first on the screen and then on a distant wall. Describe in detail what is observed. Can *geometrical optics* account for the patterns observed?

3. Shining Laser Light Through a Pinhole. Arrange a 5-mW diode laser, pinhole (50 micrometers or so in diameter), and screen along an optical bench. Carefully align the laser beam so that it falls perpendicularly on the tiny pinhole. Observe the light that passes through the pinhole on a white cardboard screen. Make minor adjustments to the relative positions of the laser and pinhole to obtain the brightest pattern on the screen. Move the screen far enough away so you can see clearly (in a darkened room) the details of the light pattern. Describe what you see. Can *geometrical optics* account for the light pattern?

Basic Concepts

I. LIGHT WAVES AND PHYSICAL OPTICS

In our study of ray optics and image formation, we represented image points as "geometrical points," without physical extent. That, of course, followed logically since light rays were used to locate the image points and light rays are *lines* that intersect clearly at geometrical points. But in reality, if you were to examine such image points with a microscope, you would see *structure* in the "point," a structure explained only when you invoke the true *wave nature* of light.

In effect, then, we are saying that, with large objects such as prisms, mirrors, and lenses—large in the sense that their dimensions are millions of times that of the wavelength of light—interference and diffraction effects are still present in the imaging process, but they occur on so small a scale as to be hardly observable to the naked eye. To a good approximation, then, with "large" objects we are able to describe light imaging quite satisfactorily with geometrical (ray) optics and obtain fairly accurate results. But when light waves pass around small objects, such as a 100-μ-diameter human hair, or through small openings, such as a 50-μ pinhole, ray optics *cannot* account for the light patterns produced on a screen beyond these objects. Only wave optics leads to the correct interpretation of such patterns.

And so now we turn to a study of the wave nature of light and to the fascinating phenomena of *interference, diffraction*, and *polarization*—and of such devices as gratings and thin-film coatings. We shall see that interference occurs when two or more light waves pass through the same region and add to or subtract from each other. Diffraction occurs when light waves pass through small openings or around small obstacles and spread, and polarization occurs due to the transverse nature of the electric field vibration in a propagating electromagnetic wave. Before we look at these phenomena, let's review briefly the nature of waves, wave fronts, and wave motion.

A. Physics of waves and wave motion

Wave optics treats light as a series of propagating electric and magnetic field oscillations. While we cannot see these extremely rapid oscillations, their wave behavior is similar to that of water waves. Thus, we find it useful to picture waves and wave motion in terms of simple water waves, such as those created by a bobbing cork on an otherwise quiet pond. See Figure 4-1a.

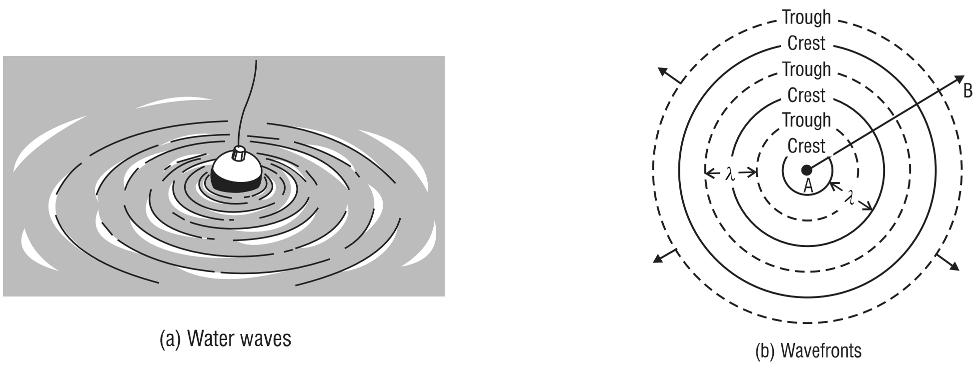
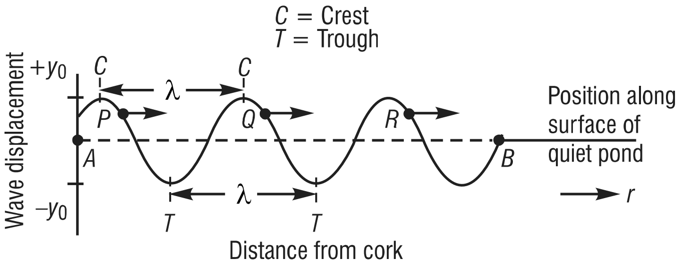


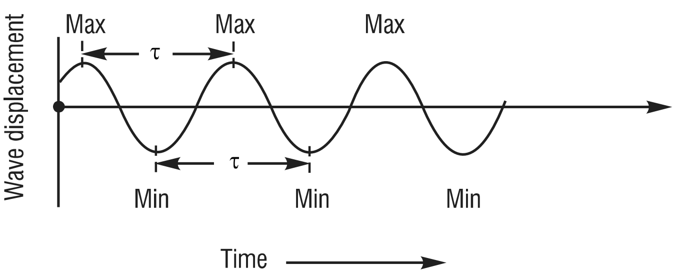
Figure 4-1 *Water waves and wave fronts*

The bobbing cork generates a series of surface disturbances that travel outward from the cork. Figure 4-1b shows the same disturbances traveling away from point *A* (the cork) as a series of successive *wave fronts* labeled *crests* and *troughs*. Recall that a *wave front* is a locus of points along which all phases and displacements are identical. The solid circles in Figure 4-1b depict the outward-moving wave crests; the dashed circles represent wave troughs. Adjacent crests are always a *wavelength* apart, as are the adjacent troughs.

If we were able to look along the surface of the pond, we would see a *sinusoid-like* profile of the traveling wave such as that shown in Figure 4-2a. The profile is a *snapshot* of the water displacement at a *certain instant of time* along a direction such as *AB*, labeled back in Figure 4-1b. The water surface rises to a maximum displacement ($+y_0$) and falls to a minimum displacement ($-y_0$) along the profile. As time varies, the "snapshot" profile in Figure 4-2a moves to the right with its characteristic wave speed. The radial distance outward from the cork at position *A*, shown in Figure 4-1b, is denoted by the variable *r* in Figure 4-2a.



(a) Wave profile along the pond at a *certain instant of time*



(b) Wave displacement *at a fixed* position on the pond as a function of time

Figure 4-2 *Two aspects of wave motion for a traveling wave*

Now suppose that—instead of looking along the surface of the pond—we look at the moving wave at one *definite position* on the pond, such as at point Q in Figure 4-2a. What happens to the wave displacement at this fixed position as the wave disturbances move away from the cork? We know from experience that the surface of the pond at Q rises and falls, repeatedly—as long as the wave disturbances move past this position. This wave displacement as a function of time—*at a fixed position*—is shown in Figure 4-2b. Note again that the shape is *sinusoid-like.*

Since we're concentrating on one position in Figure 4-2b, we cannot "see" the whole wave. All we see is the *up and down* motion of point Q. The time between successive maxima or successive minima is defined as the *period* (τ) of the wave. The number of times point Q goes from max to min to max per second is called the frequency (f) of the wave. The period τ and the frequency f are related by the simple relationship $f = 1/\tau$, as presented in Module 1-1, *Nature and Properties of Light.*

B. The mathematics of sinusoidal waveforms (optional)*

The two aspects of wave motion depicted in Figures 4-2a and 4-2b—one at a fixed time, the other at a fixed position—are addressed in a mathematical equation that describes a *sinusoidally varying* traveling wave. Refer to Equation 4-1,

$$\boxed{y(r, t) = y_0 \sin\left[\frac{2\pi}{\lambda}(r - \mathrm{v}t)\right]} \qquad (4\text{-}1)$$

where: $y(r, t)$ is the wave displacement at position r and time t

y_0 is the wave amplitude as shown in Figure 4-2a

λ is the wavelength

r is the position along the traveling wave

v is the wave speed, equal to $\lambda \times f$, and

t is the time

If we "freeze" time at some value t_0, for example, we obtain the specialized equation $y(r, t_0) = y_0 \sin\left[\frac{2\pi}{\lambda}(r - \text{constant})\right]$. This is a mathematical description of the wave profile shown in Figure 4-2a. On the other hand, if we select a fixed position r_0, we obtain another specialized equation $y(r_0, t) = y_0 \sin\left[\frac{2\pi}{\lambda}(\text{constant} - \mathrm{v}t)\right]$. This is a mathematical description of the waveform shown in Figure 4-2b.

