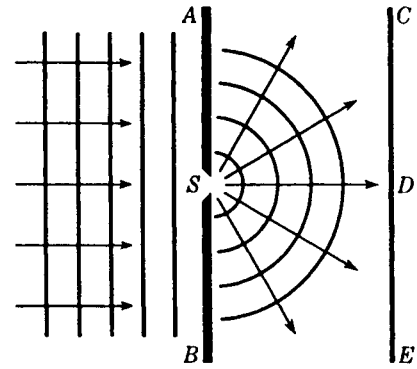


FIGURE 13A
Diffraction of waves passing through a small aperture.

13.1 HUYGENS' PRINCIPLE

When waves pass through an aperture or past the edge of an obstacle, they always spread to some extent into the region which is not directly exposed to the oncoming waves. This phenomenon is called *diffraction.* In order to explain this bending of light, Huygens nearly three centuries ago proposed the rule that *each point on a wave front may be regarded as a new source of waves.** This principle has very far-reaching applications and will be used later in discussing the diffraction of light, but we shall consider here only a very simple proof of its correctness. In Fig. 13A let a set of plane waves approach the barrier AB from the left, and let the barrier contain an opening S of width somewhat smaller than the wavelength. At all points except S the waves will be either reflected or absorbed, but S will be free to produce a disturbance behind the screen. It is found experimentally, in agreement with the above principle, that the waves spread out from S in the form of semicircles.

Huygens' principle as shown in Fig. 13A can be illustrated very successfully with water waves. An arc lamp on the floor, with a glass-bottomed tray or tank above it, will cast shadows of waves on a white ceiling. A vibrating strip of metal or a wire fastened to one prong of a tuning fork of low frequency will serve as a source of waves at one end of the tray. If an electrically driven tuning fork is used, the waves can be made apparently to stand still by placing a slotted disk on the shaft of a motor in front of the arc lamp. The disk is set rotating with the same frequency as the tuning fork to give the stroboscopic effect. This experiment can be performed for a fairly large audience and is well worth doing. Descriptions of diffraction experiments in light will be given in Chap. 15.

If the experiment in Fig. 13A is performed with light, one would naturally expect, from the fact that light generally travels in straight lines, that merely a narrow patch of light would appear at D. However, if the slit is made very narrow, an ap-

* The "waves" envisioned by Huygens were not continuous trains but a series of random pulses. Furthermore, he supposed the secondary waves to be effective only at the point of tangency to their common envelope, thus denying the possibility of diffraction. The correct application of the principle was first made by Fresnel, more than a century later.

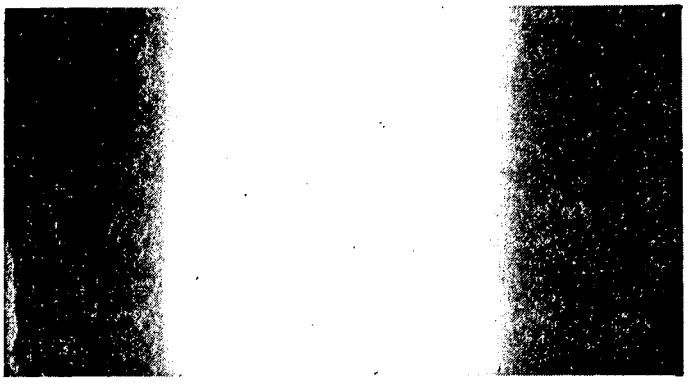


FIGURE 13B
Photograph of the diffraction of light from a slit of width 0.001 mm.

preciable broadening of this patch is observed, its breadth increasing as the slit is narrowed further. This is remarkable evidence that light does not always travel in straight lines and that waves on passing through a narrow opening spread out into a continuous fan of light rays. When the screen CE is replaced by a photographic plate, a picture like the one shown in Fig. 13B is obtained. The light is most intense in the forward direction, but its intensity decreases slowly as the angle increases. If the slit is small compared with the wavelength of light, the intensity does not come to zero even when the angle of observation becomes 90°. While this brief introduction to Huygens' principle will be sufficient for understanding the interference phenomena we are to discuss, we shall return in Chaps. 15 and 18 to a more detailed consideration of diffraction at a single opening.

13.2 YOUNG'S EXPERIMENT

The original experiment performed by Young is shown schematically in Fig. 13C. Sunlight was first allowed to pass through a pinhole S and then, at a considerable distance away, through two pinholes S_1 and S_2. The two sets of spherical waves emerging from the two holes interfered with each other in such a way as to form a symmetrical pattern of varying intensity on the screen AC. Since this early experiment was performed, it has been found convenient to replace the pinholes by narrow slits and to use a source giving monochromatic light, i.e., light of a single wavelength. In place of spherical wave fronts we now have cylindrical wave fronts, represented equally well in two dimensions by the same Fig. 13C. If the circular lines represent crests of waves, the intersections of any two lines represent the arrival at those points of two waves with the same phase or with phases differing by a multiple of 2π. Such points are therefore those of maximum disturbance or brightness. A close examination of the light on the screen will reveal evenly spaced light and dark bands or fringes, similar to those shown in Fig. 13D. Such photographs are obtained by replacing the screen AC of Fig. 13C by a photographic plate.

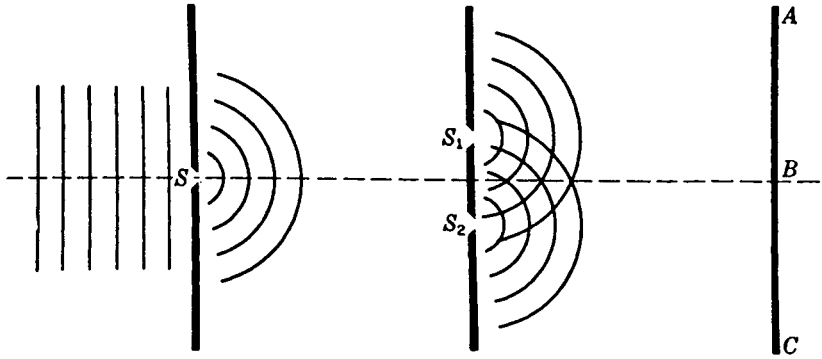


FIGURE 13C
Experimental arrangement for Young's double-slit experiment.

A very simple demonstration of Young's experiment can be accomplished in the laboratory or lecture room by setting up a single-filament lamp L (Fig. 13E) at the front of the room. The straight vertical filament S acts as the source and first slit. Double slits for each observer can be easily made from small photographic plates about 1 to 2 in. square. The slits are made in the photographic emulsion by drawing the point of a penknife across the plate, guided by a straightedge. The plates need not be developed or blackened but can be used as they are. The lamp is now viewed by holding the double slit D close to the eye E and looking at the lamp filament. If the slits are close together, for example, 0.2 mm apart, they give widely spaced fringes, whereas slits farther apart, for example, 1.0 mm, give very narrow fringes. A piece of red glass F placed adjacent to and above another of green glass in front of the lamp will show that the red waves produce wider fringes than the green, which we shall see is due to their greater wavelength.

Frequently one wishes to perform accurate experiments by using more nearly monochromatic light than that obtained by white light and a red or green glass filter. Perhaps the most convenient method is to use the sodium arc now available on the market or a mercury arc plus a filter to isolate the green line, $\lambda 5461$. A suitable filter consists of a combination of didymium glass, to absorb the yellow lines, and a light yellow glass, to absorb the blue and violet lines.

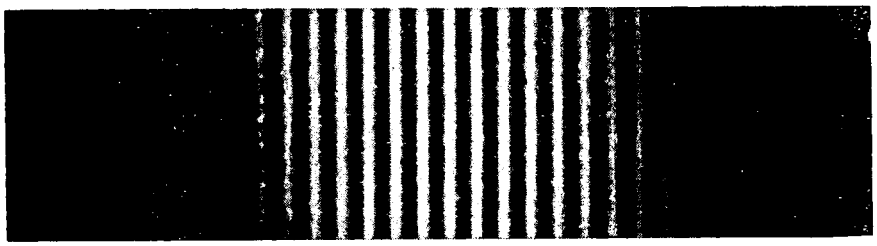


FIGURE 13D
Interference fringes produced by a double slit using the arrangement shown in Fig. 13C.

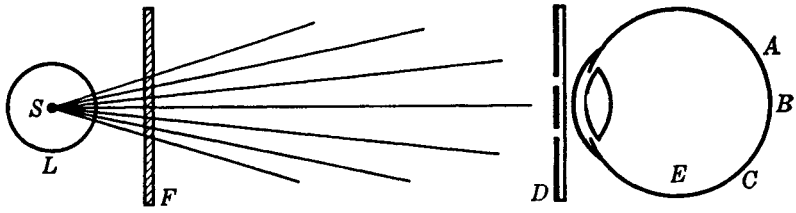


FIGURE 13E
Simple method for observing interference fringes.

13.3 INTERFERENCE FRINGES FROM A DOUBLE SOURCE

We shall now derive an equation for the intensity at any point P on the screen (Fig. 13F) and investigate the spacing of the interference fringes. Two waves arrive at P, having traversed different distances S_2P and S_1P. Hence they are superimposed with a phase difference given by

$$\delta = \frac{2\pi}{\lambda}\Delta = \frac{2\pi}{\lambda}(S_2P - S_1P) \tag{13a}$$

It is assumed that the waves start out from S_1 and S_2 in the same phase, because these slits were taken to be equidistant from the source slit S. Furthermore, the amplitudes are practically the same if (as is usually the case) S_1 and S_2 are of equal width and very close together. The problem of finding the resultant intensity at P therefore reduces to that discussed in Sec. 12.1, where we considered the addition of two simple harmonic motions of the same frequency and amplitude, but of phase difference δ. The intensity was given by Eq. (12g) as

$$I \approx A^2 = 4a^2 \cos^2 \frac{\delta}{2} \tag{13b}$$

where a is the amplitude of the separate waves and A that of their resultant.

It now remains to evaluate the phase difference in terms of the distance x on the screen from the central point P_0, the separation d of the two slits, and the distance D from the slits to the screen. The corresponding path difference is the distance S_2A in Fig. 13F, where the dashed line S_1A has been drawn to make S_1 and A equidistant from P. As Young's experiment is usually performed, D is some thousand times larger than d or x. Hence the angles θ and θ' are very small and practically equal. Under these conditions, S_1AS_2 may be regarded as a right triangle, and the path difference becomes $d \sin \theta' \approx d \sin \theta$. To the same approximation, we may set the sine of the angle equal to the tangent, so that $\sin \theta \approx x/D$. With these assumptions, we obtain

$$\Delta = d \sin \theta = d\frac{x}{D} \tag{13c}$$

This is the value of the path difference to be substituted in Eq. (13a) to obtain the phase difference δ. Now Eq. (13b) for the intensity has maximum values equal to

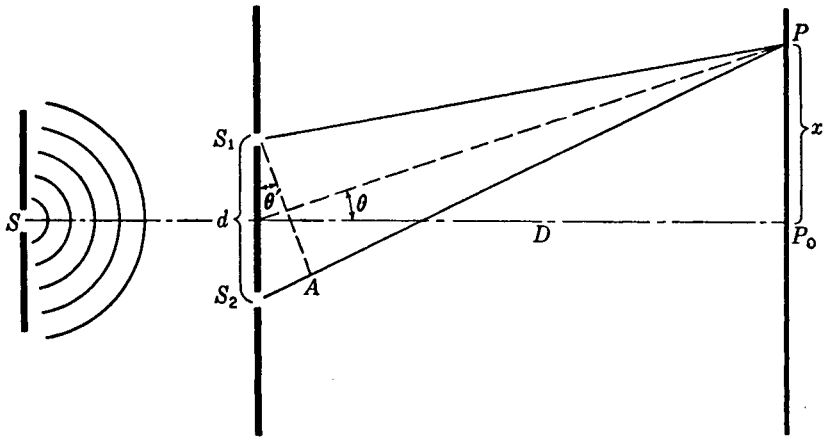


FIGURE 13F
Path difference in Young's experiment.

$4a^2$ whenever δ is an integral multiple of 2π, and according to Eq. (13a) this will occur when the path difference is an integral multiple of λ. Hence we have

$$\frac{xd}{D} = 0, \lambda, 2\lambda, 3\lambda, \ldots = m\lambda$$

or

$$x = m\lambda \frac{D}{d} \qquad \textit{Bright fringes} \qquad (13d)$$

The minimum value of the intensity is zero, and this occurs when $\delta = \pi, 3\pi, 5\pi, \ldots$. For these points

$$\frac{xd}{D} = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \ldots = \left(m + \frac{1}{2}\right)\lambda$$

or

$$x = \left(m + \frac{1}{2}\right)\lambda \frac{D}{d} \qquad \textit{Dark fringes} \qquad (13e)$$

The whole number m, which characterizes a particular bright fringe, is called the *order* of interference. Thus the fringes with $m = 0, 1, 2, \ldots$ are called the *zero, first, second,* etc., *orders.*

According to these equations the distance on the screen between two successive fringes, which is obtained by changing m by unity in either Eq. (13d) or (13e), is constant and equal to $\lambda D/d$. Not only is this equality of spacing verified by measurement of an interference pattern such as Fig. 13D, but one also finds by experiment that its magnitude *is directly proportional to the slit-screen distance D, inversely proportional to the separation of the slits d, and directly proportional to the wavelength* λ. Knowledge of the spacing of these fringes thus gives us a direct determination of λ in terms of known quantities.

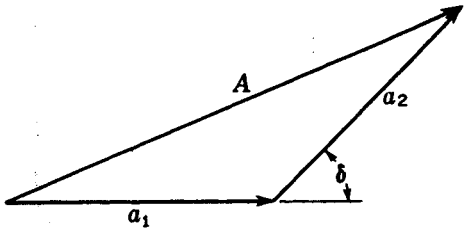


FIGURE 13G
The composition of two waves of the same frequency and amplitude but different phase.

These maxima and minima of intensity exist throughout the space behind the slits. A lens is not required to produce them, although they are usually so fine that a magnifier or eyepiece must be used to see them. Because of the approximations made in deriving Eq. (13c), careful measurements would show that, particularly in the region near the slits, the fringe spacing departs from the simple linear dependence required by Eq. (13d). A section of the fringe system in the plane of the paper of Fig. 13C, instead of consisting of a system of straight lines radiating from the midpoint between the slits, is actually a set of hyperbolas. The hyperbola, being the curve for which the difference in the distance from two fixed points is constant, obviously fits the condition for a given fringe, namely, the constancy of the path difference. Although this deviation from linearity may become important with sound and other waves, it is usually negligible when the wavelengths are as short as those of light.

13.4 INTENSITY DISTRIBUTION IN THE FRINGE SYSTEM

To find the intensity on the screen at points between the maxima, we may apply the vector method of compounding amplitudes described in Sec. 12.2 and illustrated for the present case in Fig. 13G. For the maxima, the angle δ is zero, and the component amplitudes a_1 and a_2 are parallel, so that if they are equal, the resultant $A = 2a$. For the minima, a_1 and a_2 are in opposite directions, and $A = 0$. In general, for any value of δ, A is the closing side of the triangle. The value of A^2, which measures the intensity, is then given by Eq. (13*b*) and varies according to $\cos^2(\delta/2)$. In Fig. 13H the solid curve represents a plot of the intensity against the phase difference.

In concluding our discussion of these fringes, one question of fundamental importance should be considered. If the two beams of light arrive at a point on the screen exactly out of phase, they interfere destructively and the resultant intensity is zero. One may well ask what becomes of the *energy* of the two beams, since the law of conservation of energy tells us that energy cannot be destroyed. The answer to this question is that the energy which apparently disappears at the minima actually is still present at the maxima, where the intensity is greater than would be produced by the two beams acting separately. In other words, the energy is not destroyed but merely redistributed in the interference pattern. The *average* intensity on the screen is exactly that which would exist in the absence of interference. Thus, as shown in Fig. 13H, the intensity in the interference pattern varies between $4a^2$ and zero. Now each beam acting separately would contribute a^2, and so without interference we would have a uniform intensity of $2a^2$, as indicated by the broken line. To obtain

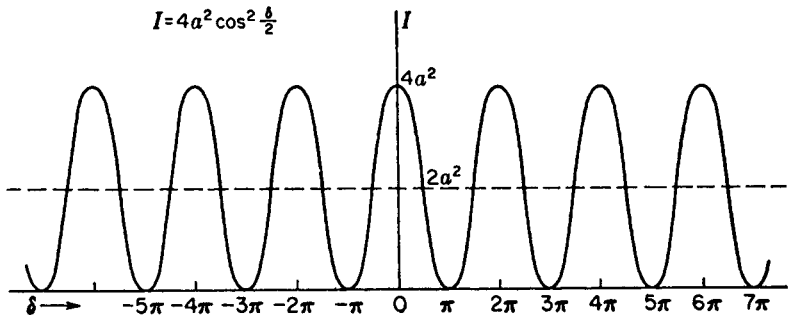


FIGURE 13H
Intensity distribution for the interference fringes from two waves of the same frequency.

the average intensity on the screen for n fringes, we note that the average value of the square of the cosine is $\frac{1}{2}$. This gives, by Eq. (13b), $I \approx 2a^2$, justifying the statement made above, and it shows that no violation of the law of conservation of energy is involved in the interference phenomenon.

13.5 FRESNEL'S BIPRISM*

Soon after the double-slit experiment was performed by Young, the objection was raised that the bright fringes he observed were probably due to some complicated modification of the light by the edges of the slits and not to true interference. Thus the wave theory of light was still questioned. Not many years passed, however, before Fresnel brought forward several new experiments in which the interference of two beams of light was proved in a manner not open to the above objection. One of these, the Fresnel biprism experiment, will be described in some detail.

A schematic diagram of the biprism experiment is shown in Fig. 13I. The thin double prism P refracts the light from the slit sources S into two overlapping beams ac and be. If screens M and N are placed as shown in the figure, interference fringes are observed only in the region bc. When the screen ae is replaced by a photographic plate, a picture like the upper one in Fig. 13J is obtained. The closely spaced fringes in the center of the photograph are due to interference, while the wide fringes at the edge of the pattern are due to diffraction. These wider bands are produced by the vertices of the two prisms, each of which acts as a straightedge, giving a pattern which will be discussed in detail in Chap. 18. When the screens M and N are removed from

* Augustin Fresnel (1788–1827). Most notable French contributor to the theory of light. Trained as an engineer, he became interested in light, and in 1814–1815 he rediscovered Young's principle of interference and extended it to complicated cases of diffraction. His mathematical investigations gave the wave theory a sound foundation.

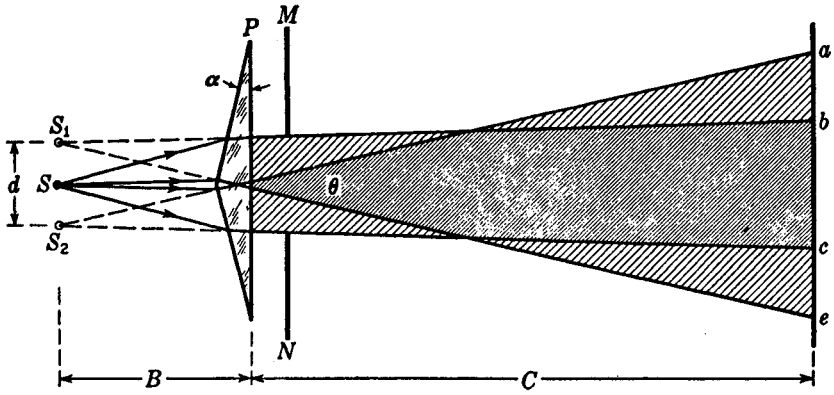


FIGURE 13I
Diagram of Fresnel's biprism experiment.

the light path, the two beams will overlap over the whole region *ae*. The lower photograph in Fig. 13J shows for this case the equally spaced interference fringes superimposed on the diffraction pattern of a wide aperture. (For the diffraction pattern above, without the interference fringes, see lowest figures in Fig. 18U.) With such an experiment Fresnel was able to produce interference without relying upon diffraction to bring the interfering beams together.