The factor in brackets in Equation 4-1 defines the *phase angle* ϕ of the wave at position r and time t. Thus,

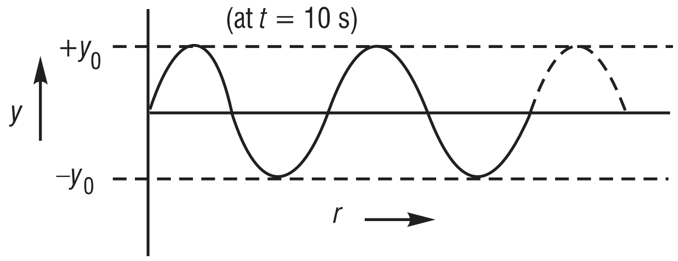
*The text material in this section, through Example 1, is optional. Depending on the background of the class, this section may or may not be covered.

$$\boxed{\phi = \left[\frac{2\pi}{\lambda}(r - \mathrm{v}t)\right]} \tag{4-2}$$

The phase angle is the *same for any point on a given wave front*, as mentioned earlier. For example, for successive wave fronts whose values of ϕ are $\left[\pi/2\right]$, $\left[\pi/2 + 2\pi\right]$, $\left[\pi/2 + 4\pi\right]$, and so on—always 2π radians (360°) apart—sin ϕ for *each of these angles* equals +1, so that $y(r, t)$ equals $+y_0$, a maximum *positive* displacement. Such wave fronts are crests. Similarly, for successive wave fronts whose values of ϕ are $\left[3\pi/2\right]$, $\left[3\pi/2 + 2\pi\right]$, $\left[3\pi/2 + 4\pi\right]$, etc., always 2π radians apart, sin ϕ for *each of these angles* equals −1, so that $y(r, t)$ equals $-y_0$, the maximum *negative* wave displacement. Such wave fronts are troughs. And so it goes for all other wave fronts between the crests and troughs. For example, points P, Q, and R in Figure 4-2a, all with the same wave displacement, represent wave fronts a wavelength apart with phase angles of values differing by 2π. Example 1 provides an application of Equations 4-1 and 4-2 to circular water waves on a quiet pond.

Example 1

Circular water waves such as those shown in Figures 4-1a and 4-1b move outward from a bobbing cork at A. The cork bobs up and down and back again—a complete cycle—once per second, and generates waves that measure 10 cm from crest to crest. Some time after the wave motion has been established, we begin to time the motion with a stopwatch. At a certain time t = 10 s on the watch, we notice that the wave profile has the shape shown below.



(a) What is the wave frequency f for this water wave?

(b) What is its wavelength λ?

(c) What is its wave speed v?

(d) What is the phase angle ϕ for a wave front at position r = 102.5 cm at time t = 10 s?

(e) What is the wave displacement y on the wave front at r = 102.5 cm?

(f) What is the phase angle ϕ for a wave front at r = 107.5 cm at t = 10 s?

(g) What is the wave displacement y on the wave front at r = 107.5 cm?

(h) If we focus on the wave motion at the position r = 105 cm and let time vary, what kind of motion do we observe?

Solution:

(a) The wave frequency is 1 cycle/s; (therefore, the period $\tau = 1/f$ is 1 second).

II. INTERACTION OF LIGHT WAVES

A. The principle of superposition

An understanding of light wave interference begins with an answer to the question, “What happens at a certain position in space when two light waves pass through that position at the same time? To answer this question, we invoke the *principle of superposition*, which states:

> *When two or more waves move simultaneously through a region of space, each wave proceeds independently as if the other were not present. The resulting wave “displacement” at any point and time is the vector sum of the “displacements” of the individual waves.*

This principle holds for water waves, mechanical waves on strings and on springs (the Slinky!), and for sound waves in gases, liquids and solids. Most important for us, it holds for all electromagnetic waves in free space. So, if we have two light waves passing through some common point P, where Wave 1 alone causes a “displacement” Y_1 and Wave 2 alone a displacement Y_2, the *principle of superposition* states that the resultant displacement Y_{RES} is given by a *vector* sum of the two displacements. If both displacements are along the same direction—as they will be for most applications in this module—we can add the two displacements *algebraically*, as in Equation 4-3.

$$\boxed{Y_{RES} = Y_1 + Y_2} \tag{4-3}$$

An application of Equation 4-3 is shown in Figure 4-4, where Wave 1 and Wave 2 are moving along the x-direction to the right. Wave 2 is drawn with ¾ the amplitude and ½ the wavelength of Wave 1. The resultant wave, obtained by applying Equation 4-3 at each point along the x-direction, is shown by the solid waveform, Y_{RES}.

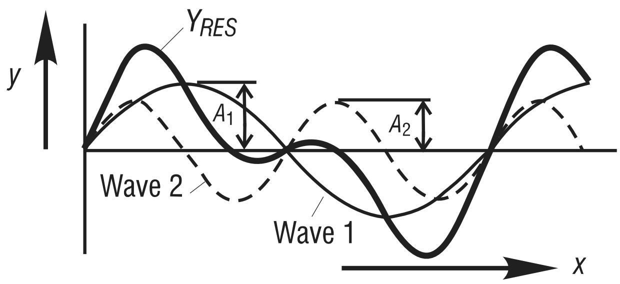


Figure 4-4 *Superposition of two waves moving along the same direction*

In Figure 4-5, we show the interference of two sinusoidal waves of the same amplitude and same frequency, traveling in the same direction. The two waves are represented by the light solid and broken curves, the resultant by the solid heavy curve. In Figure 4-5a the two waves are *exactly in phase*, with their maximum and minimum points matching perfectly. Applying the *principle of superposition* to the two waves, the resultant wave is seen to have the same amplitude and frequency but *twice* the amplitude 2A of either initial wave. This is an example of *constructive interference*. In Figure 4-5b the two curves are *exactly out of phase*, with the crest of one falling on the trough of the other, and so on. Since one wave effectively cancels the effect of the other at each point, the resultant wave has *zero displacement* everywhere, as indicated by the solid black line. This is an example of *destructive interference.* In Figure 4-5c, the two waves are neither completely in phase nor completely out of phase. The resultant wave then has an amplitude somewhere between A and 2A, as shown.

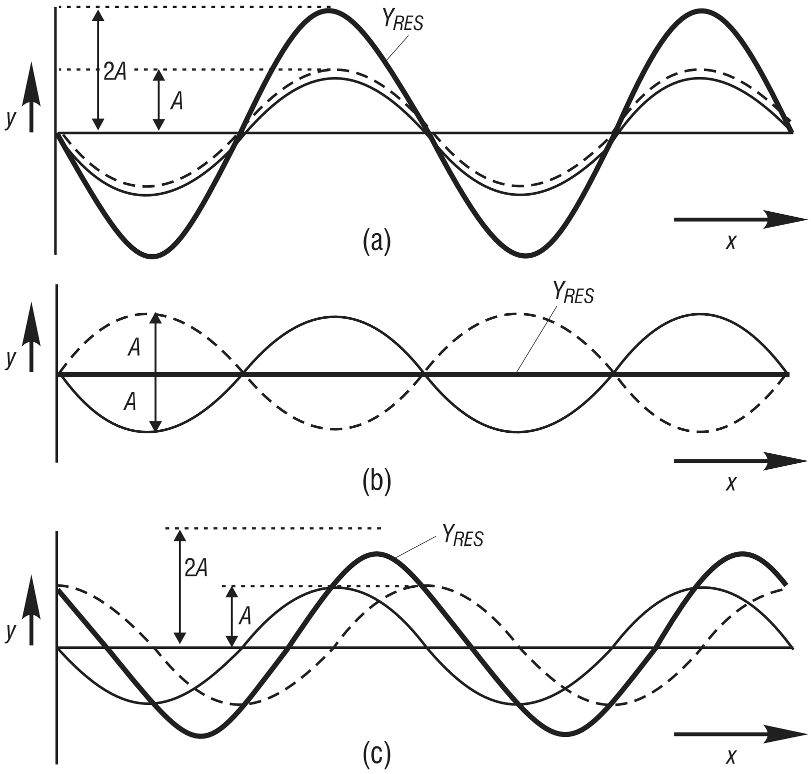


Figure 4-5 *Interference of two identical sinusoidal waves*

B. Huygens' wavelets

Long before people understood the electromagnetic character of light, Christian Huygens—a 17th-century scientist—came up with a technique for propagating waves from one position to another, determining, in effect, the shapes of the developing wave fronts. This technique is basic to a quantitative study of interference and diffraction, so we cover it here briefly. Huygens claimed that:

> *Every point on a known wave front in a given medium can be treated as a point source of secondary wavelets* (spherical waves "bubbling" out of the point, so to speak) *which spread out in all directions with a wave speed characteristic of that medium. The developing wave front at any subsequent time is the envelope of these advancing spherical wavelets.*

Figure 4-6 shows how Huygens' principle is used to demonstrate the propagation of successive (a) *plane* wave fronts and (b) *spherical* wave fronts. Huygens' technique involves the use of a series of points $P_1 \ldots P_8$, for example, on a given wave front defined at a time $t = 0$. From these points—as many as one wishes, actually—spherical wavelets are assumed to emerge, as shown in Figures 4-6a and 4-6b. Radiating outward from each of the *P*-points, with a speed v, the series of secondary wavelets of radius $r = vt$ defines a new wave front at some time t later. In Figure 4-6a the new wave front is drawn as an *envelope tangent* to the secondary wavelets at a distance $r = vt$ from the initial plane wave front. It is, of course, another *plane* wave front. In Figure 4-6b, the new wave front at time t is drawn as an *envelope tangent* to the secondary wavelets at a distance $r = vt$ from the initial spherical wave front. It is an advancing *spherical* wave front.

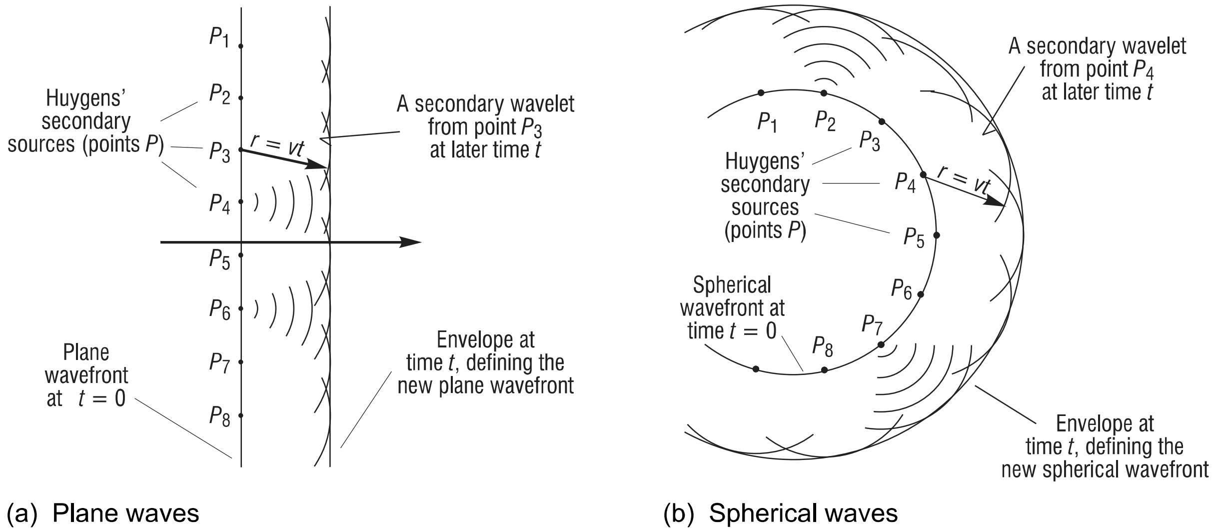


Figure 4-6 *Huygens' principle applied to the propagation of plane and spherical wave fronts*

While there seems to be no physical basis for the existence of Huygens' "secondary" point sources, Huygens' technique has enjoyed extensive use, since it does predict accurately—with waves, not rays—both the *law of reflection* and *Snell's law of refraction*. In addition, Huygens' principle forms the basis for calculating, for example, the diffraction pattern formed with multiple slits. We shall soon make use of Huygens' secondary sources when we set up the problem for diffraction from a single slit.

III. INTERFERENCE

Today we produce interference effects with little difficulty. In the days of Sir Isaac Newton and Christian Huygens, however, light interference was not easily demonstrated. There were several reasons for this. One was based on the extremely short wavelength of visible light—around 20 millionths of an inch—and the obvious difficulty associated with seeing or detecting interference patterns formed by overlapping waves of so short a wavelength, and so rapid a vibration—around a million billion cycles per second! Another reason was based on the difficulty—before the laser came along—of creating *coherent* waves, that is, waves with a phase relationship with each other that remained *fixed* during the time when interference was observed.

It turns out that we *can* develop phase coherence with *nonlaser* light sources to demonstrate interference, but we must work at it. We must "prepare" light from readily available incoherent light sources—which typically emit individual, uncoordinated, short wave trains of fixed phase of no longer than 10^{-8} seconds—so that the light from such sources remains coherent over periods of time long enough to overlap and produce visible interference patterns. There are generally two ways to do this.

- Develop several coherent *virtual* sources from a single incoherent "point" source with the help of mirrors. Allow light from the two virtual sources to overlap and interfere. (This method is used, for example, in the Loyd's mirror experiment.)
- Take monochromatic light from a single "point" source and pass it through two small openings or slits. Allow light from the two slits to overlap on a screen and interfere.

We shall use the second of these two methods to demonstrate Thomas Young's famous *double-slit experiment,* worked out for the first time at the very beginning of the 19th century. But first, let's consider the basics of interference from two point sources.

A. Constructive and destructive interference

Figure 4-7 shows two "point" sources of light, S and S', whose radiating waves maintain a fixed phase relationship with each other as they travel outward. The emerging waves are in effect spherical, but we show them as circular in the two-dimensional drawing. The solid circles represent crests, the dashed circles, troughs.

Earlier, in Figure 4-5a, we saw the effect of *constructive interference* for waves perfectly in phase and, in Figure 4-5b, the effect of *destructive interference* for waves perfectly out of phase. In Figure 4-7, along directions OP, OP_2, and OP_2' (emphasized by solid dots) crests from S and S' meet (as do the troughs), thereby creating a condition of *constructive interference*. As a result, light striking the screen at points P, P_2, and P_2' is at a maximum intensity and a bright spot appears. By contrast, along directions OP_1 and OP_1' (emphasized by open circles) crests and troughs meet each other, creating a condition of *destructive interference*. So at points P_1 and P_1' on the screen, no light appears, leaving a dark spot.

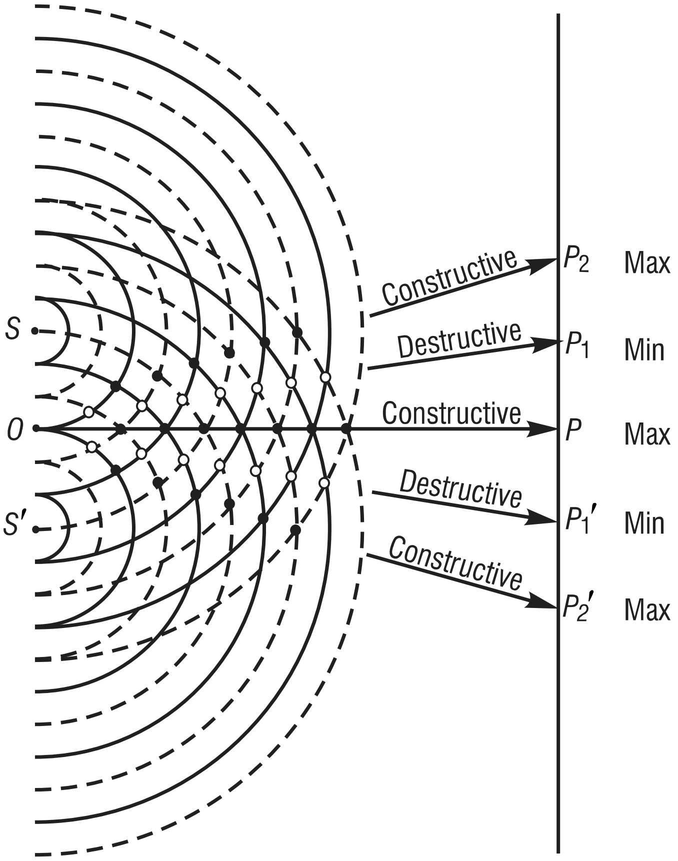


Figure 4-7 *Wave interference created by overlapping waves from coherent sources* S *and* S′

The requirement of *coherent sources* is a stringent requirement if interference is to be observed. To see this clearly, suppose for a moment that sources S and S' in Figure 4-7 are, in fact, *two corks* bobbing up and down on a quiet pond. As long as the two corks maintain a *fixed* relationship between their vertical motions, each will produce a series of *related* crests and troughs, and observable interference patterns in the overlap region will occur. But if the two corks bob up and down in a *random, disorganized manner,* no series of related, fixed-phase crests and troughs will form and no interference patterns of sufficiently long duration can develop, and so interference will not be observed.

B. Young's double-slit interference experiment

Figure 4-8a shows the general setup for producing *interference* with coherent light from two slits S_1 and S_2. The source S_0 is a monochromatic point source of light whose spherical wave fronts (circular in the drawing) fall on the two slits to create secondary sources S_1 and S_2. Spherical waves radiating out from the two secondary sources S_1 and S_2 maintain a fixed phase relationship with each other as they spread out and overlap on the screen, to produce a series of alternate bright and dark regions, as we saw in Figure 4-7. The alternate regions of bright and dark are referred to as *interference fringes*. Figure 4-8b shows such interference fringes, greatly expanded, for a small central portion of the screen shown in Figure 4-8a.