Just as in Young's double-slit experiment, the wavelength of light can be determined from measurements of the interference fringes produced by the biprism. Calling B and C the distances of the source and screen, respectively, from the prism P, d the distance between the virtual images S_1 and S_2, and Δx the distance between the successive fringes on the screen, the wavelength of the light is given from Eq. (13d) as

$$\lambda = \frac{\Delta x\, d}{B + C} \qquad (13f)$$

Thus the virtual images S_1 and S_2 act like the two slit sources in Young's experiment.

In order to find d, the linear separation of the virtual sources, one can measure their angular separation θ on a spectrometer and assume, to sufficient accuracy, that $d = B\theta$. If the parallel light from the collimator covers both halves of the biprism, two images of the slit are produced and the angle θ between these is easily measured with the telescope. An even simpler measurement of this angle can be made by holding the prism close to one eye and viewing a round frosted light bulb. At a certain distance from the light the two images can be brought to the point where their inner edges just touch. The diameter of the bulb divided by the distance from the bulb to the prism then gives θ directly.

Fresnel biprisms are easily made from a small piece of glass, such as half a microscope slide, by beveling about $\frac{1}{8}$ to $\frac{1}{4}$ in. on one side. This requires very little grinding with ordinary abrasive materials and polishing with rouge, since the angle required is only about 1°.

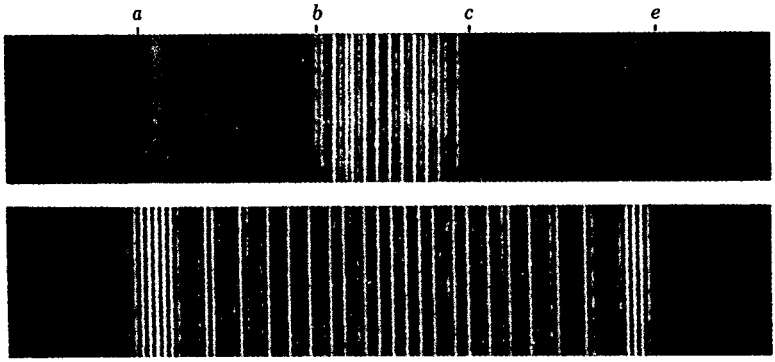


FIGURE 13J
Interference and diffraction fringes produced in the Fresnel biprism experiment.

13.6 OTHER APPARATUS DEPENDING ON DIVISION OF THE WAVE FRONT

Two beams can be brought together in other ways to produce interference. In the arrangement known as *Fresnel's mirrors*, light from a slit is reflected in two plane mirrors slightly inclined to each other. The mirrors produce two virtual images of the slit, as shown in Fig. 13K. They act in every respect like the images formed by the biprism, and interference fringes are observed in the region *bc*, where the reflected beams overlap. The symbols in this diagram correspond to those in Fig. 13I, and Eq. (13f) is again applicable. It will be noted that the angle 2θ subtended at the point of intersection M by the two sources is twice the angle between the mirrors.

The Fresnel double-mirror experiment is usually performed on an optical bench, with the light reflected from the mirrors at nearly grazing angles. Two pieces of ordinary plate glass about 2 in. square make a very good double mirror. One plate should have an adjusting screw for changing the angle θ and the other a screw for making the edges of the two mirrors parallel.

An even simpler device, shown in Fig. 13L, produces interference between the light reflected in one long mirror and the light coming directly from the source without reflection. In this arrangement, known as *Lloyd's mirror*, the quantitative relations are similar to those in the foregoing cases, with the slit and its virtual image constituting the double source. An important feature of the Lloyd's-mirror experiment lies in the fact that when the screen is placed in contact with the end of the mirror (in the position *MN*, Fig. 13L), the edge O of the reflecting surface comes at the center of a *dark* fringe, instead of a bright one as might be expected. This means that one of the two beams has undergone a phase change of π. Since the direct beam could not change phase, this experimental observation is interpreted to mean that the reflected light has changed phase at reflection. Two photographs of the Lloyd's-mirror fringes taken in this way are reproduced in Fig. 13M, one taken with visible light and the other with X rays.

If the light from source S_1 in Fig. 13L is allowed to enter the end of the glass

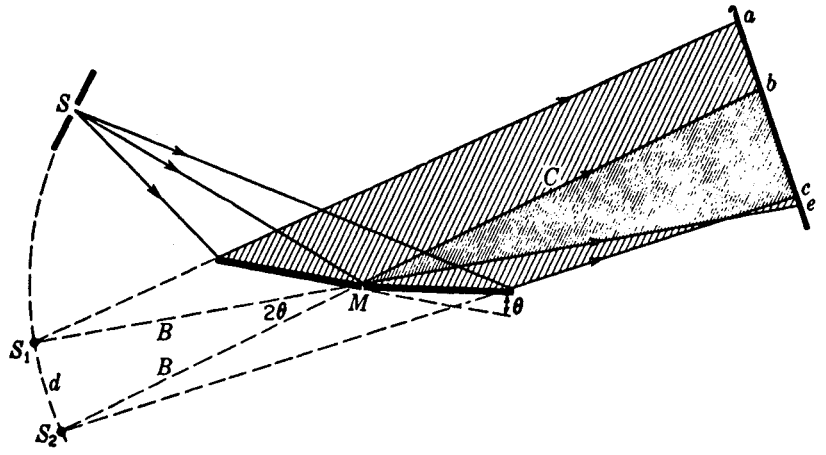


FIGURE 13K
Geometry of Fresnel's mirrors.

plate by moving the latter up, and to be internally reflected from the upper glass surface, fringes will again be observed in the interval OP, with a dark fringe at O. This shows that there is again a phase change of π at reflection. As will be shown in Chap. 25, this is not a contradiction of the discussion of phase change given in Sec. 14.1. In this instance the light is incident at an angle greater than the critical angle for total reflection.

Lloyd's mirror is readily set up for demonstration purposes as follows. A carbon arc, followed by a colored glass filter and a narrow slit, serves as a source. A strip of ordinary plate glass 1 to 2 in. wide and 1 ft or more long makes an excellent mirror. A magnifying glass focused on the far end of the mirror enables one to observe the fringes shown in Fig. 13M. Internal fringes can be observed by polishing the ends of

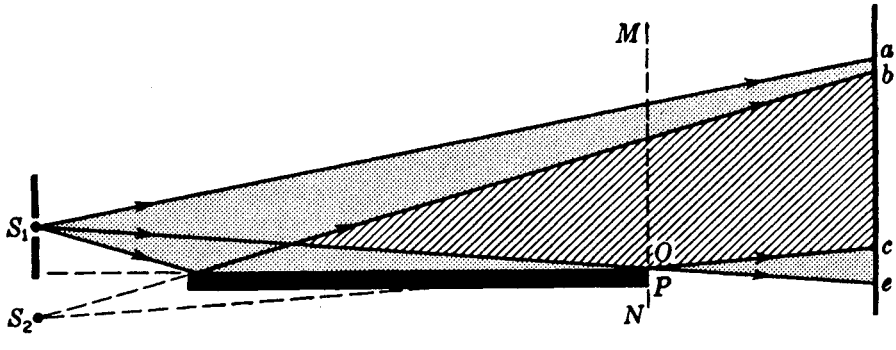
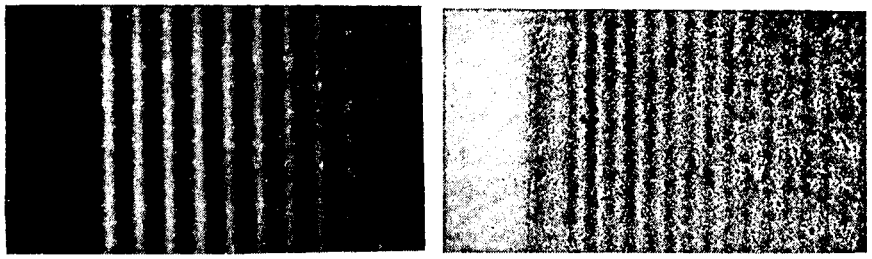


FIGURE 13L
Lloyd's mirror.



(*a*) (*b*)

FIGURE 13M
Interference fringes produced with Lloyd's mirror. (*a*) Taken with visible light, $\lambda = 4358$ Å. (*After White.*) (*b*) Taken with X rays, $\lambda = 8.33$ Å. (*After Kellstrom.*)

the mirror to allow the light to enter and leave the glass, and by roughening one of the glass faces with coarse emery.

Other ways exist* for dividing the wave front into two segments and subsequently recombining these at a small angle with each other. For example, one can cut a lens into two halves on a plane through the lens axis and separate the parts slightly, to form two closely adjacent real images of a slit. The images produced in this device, known as *Billet's split lens*, act like the two slits in Young's experiment. A single lens followed by a biplate (two plane-parallel plates at a slight angle) will accomplish the same result.

13.7 COHERENT SOURCES

It will be noticed that the various methods of demonstrating interference so far discussed have one important feature in common: the two interfering beams are always derived from the same source of light. We find by experiment that it is impossible to obtain interference fringes from two separate sources, such as two lamp filaments set side by side. This failure is caused by the fact that the light from any one source is not an infinite train of waves. On the contrary, there are sudden changes in phase occurring in very short intervals of time (of the order of 10^{-8} s). This point has already been mentioned in Secs. 11.1 and 12.6. Thus, although interference fringes may exist on the screen for such a short interval, they will shift their position each time there is a phase change, with the result that no fringes at all will be seen. In Young's experiment and in Fresnel's mirrors and biprism, the two sources S_1 and S_2 always have a point-to-point correspondence of phase, since they are both derived from the same source. If the phase of the light from a point in S_1 suddenly shifts, that of the light from the corresponding point in S_2 will shift simultaneously. The result is that the *difference* in phase between any pair of points in the two sources always remains constant, and so the interference fringes are stationary. It is a charac-

* Good descriptions will be found in T. Preston, "Theory of Light," 5th ed., chap. 7, The Macmillan Company, New York, 1928.

teristic of any interference experiment with light that the sources must have this point-to-point phase relation, and sources that have this relation are called *coherent sources.*

While special arrangements are necessary for producing coherent sources of light, the same is not true of *microwaves,* which are radio waves of a few centimeters wavelength. These are produced by an oscillator which emits a continuous wave, the phase of which remains constant over a time long compared with the duration of an observation. Two independent microwave sources of the same frequency are therefore coherent and can be used to demonstrate interference. Because of the convenient magnitude of their wavelength, microwaves are used to illustrate many common optical interference and diffraction effects.*

If in Young's experiment the source slit S (Fig. 13C) is made too wide or the angle between the rays which leave it too large, the double slit no longer represents two coherent sources and the interference fringes disappear. This subject will be discussed in more detail in Chap. 16.

13.8 DIVISION OF AMPLITUDE. MICHELSON† INTERFEROMETER

Interference apparatus may be conveniently divided into two main classes, those based on *division of wave front* and those based on *division of amplitude.* The previous examples all belong to the former class, in which the wave front is divided laterally into segments by mirrors or diaphragms. It is also possible to divide a wave by partial reflection, the two resulting wave fronts maintaining the original width but having reduced amplitudes. The Michelson interferometer is an important example of this second class. Here the two beams obtained by amplitude division are sent in quite different directions against plane mirrors, whence they are brought together again to form interference fringes. The arrangement is shown schematically in Fig. 13N. The main optical parts consist of two highly polished plane mirrors M_1 and M_2 and two plane-parallel plates of glass G_1 and G_2. Sometimes the rear side of the plate G_1 is lightly silvered (shown by the heavy line in the figure) so that the light coming from the source S is divided into (1) a reflected and (2) a transmitted beam of equal intensity. The light reflected normally from mirror M_1 passes through G_1 a third time and reaches the eye as shown. The light reflected from the mirror M_2 passes back through G_2 for the second time, is reflected from the surface of G_1 and into the

* The technique of such experiments is discussed by G. F. Hull, Jr., *Am. J. Phys.*, 17:599 (1949).

† A. A. Michelson (1852–1931). American physicist of genius. He early became interested in the velocity of light and began experiments while an instructor in physics and chemistry at the Naval Academy, from which he graduated in 1873. It is related that the superintendent of the Academy asked young Michelson why he wasted his time on such useless experiments. Years later Michelson was awarded the Nobel prize (1907) for his work on light. Much of his work on the speed of light (Sec. 19.3) was done during 10 years spent at the Case Institute of Technology. During the latter part of his life he was professor of physics at the University of Chicago, where many of his famous experiments on the interference of light were done.

14

INTERFERENCE INVOLVING MULTIPLE REFLECTIONS

Some of the most beautiful effects of interference result from the multiple reflection of light between the two surfaces of a thin film of transparent material. These effects require no special apparatus for their production or observation and are familiar to anyone who has noticed the colors shown by thin films of oil on water, by soap bubbles, or by cracks in a piece of glass. We begin our investigation of this class of interference by considering the somewhat idealized case of reflection and refraction from the boundary separating different optical media. In Fig. 14A(*a*) a ray of light in air or vacuum incident on a plane surface of a transparent medium like water is indicated by *a*. The reflected and refracted rays are indicated by *ar* and *at*, respectively.

A question of particular interest from the standpoint of physical optics is that of a possible abrupt *change of phase* of waves when they are reflected from a boundary. For a given boundary the result will differ, as we shall now show, according to whether the waves approach from the side of higher velocity or from that of lower velocity. Thus, let the symbol *a* in the left-hand part of Fig. 14A represent the amplitude (not the intensity) of a set of waves striking the surface, let *r* be the fraction of the amplitude reflected, and let *t* be the fraction transmitted. The amplitudes of the two sets of waves will then be *ar* and *at*, as shown. Now, following a treatment given by

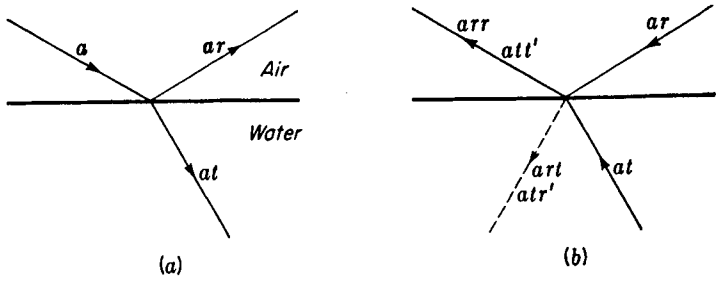


FIGURE 14A
Stokes' treatment of reflection.

Stokes*, imagine the two sets reversed in direction, as in part (*b*) of the figure. Provided there is no dissipation of energy by absorption, a wave motion is a strictly reversible phenomenon. It must conform to the law of mechanics known as the *principle of reversibility*, according to which the result of an instantaneous reversal of all the velocities in a dynamic system is to cause the system to retrace its whole previous motion. That the paths of light rays are in conformity with this principle has already been stated in Sec. 1.8. The two reversed trains, of amplitude *ar* and *at*, should accordingly have as their net effect after striking the surface a wave in air equal in amplitude to the incident wave in part (*a*) but traveling in the opposite direction. The wave of amplitude *ar* gives a reflected wave of amplitude *arr* and a refracted wave of amplitude *art*. If we call r' and t' the fractions of the amplitude reflected and refracted when the reversed wave *at* strikes the boundary from below, this contributes amplitudes *att'* and *atr'* to the two waves, as indicated. Now, since the resultant effect must consist only of a wave in air of amplitude *a*, we have

$$att' + arr = a \qquad (14a)$$

and

$$art + atr' = 0 \qquad (14b)$$

The second equation states that the two incident waves shall produce no net disturbance on the water side of the boundary. From Eq. (14a) we obtain

$$tt' = 1 - r^2 \qquad (14c)$$

and from Eq. (14b)

$$r' = -r \qquad (14d)$$

It might at first appear that Eq. (14c) could be carried further by using the fact that intensities are proportional to squares of amplitudes and by writing, by conservation of energy, $r^2 + t^2 = 1$. This would immediately yield $t = t'$. The result is not correct, however, for two reasons: (1) although the proportionality of intensity with square of amplitude holds for light traveling in a single medium, passage into

* Sir George Stokes (1819–1903), versatile mathematician and physicist of Pembroke College, Cambridge, and pioneer in the study of the interaction of light with matter. He is known for his laws of fluorescence (Sec. 22.6) and of the rate of fall of spheres in viscous fluids. The treatment referred to here was given in his "Mathematical and Physical Papers," vol. 2, pp. 89ff., especially p. 91.

a different medium brings in the additional factor of the index of refraction in determining the intensity; (2) it is not to the intensities that the conservation law is to be applied but to the total energies of the beams. When there is a change in width of the beam, as in refraction, it must also be taken into account.

The second of Stokes' relations, Eq. (14d), shows that the *reflectance*, or fraction of the intensity reflected, is the same for a wave incident from either side of the boundary, since the negative sign disappears upon squaring the amplitudes. It should be noted, however, that the waves must be incident at angles such that they correspond to angles of incidence and refraction. The difference in sign of the amplitudes in Eq. (14d) indicates a difference of phase of π between the two cases, since a reversal of sign means a displacement in the opposite sense. If there is no phase change on reflection from above, there must be a phase change of π on reflection from below; or correspondingly, if there is no change on reflection from below, there must be a change of π on reflection from above.

The principle of reversibility as applied to light waves is often useful in optical problems; for example, it proves at once the interchangeability of object and image. The conclusion reached above about the change of phase is not dependent on the applicability of the principle, i.e., on the absence of absorption, but holds for reflection from any boundary. It is a matter of experimental observation that in the reflection of light under the above conditions, the phase change of π occurs when the light strikes the boundary from the side of higher velocity,* so that the second of the two alternatives mentioned is the correct one in this case. A change of phase of the same type is encountered in the reflection of simple mechanical waves, such as transverse waves in a rope. Reflection with change of phase where the velocity decreases in crossing the boundary corresponds to the reflection of waves from a fixed end of a rope. Here the elastic reaction of the fixed end of the rope immediately produces a reflected train of opposite phase traveling back along the rope. The case where the velocity increases in crossing the boundary has its parallel in reflection from a free end of a rope. The end of the rope undergoes a displacement of twice the amount it would have if the rope were continuous, and it immediately starts a wave in the reverse direction having the same phase as the incident wave.

14.1 REFLECTION FROM A PLANE-PARALLEL FILM

Let a ray of light from a source S be incident on the surface of such a film at A (Fig. 14B). Part of this will be reflected as ray 1 and part refracted in the direction AF. Upon arrival at F, part of the latter will be reflected to B and part refracted toward H. At B the ray FB will be again divided. A continuation of this process yields two sets of parallel rays, one on each side of the film. In each of these sets, of course, the intensity decreases rapidly from one ray to the next. If the set of parallel reflected rays is now collected by a lens and focused at the point P, each ray will have traveled a different distance, and the phase relations may be such as to produce destructive or constructive interference at that point. It is such interference that produces the

* See the discussion in Sec. 13.6 in connection with Lloyd's mirror.

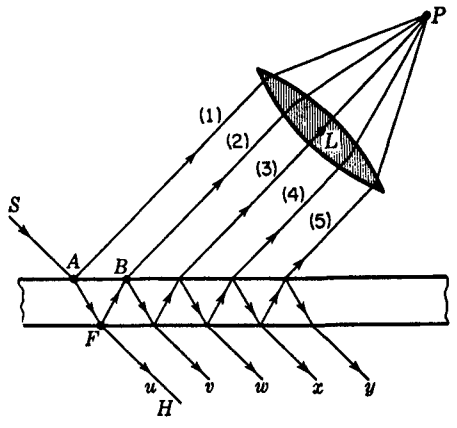


FIGURE 14B
Multiple reflections in a plane-parallel film.

colors of thin films when they are viewed by the naked eye. In such a case L is the lens of the eye, and P lies on the retina.