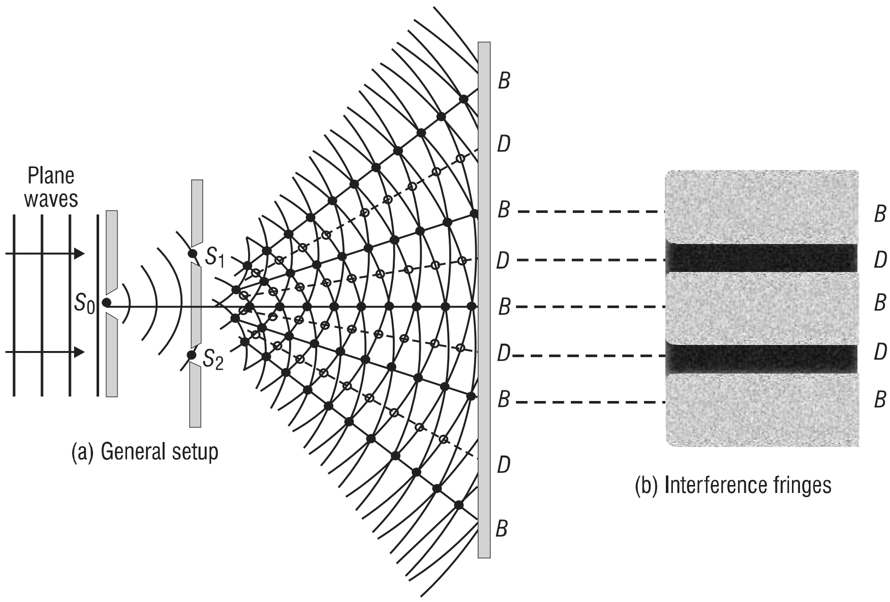


Figure 4-8 *Young's double-slit interference experiment showing (a) general setup and (b) typical interference fringes*

1. Detailed analysis of interference from a double slit: With the help of the *principle of superposition,* we can calculate the positions of the alternate maxima (bright regions) and minima (dark regions) shown in Figure 4-8. To do this we shall make use of Figure 4-9 and the following conditions:

(a) Light from slits S_1 and S_2 is coherent; that is, there exists a fixed phase relationship between the waves from the two sources.

(b) Light from slits S_1 and S_2 is of the same wavelength.

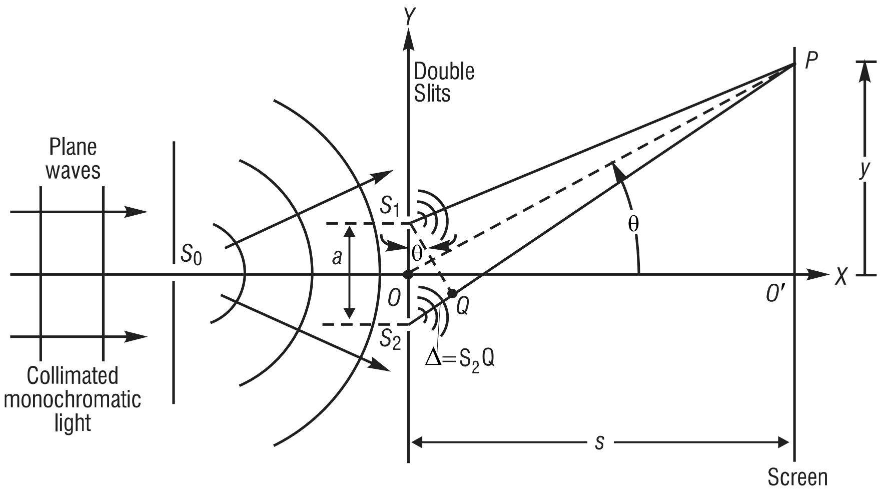


Figure 4-9 *Schematic for double-slit interference calculations. Source* S_0 *is generally a small hole or narrow slit; sources* S_1 *and* S_2 *are generally long, narrow slits perpendicular to the page.*

In Figure 4-9, light waves from S_1 and S_2 spread out and overlap at an arbitrary point P on the screen. If the overlapping waves are in phase, we expect a bright spot at P; if they are out of phase, we expect a dark spot. So the *phase difference* between the two waves arriving at point P is a key factor in determining what happens there. We shall express the phase difference in terms of the *path difference*, which we can relate to the wavelength λ.

For clarity, Figure 4-9 is not drawn to scale. It will be helpful in viewing the drawing to know that, in practice, the distance s from the slits to the screen is about *one meter*, the distance a between slits is *less than a millimeter*, so that the angle θ in triangle S_1S_2Q, or triangle OPO', is quite small. And on top of all this, the wavelength of light is a fraction of a micrometer.

The path difference Δ between S_1P and S_2P, as seen in Figure 4-9, is given by Equation 4-4, since the distances PS_1 and PQ are equal and since $\sin\theta = \Delta/a$ in triangle S_1S_2Q.

$$\boxed{\Delta = S_2P - S_1P = S_2Q = a\sin\theta} \qquad (4\text{-}4)$$

If the path difference Δ is equal to λ or some integral multiple of λ, the two waves arrive at P in phase and a bright fringe appears there (constructive interference). The condition for bright (B) fringes is, then,

$$\boxed{\Delta_B = a\sin\theta = m\lambda \text{ where } m = 0, \pm 1, \pm 2, \ldots} \qquad (4\text{-}5)$$

The number m is called the *order number*. The central bright fringe at $\theta = 0$ (point 0′ on the screen) is called the zeroth-order maximum ($m = 0$). The first maximum on either side, for which $m = \pm 1$, is called the *first-order maximum*, and so on.

If, on the other hand, the path difference at P is an odd multiple of λ/2, the two waves arrive out of phase and create a dark fringe (destructive interference). The condition for dark (D) fringes is given by Equation 4-6.

$$\boxed{\Delta_D = a\sin\theta = (m + \tfrac{1}{2})\lambda \text{ where } m = 0, \pm 1, \pm 2, \ldots} \qquad (4\text{-}6)$$

Since the angle θ exists in both triangles S_1S_2Q and OPO', we can find an expression for the *positions* of the bright and dark fringes along the screen. Because θ is small, as mentioned above, we know that $\sin\theta \cong \tan\theta$, so that for triangle OPO' we can write

$$\sin\theta \cong \tan\theta = \frac{y}{\lambda s} \qquad (4\text{-}7)$$

Combining Equation 4-7 with Equations 4-5 and 4-6 in turn, by substituting for sin θ in each, we obtain expressions for the position y of bright and dark fringes on the screen.

$$\boxed{y_B = \frac{\lambda s}{a} m \text{ where } m = 0, \pm 1, \pm 2, \ldots} \qquad (4\text{-}8)$$

and

$$\boxed{y_D = \frac{\lambda s}{a}\left(m + \tfrac{1}{2}\right) \text{ where } m = 0, \pm 1, \pm 2, \ldots} \tag{4-9}$$

In Example 2, through the use of Equation 4-8, we recreate the method used by Thomas Young to make the *first* measurement of the wavelength of light.

Example 2

A double-slit source with slit separation 0.2 mm is located 1.2 m from a screen. The distance between successive bright fringes on the screen is measured to be 3.30 mm. What is the wavelength of the light?

Solution: Using Equation 4-8 for any two adjacent bright fringes, we can obtain an equation for Δy, the fringe separation. Thus,

$$\Delta = \left(y_B\right)_{m+1} - \left(y_B\right)_m = \frac{\lambda s(m+1)}{a} - \frac{\lambda s(m)}{a} = \frac{\lambda s}{a}$$

$$\therefore\ \Delta y = \frac{\lambda s}{a}, \text{so that } \lambda = \frac{(\Delta y)a}{s}, \text{ giving}$$

$$\lambda = \frac{\left(3.30 \times 10^{-3}\,\text{m}\right)\left(2 \times 10^{-4}\,\text{m}\right)}{1.2\ \text{m}} = 5.5 \times 10^{-7}\,\text{m} = 550 \times 10^{-9}\,\text{m}$$

So the wavelength is about 550 nm and the light is yellowish green in color.

2. Intensity variation in the interference pattern. Knowing how to locate the positions for the fringes on a screen, we might now ask, "How does the brightness (intensity) of the fringes vary as we move, in either direction, from the central bright fringe ($m = 0$)?" We obtain a satisfactory answer to this question by representing the two separate electric fields at point P, the one coming from S_1 as $E_1 = E_0 \sin 2\pi ft$ and the one from S_2 as $E_2 = E_0 \sin(2\pi ft + \delta)$. The waves are assumed to have the same amplitude E_0. Here δ is the *phase angle* difference between the two waves arriving at P. The *path difference* Δ is related to the *phase angle* δ by the relationship

$$\boxed{\frac{\delta}{\Delta} = \frac{2\pi}{\lambda}} \tag{4-10}$$

so that if $\Delta = \lambda$, $\delta = 2\pi$ rad = 360°, if $\Delta = \lambda/2$, $\delta = \pi$ rad = 180°, and so on.

Then, by using the *principle of superposition,* we can add the two electric fields at point P to obtain $E_{RES} = E_1 + E_2$. (Carrying out this step involves some trigonometry, the details of which can be found in most optics texts.) Since the intensity I of the light goes as the *square* of the electric field E, we square E_{RES} and *average* the result over one cycle of wave oscillation at P, obtaining, finally, an expression for the average intensity, I_{AV}.

$$\boxed{I_{AV} = I_0 \cos^2 \frac{\delta}{2}} \tag{4-11}$$

Here δ is the critical *phase angle difference* at point P. For all points P for which $\delta = 0, 2\pi, 4\pi$, and so on, corresponding to $\Delta = 0, \lambda, 2\lambda$, etc., $\cos^2\left(\frac{\delta}{2}\right) = 1$ and $I_{AV} = I_0$, the maximum possible "brightness." At these points, bright fringes form. For $\delta = \pi, 3\pi, 5\pi$, and so on, corresponding to $\Delta = \lambda/2, 3\,\lambda/2, 5\,\lambda/2$, etc., $\cos^2\left(\frac{\delta}{2}\right) = 0$, and dark fringes form.

The maximum intensity I_0 is equal to $(E_0 + E_0)^2$ or $4E_0^2$, since each wave has amplitude E_0. Further, from Equations 4-10 and 4-4, we see that

$$\delta = \frac{2\pi}{\lambda}\Delta = \frac{2\pi}{\lambda}a\sin\theta \tag{4-12}$$

so that the phase angle δ is connected clearly through the angle θ to different points P on the screen. Going one step further, replacing $\sin\theta$ by $\frac{y}{s}$ in Equation 4-12, we have the connection between δ and any position y on the screen, such that

$$\delta = \frac{2\pi a}{\lambda s}y \tag{4-13}$$

With Equation 4-13 and $I_0 = 4\,E_0^2$, we can rewrite Equation 4-11 in a form that relates I_{AV} directly to a position y on the screen.

$$\boxed{I_{AV} = 4E_0^2\cos^2\left(\frac{\pi a}{\lambda s}y\right)} \tag{4-14}$$

where: I_{AV} = intensity of light along screen at position y

E_0 = amplitude of light wave from S_1 or S_2

s = distance from the plane of the double slit to the screen

a = slit separation

λ = wavelength of monochromatic light

y = distance above (or below) central bright fringe on the screen

Example 3

Using Equation 4-14 and the double-slit arrangement described in Example 2, determine how I_{AV} varies along the screen as a function of y.

Solution:

$I_{AV} = 4E_0^2\cos^2\left(\frac{\pi a}{\lambda s}y\right)$, where $a = 2 \times 10^{-4}$ m, $\lambda = 550 \times 10^{-9}$ m, and $s = 1.2$ m

$$I_{AV} = 4E_0^{\ 2} \cos^2\left(\frac{\pi\left(2\times 10^{-4}\right)y}{550\times 10^{-9}(1.2)}\right)$$

$$I_{AV} = 4E_0^{\ 2} \cos^2(303\ \pi y)$$

Note that, when $y = \frac{1}{303}, \frac{2}{303}, \frac{3}{303}$ and so on, the angle (303 πy) becomes π rad, 2π rad, 3π rad, and so on, for which $\cos^2(303\ \pi y)$ is always 1. At these values of y, we have the first order, second order and third order of bright fringes—each of intensity $I_{AV} = 4E_0^2$. Since the interval Δy between successive fringes is $\frac{1}{303}$ meter, we get $\Delta y = 3.3 \times 10^{-3}$ m or 3.3 mm, in agreement with the value of Δy given in Example 2.

C. Thin-film interference

Interference effects provide us with the rainbow of colors we often see on thin-film soap bubbles and "oil slicks." Each is an example of the interference of white light reflecting from opposite surfaces of the thin film. When thin films of different refractive indexes and thicknesses are judiciously stacked, coatings can be created that either enhance reflection greatly (HR coats) or suppress reflection (AR coats). A basic appreciation of these phenomena begins with an understanding of interference in a *single thin film*.

1. Single-film interference. The geometry for thin-film interference is shown in Figure 4-10. We assume that the light strikes the film—of thickness t and refractive index n_f—at near-perpendicular incidence. In addition we take into account the following established facts:

- A light wave traveling from a medium of lower refractive index to a medium of higher refractive index *automatically undergoes a phase change of* π *(180°)* upon reflection. A light wave traveling from a medium of higher index to one of lower index *undergoes no phase change* upon reflection. (We state this without proof.)
- The wavelength of light λ_n in a medium of refractive index n is given by $\lambda_n = \lambda_0/n$, where λ_0 is the wavelength in a vacuum or, approximately, in air.

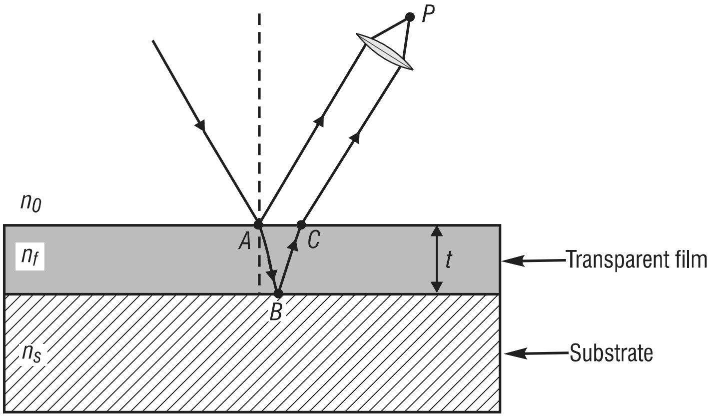


Figure 4-10 *Two-beam interference from a thin film. Rays reflected from the film's top and bottom plane surfaces are brought together at* P *by a lens.*

In Figure 4-10, we show a light beam in medium of index n_0 incident on the transparent film of index n_f. The film itself rests on a substrate of index n_s. Generally, the initial medium is air, so that $n_0 = 1$. The beam incident on the film surface at *A* divides into reflected and refracted portions. The refracted beam reflects again at the film-substrate interface at *B* and leaves the film at *C*, in the same direction as the beam reflected at *A*. Part of the beam may reflect internally again at *C* and continue to experience multiple reflections within the film layer until it has lost its intensity. There will thus exist multiple parallel beams emerging from the top surface, although with rapidly diminishing amplitudes.

Unless the reflectance of the film is large, a *good approximation* to the more complex situation of multiple reflection is to consider only *the first two emerging beams*. The two parallel beams leaving the film at *A* and *C* can be brought together by a converging lens, the eye, for example. The two beams intersecting at *P* overlap and interfere. Since the two beams travel different paths from point *A* onward, one in air, the other partly in the film, a relative phase difference develops that can produce *constructive* or *destructive* interference at *P*. The optical path difference Δ—in the case of *normal incidence*—is the additional path length *ABC* traveled by the refracted ray. The *optical path difference* in the film is equal to the product of the geometrical path difference (*AB* + *BC*) times the refractive index of the film. If the incident ray is nearly perpendicular to the surface, the path difference (*AB* +*BC*) is approximately equal to twice the film thickness 2*t*. Then,

$$\boxed{\Delta = n(AB + BC) = n(2t)} \qquad (4\text{-}15)$$

where *t* is the film thickness. For example, if $2nt = \lambda_0$, the wavelength of the light in air, the two interfering beams—on the basis of optical path difference alone—would be in phase and produce constructive interference.

However, an additional phase difference, due to the phenomenon mentioned above—*phase change on reflection*—must be considered. Suppose that $n_f > n_0$ and $n_f > n_s$. Often, in practice, $n_0 = n_s$, because the two media bounding the film are identical, as in the case of a water film (soap bubble) in air. Then the reflection at *A* occurs with light going from a lower index n_0 (air) toward the higher index n_f (film). The reflection at *B*, on the other hand, occurs for light going from a higher index n_f (film) toward a lower index n_s (air). Thus, the light reflecting at *A* shifts phase by 180° (equivalent to one-half wavelength) while the light reflecting at B does not. As a result, if $2nt = \lambda_0$ and we add to this the additional $\lambda_0/2$ phase shift for the beam reflecting at A, we have a total optical path difference of $(\lambda_0 + \lambda_0/2)$, leading to *destructive*—rather than constructive—interference. So, in addition to the phase change introduced by path differences, we must always consider the *possible phase change upon reflection* at the interfaces.

If we denote Δ_p as the optical path difference due to the film and Δ_r as the equivalent path difference introduced upon reflection, the condition for *constructive interference* becomes

$$\boxed{\Delta_p + \Delta_r = m\lambda, \quad (m = 1, 2, 3, \ldots)} \tag{4-17}$$

where m equals the order of interference.

For a thin film of thickness t and refractive index n_f, *located in air*, $\Delta_p = 2n_f t$ (according to Equation 4-15), and $\Delta_r = \lambda_0/2$. Thus, Equation 4-17—for *constructive interference*—becomes

$$\textit{normal incidence}: \quad \boxed{2n_f t + \frac{\lambda_0}{2} = m\lambda_0, \quad (m = 1, 2, 3, \ldots)} \tag{4-18}$$

where λ_0 is the wavelength in air. For *destructive interference*, Equation 4-18 changes slightly to

$$\textit{normal incidence}: \quad \boxed{2n_f t + \frac{\lambda_0}{2} = \left(m + \tfrac{1}{2}\right)\lambda_0, \quad (m = 1, 2, 3, \ldots)} \tag{4-19}$$

Let's apply these ideas to the results of interference seen in soap-bubble films.