In order to find the phase difference between these rays, we must first evaluate the difference in the optical path traversed by a pair of successive rays, such as rays 1 and 2. In Fig. 14C let d be the thickness of the film, n its index of refraction, λ the wavelength of the light, and ϕ and ϕ' the angles of incidence and refraction. If BD is perpendicular to ray 1, the optical paths from D and B to the focus of the lens will be equal. Starting at A, ray 2 has the path AFB in the film and ray 1 the path AD in air. The difference in these optical paths is given by

$$\Delta = n(AFB) - AD$$

If BF is extended to intersect the perpendicular line AE at G, $AF = GF$ because of the equality of the angles of incidence and reflection at the lower surface. Thus we have

$$\Delta = n(GB) - AD = n(GC + CB) - AD$$

Now AC is drawn perpendicular to FB; so the broken lines AC and DB represent two successive positions of a wave front reflected from the lower surface. The optical paths must be the same by any ray drawn between two wave fronts; so we may write

$$n(CB) = AD$$

The path difference then reduces to

$$\Delta = n(GC) = n(2d \cos \phi') \qquad (14e)$$

If this path difference is a whole number of wavelengths, we might expect rays 1 and 2 to arrive at the focus of the lens in phase with each other and produce a maximum of intensity. However, we must take account of the fact that ray 1 undergoes a phase change of π at reflection, while ray 2 does not, since it is internally reflected. The condition

$$2nd \cos \phi' = m\lambda \qquad \textit{Minima} \qquad (14f)$$

then becomes a condition for *destructive* interference as far as rays 1 and 2 are concerned. As before, $m = 0, 1, 2, \ldots$ is the order of interference.

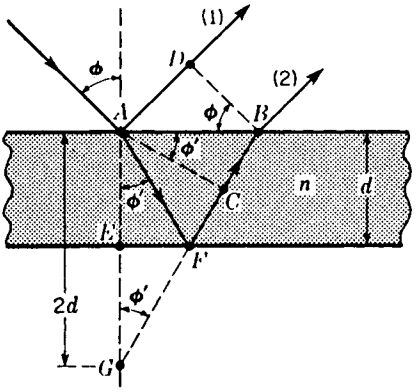


FIGURE 14C
Optical-path difference between two consecutive rays in multiple reflection (see Fig. 14A).

Next we examine the phases of the remaining rays, 3, 4, 5, ... Since the geometry is the same, the path difference between rays 3 and 2 will also be given by Eq. (14e), but here there are only internal reflections involved, so that if Eq. (14f) is fulfilled, ray 3 will be in the same phase as ray 2. The same holds for all succeeding pairs, and so we conclude that under these conditions rays 1 and 2 will be out of phase, but rays 2, 3, 4, ..., will be in phase with each other. On the other hand, if conditions are such that

$$2nd \cos \phi' = (m + \tfrac{1}{2})\lambda \qquad \textit{Maxima} \qquad (14g)$$

ray 2 will be in phase with 1, but 3, 5, 7, ... will be out of phase with 2, 4, 6, ... Since 2 is more intense than 3, 4 more intense than 5, etc., these pairs cannot cancel each other, and since the stronger series combines with 1, the strongest of all, there will be a maximum of intensity.

For the minima of intensity, ray 2 is out of phase with ray 1, but 1 has a considerably greater amplitude than 2, so that these two will not completely annul each other. We can now prove that the addition of 3, 4, 5, ..., which are all in phase with 2, will give a net amplitude just sufficient to make up the difference and to produce complete darkness at the minima. Using a for the amplitude of the incident wave, r for the fraction of this reflected, and t or t' for the fraction transmitted in going from rare to dense or dense to rare, as was done in Stokes' treatment of reflection, Fig. 14D is constructed and the amplitudes labeled as shown. In accordance with Eq. (14d), we have taken the fraction reflected internally and externally to be the same. Adding the amplitudes of all the reflected rays but the first on the upper side of the film, we obtain the resultant amplitude,

$$\begin{aligned} A &= atrt' + atr^3t' + atr^5t' + atr^7t' + \cdots \\ &= atrt'(1 + r^2 + r^4 + r^6 + \cdots) \end{aligned}$$

Since r is necessarily less than 1, the geometrical series in parentheses has a finite sum equal to $1/(1 - r^2)$, giving

$$A = atrt' \frac{1}{1 - r^2}$$

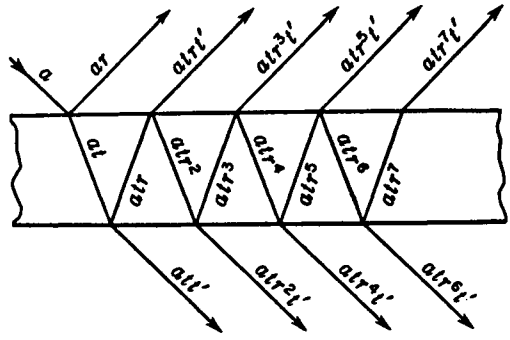


FIGURE 14D
Amplitudes of successive rays in multiple reflection.

But from Stokes' treatment, Eq. (14c), $tt' = 1 - r^2$; so we obtain finally

$$A = ar \qquad (14h)$$

This is just equal to the amplitude of the first reflected ray, so we conclude that under the conditions of Eq. (14f) there will be complete destructive interference.

14.2 FRINGES OF EQUAL INCLINATION

If the image of an *extended* source reflected in a thin plane-parallel film is examined, it will be found to be crossed by a system of distinct interference fringes, provided the source emits monochromatic light and provided the film is sufficiently thin. Each bright fringe corresponds to a particular path difference giving an integral value of m in Eq. (14g). For any fringe, the value of ϕ is fixed; so the fringe will have the form of the arc of a circle whose center is at the foot of the perpendicular drawn from the eye to the plane of the film. Evidently we are here concerned with fringes of equal inclination, and the equation for the path difference has the same form as for the circular fringes in the Michelson interferometer (Sec. 13.9).

Note that if m is the order of interference for light incident on the film at $\phi = 0°$, Eq. (14f) gives

$$m = \frac{2nd}{\lambda}$$

which would be a dark fringe. Since the path difference for the first, second, and third, etc., bright fringes will be at progressively larger angles of ϕ and ϕ' [Eq. (14g)], the successive path differences, $2nd \cos \phi'$, will be successively shorter and bright-light fringes will be at angles where $2nd \cos \phi'$ is equal $(m - \frac{1}{2})\lambda$, $(m - \frac{3}{2})\lambda$, $(m - \frac{5}{2})\lambda$, etc.

The necessity of using an extended source will become clear upon consideration of Fig. 14B. If a very distant point source S is used, the parallel rays will necessarily reach the eye at only one angle (that required by the law of reflection) and will be

focused to a point P. Thus only one point will be seen, either bright or dark, according to the phase difference at this particular angle. It is true that if the source is not very far away, its image on the retina will be slightly blurred, because the eye must be focused for parallel rays to observe the interference. The area illuminated is small, however, and in order to see an extended system of fringes, we must obviously have many points S, spread out in a broad source so that the light reaches the eye from various directions.

These fringes are seen by the eye only if the film is very thin, unless the light is reflected practically normal to the film. At other angles, since the pupil of the eye has a small aperture, increasing the thickness of the film will cause the reflected rays to get so far apart that only one enters the eye at a time. Obviously no interference can occur under these conditions. Using a telescope of large aperture, the lens may include enough rays for the fringes to be visible with thick plates, but unless viewed nearly normal to the plate, they will be so finely spaced as to be invisible. The fringes seen with thick plates near normal incidence are often called *Haidinger* fringes.*

14.3 INTERFERENCE IN THE TRANSMITTED LIGHT

The rays emerging from the lower side of the film, shown in Fig. 14B and 14D, can also be brought together with a lens and caused to interfere. Here, however, there are no phase changes at reflection for any of the rays, and the relations are such that Eq. (14f) now becomes the condition for maxima and Eq. (14g) the condition for minima. For maxima, the rays $u\ v,\ w, \ldots$ of Fig. 14B are all in phase, while for minima, $v,\ x, \ldots$ are out of phase with $u,\ w, \ldots$. When the reflectance r^2 has a low value, as with the surfaces of unsilvered glass, the amplitude of u is much the greatest in the series and the minima are not by any means black. Figure 14E shows quantitative curves for the intensity transmitted I_T and reflected I_R plotted in this instance for $r = 0.2$ according to Eqs. (14n) and (14o), ahead. The corresponding reflectance of 4 percent is close to that of glass at normal incidence. The abscissas δ in the figure represent the phase difference between successive rays in the transmitted set or between all but the first pair in the reflected set, which by Eq. (14e) is

$$\delta = k\Delta = \frac{2\pi}{\lambda}\Delta = \frac{4\pi}{\lambda} nd \cos \phi' \qquad (14i)$$

It will be noted that the curve for I_R looks very much like the $\cos^2$ contour obtained from the interference of two beams. It is not exactly the same, however, and the resemblance holds only when the reflectance is small. Then the first two reflected beams are so much stronger than the rest that the latter have little effect. The important changes that come in at higher values of the reflectance will be discussed in Sec. 14.7.

* W. K. von Haidinger (1795–1871). Austrian mineralogist and geologist, for 17 years director of the Imperial Geological Institute in Vienna.

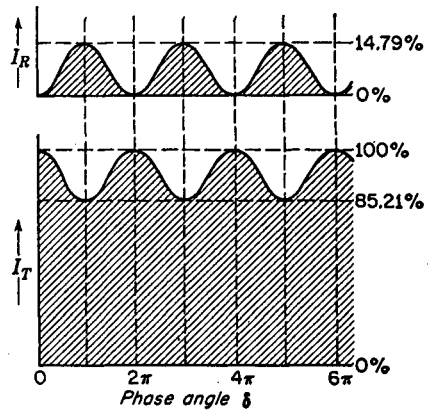


FIGURE 14E
Intensity contours of the reflected and transmitted fringes from a film having a reflectance of 4 percent.

14.4 FRINGES OF EQUAL THICKNESS

If the film is not plane-parallel, so that the surfaces make an appreciable angle with each other as in Fig. 14F(*a*), the interfering rays do not enter the eye parallel to each other but appear to diverge from a point near the film. The resulting fringes resemble the localized fringes in the Michelson interferometer and appear to be formed in the film itself. If the two surfaces are plane, so that the film is wedge-shaped, the fringes will be practically straight, following the lines of equal thickness. In this case the path difference for a given pair of rays is practically that given by Eq. (14e). Provided that observations are made almost normal to the film, the factor cos ϕ' may be considered equal to 1, and the condition for bright fringes becomes

$$2nd = (m + \tfrac{1}{2})\lambda \qquad (14j)$$

In going from one fringe to the next m increases by 1, and this requires the optical thickness of the film nd to change by $\lambda/2$.

Fringes formed in thin films are easily shown in the laboratory or lecture room by using two pieces of ordinary plate glass. If they are laid together with a thin strip of paper along one edge, we obtain a wedge-shaped film of air between the plates. When a sodium flame or arc is viewed as in Fig. 14F, yellow fringes are clearly seen. If a carbon arc and filter are used, the fringes may be projected on a screen with a lens. On viewing the reflected image of a monochromatic source, one will find it to be crossed by more or less straight fringes, such as those in Fig. 14F(*b*).

This class of fringes has important practical applications in the testing of optical surfaces for planeness. If an air film is formed between two surfaces, one of which is perfectly plane and the other not, the fringes will be irregular in shape. Any fringe is characterized by a particular value of m in Eq. (14j), and hence will follow those parts of the film where d is constant. That is, the fringes form the equivalent of *contour lines* for the uneven surface. The contour interval is $\lambda/2$, since for air $n = 1$, and going from one fringe to the next corresponds to increasing d by this amount. The standard method of producing optically plane surfaces uses repeated observation of the fringes

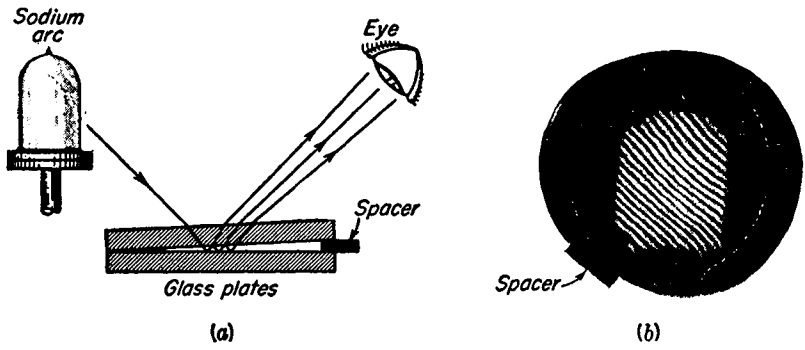


FIGURE 14F
Fringes of equal thickness: (*a*) method of visual observation; (*b*) photograph taken with a camera focused on the plates.

formed between the working surface and an optical flat, the polishing being continued until the fringes are straight. In Fig. 14F(*b*) it will be noticed that there is considerable distortion of one of the plates near the bottom.

14.5 NEWTON'S RINGS

If the fringes of equal thickness are produced in the air film between a convex surface of a long-focus lens and a plane glass surface, the contour lines will be circular. The ring-shaped fringes thus produced were studied in detail by Newton,* although he was not able to explain them correctly. For purposes of measurement, the observations are usually made at normal incidence by an arrangement such as that in Fig. 14G, where the glass plate G reflects the light down on the plates. After reflection, it is transmitted by G and observed in the low-power microscope T. Under these conditions the positions of the maxima are given by Eq. (14j), where d is the thickness of the air film. Now if we designate by R the radius of curvature of the surface A and assume that A and B are just touching at the center, the value of d for any ring of radius r is the sagitta of the arc, given by

$$d = \frac{r^2}{2R} \qquad (14k)$$

Substitution of this value in Eq. (14j) will then give a relation between the radii of the rings and the wavelength of the light. For quantitative work, one may not assume

* Sir Isaac Newton (1642–1727). Besides laying foundations of the science of mechanics, Newton devoted considerable time to the study of light and embodied the results in his famous "Opticks." It seems strange that one of the most striking demonstrations of the interference of light, Newton's rings, should be credited to the chief proponent of the corpuscular theory of light. Newton's advocacy of the corpuscular theory was not so uncompromising as it is generally represented. This is evident to anyone consulting his original writings. The original discovery of Newton's rings is now attributed to Robert Hooke.

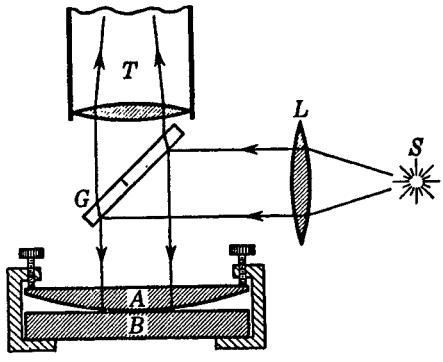


FIGURE 14G
Experimental arrangement used in viewing and measuring Newton's rings.

the plates to barely touch at the point of contact, since there will always be either some dust particles or distortion by pressure. Such disturbances will merely add a small constant to Eq.(14k), however, and their effect can be eliminated by measuring the diameters of at least two rings.

Because the ring diameters depend on wavelength, white light will produce only a few colored rings near the point of contact. With monochromatic light, however, an extensive fringe system such as that shown in Fig. 14H is observed. When the contact is perfect, the central spot is black. This is direct evidence of the relative phase change of π between the two types of reflection, air-to-glass and glass-to-air, mentioned in Sec. 14.1. If there were no such phase change, the rays reflected from the two surfaces in contact should be in the same phase and produce a bright spot at the center. In an interesting modification of the experiment, due to Thomas Young, the lower plate has a higher index of refraction than the lens, and the film between is filled with an oil of intermediate index. Then both reflections are at "rare-to-dense" surfaces, no relative phase change occurs, and the central fringe of the reflected system is bright. The experiment does not tell us at which surface the phase change in the ordinary arrangement occurs, but it is now definitely known (Sec. 25.4) that it occurs at the lower (air-to-glass) surface.

A ring system is also observed in the light transmitted by the Newton's-ring plates. These rings are exactly complementary to the reflected ring system, so that the center spot is now bright. The contrast between bright and dark rings is small, for reasons already discussed in Sec. 14.3.

14.6 NONREFLECTING FILMS

A simple and very important application of the principles of interference in thin films has been the production of *coated* surfaces. If a film of a transparent substance of refractive index n' is deposited on glass of a larger index n to a thickness of one-quarter of the wavelength of light in the film, so that

$$d = \frac{\lambda}{4n'}$$

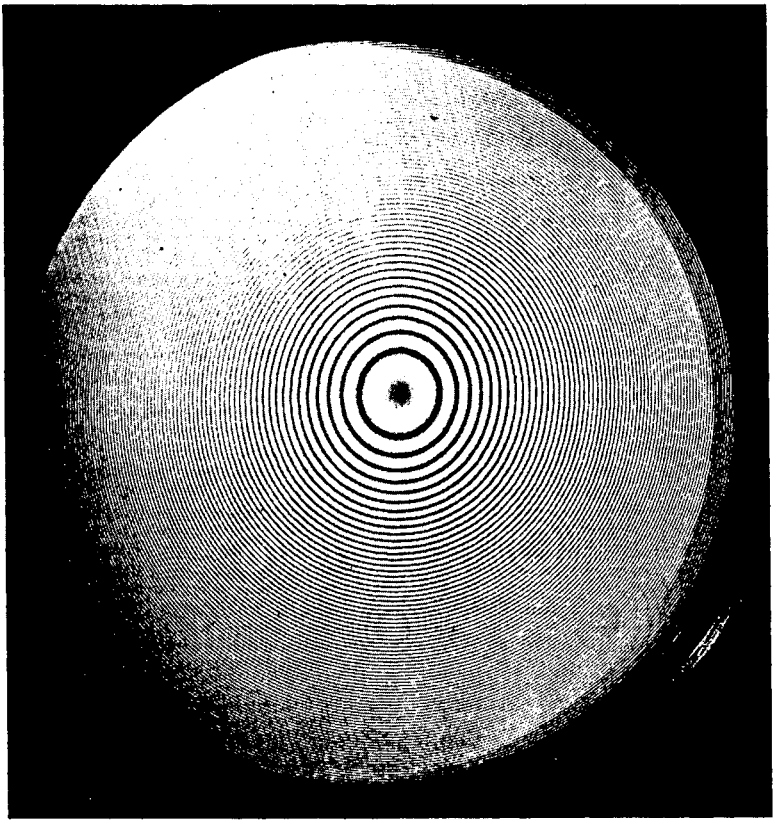


FIGURE 14H
Newton's rings. (*By permission from Bausch & Lomb Incorporated.*)

the light reflected at normal incidence is almost completely suppressed by interference. This corresponds to the condition $m = 0$ in Eq. (14g), which here becomes a condition for *minima* because the reflections at both surfaces are "rare-to-dense." The waves reflected from the lower surface have an extra path of one-half wavelength over those from the upper surface, and the two, combined with the weaker waves from multiple reflections, therefore interfere destructively. For the destruction to be complete, however, the fraction of the amplitude reflected at each of the two surfaces must be exactly the same, since this specification is made in proving the relation of Eq. (14h). It will be true for a film in contact with a medium of higher index only if the index of the film obeys the relation

$$n' = \sqrt{n}$$

This can be proved from Eq. (25*e*) of Chap. 25 by substituting n' for the refractive index of the upper surface and n/n' for that of the lower. Similar considerations will show that such a film will give zero reflection from the glass side as well as from the air side. Of course no light is destroyed by a nonreflecting film; there is merely

a redistribution such that a decrease of reflection carries with it a corresponding increase of transmission.

The practical importance of these films is that by their use one can greatly reduce the loss of light by reflection at the various surfaces of a system of lenses or prisms. Stray light reaching the image as a result of these reflections is also largely eliminated, with a resulting increase in contrast. Almost all optical parts of high quality are now coated to reduce reflection. The coatings were first made by depositing several monomolecular layers of an organic substance on glass plates. More durable ones are now made by evaporating calcium or magnesium fluoride on the surface in vacuum or by chemical treatment with acids which leave a thin layer of silica on the surface of the glass. Properly coated lenses have a purplish hue by reflected light. This is a consequence of the fact that the condition for destructive interference can be fulfilled for only one wavelength, which is usually chosen to be one near the middle of the visible spectrum. The reflection of red and violet light is then somewhat larger. Furthermore, coating materials of sufficient durability have too high a refractive index to fulfill the condition stated above. Considerable improvement in these respects can be achieved by using two or more superimposed layers, and such films are capable of reducing the total reflected light to one-tenth of its value for the uncoated glass. This refers, of course, to light incident perpendicularly on the surface. At other angles, the path difference will change because of the factor $\cos \phi'$ in Eq. (14e). Since, however, the cosine does not change rapidly in the neighborhood of 0°, the reflection remains low over a fairly large range of angles about the normal. The multiple films, now called *multilayers*, may also be used, with suitable thickness, to accomplish the opposite purpose, namely to increase the reflectance. They may be used, for example, as beam-splitting mirrors to divide a beam of light into two parts of a given intensity ratio. The division can thus be accomplished without the losses of energy by absorption that are inherent in the transmission through, and reflection from, a thin metallic film.

14.7 SHARPNESS OF THE FRINGES

As the reflectance of the surfaces is increased, either by the above method or by lightly silvering them, the fringes due to multiple reflections become much narrower. The striking changes that occur are shown in Fig. 14I, which is plotted for $r^2 = 0.04$, 0.50, and 0.80 according to the theoretical equations to be derived below. The curve labeled 4% is just that for unsilvered glass which was given in Fig. 14E. Since, in the absence of any absorption, the intensity transmitted must be just the complement of that reflected, the same plot will represent the contour of either set. One is obtained from the other by merely turning the figure upside down or by inverting the scale of ordinates, as shown by the down arrow at the right in Fig. 14I.

In order to understand the reason for the narrowness of the transmitted fringes when the reflectance is high, we use the graphical method of compounding amplitudes already discussed in Secs. 12.2 and 13.4. Referring back to Fig. 14D, we notice that the amplitudes of the transmitted rays are given by att', $att'r^2$, $att'r^4$, . . . , or in general for the mth ray by $att'r^{2m}$. We thus have to find the resultant of an infinite

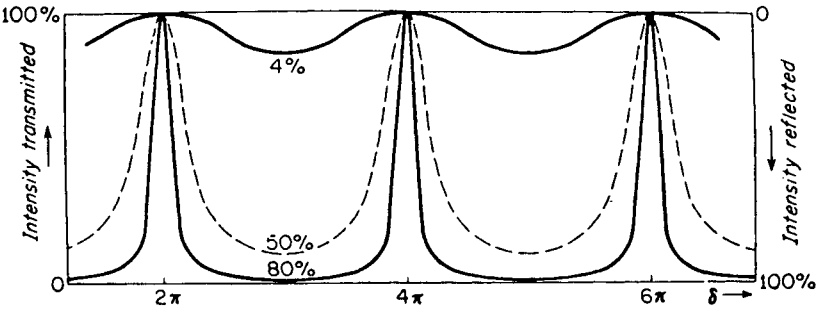


FIGURE 14I
Intensity contours of fringes due to multiple reflections, showing how the sharpness depends on reflectance.

number of amplitudes which decrease in magnitude more rapidly the smaller the fraction r. In Fig. 14J(a) the magnitudes of the amplitudes of the first 10 transmitted rays are drawn to scale for the 50 and 80 percent cases in Fig. 14I, that is, essentially for $r = 0.7$ and 0.9. Starting at any principal maximum, with $\delta = 2\pi m$, these individual amplitudes will all be in phase with each other, so the vectors are all drawn parallel to give a resultant that has been made equal for the two cases. If we now go slightly to one side of the maximum, where the phase difference introduced between successive rays is $\pi/10$, each of the individual vectors must be drawn making an angle of $\pi/10$ with the preceding one and the resultant found by joining the tail of the first to the head of the last. The result is shown in diagram (b). It will be seen that in the case $r = 0.9$, in which the individual amplitudes are much more nearly equal to each other, the resultant R is already considerably less than in the other case. In diagram (c), where the phase has changed by $\pi/5$, this effect is much more pronounced; the

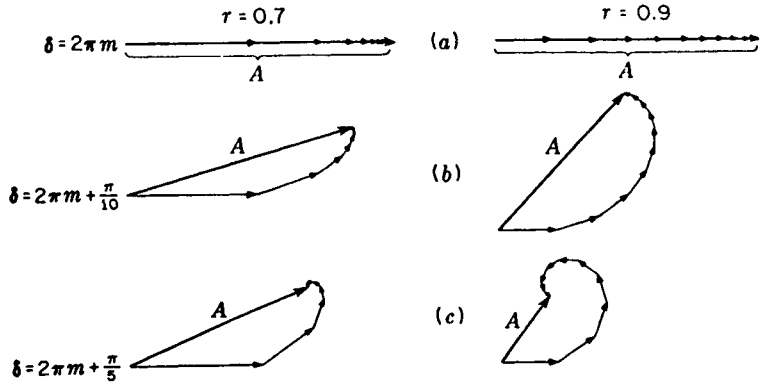


FIGURE 14J
Graphical composition of amplitudes for the first 10 multiply reflected rays, with two difference reflectances.

resultant has fallen to a considerably smaller value in the right-hand picture. Although a correct picture would include an infinite number of vectors, the later ones will have vanishing amplitudes, and we would reach a result similar to that found with the first 10.

These qualitative considerations can be made more precise by deriving an exact equation for the intensity. To accomplish this, we must find an expression for the resultant amplitude A, the square of which determines the intensity. Now A is the vector sum of an infinite series of diminishing amplitudes having a certain phase difference δ given by Eq. (14i). Here we can apply the standard method of adding vectors by first finding the sum of the horizontal components, then that of the vertical components, squaring each sum, and adding to get A^2. In doing this, however, the use of trigonometric functions as in Sec. 12.1 becomes too cumbersome. Hence an alternative way of compounding vibrations, which is mathematically simpler for complicated cases, will be used.

14.8 METHOD OF COMPLEX AMPLITUDES

In place of using the sine or the cosine to represent a simple harmonic wave, one may write the equation in the exponential form*

$$y = ae^{i(\omega t - kx)} = ae^{i\omega t}e^{-i\delta}$$

where $\delta = kx$ and is constant at a particular point in space. The presence of $i = \sqrt{-1}$ in this equation makes the quantities complex. We can nevertheless use this representation and at the end of the problem take either the real (cosine) or the imaginary (sine) part of the resulting expression. The time-varying factor exp $(i\omega t)$ is of no importance in combining waves of the same frequency, since the amplitudes and relative phases are independent of time. The other factor, a exp $(-i\delta)$, is called the *complex amplitude.* It is a complex number whose modulus a is the real amplitude and whose argument δ gives the phase relative to some standard phase. The negative sign merely indicates that the phase is behind the standard phase. In general, the vector $\mathbf{a}$ is given by

$$\mathbf{a} = ae^{i\delta} = x + iy = a(\cos\delta + i\sin\delta)$$

Then it will be seen that

$$a = \sqrt{x^2 + y^2} \qquad \tan\delta = \frac{y}{x}$$

Thus if $\mathbf{a}$ is represented as in Fig. 14K, plotting horizontally its real part and vertically its imaginary part, it will have the magnitude a and will make the angle δ with the x axis, as we require for vector addition.

The advantage of using complex amplitudes lies in the fact that the algebraic

* For the mathematical background of this method, see E. T. Whittaker and G. N. Watson, "Modern Analysis," chap. 1, Cambridge University Press, New York, 1935.

15

FRAUNHOFER DIFFRACTION BY A SINGLE OPENING

When a beam of light passes through a narrow slit, it spreads out to a certain extent into the region of the geometrical shadow. This effect, already noted and illustrated at the beginning of Chap. 13, Fig. 13B, is one of the simplest examples of *diffraction*, i.e., of the failure of light to travel in straight lines. It can be satisfactorily explained only by assuming a wave character for light, and in this chapter we shall investigate quantitatively the *diffraction pattern*, or distribution of intensity of the light behind the aperture, using the principles of wave motion already discussed.

15.1 FRESNEL AND FRAUNHOFER DIFFRACTION

Diffraction phenomena are conveniently divided into two general classes, (1) those in which the source of light and the screen on which the pattern is observed are effectively at infinite distances from the aperture causing the diffraction and (2) those in which either the source or the screen, or both, are at finite distances from the aperture. The phenomena coming under class (1) are called, for historical reasons, *Fraunhofer diffraction*, and those coming under class (2) *Fresnel diffraction*. Fraunhofer diffraction is much simpler to treat theoretically. It is easily observed in practice by rendering

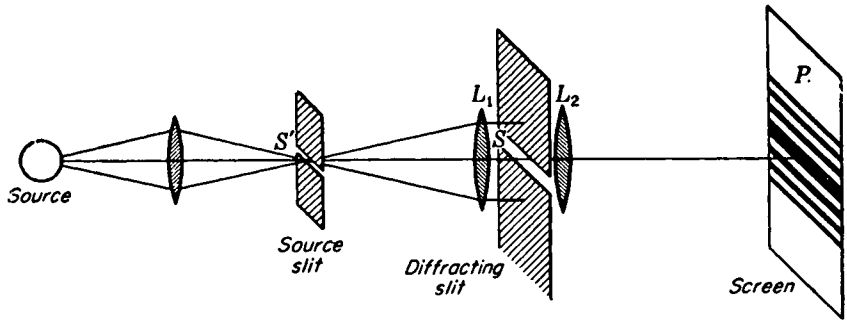


FIGURE 15A
Experimental arrangement for obtaining the diffraction pattern of a single slit; Fraunhofer diffraction.

the light from a source parallel with a lens and focusing it on a screen with another lens placed behind the aperture, an arrangement which effectively removes the source and screen to infinity. In the observation of Fresnel diffraction, on the other hand, no lenses are necessary, but here the wave fronts are divergent instead of plane, and the theoretical treatment is consequently more complex. Only Fraunhofer diffraction will be considered in this chapter, and Fresnel diffraction in Chap. 18.

15.2 DIFFRACTION BY A SINGLE SLIT

A slit is a rectangular aperture of length large compared to its breadth. Consider a slit S to be set up as in Fig. 15A, with its long dimension perpendicular to the plane of the page, and to be illuminated by parallel monochromatic light from the narrow slit S', at the principal focus of the lens L_1. The light focused by another lens L_2 on a screen or photographic plate P at its principal focus will form a diffraction pattern, as indicated schematically. Figure 15B(b) and (c) shows two actual photographs, taken with different exposure times, of such a pattern, using violet light of wavelength 4358 Å. The distance $S'L_1$ was 25.0 cm, and L_2P was 100 cm. The width of the slit S was 0.090 mm, and of S', 0.10 mm. If S' was widened to more than about 0.3 mm, the details of the pattern began to be lost. On the original plate, the half width d of the central maximum was 4.84 mm. It is important to notice that the width of the central maximum is *twice* as great as that of the fainter side maxima. That this effect comes under the heading of diffraction as previously defined is clear when we note that the strip drawn in Fig. 15B(a) is the width of the geometical image of the slit S', or practically that which would be obtained by removing the second slit and using the whole aperture of the lens. This pattern can easily be observed by ruling a single transparent line on a photographic plate and using it in front of the eye as explained in Sec. 13.2, Fig. 13E.

The explanation of the single-slit pattern lies in the interference of the Huygens secondary wavelets which can be thought of as sent out from every point on the wave

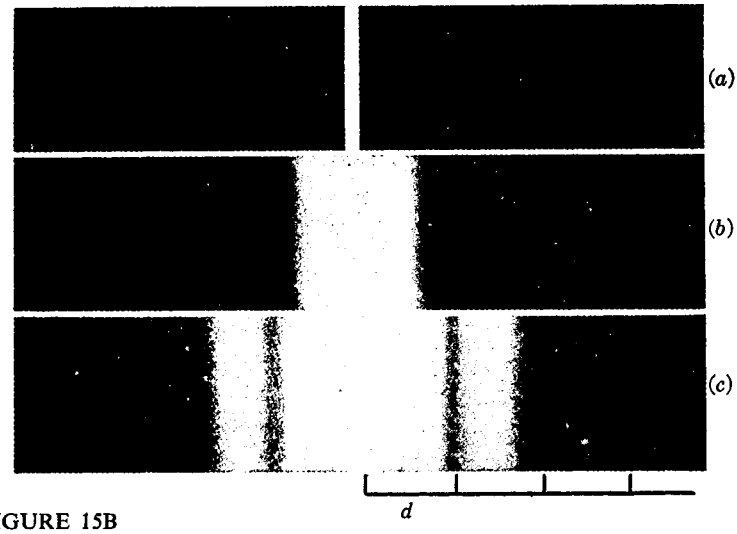


FIGURE 15B
Photographs of the single-slit diffraction pattern.

front at the instant that it occupies the plane of the slit. To a first approximation, one may consider these wavelets to be uniform spherical waves, the emission of which stops abruptly at the edges of the slit. The results obtained in this way, although they give a fairly accurate account of the observed facts, are subject to certain modifications in the light of the more rigorous theory.

Figure 15C represents a section of a slit of width b, illuminated by parallel light from the left. Let ds be an element of width of the wave front in the plane of the slit, at a distance s from the center O, which we shall call the origin. The parts of each secondary wave which travel normal to the plane of the slit will be focused at P_0, while those which travel at any angle θ will reach P. Considering first the wavelet emitted by the element ds situated at the origin, its amplitude will be directly proportional to the length ds and inversely proportional to the distance x. At P it will produce an infinitesimal displacement which, for a spherical wave, may be expressed as

$$dy_0 = \frac{a\,ds}{x} \sin(\omega t - kx)$$

As the position of ds is varied, the displacement it produces will vary in phase because of the different path length to P. When it is at a distance s below the origin, the contribution will be

$$dy_s = \frac{a\,ds}{x} \sin[\omega t - k(x + \Delta)]$$

$$= \frac{a\,ds}{x} \sin(\omega t - kx - ks \sin\theta) \qquad (15a)$$

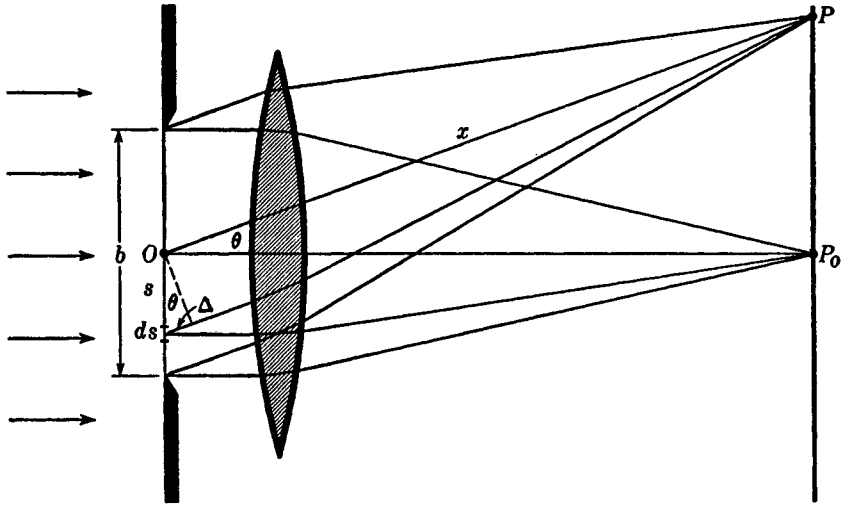


FIGURE 15C
Geometrical construction for investigating the intensity in the single-slit diffraction pattern.

We now wish to sum the effects of all elements from one edge of the slit to the other. This can be done by integrating Eq. (15a) from $s = -b/2$ to $b/2$. The simplest way* is to integrate the contributions from pairs of elements symmetrically placed at s and $-s$, each contribution being

$$dy = dy_{-s} + dy_s$$
$$= \frac{a\,ds}{x}[\sin(\omega t - kx - ks\sin\theta) + \sin(\omega t - kx + ks\sin\theta)]$$

By the identity $\sin\alpha + \sin\beta = 2\cos\tfrac{1}{2}(\alpha - \beta)\sin\tfrac{1}{2}(\alpha + \beta)$, we have

$$dy = \frac{a\,ds}{x}[2\cos(ks\sin\theta)\sin(\omega t - kx)]$$

which must be integrated from $s = 0$ to $b/2$. In doing so, x may be regarded as constant, insofar as it affects the amplitude. Thus

$$y = \frac{2a}{x}\sin(\omega t - kx)\int_0^{b/2}\cos(ks\sin\theta)\,ds$$
$$= \frac{2a}{x}\left[\frac{\sin(ks\sin\theta)}{k\sin\theta}\right]_0^{b/2}\sin(\omega t - kx)$$
$$= \frac{ab}{x}\frac{\sin(\frac{1}{2}kb\sin\theta)}{\frac{1}{2}kb\sin\theta}\sin(\omega t - kx) \qquad (15b)$$

* The method of complex amplitudes (Sec. 14.8) starts with $(ab/x)\int\exp(iks\sin\theta)\,ds$, and yields the real amplitude upon multiplication of the result by its complex conjugate. No simplification results from using the method here.

The resultant vibration will therefore be a simple harmonic one, the amplitude of which varies with the position of P, since the latter is determined by θ. We may represent its amplitude as

$$A = A_0 \frac{\sin \beta}{\beta} \qquad (15c)$$

where $\beta = \frac{1}{2}kb \sin \theta = (\pi b \sin \theta)/\lambda$ and $A_0 = ab/x$. The quantity β is a convenient variable, which signifies one-half the phase difference between the contributions coming from opposite edges of the slit. The intensity on the screen is then

$$I \approx A^2 = A_0{}^2 \frac{\sin^2 \beta}{\beta^2} \qquad (15d)$$

If the light, instead of being incident on the slit perpendicular to its plane, makes an angle i, a little consideration will show that it is merely necessary to replace the above expression for β by the more general expression

$$\beta = \frac{\pi b(\sin i + \sin \theta)}{\lambda} \qquad (15e)$$

15.3 FURTHER INVESTIGATION OF THE SINGLE-SLIT DIFFRACTION PATTERN

In Fig. 15D(a) graphs are shown of Eq. (15c) for the *amplitude* (dotted curve) and Eq. (15d) for the *intensity*, taking the constant A_0 in each case as unity. The intensity curve will be seen to have the form required by the experimental result in Fig. 15B. The maximum intensity of the strong central band comes at the point P_0 of Fig. 15C, where evidently all the secondary wavelets will arrive in phase because the path difference $\Delta = 0$. For this point $\beta = 0$, and although the quotient $(\sin \beta)/\beta$ becomes indeterminate for $\beta = 0$, it will be remembered that $\sin \beta$ approaches β for small angles and is equal to it when β vanishes. Hence for $\beta = 0$, $(\sin \beta)/\beta = 1$. We now see the significance of the constant A_0. Since for $\beta = 0$, $A = A_0$, it represents the amplitude when all the wavelets arrive in phase. $A_0{}^2$ is then the value of the maximum intensity, at the center of the pattern. From this *principal maximum* the intensity falls to zero at $\beta = \pm\pi$, then passes through several *secondary maxima*, with equally spaced points of zero intensity at $\beta = \pm\pi, \pm 2\pi, \pm 3\pi, \ldots$, or in general $\beta = m\pi$. The secondary maxima do not fall halfway between these points, but are displaced toward the center of the pattern by an amount which decreases with increasing m. The exact values of β for these maxima can be found by differentiating Eq. (15c) with respect to β and equating to zero. This yields the condition

$$\tan \beta = \beta$$

The values of β satisfying this relation are easily found graphically as the intersections of the curve $y = \tan \beta$ and the straight line $y = \beta$. In Fig. 15D(b) these points of intersection lie directly below the corresponding secondary maxima.