Example 4

White light is incident normally on the surface of a soap bubble. A portion of the surface reflects green light of wavelength $\lambda_0 = 540$ nm. Assume that the refractive index of the soap film is near that of water, so that $n_f = 1.33$. Estimate the thickness (in nanometers) of the soap bubble surface that appears green in second order.

Solution: Since the soap-bubble film is surrounded by air, Equation 4-18 applies. Rearranging Equation 4-18 to solve for the thickness t gives

$$t = \frac{\left(m\lambda_0 - \dfrac{\lambda_0}{2}\right)}{2n_f}$$

where $m = 2$, $n_f = 1.33$, and $\lambda_0 = 540$ nm. Thus,

$$t = \frac{\tfrac{3}{2}\lambda_0}{2n_f} = \frac{1.5(540 \text{ nm})}{2(1.33)} \cong 305 \text{ nm}$$

The soap film thickness is about 0.3 thousandths of a millimeter.

2. Single-layer antireflection (AR) coat. A common use of single-layer films deposited on glass substrates occurs in the production of *antireflecting (AR) coatings* on optical surfaces, often found in lenses for cameras and binoculars. The arrangement of a single-layer AR coat is shown in Figure 4-11, with the film made of magnesium fluoride (M_gF_2) coated on top of a glass substrate.

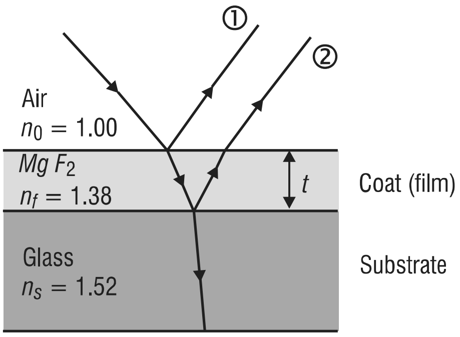


Figure 4-11 *Single-layer AR coat on glass substrate*

According to the rules for phase change upon reflection, *both* rays 1 and 2 undergo 180° shifts equal to $\lambda_0/2$, since both reflections occur at interfaces separating lower-to-higher refractive indexes. So the difference in phase between rays 1 and 2 comes from only the optical path difference due to the coating thickness *t*. If the thickness *t* is such that ray 2 falls behind ray 1 by $\lambda_{coat}/2$, the two rays interfere destructively, minimizing the reflected light. At *near-normal incidence* this requires that the distance 2*t*, down and back, equal $\lambda_{coat}/2$. The mathematical condition for antireflection is then given by $2t = \frac{\lambda_{coat}}{2}$, and, since $\lambda_{\text{coat}} = \frac{\lambda_{air}}{n_f}$, we have finally

$$\boxed{t = \frac{\lambda_{air}}{4n_f}} \tag{4-20}$$

Example 5

Determine the minimum thickness of an AR coat of magnesium fluoride, M_gF_2, deposited on a glass substrate ($n_s = 1.52$) if the coating is to be highly antireflective for the center of the white light spectrum, say at $\lambda_{air} = 550$ nm. The refractive index for M_gF_2 is near 1.38.

Solution: Application of Equation 4-20 gives

$$t_{min} = \frac{\lambda_{air}}{4n_f} = \frac{550\text{ nm}}{4(1.38)}$$

$$t_{min} = 99.6\text{ nm, about 100 nm}$$

Without a coating (bare lens surface) the amount of light reflected is around 30% of the incident light. With a single-layer *AR* coat of 100 nm of M_gF_2 on the lens surface, the light reflected drops to around 10%. Thus, the transmission of light through the lens increases from 70% to 90%.

3. Interference with multilayer films. As an extension of single-layer interference, consider the multilayer stack shown in Figure 4-12.

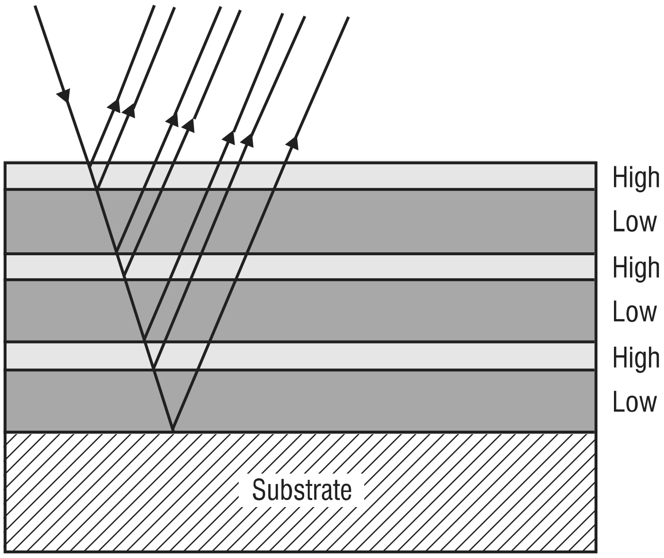


Figure 4-12 *Multilayer stack of quarter-wave thin films of alternating high and low refractive indexes. Each film has an optical thickness of* $\lambda_f/4$.

The stack is composed of alternate layers of identical *high* index and *low* index films. If each film has an optical thickness of $\lambda_f/4$, a little analysis shows that all emerging beams are in phase. Multiple reflections in the region of λ_0 increase the total reflected intensity, and the quarter-wave stack performs as an efficient mirror. Such multilayer stacks can be designed to satisfy extinction of reflected light—AR effect—or enhancement of reflected light—HR effect—over a greater portion of the spectrum than with a single-layer film. Such multilayer stacks are used in the design of *narrow-band interference filters* that filter out unwanted light, transmitting only light of the desired wavelength. For antireflection over *broader-wavelength* regions, the optical industry produces *HEBBAR™ coatings* (High Efficiency Broadband Anti Reflection) for regions of ultraviolet and infrared light, as well as for visible light. The coating industry also produces *V-coatings,* which reduce reflectance to near zero at one specific wavelength for an optical component. *High-reflection* coatings are produced over broadbands with multilayer stacks of thin films—just as for the antireflection coatings. In addition *HR* coats are used as overcoatings on *metallic reflectors,* which typically use aluminum, silver, and gold as the base metals. The overcoats protect the metals from oxidation and scratching.

IV. DIFFRACTION

The ability of light to bend around corners, a consequence of the wave nature of light, is fundamental to both interference and diffraction. *Diffraction* is simply any deviation from geometrical optics resulting from the *obstruction* of a wave front of light by some obstacle or some opening. Diffraction occurs when light waves pass through small openings, around obstacles, or by sharp edges.

Several common diffraction patterns—as sketched by an artist—are shown in Figure 4-13. Figure 4-13a is a typical diffraction pattern for HeNe laser light passing through a circular pinhole. Figure 4-13b is a typical diffraction pattern for HeNe laser light passing through a narrow (vertical) slit. And Figure 4-13c is a typical pattern for diffraction by a sharp edge.

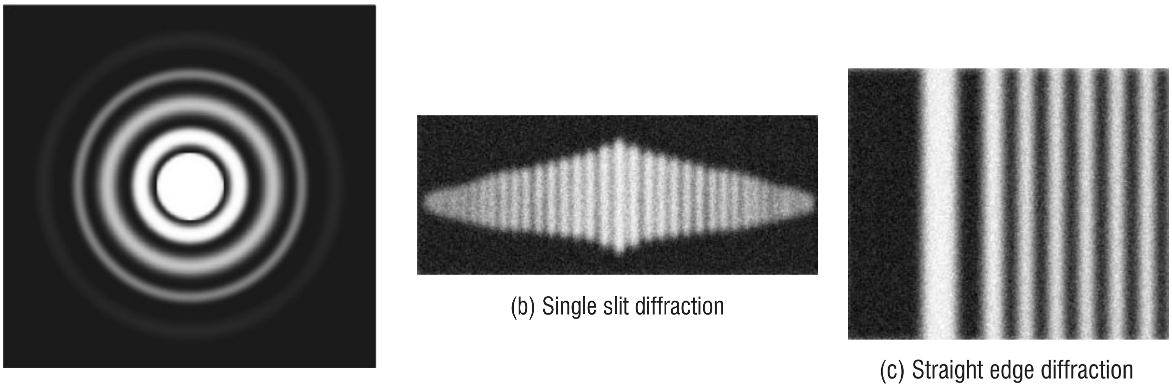


Figure 4-13 *Sketches of several common diffraction patterns*

The intricacy of the patterns should convince us—once and for all—that geometrical ray optics is incapable of dealing with diffraction phenomena. To demonstrate how wave theory does account for such patterns, we now examine the phenomenon of diffraction of waves by a single slit.