The intensities of the secondary maxima can be calculated to a very close approximation by finding the values of $(\sin^2 \beta)/\beta^2$ at the halfway positions, i.e.,

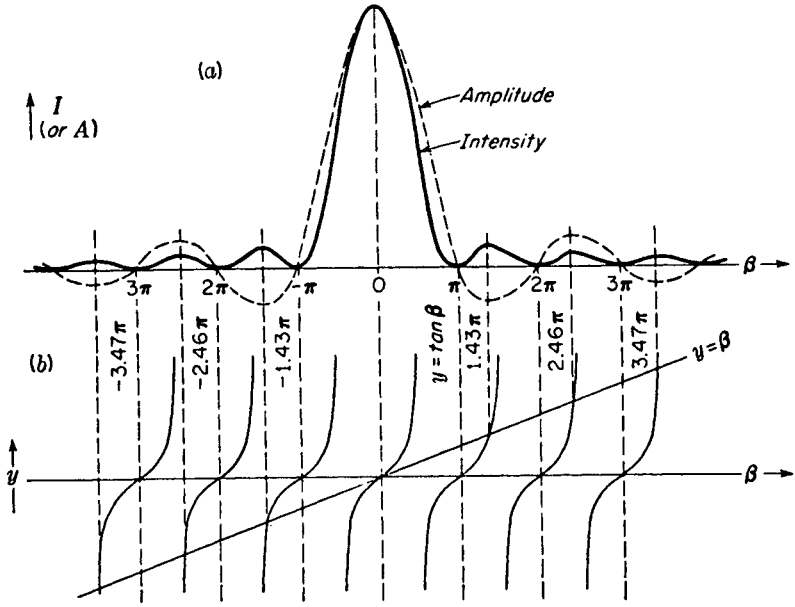


FIGURE 15D
Amplitude and intensity contours for Fraunhofer diffraction of a single slit, showing positions of maxima and minima.

where $\beta = 3\pi/2, 5\pi/2, 7\pi/2, \ldots$. This gives $4/9\pi^2$, $4/25\pi^2$, $4/49\pi^2, \ldots$, or 1/22.2, 1/61.7, 1/121, . . . , of the intensity of the principal maximum. Reference to Table 15A ahead are the exact values of the intensity for every 15° intervals for the central maximum. These values are useful in plotting graphs. The first *secondary maximum* is only 4.72 percent the intensity of the central maximum, while the *second* and *third secondary maxima* are only 1.65 and 0.83 percent respectively.

A very clear idea of the origin of the single-slit pattern is obtained by the following simple treatment. Consider the light from the slit of Fig. 15E coming to the point P_1 on the screen, this point being just one wavelength farther from the upper

Table 15A VALUES FOR CENTRAL MAXIMUM FOR FRAUNHOFER DIFFRACTION OF A SINGLE SLIT

β deg	rad	sin β	A^2	β deg	rad	sin β	A^2
0	0	0	1	105	1.8326	0.9659	0.2778
15	0.2618	0.2588	0.9774	120	2.0944	0.8660	0.1710
30	0.5236	0.5000	0.9119	135	2.3562	0.7071	0.0901
45	0.7854	0.7071	0.8106	150	2.6180	0.5000	0.0365
60	1.0472	0.8660	0.6839	165	2.8798	0.2588	0.0081
75	1.3090	0.9659	0.5445	180	3.1416	0	0
90	1.5708	1.0000	0.4053	195	3.4034	0.2588	0.0058

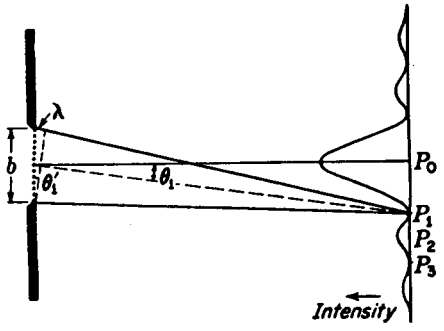


FIGURE 15E
Angle of the first minimum of the single-slit diffraction pattern.

edge of the slit than from the lower. The secondary wavelet from the point in the slit adjacent to the upper edge will travel approximately $\lambda/2$ farther than that from the point at the center, and so these two will produce vibrations with a phase difference of π and will give a resultant displacement of zero at P_1. Similarly the wavelet from the next point below the upper edge will cancel that from the next point below the center, and we can continue this pairing off to include all points in the wave front, so that the resultant effect at P_1 is zero. At P_3 the path difference is 2λ, and if we divide the slit into four parts, the pairing of points again gives zero resultant, since the parts cancel in pairs. For the point P_2, on the other hand, the path difference is $3\lambda/2$, and we divide the slit into thirds, two of which will cancel, leaving one third to account for the intensity at this point. The resultant amplitude at P_2 is, of course, not even approximately one-third that at P_0, because the phases of the wavelets from the remaining third are not by any means equal.

The above method, though instructive, is not exact if the screen is at a finite distance from the slit. As Fig. 15E is drawn, the shorter broken line is drawn to cut off equal distances on the rays to P_1. It will be seen from this that the path difference to P_1 between the light coming from the upper edge and that from the center is slightly greater than $\lambda/2$ and that between the center and lower edge slightly less than $\lambda/2$. Hence the resultant intensity will not be zero at P_1 and P_3, but it will be more nearly so the greater the distance between slit and screen or the narrower the slit. This corresponds to the transition from Fresnel diffraction to Fraunhofer diffraction. Obviously, with the relative dimensions shown in the figure, the geometrical shadow of the slit would considerably widen the central maximum as drawn. Just as was true with Young's experiment (Sec. 13.3), when the screen is at infinity, the relations become simpler. Then the two angles θ_1 and θ_1' in Fig. 15E become exactly equal, i.e., the two broken lines are perpendicular to each other, and $\lambda = b \sin \theta_1$ for the first minimum corresponding to $\beta = \pi$. This gives

$$\sin \theta_1 = \frac{\lambda}{b} \qquad (15f)$$

In practice θ_1 is usually a very small angle, so we may put the sine equal to the angle. Then

$$\theta_1 = \frac{\lambda}{b} \qquad (15g)$$

a relation which shows at once how the dimensions of the pattern vary with λ and b. The *linear* width of the pattern on a screen will be proportional to the slit-screen distance, which is the focal length f of a lens placed close to the slit. The linear distance d between successive minima corresponding to the angular separation $\theta_1 = \lambda/b$ is thus

$$d = \frac{f\lambda}{b}$$

The width of the pattern increases in proportion to the wavelength, so that for red light it is roughly twice as wide as for violet light, the slit width, etc., being the same. If white light is used, the central maximum is white in the middle but reddish on its outer edge, shading into a purple and other impure colors farther out.

The angular width of the pattern for a given wavelength is inversely proportional to the slit width b, so that as b is made larger, the pattern shrinks rapidly to a smaller scale. In photographing Fig. 15B, if the slit S had been 9.0 mm wide, the whole visible pattern (of five maxima) would have been included in a width of 0.24 mm on the original plate instead of 2.4 cm. This fact (that when the width of the aperture is large compared to a wavelength, the diffraction is practically negligible) led the early investigators to conclude that light travels in straight lines and that it could not be a wave motion. Sound waves will be diffracted through large angles in passing through an aperture of ordinary size, such as an open window.

15.4 GRAPHICAL TREATMENT OF AMPLITUDES. THE VIBRATION CURVE

The addition of the amplitude contributions from all the secondary wavelets originating in the slit can be carried out by a graphical method based on the vector addition of amplitudes discussed in Sec. 12.2. It will be worthwhile to consider this method in some detail, because it may be applied to advantage in other more complicated cases to be treated in later chapters, and because it gives a very clear physical picture of the origin of the diffraction pattern. Let us divide the width of the slit into a fairly large number of equal parts, say nine. The amplitude r contributed at a point on the screen by any one of these parts will be the same, since they are of equal width. The phases of these contributions will differ, however, for any point except that lying on the axis, i.e., on the normal to the slit at its center (P_0, Fig. 15E). For a point off the axis, each of the nine segments will contribute vibrations differing in phase, because the segments are at different average distances from the point. Furthermore the difference in phase δ between the contributions from adjacent segments will be constant, since each element is on the average the same amount farther away (or nearer) than its neighbor.

Now, using the fact that the resultant amplitude and phase may be found by the vector addition of the individual amplitudes making angles with each other equal to the phase difference, a vector diagram like that shown in Fig. 15F(b) may be drawn. Each of the nine equal amplitudes a is inclined a tan angle δ with the preceding one, and their vector sum A is the resultant amplitude required. Now suppose that instead of dividing the slit into nine elements, we had divided it into many thousand or, in the

16
THE DOUBLE SLIT

The interference of light from two narrow slits close together, first demonstrated by Young, has already been discussed (Sec. 13.2) as a simple example of the interference of two beams of light. In our discussion of the experiment, the slits were assumed to have widths not much greater than a wavelength of light, so that the central maximum in the diffraction pattern from each slit separately was wide enough to occupy a large angle behind the screen (Figs. 13A and 13B). It is important to understand the modifications of the interference pattern which occur when the width of the individual slits is made greater, until it becomes comparable with the distance between them. This corresponds more nearly to the actual conditions under which the experiment is usually performed. In this chapter we shall discuss the *Fraunhofer diffraction by a double slit*, and some of its applications.

16.1 QUALITATIVE ASPECTS OF THE PATTERN

In Fig. 16A(*b*) and (*c*) photographs are shown of the patterns obtained from two different double slits in which the widths of the individual slits were equal in each pair but the two pairs were different. Figure 16B shows the experimental arrangement for

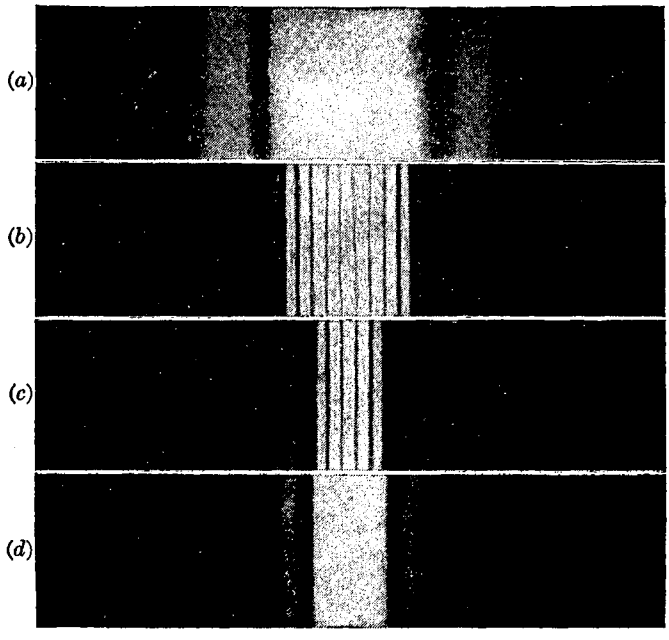


FIGURE 16A
Diffraction patterns from (*a*) a single narrow slit, (*b*) two narrow slits, (*c*) two wider slits, (*d*) one wider slit.

photographing these patterns; the *slit width b* of each slit was greater for Fig. 16A(*c*) than for Fig. 16A(*b*), but the distance between centers $d = b + c$, or the *separation* of the slits, was the same in the two cases. In the central part of Fig. 16A(*b*) are seen a number of interference maxima of approximately uniform intensity, resembling the interference fringes described in Chap. 13 and shown in Fig. 13D. The intensities of these maxima are not actually constant, however, but fall off slowly to zero on either side and then reappear with low intensity two or three times before becoming too faint to observe without difficulty. The same changes are seen to occur much more rapidly in Fig. 16A(*c*), which was taken with the slit widths *b* somewhat larger.

16.2 DERIVATION OF THE EQUATION FOR THE INTENSITY

Following the same procedure as that used for the single slit in Sec. 15.2, it is merely necessary to change the limits of integration in Eq. (15b) to include the two portions of the wave front transmitted by the double slit.* Thus if we have, as in Fig. 16B, two equal slits of width *b*, separated by an opaque space of width *c*, the origin may

* The result of this derivation is obviously a special case of the general formula for *N* slits, which will be obtained by the method of complex amplitudes in the following chapter.

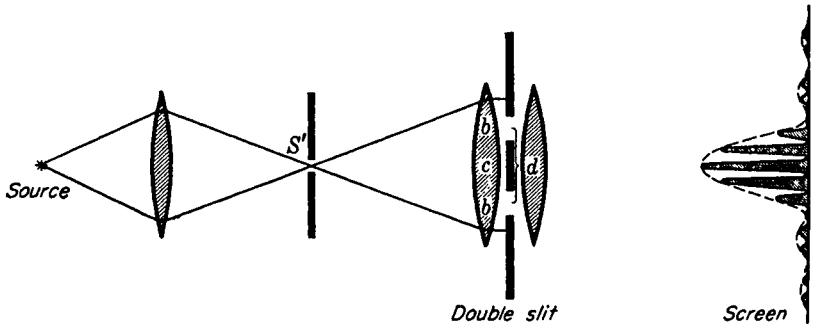


FIGURE 16B
Apparatus for observing Fraunhofer diffraction from a double slit. Drawn for $2b = c$, that is, $d = 3b$.

be chosen at the center of c, and the integration extended from $s = d/2 - b/2$ to $s = d/2 + b/2$. This gives

$$y = \frac{2a}{xk \sin\theta} \{\sin [\tfrac{1}{2}k(d + b) \sin\theta] - \sin [\tfrac{1}{2}k(d - b) \sin\theta]\} [\sin (\omega t - kx)]$$

The quantity in braces is of the form $\sin (A + B) - \sin (A - B)$, and when it is expanded, we obtain

$$y = \frac{2ba}{x} \frac{\sin\beta}{\beta} \cos\gamma \sin (\omega t - kx) \qquad (16a)$$

where, as before,

$$\beta = \tfrac{1}{2}kb \sin\theta = \frac{\pi}{\lambda} b \sin\theta$$

and where

$$\gamma = \tfrac{1}{2}k(b + c) \sin\theta = \frac{\pi}{\lambda} d \sin\theta \qquad (16b)$$

The intensity is proportional to the square of the amplitude of Eq. (16a), so that, replacing ba/x by A_0 as before, we have

$$I = 4A_0^2 \frac{\sin^2\beta}{\beta^2} \cos^2\gamma \qquad (16c)$$

The factor $(\sin^2\beta)/\beta^2$ in this equation is just that derived for the single slit of width b in the previous chapter [Eq. (15d)]. The second factor $\cos^2\gamma$ is characteristic of the interference pattern produced by two beams of equal intensity and phase difference δ, as shown in Eq. (13b) of Sec. 13.3. There the resultant intensity was found to be proportional to $\cos^2(\delta/2)$, so that the expressions correspond if we put $\gamma = \delta/2$. The resultant intensity will be zero when either of the two factors is zero. For the first factor this will occur when $\beta = \pi, 2\pi, 3\pi, \ldots$, and for the second factor when $\gamma = \pi/2, 3\pi/2, 5\pi/2, \ldots$. That the two variables β and γ are not independent will be seen from Fig. 16C. The difference in path from the two edges of a given slit to the screen is, as indicated, $b \sin\theta$. The corresponding phase difference is, by Eq. (15c),

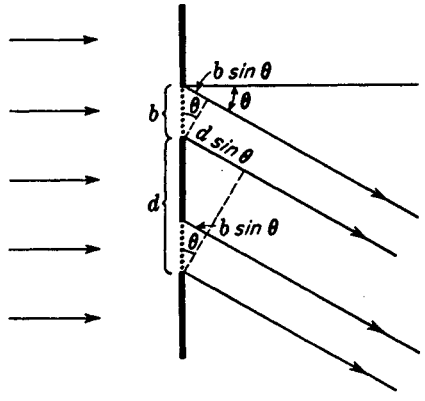


FIGURE 16C
Path differences of parallel rays leaving a double slit.

$(2\pi/\lambda)b \sin\theta$, which equals 2β. The path difference from any two corresponding points in the two slits, as illustrated for the two points at the lower edges of the slits, is $d \sin\theta$, and the phase difference is $\delta = (2\pi/\lambda)d \sin\theta = 2\gamma$. Therefore, in terms of the dimensions of the slits,

$$\frac{\delta}{2\beta} = \frac{\gamma}{\beta} = \frac{d}{b} \qquad (16d)$$

16.3 COMPARISON OF THE SINGLE-SLIT AND DOUBLE-SLIT PATTERNS

It is instructive to compare the double-slit pattern with that given by a single slit of width equal to that of either of the two slits. This amounts to comparing the effect obtained with the two slits in the arrangement shown in Fig. 16B with that obtained when one of the slits is entirely blocked off with an opaque screen. If this is done, the corresponding single-slit diffraction patterns are observed, and they are related to the double-slit patterns as shown in Fig. 16A(*a*) and (*d*). It will be seen that the intensities of the interference fringes in the double-slit pattern correspond to the intensity of the single-slit pattern at any point. If one or other of the two slits is covered, we obtain exactly the same single-slit pattern in the same position, while if both slits are uncovered, the pattern, instead of being a single-slit one with twice the intensity, breaks up into the narrow maxima and minima called *interference fringes.* The intensity at the maximum of these fringes is 4 times the intensity of either single-slit pattern at that point, while it is zero at the minima (see Sec. 13.4).

16.4 DISTINCTION BETWEEN INTERFERENCE AND DIFFRACTION

One is quite justified in explaining the above results by saying that the light from the two slits undergoes interference to produce fringes of the type obtained with two beams but that the intensities of these fringes are limited by the amount of light arriving

at the given point on the screen by virtue of the diffraction occurring at each slit. The relative intensities in the resultant pattern as given by Eq. (16c) are just those obtained by multiplying the intensity function for the interference pattern from two infinitely narrow slits of separation d [Eq. (13b)] by the intensity function for diffraction from a single slit of width b [Eq. (15d)]. Thus, the result may be regarded as due to the joint action of interference between the rays coming from corresponding points in the two slits and of diffraction, which determines the amount of light emerging from either slit at a given angle. But diffraction is merely the result of the interference of all the secondary wavelets originating from the different elements of the wave front. Hence it is proper to say that the whole pattern is an interference pattern. It is just as correct to refer to it as a diffraction pattern, since, as we saw from the derivation of the intensity function in Sec. 16.2, it is obtained by direct summing the effects of all of the elements of the exposed part of the wave front. However, if we reserve the term *interference* for those cases in which a modification of amplitude is produced by the superposition of a finite (usually small) number of beams, and *diffraction* for those in which the amplitude is determined by an integration over the infinitesimal elements of the wave front, the double-slit pattern can be said to be due to a combination of interference and diffraction. Interference of the beams from the two slits produces the narrow maxima and minima given by the $\cos^2 \gamma$ factor, and diffraction, represented by $(\sin^2 \beta)/\beta^2$, modulates the intensities of these interference fringes. The student should not be misled by this statement into thinking that diffraction is anything other than a rather complicated case of interference.