A. Diffraction by a single slit

The overall geometry for diffraction by a single slit is shown in Figure 4-14. The slit opening, seen in cross section, is in fact a long, narrow slit, perpendicular to the page. The shaded "humps" shown along the screen give a rough idea of intensity variation in the pattern, and the sketch of bright and dark regions to the right of the screen simulates the actual fringe pattern seen on the screen. We observe a wide central bright fringe, bordered by narrower regions of dark and bright. The angle θ shown connects a point *P* on the screen to the center of the slit.

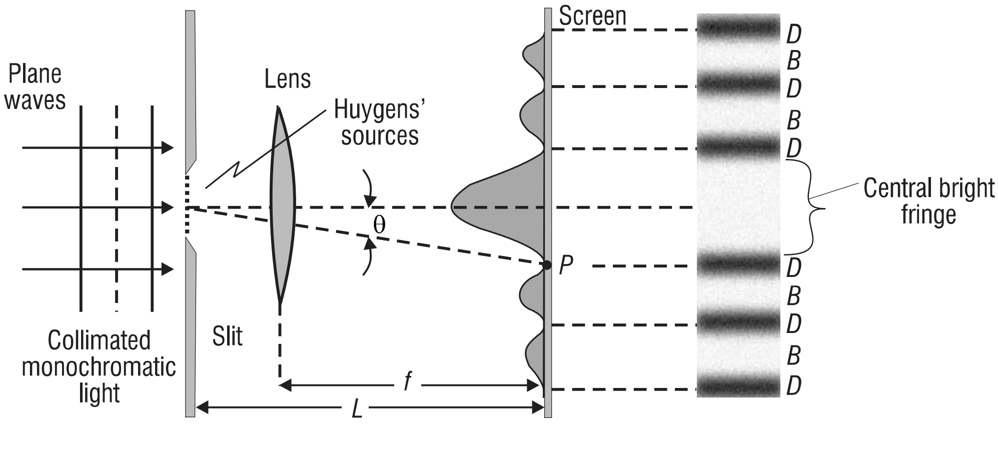


Figure 4-14 *Diffraction pattern from a single slit*

Since plane waves are incident on the screen, the diffraction pattern—in the absence of the focusing lens—would be formed far away from the slit and be much more spread out than that shown in Figure 4-14. The lens serves to focus the light passing through the slit onto the screen, just a focal length *f* away from the lens, while preserving faithfully the relative details of the diffraction pattern that would be formed on a distant screen without the lens.

To determine the location of the minima and maxima on the screen, we divide the slit opening through which a plane wave is passing into many point sources (Huygens' sources), as shown by the series of tiny dots in the slit opening of Figure 4-14. These numerous point sources send out Huygens' spherical waves, all in phase, toward the screen. There, at a point such as *P*, light waves from the various Huygens' sources overlap and interfere, forming the variation in light intensity shown in Figure 4-14. Thus, diffraction considers the contribution from *every part of the wave front* passing through the aperture. By contrast, when we looked at interference from Young's double slit, we considered *each slit* as a point source, ignoring details of the portions of the wave fronts in the slit openings themselves.

The mathematical details involved in adding the contributions at point *P* from each of the Huygens' sources can be found in basic texts on *physical optics*. Here we give only the end result of the calculation. Equation 4-21 locates the minima, y_{min}, on the screen, in terms of the slit width b, slit-to-screen distance L, wavelength λ, and order m.

$$\boxed{y_{min} = \frac{m\lambda L}{b} \quad \text{where } m = 1, 2, 3, \ldots} \tag{4-21}$$

Figure 4-15 shows the positions of several orders of minima and the essential parameters associated with the single-slit diffraction pattern. (The positions of the *maxima* are mathematically more complicated to express, so we typically work with the positions of the well-defined minima.)

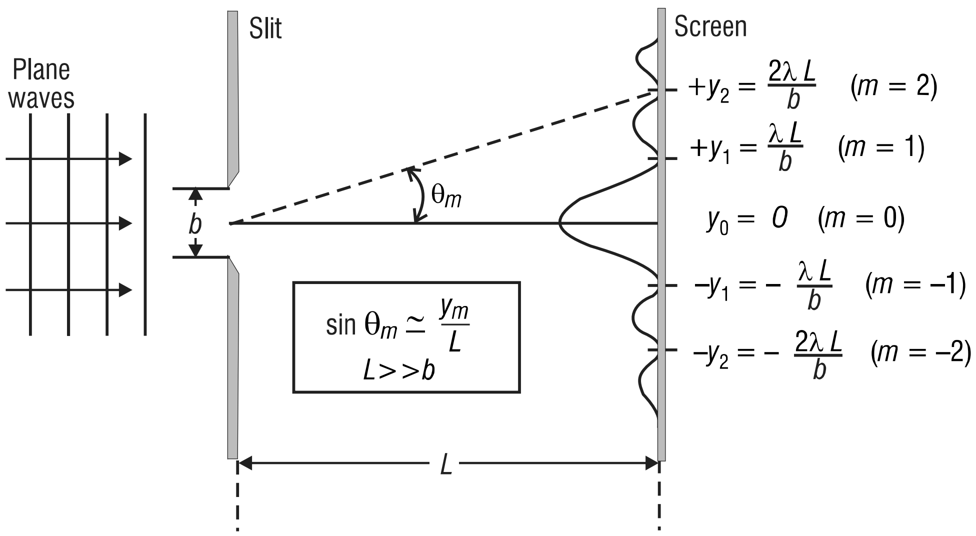


Figure 4-15 *Positions of adjacent minima in the diffraction patterns* (Drawing is not to scale.)

Now let's use Equation 4-21 to work several sample problems.

Example 6

Coherent laser light of wavelength 633 nm is incident on a single slit of width 0.25 mm. The observation screen is 2.0 m from the slit. (a) What is the width of the central bright fringe? (b) What is the width of the bright fringe between the 5th and 6th minima?

Solution:

(a) The width of the central bright fringe is $2y_1$, where y_1 is the distance to the first minimum (m = 1) on either side. Thus, using Equation 4-21,

$$Width = 2y_1 = 2\left(\frac{m\lambda L}{b}\right) = \frac{(2)(1)\left(633 \times 10^{-9}\ \text{m}\right)(2.0\ \text{m})}{2.5 \times 10^{-4}\ \text{m}} = 0.01\ \text{m}$$

The width of the central bright fringe is about 1 cm.

(b) $\text{Width} = y_6 - y_5 = \dfrac{6\lambda L}{b} - \dfrac{5\lambda L}{b} = \dfrac{\lambda L}{b}$

$$\text{Width} = \frac{\left(633 \times 10^{-9}\ \text{m}\right)(2.0\ \text{m})}{2.5 \times 10^{-4}\ \text{m}} = 5.06 \times 10^{-3}\ \text{m} \cong 0.5\ \text{cm}$$

The width of bright fringe between the 5th and 6th minima is about half the width of the central bright fringe.

Example 7

Monochromatic light is incident on a single slit of width 0.30 mm. On a screen located 2.0 m away, the width of the central bright fringe is measured and found to be near 7.8 mm. What is the wavelength of the incident light?

Solution: Since the width of the central bright fringe is 7.8 mm, equal to $2y_1$, we see that y_1 = 3.9 mm. Then, rearranging Equation 4-21 to find λ, we have $\lambda = \dfrac{y_{min}\, b}{mL}$, where $y_{min} = y_1$ = 3.9 mm, m = 1, L = 2.0 m, and b = 0.30 mm. Thus,

$$\lambda = \frac{\left(3.9 \times 10^{-3}\right)\left(3 \times 10^{-4}\right)}{(1)(2.0)} = 5.85 \times 10^{-7}\ \text{m}$$

λ ≅ 585 nm, very near the principal wavelengths of light from sodium lamps.

B. Fraunhofer and Fresnel diffraction

In general, if the observation screen is far removed from the slit on which plane waves fall (as in Figure 4-15) or a lens is used to focus the collimated light passing through the slit onto the screen (as in Figure 4-14), the diffraction occurring is described as *Fraunhofer diffraction*, after Joseph von Fraunhofer (1787-1826), who first investigated and explained this type of so-called *far-field* diffraction. If however, no lens is used and the observation screen is near to the slit, for either incident plane or spherical waves, the diffraction is called *Fresnel diffraction*, after Augustin Fresnel (1788-1829), who explained this type of *near-field* diffraction. The mathematical calculations required to determine the details of a diffraction pattern and account for the variations in intensity on the pattern are considerably more complicated for Fresnel diffraction than for Fraunhofer diffraction, so typically one studies first the Fraunhofer diffraction patterns, as we have.

Without going into the details of how to distinguish mathematically between Fresnel and Fraunhofer diffraction we can give results that help you decide whether the diffraction pattern formed is Fraunhofer or Fresnel in origin. Knowing this distinction helps you choose which equations to use in describing a particular diffraction pattern arising from a particular optical setup.

1. Criteria for far-field and near-field diffraction. Figure 4-16 shows the essential features of a general diffraction geometry, involving a source of light of wavelength λ, an opening to "obstruct" the light, and a screen to form the diffraction pattern.