16.5 POSITIONS OF THE MAXIMA AND MINIMA. MISSING ORDERS

As shown in Sec. 16.2, the intensity will be zero wherever $\gamma = \pi/2, 3\pi/2, 5\pi/2, \ldots$ and also when $\beta = \pi, 2\pi, 3\pi, \ldots$. The first of these two sets are the minima for the interference pattern, and since by definition $\gamma = (\pi/\lambda)d \sin \theta$, they occur at angles θ such that

$$d \sin \theta = \frac{\lambda}{2}, \frac{3\lambda}{2}, \frac{5\lambda}{2}, \ldots = (m + \tfrac{1}{2})\lambda \qquad \textit{Minima} \qquad (16e)$$

m being any whole number starting with zero. The second series of minima are those for the diffraction pattern, and these, since $\beta = (\pi/\lambda)a \sin \theta$, occur where

$$b \sin \theta = \lambda, 2\lambda, 3\lambda, \ldots = p\lambda \qquad \textit{Minima} \qquad (16f)$$

the smallest value of p being 1. The exact positions of the *maxima* are not given by any simple relation, but their approximate positions can be found by neglecting the variation of the factor $(\sin^2 \beta)/\beta^2$, a justified assumption only when the slits are very narrow and when the maxima near the center of the pattern are considered [Fig. 16A(*b*)]. The positions of the maxima will then be determined solely by the $\cos^2 \gamma$ factor, which has maxima for $\gamma = 0, \pi, 2\pi, \ldots$, that is, for

$$d \sin \theta = 0, \lambda, 2\lambda, 3\lambda, \ldots = m\lambda \qquad \textit{Maxima} \qquad (16g)$$

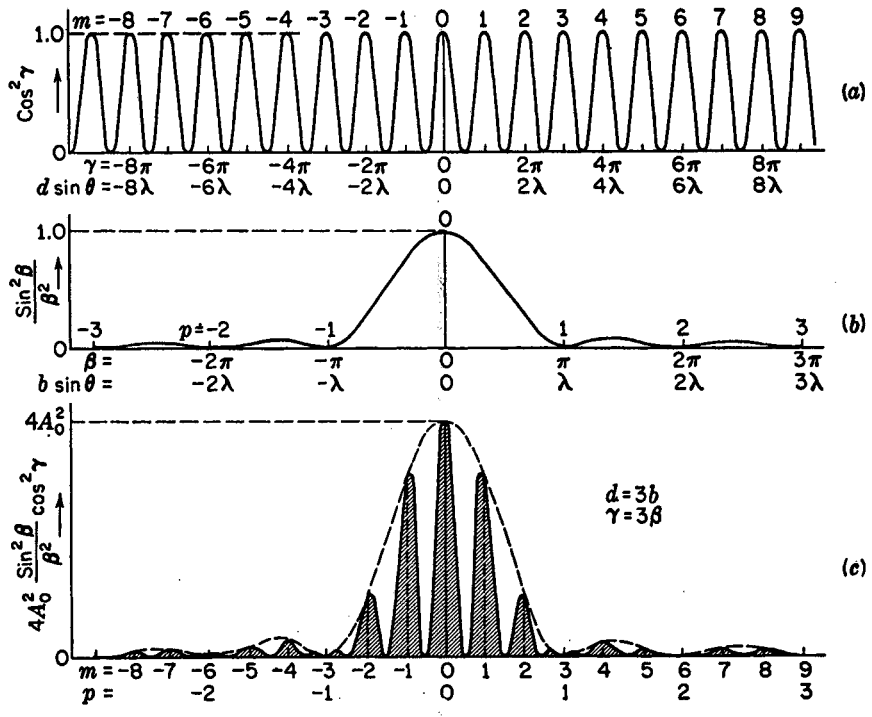


FIGURE 16D
Intensity curves for a double slit where $d = 3b$.

The whole number m represents physically the number of wavelengths in the path difference from corresponding points in the two slits (see Fig. 16C) and represents the *order* of interference.

Figure 16D(a) is a plot of the factor $\cos^2 \gamma$, and here the values of the order, of half the phase difference $\gamma = \delta/2$, and of the path difference are indicated for the various maxima. These are all of equal intensity and equidistant on a scale of $d \sin \theta$, or practically on a scale of θ, since when θ is small, $\sin \theta \approx \theta$ and the maxima occur at angles $\theta = 0, \lambda/d, 2\lambda/d, \ldots$. With a finite slit width b the variation of the factor $(\sin^2 \beta)/\beta^2$ must be taken into account. This factor alone gives just the single-slit pattern discussed in the last chapter, and is plotted in Fig. 16D(b). The complete double-slit pattern as given by Eq. (16c) is the product of these two factors and therefore is obtained by multiplying the ordinates of curve (a) by those of curve (b) and the constant $4A_0^2$. This pattern is shown in Fig. 16D(c). The result will depend on the relative scale of the abscissas β and γ, which in the figure are chosen so that for a given abscissa $\gamma = 3\beta$. But the relation between β and γ for a given angle θ is determined, according to Eq. (16d), by the ratio of the slit width to the slit separation. Hence if $d = 3b$, the two curves (a) and (b) are plotted to the same scale of θ. For the particular case of two slits of width b separated by an opaque space of width $c = 2b$, the curve (c), which is the product of (a) and (b), then gives the resultant pattern. The positions of the maxima in this curve are slightly different from those in

curve (*a*) for all except the central maximum ($m = 0$), because when the ordinates near one of the maxima of curve (*a*) are multiplied by a factor which is decreasing or increasing, the ordinates on one side of the maximum are changed by a different amount from those of the other, and this displaces the resultant maximum slightly in the direction in which the factor is increasing. Hence the positions of the maxima in curve (*c*) are not exactly those given by Eq. (16g) but in most cases will be very close to them.

Let us now return to the explanation of the differences in the two patterns (*b*) and (*c*) of Fig. 16A, taken with the same slit separation d but different slit widths b. Pattern (*c*) was taken for the case $d = 3b$, and is seen to agree with the description given above. For pattern (*b*), the slit separation d was the same, giving the same spacing for the interference fringes, but the slit width b was smaller, such that $d = 6b$. In Fig. 13D, $d = 14b$. This greatly increases the scale for the single-slit pattern relative to the interference pattern, so that many interference maxima now fall within the central maximum of the diffraction pattern. Hence the effect of decreasing b, keeping d unchanged, is merely to broaden out the single-slit pattern, which acts as an envelope of the interference pattern as indicated by the dotted curve of Fig. 16D(*c*).

If the slit-width b is kept constant and the separation of the slits d is varied, the scale of the interference pattern varies, but that of the diffraction pattern remains the same. A series of photographs taken to illustrate this is shown in Fig. 16E. For each pattern three different exposures are shown, to bring out the details of the faint and the strong parts of the pattern. The maxima of the curves are labeled by the order m, and underneath the upper one is a given scale of angular positions θ. A study of these figures shows that certain orders are missing, or at least reduced to two maxima of very low intensity. These *missing orders* occur where the condition for a maximum of the interference, Eq. (16g), and for a minimum of the diffraction, Eq. (16f), are both fulfilled for the same value of θ, that is for

$$d \sin \theta = m\lambda \qquad \text{and} \qquad b \sin \theta = p\lambda$$

so that

$$\frac{d}{b} = \frac{m}{p} \tag{16h}$$

Since m and p are both integers, d/b must be in the ratio of two integers if we are to have missing orders. This ratio determines the orders which are missing, in such a way that when $d/b = 2$, orders $2, 4, 6, \ldots$ are missing; when $d/b = 3$, orders $3, 6, 9, \ldots$ are missing; etc. When $d/b = 1$, the two slits exactly join, and all orders should be missing. However, the two faint maxima into which each order is split can then be shown to correspond exactly to the subsidiary maxima of a single-slit pattern of width $2b$.

Our physical picture of the cause of missing orders is as follows. Considering, for example, the missing order $m = +3$ in Fig. 16D(*c*), this point on the screen is just three wavelengths farther from the center of one slit than from the center of the other. Hence we might expect the waves from the two slits to arrive in phase and to produce a maximum. However, this point is at the same time one wavelength farther from the edge of one slit than from the other edge of that slit. Addition of the second-

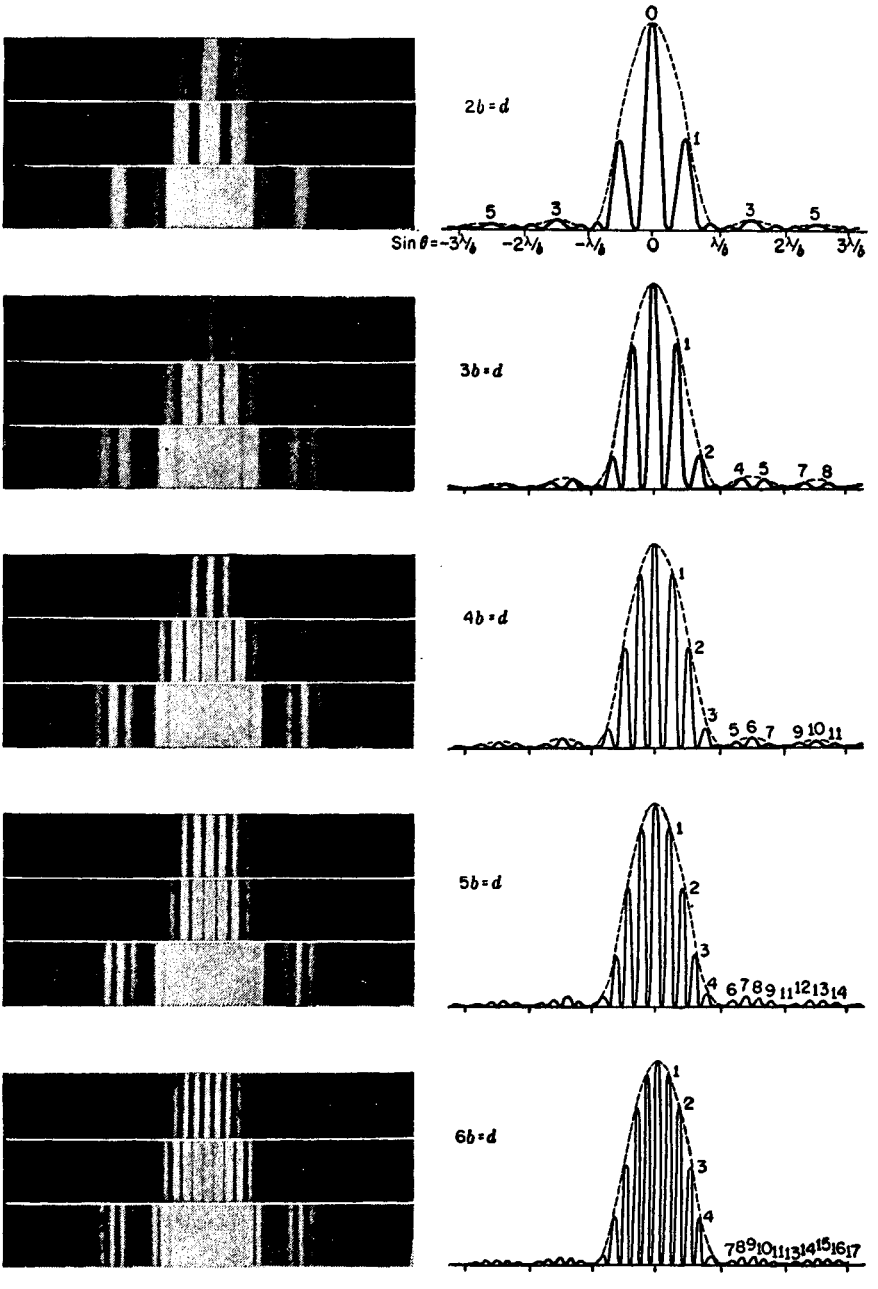


FIGURE 16E
Photographs and intensity curves for double-slit diffraction patterns.

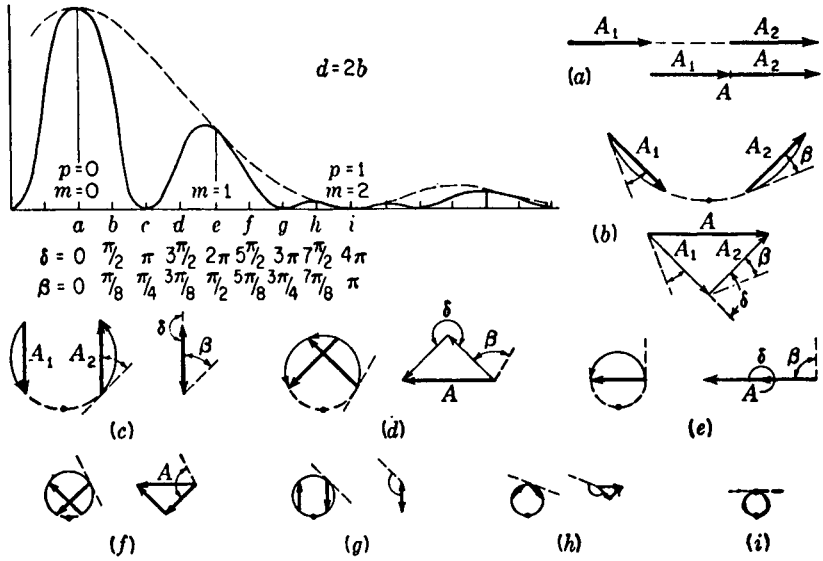


FIGURE 16F
How the intensity curve for a double slit is obtained by the graphical addition of amplitudes.

ary wavelets from one slit gives zero intensity under these conditions. The same holds true for either slit, so that although we may add the contributions from the two slits, both contributions are zero and must therefore give zero resultant.

16.6 VIBRATION CURVE

The same method as that applied in Sec. 15.4 for finding the resultant amplitude graphically in the case of the single slit is applicable to the present problem. For illustration we take a double slit in which the width of each slit equals that of the opaque space between the two, so that $d = 2b$. A photograph of this pattern appears in Fig. 16E at the top. A vector diagram of the amplitude contributions from one slit gives the arc of a circle, as before, the difference between the slopes of the tangents to the arc at the two ends being the phase difference 2β between the contributions from the two edges of the slit. Such an arc must now be drawn for each of the two slits, and the two arcs must be related in such a way that the phases (slopes of the tangents) differ for corresponding points on the two slits by 2γ, or δ. In the present case, since $d = 2b$, we must have $\gamma = 2\beta$ or $\delta = 4\beta$. Thus in Fig. 16F(b) showing the vibration curve for $\beta = \pi/8$, both arcs subtend an angle of $\pi/4$ ($= 2\beta$), the phase difference for the two edges of each slit, and the arcs are separated by $\pi/4$ so that corresponding points on the two arcs differ by $\pi/2 (= \delta)$. Now the resultant contributions from the two slits are represented in amplitude and phase by the chords of these two arcs, that is by A_1 and A_2. Diagrams (a) to (i) give the construction for the points similarly

24

THE POLARIZATION OF LIGHT

From the properties of interference and diffraction we are led to conclude that light is a wave phenomenon, and we utilize these properties to measure the wavelength. These effects tell us nothing about the types of waves with which we are dealing—whether they are longitudinal or transverse, or whether the vibrations are linear, circular, or torsional. The electromagnetic theory, however, specifically requires that the vibrations be traverse, being therefore entirely confined to the plane of the wave front. The most general type of vibration is elliptical, of which linear and circular vibrations are extreme cases. Experiments which bring out these characteristics are those dealing with the polarization of light. Although a longitudinal wave like a sound wave must necessarily be symmetrical about the direction of its propagation, transverse waves may show dissymmetries, and if any beam of light shows such a dissymmetry, we say it is polarized.

The present chapter, by way of introduction to the subject of polarization, gives a brief account of the principal ways of producing plane-polarized light from ordinary unpolarized light. Most of the phenomena to be discussed here will be covered in more detail in later chapters. It will be helpful, however, to have a preliminary acquaintance with the experimental methods and a mental picture of how

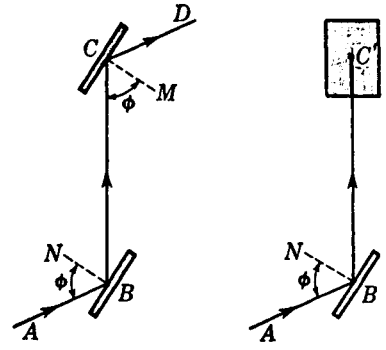


FIGURE 24A
Polarization by reflection from glass surfaces.

the various polarizing devices act to separate ordinary light into its polarized components. The common methods used in producing and demonstrating the polarization of light may be grouped under the following heads: (1) reflection, (2) transmission through a pile of plates, (3) dichroism, (4) double refraction, and (5) scattering.

24.1 POLARIZATION BY REFLECTION

Perhaps the simplest method of polarizing light is the one discovered by Malus in 1808. If a beam of white light is incident at one certain angle on the polished surface of a plate of ordinary glass, it is found upon reflection to be plane-polarized. By plane-polarized is meant that all the light is vibrating parallel to a plane through the axis of the beam (Sec. 11.6). Although this light appears to the eye to be no different from the incident light, its polarization or asymmetry is easily shown by reflection from a second plate of glass as follows. A beam of unpolarized light, AB in Fig. 24A, is incident at an angle of about 57° on the first glass surface at B. This light is again reflected at 57° by a second glass plate C placed parallel to the first as shown at the left. If now the upper plate is rotated about BC as an axis, the intensity of the reflected beam is found to decrease, reaching zero for a rotation of 90°. Rotation about BC keeps the angle of incidence constant. The experiment is best performed with the back surfaces of the glass painted black. The first reflected beam BC' then appears to be cut off and to vanish at C'. As the upper mirror is rotated further about BC the reflected beam CD reappears, increasing in intensity to reach a maximum at 180°. Continued rotation produces zero intensity again at 270°, and a maximum again at 360°, the starting point.

If the angle of incidence on either the lower or upper mirror is not 57°, the twice-reflected beam will go through maxima and minima as before, but the minima will not have zero intensity. In other words there will always be a reflected beam from C. Calling the angle of incidence ϕ in general, the critical value $\bar{\phi}$ which produces a zero minimum for the second reflection is called the *polarizing angle* and varies with the kind of glass used. Before undertaking the explanation of this experiment, it will be worthwhile to consider briefly the accepted ideas concerning the nature of the vibrations in ordinary and polarized light.

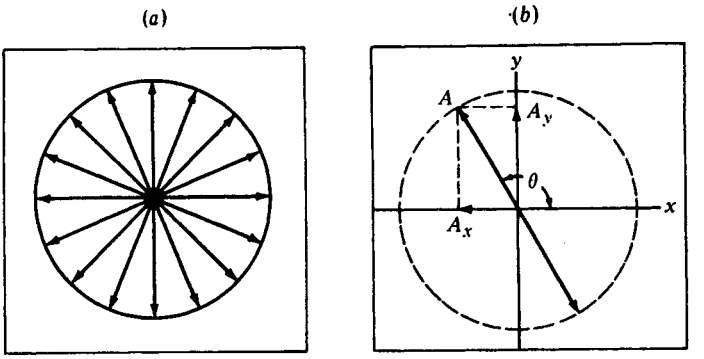


FIGURE 24B
Vibrations in unpolarized light viewed end-on. (*a*) All planes are equally probable. (*b*) Each vibration can be resolved into two components in the x and y directions.

24.2 REPRESENTATION OF THE VIBRATIONS IN LIGHT

According to the electromagnetic theory, any type of light consists of transverse waves, in which the oscillating magnitudes are the electric and magnetic vectors. The question as to which of these is to be chosen as constituting the "vibrations" will be deferred until later (Sec. 25.12), but it is immaterial for our present purpose. Let us assume that in a beam of light traveling toward the observer, along the $+z$ axis in Fig. 24B, the electric vector at some instant is executing a linear vibration with the direction and amplitude indicated. If this vibration continues unchanged, we say that the light is *plane-polarized*, since its vibrations are confined to the plane containing the z axis and oriented at the angle θ. If, on the other hand, the light is *unpolarized*, like most natural light, one may imagine that there are sudden, random changes in θ, occurring in time intervals of the order of 10^{-8} s. Every orientation of A is to be regarded as equally probable, so that, as indicated by the solid circle in Fig. 24B(*a*) the average effect is completely symmetrical about the direction of propagation.

This picture of unpolarized light, although a legitimate one, is oversimplified because if there are fluctuations in orientation, there should be fluctuations in amplitude as well. Furthermore, linear vibrations are a special case of elliptical ones, and there is no reason for this special type to be preferred. Hence a truer picture is one of elliptical vibrations changing frequently in size, eccentricity, and orientation, but confined to the xy plane. This complexity presents little difficulty, however, since because all azimuths are equivalent, the simpler representation in terms of linear vibrations of constant amplitude and rapidly shifting orientation completely describes the facts. Also, since motion in an ellipse can be regarded as made up of two linear motions at right angles (Sec. 12.9), the two descriptions are in fact mathematically the same.

Still another representation of unpolarized light is perhaps the most useful. If we resolve the vibration of Fig. 24B(*b*) into linear components $A_x = A \cos \theta$ and $A_y = A \sin \theta$, they will in general be unequal [see Sec. 24.5 and Eq. (24d)]. But

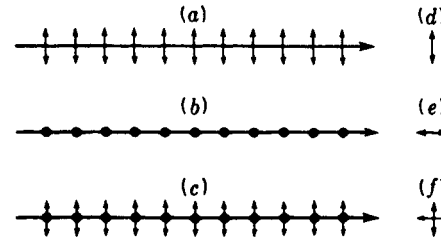


FIGURE 24C
Pictorial representations of side and end views of plane-polarized and ordinary light beams.

when θ is allowed to assume all values at random, the net result is as though we had two vibrations at right angles with equal amplitudes but no coherence of phase. Each is the resultant of a large number of individual vibrations with random phases (Sec. 12.4) and because of this randomness a complete incoherence is produced. Figure 24C shows a common way of picturing these vibrations, parts (*a*) and (*b*) representing the two plane-polarized components, and part (*c*) the two together in an unpolarized beam. Dots represent the end-on view of linear vibrations, and double-pointed arrows represent vibrations confined to the plane of the paper. Thus (*d*), (*e*), and (*f*) of the figure show how the vibrations in (*a*), (*b*), and (*c*) would appear if one were looking along the direction of the rays.

24.3 POLARIZING ANGLE AND BREWSTER'S LAW

Consider unpolarized light to be incident at an angle ϕ on a dielectric like glass, as shown in Fig. 24D(*a*). There will always be a reflected ray *OR* and a refracted ray *OT*. An experiment like the one described in Sec. 24.1 and shown in Fig. 24A shows that the reflected ray *OR* is partially plane-polarized and that only at a certain definite angle, about 57° for ordinary glass, is it plane-polarized. It was Brewster who first discovered that at this polarizing angle $\bar{\phi}$ the reflected and refracted rays are just 90° apart. This remarkable discovery enables one to correlate polarization with the refractive index

$$\frac{\sin \phi}{\sin \phi'} = n \qquad (24a)$$

Since at $\bar{\phi}$ the angle $ROT = 90°$, we have $\sin \bar{\phi}' = \cos \bar{\phi}$, giving

$$\frac{\sin \bar{\phi}}{\sin \bar{\phi}'} = \frac{\sin \bar{\phi}}{\cos \bar{\phi}} = n$$

$$n = \tan \bar{\phi} \qquad (24b)$$

This is *Brewster's law*, which shows that the angle of incidence for maximum polarization depends only on the refractive index. It therefore varies somewhat with wavelength, but for ordinary glass the dispersion is such that the polarizing angle $\bar{\phi}$ does not change much over the whole visible spectrum. This fact is readily verified by calculating $\bar{\phi}$ for several wavelengths, using the values of n from Table 23B, as suggested in Prob. 24.1 at the end of this chapter.

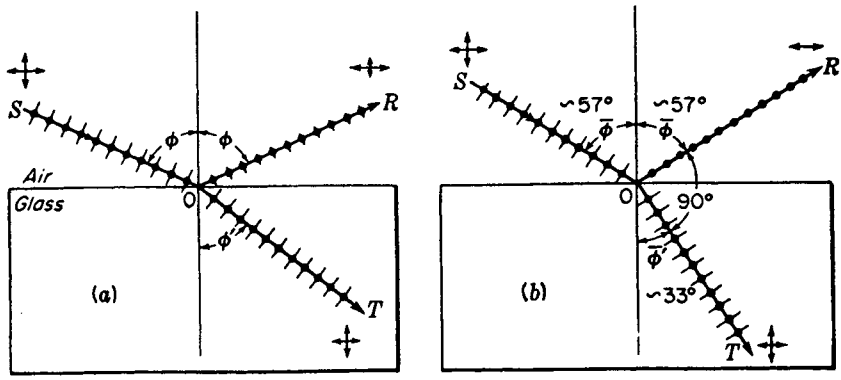


FIGURE 24D
(*a*) Polarization by reflection and refraction. (*b*) Brewster's law for the polarizing angle.

It is not difficult to understand the physical reason why the light vibrating in the place of incidence is not reflected at Brewster's angle. The incident light sets the electrons in the atoms of the material into oscillation, and it is the reradiation from these that generates the reflected beam. When the latter is observed at 90° to the refracted beam, only the vibrations that are perpendicular to the plane of incidence can contribute. Those in the plane of incidence have no component traverse to the 90° direction and hence cannot radiate in that direction. The reason is the same as that which causes the radiation from a horizontal radio-transmitter antenna to drop to zero along the direction of the wires. If the student keeps this picture in mind and remembers that light waves are strictly transverse, he will have no trouble remembering which of the two components is reflected at the polarizing angle.

24.4 POLARIZATION BY A PILE OF PLATES

Upon examining the refracted light in Fig. 24D(*a*) for polarization, it is found to be partially polarized for all angles of incidence, there being no angle at which the light is completely plane-polarized. The action of the reflecting surface may be described somewhat as follows. Let the ordinary incident light be thought of as being made up of two mutually perpendicular plane-polarized beams of light as shown in Sec. 24.2. Of those waves vibrating in the plane of incidence, i.e., in the plane of the page, part are reflected and part refracted for all angles with the single exception of the polarizing angle $\bar{\phi}$, for which all of the light is refracted. Of the waves vibrating perpendicular to the plane of incidence, some of the energy is reflected and the rest refracted for any angle of incidence. Thus the refracted ray always contains some of both planes of polarization. For a single surface of glass with $n = 1.50$, it will be shown later [Sec. 25.1 and Fig. 25B(*a*)] that at the polarizing angle 100 percent of the light vibrating parallel to the plane of incidence is transmitted, whereas for the perpendicular vibrations only 85 percent is transmitted, the other 15 percent being re-

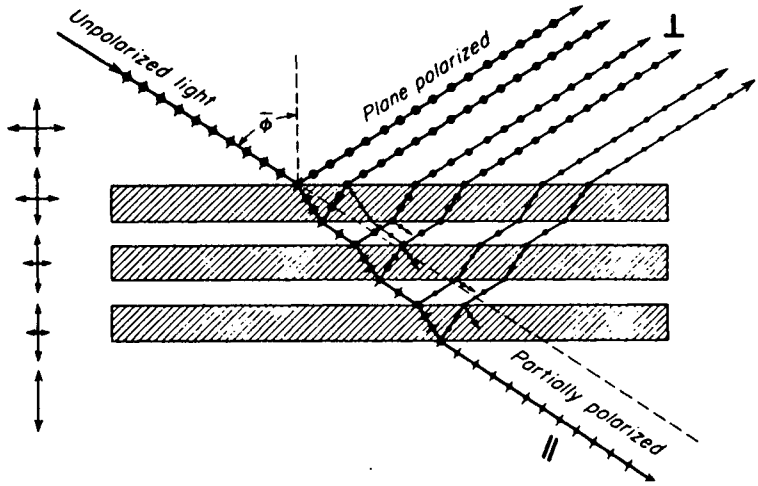


FIGURE 24E
Polarization of light by a pile of glass plates.

flected. Obviously the degree of polarization of the transmitted beam is small for a single surface.

If a beam of ordinary light is incident at the polarizing angle on a pile of plates, as shown in Fig. 24E, some of the vibrations perpendicular to the plane of incidence are reflected at each surface and all those parallel to it are refracted. The net result is that the reflected beams are all plane-polarized in the same plane, and the refracted beam, having lost more and more of its perpendicular vibrations, is partially plane-polarized. The larger the number of surfaces, the more nearly plane-polarized this transmitted beam is. This is illustrated by the vibration figures at the left in Fig. 24E. In a more detailed treatment of polarization by reflection and refraction (see Chap. 25), the polarizing angle for *internal* reflection is shown to correspond exactly to the angle of refraction $\bar{\phi}'$ in Fig. 24D(*b*). This means that light internally reflected at the angle $\bar{\phi}'$ will also be plane-polarized.

The degree of polarization P of the transmitted light can be calculated by summing the intensities of the parallel and perpendicular components. If these intensities are called I_p and I_s, respectively, it has been shown* that

$$P = \frac{I_p - I_s}{I_p + I_s} = \frac{m}{m + [2n^2/(1 - n^2)]} \tag{24c}$$

where m is the number of plates, that is, $2m$ surfaces, and n their refractive index. This equation shows that by the use of enough plates the degree of polarization can be made to approach unity, or ~ 100 percent. Better methods of producing a wide

* F. Provostaye and P. Desains, *Ann. Chim. Phys.*, 30:159 (1850). The calculation takes into account not only the ray going directly through but also those internally reflected two or more times (see Fig. 24E). It does not, however, include any effects of absorption, which would increase P somewhat above the value given by Eq. (24c).

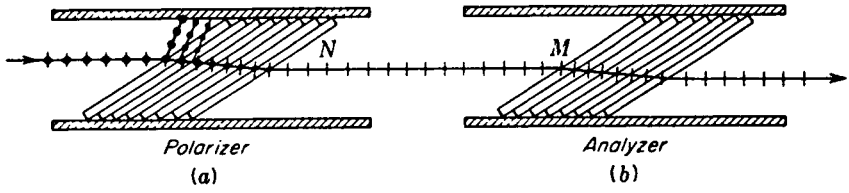


FIGURE 24F
Glass plates mounted at the polarizing angle $\bar{\phi}$.

beam of polarized light now available will be described below. The pile of plates may be used, however, to illustrate a common arrangement in the production and analysis of polarized light.

Figure 24F shows two such piles, the polarizer (*a*) and the analyzer (*b*), with their planes of incidence parallel. The light emerging at N is nearly plane-polarized and will be transmitted freely by the analyzer. Rotation of the latter by 90° about the line NM as an axis will cause the transmitted light to be nearly extinguished, since the vibrations are now perpendicular to the plane of incidence of the analyzer and will be reflected to the side. A further rotation of 90° will restore the light, and in a complete revolution there will be two maxima and two minima. Any arrangement of polarizer and analyzer in tandem is called a *polariscope* and has numerous uses.

24.5 LAW OF MALUS*

This law tells us how the intensity transmitted by the analyzer varies with the angle that its plane of transmission makes with that of the polarizer. In the case of two piles of plates, the plane of transmission is the plane of incidence, and for the law of Malus to hold we must assume that the transmitted light is completely plane-polarized. A better illustration would be the double reflection experiment of Sec. 24.1, or a combination of two polaroids or nicol prisms (see below), for which the polarization is complete. Then the law of Malus states that the transmitted intensity varies as the *square of the cosine* of the angle between the two planes of transmission.

The proof of the law rests on the simple fact that any plane-polarized vibration—let us say the one produced by our polarizer—can be resolved into two components, one parallel to the transmission plane of the analyzer and the other at right angles to it. Only the first of these gets through. In Fig. 24G, let A represent the amplitude transmitted by the polarizer for which the plane of transmission intersects the plane of the figure in the vertical dashed line. When this light strikes the analyzer, set at the angle θ, one can resolve the incident amplitude into components A_1 and A_2, the latter of which is eliminated in the analyzer. In the pile of plates, it is reflected to one side. The amplitude of the light that passes through the analyzer is therefore

$$A_1 = A \cos \theta \qquad (24d)$$

* Étienne Louis Malus (1775–1812). French army engineer. His discovery of polarization by reflection was made by accident when looking through a calcite crystal at the light reflected from the windows of the Luxembourg Palace.

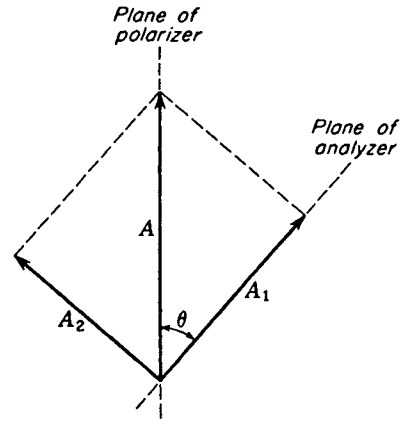


FIGURE 24G
Resolution into components of the amplitude of plane-polarized light.

and its intensity

$$I_1 = A_1^2 = A^2 \cos^2 \theta = I_0 \cos^2 \theta \qquad (24e)$$

Here I_0 signifies the intensity of the incident polarized light. This is, of course, one-half of the intensity of the unpolarized light striking the polarizer, provided one neglects losses of light by absorption in traversing it. There will also be losses in the analyzer. For Polaroids or nicols some light will be removed from the beam by reflection at the surfaces. Although these effects are neglected in deriving Eq. (24e), it will be noticed that they change only the constant in the equation and do not spoil the dependence of the *relative* intensity on $\cos^2 \theta$. Thus Malus' law is rigorously true and applies, for example, to the intensity of the twice-reflected light in the experiment of Sec. 24.1 even though its maximum value is only a small fraction of the original intensity. In such cases, the I_0 in Eq. (24e) is merely the intensity when the analyzer is parallel to the polarizer.

24.6 POLARIZATION BY DICHROIC CRYSTALS

These crystals have the property of selectivity absorbing one of the two rectangular components of ordinary light. Dichroism is exhibited by a number of minerals and by some organic compounds. Perhaps the best known of the mineral crystals is *tourmaline*. When a pencil of ordinary light is sent through a thin slab of tourmaline like T_1, shown in Fig. 24H, the transmitted light is polarized. This can be verified by a second crystal T_2. With T_1 and T_2 parallel to each other the light transmitted by the first crystal is also transmitted by the second. When the second crystal is rotated through 90°, no light gets through. The observed effect is due to a selective absorption by tourmaline of all light rays vibrating in one particular plane (called, for reasons explained below, the O vibrations) but not those vibrating in a plane at right angles (called the E vibrations). Thus in the figures shown, only the E vibrations parallel to the long edges of the crystals are transmitted, so that no light will emerge from the crossed crystals. Since tourmaline crystals are somewhat colored, they are not used in optical instruments as polarizing or analyzing devices.

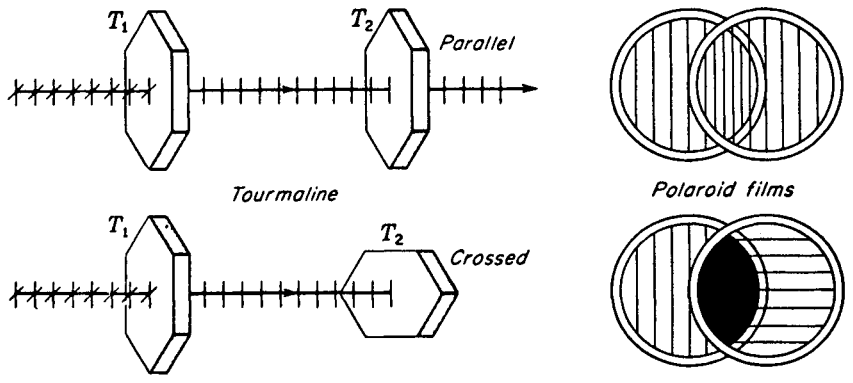


FIGURE 24H
Dichroic crystals and polarizing films in the *parallel* and *crossed* positions.

Attempts to produce polarizing crystals of large aperture were made by Herapath* in 1852. He was successful in producing good but small crystals of the organic compound quinine iodosulfate (now known as herapathite), which completely absorbs one component of polarization and transmits the other with little loss. One variety of Polaroid contains crystals of this substance. Polaroid was invented in 1932 by Land,† and has found uses in many different kinds of optical instruments. These films consist of thin sheets of nitrocellulose packed with ultramicroscopic polarizing crystals with their optic axes all parallel. In more recent developments the lining-up process is accomplished somewhat as follows. Polyvinyl alcohol films are stretched to line up the complex molecules and then are impregnated with iodine. From X-ray diffraction studies of these dichroic films, it can be seen that the iodine is present in polymeric form, i.e., as independent long strings of iodine atoms all lying parallel to the fiber axis, with a periodicity in this direction of about 3.10 Å. Films prepared in this way are called H-Polaroid. Land and Rogers found further that when an oriented transparent film of polyvinyl alcohol is heated in the presence of an active dehydrating catalyst such as hydrogen chloride, the film darkens slightly and becomes strongly dichroic. Such a film becomes very stable and, having no dyestuffs, is not bleached by strong sunlight. This so-called K-Polaroid is very suitable for polarizing uses such as automobile headlights and visors. Polarizing films are usually mounted between two thin plates of optical glass.

24.7 DOUBLE REFRACTION

The production and study of polarized light over a wider range of wavelengths than is afforded by Polaroid use the phenomenon of double refraction in crystals of *calcite* and *quartz*. Both these crystals are transparent to visible as well as ultraviolet light.

* W. B. Herapath, *Phil. Mag.*, 3:161 (1852).

† A good summary of the development of sheet polarizers is given by E. H. Land, *J. Opt. Soc. Am.*, 41:957 (1951).

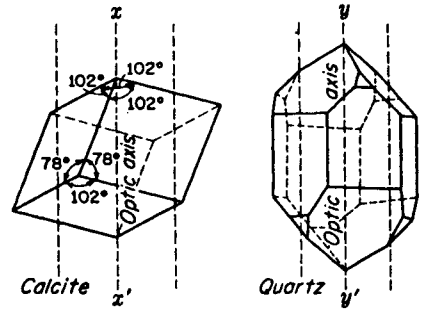


FIGURE 24I
Calcite and quartz crystal forms. The direction of the optic axis is indicated by broken lines.

Calcite, which chemically is calcium carbonate ($CaCO_3$), occurs in nature in a great variety of crystal forms (in the rhombohedral class of the hexagonal system), but it breaks readily into simple cleavage rhombohedrons of the form shown at the left in Fig. 24I. Each face of the crystal is a parallelogram whose angles are 78° 5′ and 101° 55′. If struck a blow with a sharp instrument, each crystal can be made to cleave or break along cleavage planes into two or more smaller crystals which always have faces that are parallelograms with angles shown in Fig. 24J.

Quartz crystals, on the other hand, are found in their natural state to have many different forms, one of the more complicated of which is shown at the right in Fig. 24I. Unlike calcite, quartz crystals will not cleave along crystal planes but will break into irregular pieces when given a sharp blow. Quartz is pure silica (SiO_2). Further details concerning these crystals will be given in this and the following chapters.

When a beam of ordinary unpolarized light is incident on a calcite or quartz crystal, there will be, in addition to the reflected beam, two refracted beams in place of the usual single one observed, for example, in glass. This phenomenon, shown in Fig. 24J for calcite, is called *double refraction*, or *birefringence*. Upon measuring the angles of refraction ϕ' for different angles of incidence ϕ, one finds that Snell's law of refraction

$$\frac{\sin \phi}{\sin \phi'} = n \qquad (24f)$$

holds for one ray but not for the other. The ray for which the law holds is called the *ordinary* or *O ray*, and the other is called the *extraordinary* or *E ray*.

Since the two opposite faces of a calcite crystal are always parallel, the two refracted rays emerge parallel to the incident beam and therefore parallel to each other. Inside the crystal the ordinary ray is always to be found in the plane of incidence. Only for special directions through the crystal is this true for the extraordinary ray. If the incident light is normal to the surface, the extraordinary ray will be refracted at some angle that is not zero and will come out parallel to, but displaced from, the incident beam; the ordinary ray will pass straight through without deviation. A rotation of the crystal about the *O* ray will in this case cause the *E* ray to rotate around the fixed *O* ray.

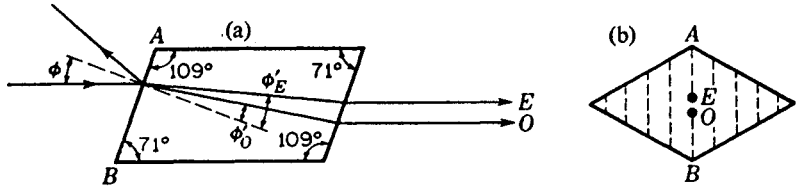


FIGURE 24J
Side and end views of the double refraction of light by a calcite crystal. (a) Cross-section of a principal plane. (b) End view.

24.8 OPTIC AXIS

Calcite and quartz are examples of *anisotropic* crystals, or ones in which the physical properties vary with direction. All crystals except those belonging to the cubic system are anisotropic to a greater or less degree. Furthermore, the two examples chosen show the simple type of anisotropy which characterizes *uniaxial* crystals. In these there is a single direction called the *optic axis*, which is an axis of symmetry with respect to both the crystal form and the arrangement of atoms. If any property, such as the heat conductivity, is measured for different directions, it is found to be the same along any line perpendicular to the optic axis. At other angles it changes, reaching a maximum or a minimum along the axis. The directions of the optic axes in calcite and quartz are shown in Fig. 24I.

The double refraction in uniaxial crystals disappears when the light is made to enter the crystal so that it travels along the optic axis. That is, there is no separation of the *O* and *E* rays in this case. This is also true in directions at right angles to the axis, but here the *O* and *E* rays behave differently in a less obvious respect, namely they differ in velocity. The consequences of the latter difference will be examined in Chap. 27.

The direction of the optic axis in calcite is determined by drawing a line like *xx'* through a *blunt corner* of the crystal, so that it makes equal angles with all faces. A blunt corner is one where three obtuse face angles come together, and there are only two such corners more or less opposite each other. In quartz the optic axis *yy'* runs lengthwise of the crystal, its direction being parallel to the six side faces, as shown. It should be emphasized that the optic axis is not a particular line through the crystal but a *direction*. That is, for any given point in the crystal an optic axis may be drawn which will be parallel to that for any other point.

24.9 PRINCIPAL SECTIONS AND PRINCIPAL PLANES

In specifying the positions of crystals, and also the directions of rays and vibrations, it is convenient to use the principal section, made by a plane containing the optic axis and normal to any cleavage face. For a point in calcite, there are therefore three principal sections, one for each pair of opposite crystal faces. A principal section always cuts the surfaces of a calcite crystal in a parallelogram with angles of 71° and 109°, as shown at the left in Fig. 24J. An end view of a principal section cuts the

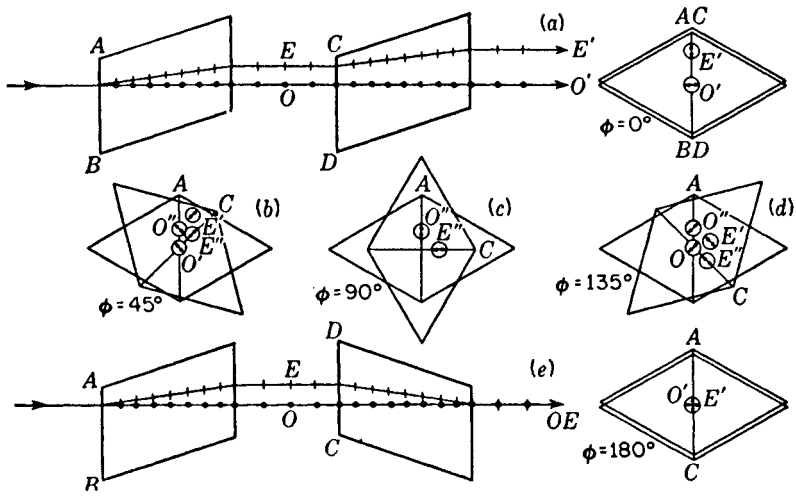


FIGURE 24K
Double refraction and polarization in two calcite crystals with their principal sections making different angles.

surface in a line parallel to AB, shown as a dotted line in the right-hand figure. All other planes through the crystal parallel to the plane represented by AB are also principal sections. These are represented by the other dotted lines.

The principal section, as so defined, does not always suffice in describing the directions of vibrations. Here we make use of the two other planes, the *principal plane of the ordinary ray*, a plane containing the optic axis and the ordinary ray, and the *principal plane of the extraordinary ray*, one containing the optic axis and the extraordinary ray. The ordinary ray always lies in the plane of incidence. This is not generally true for the extraordinary ray. The principal planes of the two refracted rays do not coincide except in special cases. The special cases are those for which the plane of incidence is a principal section as shown in Fig. 24J. Under these conditions the plane of incidence, the principal section, and the principal planes of the O and E rays all coincide.

24.10 POLARIZATION BY DOUBLE REFRACTION

The polarization of light by double refraction in calcite was discovered by Huygens in 1678. He sent a beam of light through two crystals as shown at the top of Fig. 24K. If the principal sections are parallel, the two rays O' and E' are separated by a distance equal to the sum of the two displacements found in each crystal if used separately. Upon rotation of the second crystal each of the two rays O and E is refracted into two parts, making four as shown by an end-on view in (*b*). At the 90°

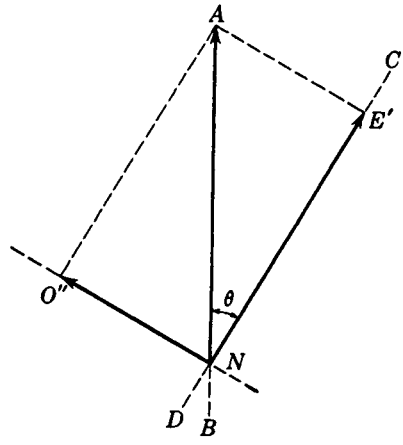


FIGURE 24L
Resolution of polarized light into components by double refraction.

position (*c*) the original O' and E' rays have faded and vanished and the new rays O'' and E'' have reached a maximum of intensity. Further rotation finds the original rays appearing, and eventually, if the crystals are of equal thickness, these rays come together into a single beam in the center for the 180° position shown at the bottom, the rays O'' and E'' having now vanished.

Thus, merely by using two natural crystals of calcite, Huygens was able to demonstrate the polarization of light. The explanation of the movement of the light rays is one simply of deviation by refraction and easily understood. The varying intensity of the spots, however, involves the polarization of the two light beams leaving the first crystal. In brief the explanation is somewhat as follows. Ordinary light upon entering the first calcite crystal is broken up into two plane-polarized rays, one, the *O* ray, vibrating perpendicular to the principal plane, which is here the same as the principal section, and the other, the *E* ray, vibrating in the principal section. In other words, the crystal resolves the light into two components by causing one type of vibration to travel one path and the other vibration to travel another path.

Consider more in detail now what happens to one of the plane-polarized beams from the first crystal when it passes through the second crystal oriented at some arbitrary angle θ. Let A in Fig. 24L represent the amplitude of the *E* ray vibrating parallel to the principal section of the first crystal just as it strikes the face of the second crystal. This second crystal, like the first, transmits light vibrating in its principal section along one path and light vibrating at right angles along another path. The *E* ray is therefore split up into two components E' with an amplitude $A \cos \theta$ and O'' with an amplitude $A \sin \theta$. These emerge from the second crystal with relative intensities given by $A^2 \cos^2 \theta$ and $A^2 \sin^2 \theta$. At $\theta = 90°$ E' vanishes and O'' reaches a maximum intensity of A^2. At all positions the sum of the two components, $A^2 \sin^2 \theta + A^2 \cos^2 \theta$, is just equal to A^2, the intensity of the incident beam.

The same treatment holds for the splitting up of the *O* beam from the first crystal into two plane-polarized beams O' and E''.

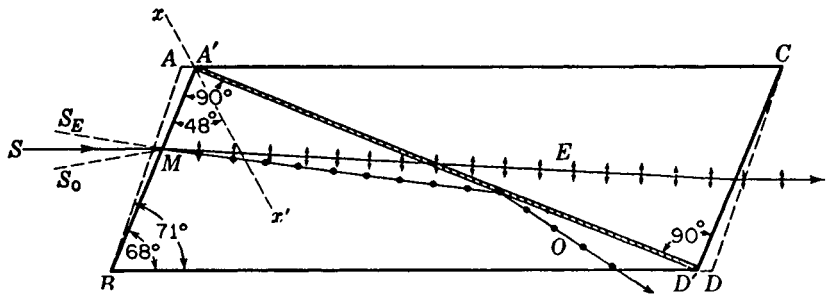


FIGURE 24M
Detailed diagram of a nicol prism, showing how it is made from a calcite crystal.

24.11 NICOL PRISM

This very useful polarizing device is made from a calcite crystal, and derives its name from its inventor.* The nicol prism is made in such a way that it removes one of the two refracted rays by total reflection, as is illustrated in Fig. 24M. There are several different forms of nicol prism,† but we shall describe here one of the commonest ones. First a crystal about 3 times as long as it is wide is taken and the ends cut down from 71° in the principal section to a more acute angle of 68°. The crystal is then cut apart along the plane $A'D'$ perpendicular to both the principal section and the end faces. The two cut surfaces are ground and polished optically flat and then cemented together with canada balsam. Canada balsam is used because it is a clear transparent substance with an index of refraction about midway between the index of the O and E rays. For sodium light,

Index of O ray	$n_O = 1.65836$
Index of canada balsam	$n_B = 1.55$
Index of E ray	$n_E = 1.48641$

Optically the balsam is denser than the calcite for the E ray and less dense for the O ray. The E ray therefore will be refracted into the balsam and on through the calcite crystal, whereas the O ray for large angles of incidence will be totally reflected. The critical angle for total reflection of the O ray at the first calcite to balsam surface is about 69° and corresponds to a limiting angle SMS_0 in Fig. 24M of about 14°. At greater angles than this, some of the O ray will be transmitted. This means that a nicol should not be used in light which is highly convergent or divergent.

The E ray in a nicol also has an angular limit, beyond which it will be totally

* William Nicol (1768–1851). Scotch physicist, who became very skillful in cutting and polishing gems and crystals. He devised his prism in 1828 and reportedly did not himself completely understand how it worked.

† Complete descriptions of polarizing prisms will be found in A. Johannsen, "Manual of Petrographic Methods," 2d ed., pp. 158–164, McGraw-Hill Book Company, New York, 1918.

reflected by the balsam. This is due to the fact that the index of refraction of calcite is different for different directions through the crystal. In the next chapter it will be seen that the index $n_E = 1.486$, as it is usually given, applies only in the special case of light traveling at right angles to the optic axis. Along the optic axis the E ray travels with the same speed as the O ray, and it therefore has the same index of 1.658. For intermediate angles the effective index lies between these two limits 1.486 and 1.658. There will therefore be a maximum angle SMS_E beyond which the balsam will be optically less dense than the calcite, and there will be total reflection of the E vibrations. The prism is so cut that this angle likewise is in the neighborhood of 14°. The direction of the incident light on a nicol therefore is limited on the one side to avoid having the O ray transmitted and on the other side to avoid having the E ray totally reflected. In practice, it is important to keep this limitation in mind.

Polarizing prisms are sometimes made with end faces cut perpendicular to the sides so that the light enters and leaves normal to the surface. The most popular one of this type, the *Glan-Thompson prism*, has an angular tolerance or aperture approaching 40°, hence much larger than that of the nicol. But this prism must be cut with the optic axis parallel to the end faces and is wasteful of calcite, large crystals of which are expensive and difficult to obtain. In another type the halves are held together so that there is a film of air between them instead of balsam. This device, called the *Foucault prism*, will transmit ultraviolet light. It has an angular aperture of only about 8°, however, and some difficulty is experienced with interference occurring in the air film.

24.12 PARALLEL AND CROSSED POLARIZERS

When two nicol prisms are lined up one behind the other, as shown in Fig. 24N, they form a good polariscope (Sec. 24.4). Positions (*a*) and (*c*) are referred to as *parallel polarizers*, and for them the E ray is transmitted. A loss of some 10 percent of the incident light is caused by reflection at the prism faces and absorption in the balsam layer, so that the overall transmission of a nicol for incident unpolarized light is about 40 percent. Position (*b*) in the figure represents one of the two positions called crossed polarizers. Here the E ray from the first nicol becomes an O ray in the second, and is totally reflected to the side. For intermediate angles, the incident E vibrations from the first nicol are broken up into components as shown by the vector diagram in Fig. 24L, where θ is the angle between the principal sections of the two nicols. The E' component is transmitted by the second nicol with the intensity $A^2 \cos^2 \theta$ and the O'' component is totally reflected. Parallel and crossed Polaroid filters are shown in Fig. 24H.

24.13 REFRACTION BY CALCITE PRISMS

Calcite prisms are sometimes cut from crystals for the purpose of illustrating double refraction and dispersion simultaneously as well as single refraction along the optic axis. Two regular prisms of calcite are shown in Fig. 24O, the first cut with the optic

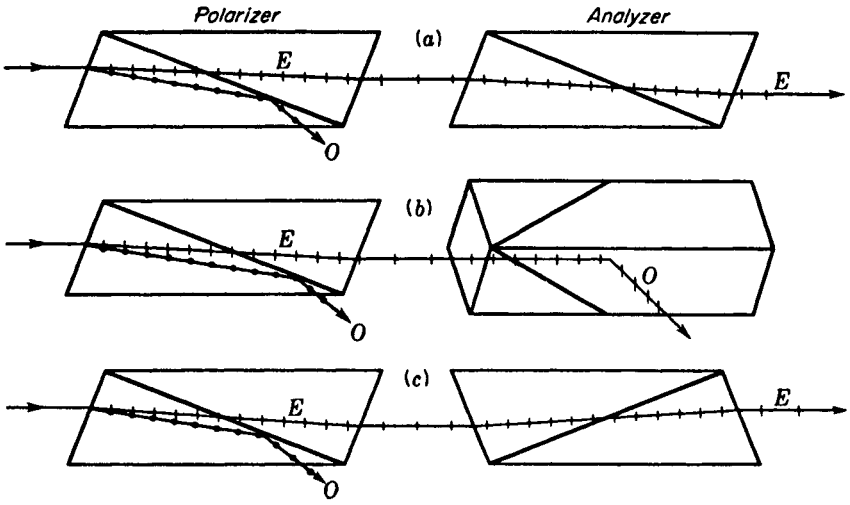


FIGURE 24N
Two nicol prisms mounted as polarizer and analyzer.

axis parallel to the refracting edge A, and the other with the axis also parallel to the base and perpendicular to the refracting edge. In the first prism there is double refraction for all wavelengths and hence two complete spectra of plane-polarized light, one with the electric vector parallel to the plane of incidence and the other with the electric vector perpendicular to it. An interesting demonstration of this polarization is accomplished by inserting a polarizer* into the incident or refracted beams. Upon rotation of the polarizer, first one spectrum is extinguished and then the other.

In the second prism, Fig. 24O(b), only one spectrum is observed, as with glass prisms. Here the light travels along the optic axis, or very nearly so, so that the two spectra are superposed. In this case a polarizer, when rotated, will not affect the

* Although nicol prisms give the most complete polarization of any of the devices commonly found in laboratories, polaroid films or a pile of glass plates mounted as in Fig. 24F are quite suitable for nearly all experimental demonstrations.

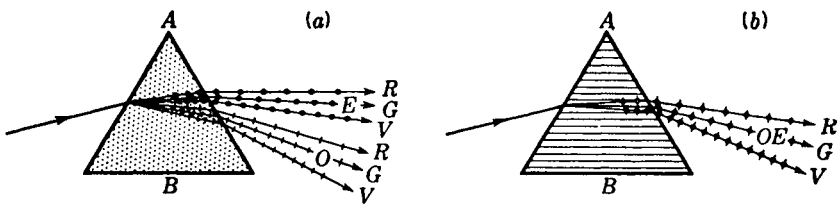


FIGURE 24O
Double and single refraction of white light by prisms cut at different angles from calcite crystals.

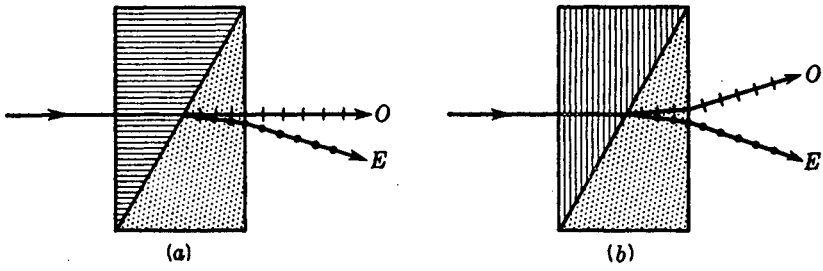


FIGURE 24P
Diagrams of (*a*) Rochon and (*b*) Wollaston prisms made from quartz.

intensity as it does with the first prism. The more detailed treatment of double refraction in Chap. 26 will clarify these experimental observations.

24.14 ROCHON AND WOLLASTON PRISMS

It is often desirable to split a light beam into two plane-polarized components, retaining both of them for a later comparison of their intensities. For this purpose other types of prisms have been designed, the most satisfactory of which are the Rochon and Wollaston prisms. These optical devices, sometimes called double-image prisms, are made of quartz or calcite cut at certain definite angles and cemented together with glycerin or castor oil.

In the Rochon prism [Fig. 24P(*a*)] the light, entering normal to the surface, travels along the optic axis of the first prism and then undergoes double refraction at the boundary of the second prism. The optic axis of the second prism is perpendicular to the plane of the page, as is indicated by the dots. In the Wollaston prism [Fig. 24P(*b*)] the light enters normal to the surface and travels perpendicular to the optic axis until it strikes the second prism, where double refraction takes place. The essential difference between the two is shown in the figures by the directions of the two refracted rays. The Rochon prism transmits the *O* vibrations without deviation, the beam being achromatic. This is frequently desired in optical instruments where only one plane-polarized beam is desired. The *E* beam, which is chromatic, is readily screened off at a sufficiently large distance from the prism.

The Wollaston prism deviates both rays and consequently yields greater separation of the two slightly chromatic beams. It is commonly used where a comparison of intensities is desired. These intensities will of course be equal for unpolarized light but will differ if the light is polarized in any way. It should be noted that in the Rochon prism the light should always enter from the left, in order for it to travel first along the optic axis, as shown in the figure. If it is sent in the other direction, the different wavelengths will emerge vibrating in different planes because of a phenomenon known as rotatory dispersion (see Sec. 28.2). This phenomenon, as well as the directions taken by the doubly refracted beams in quartz, will be treated in detail in the following chapters.