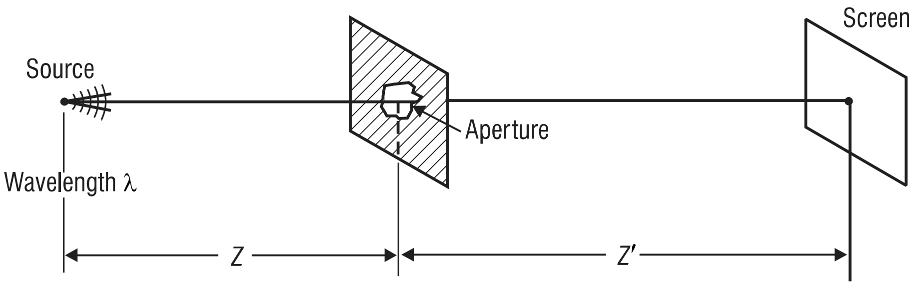


Figure 4-16 *General diffraction geometry involving source, aperture, and screen*

The distance from source to aperture is denoted as Z and that from aperture to screen as Z'. Calculations based on geometries that give rise to Fraunhofer and Fresnel diffraction patterns verify the following:

- If the distance Z from source to aperture ***and*** the distance Z' from aperture to screen are ***both*** greater than the ratio $\left(\dfrac{\text{aperture area}}{\lambda}\right)$ by a factor of 100 or so, the diffraction pattern on the screen is characteristic of *Fraunhofer diffraction*—and the screen is said to be in the *far field*. For this situation, all Fraunhofer-derived equations apply to the details of the diffraction pattern.

- If ***either*** distance—Z or Z'—is of the order of, or less than, the ratio $\left(\dfrac{\text{aperture area}}{\lambda}\right)$, the diffraction pattern on the screen is characteristic of *Fresnel diffraction* and is said to be in the *near field.* For this situation, all Fresnel-derived equations apply to the details of the diffraction pattern.

- Equation 4-22 indicates the "rule-of-thumb" conditions to be satisfied for both Z and Z' for *Fraunhofer diffraction*.

$$\text{Far - field condition: (Fraunhofer)} \quad \left\{ \begin{aligned} Z &> 100 \left[\frac{\text{aperture area}}{\lambda} \right] \\ Z' &> 100 \left[\frac{\text{aperture area}}{\lambda} \right] \end{aligned} \right\} \tag{4-22}$$

Figure 4-18 illustrates these conditions and shows the locations of the *near field, far field*, and a *gray area* in between. If the screen is in the *gray area* and accuracy is important, a Fresnel analysis is usually applied. If the screen is in the *gray area* and approximate results are acceptable, a Fraunhofer analysis (significantly simpler than a Fresnel analysis) can be applied.

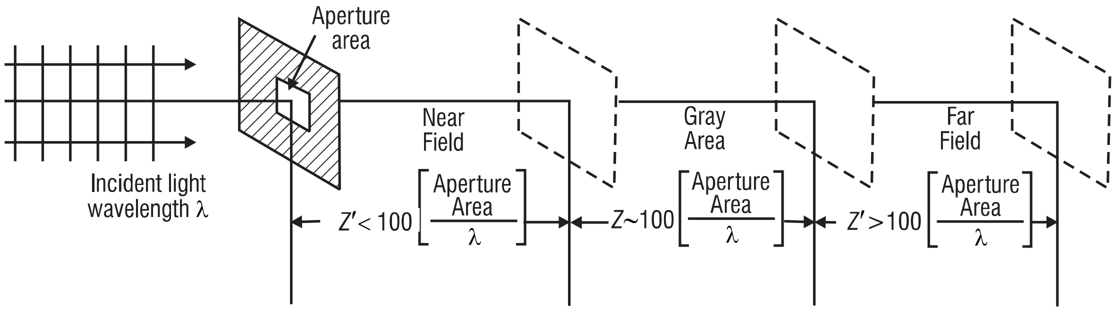


Figure 4-17 *Defining near-field and far-field regions for diffraction*

Figure 4-18 shows how we can satisfy the conditions for Fraunhofer diffraction, as spelled out in Equation 4-22, through the use of focusing lenses *on both sides* of the aperture (Figure 4-18a)—or with a *laser illuminating the aperture* and a *focusing lens* located on the screen side of the aperture (Figure 4-18b). Either optical arrangement has plane waves approaching and leaving the aperture, guaranteeing that the diffraction patterns formed are truly Fraunhofer in nature.

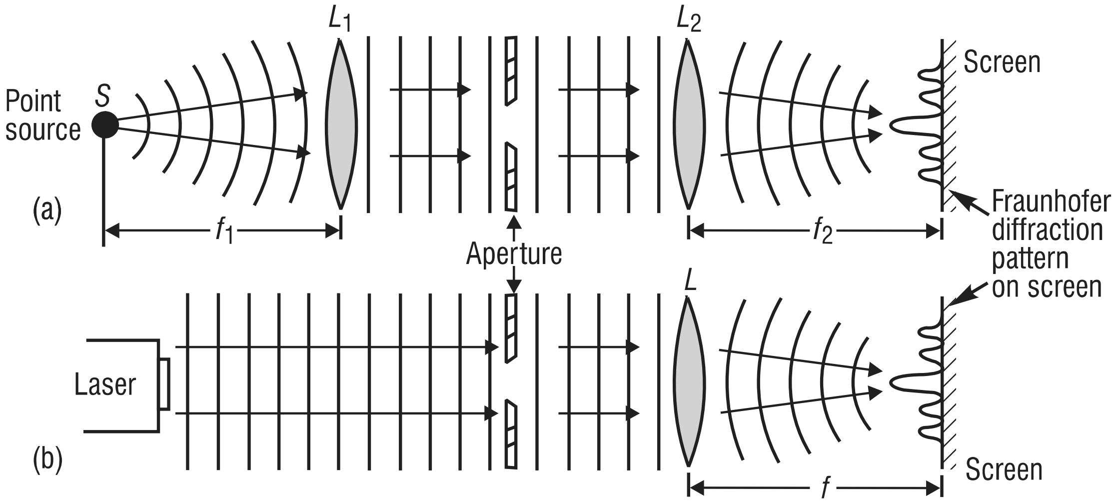


Figure 4-18 *Optical arrangements for Fraunhofer diffraction*

Now let's see how Equation 4-22 and Figure 4-18 are applied in a real situation.

Example 8

Minati, a photonics technician, has been asked to produce a Fraunhofer diffraction pattern formed when light from a HeNe laser (λ = 633 nm) passes through a pinhole of 150-μm diameter. In order to set up the correct geometry for Fraunhofer diffraction, Minati needs to know (a) the distance Z from the laser to the pinhole and (b) the distance Z' from the pinhole to the screen.

Solution: Minati needs first to test the conditions given in Equation 4-22 so she calculates the ratio of $\left(\dfrac{\text{aperture area}}{\lambda}\right)$ assuming the pinhole to be circular.

$$\text{Ratio} = \frac{\text{aperture area}}{\lambda} = \frac{\pi D^2}{4\lambda} = \frac{(3.14)(150\times 10^{-6})^2}{(4)(633\times 10^{-9})}$$

$$\text{Ratio} = 0.0279 \text{ m}$$

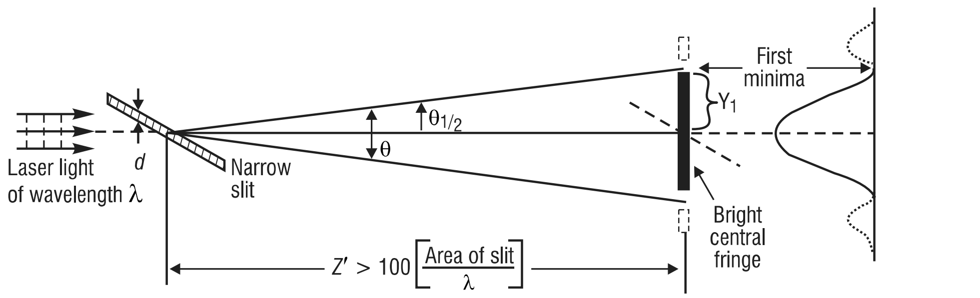
(a) Minati knows that light from the HeNe laser is fairly well collimated, so that nearly plane waves are incident on the pinhole, as illustrated in Figure 4-18b. She knows that plane waves are those that come—or appear to come—from very distant sources. So she concludes that, with the laser, the distance Z is much greater than 100 (0.0279 m)—that is, greater than about 2.8 m—and so the "Z-condition" for Fraunhofer diffraction is automatically satisfied.

(b) From her calculation of the ratio $\left(\dfrac{\text{aperture area}}{\lambda}\right)$ she knows also that the distance Z' must be greater than 2.8 m. So she can place the screen 3 meters or so from the aperture and form a Fraunhofer diffraction pattern—**OR** she can place a positive lens just beyond the aperture—as in Figure 4-18b—and focus the diffracting light on a screen a focal length away. With the focusing lens in place she obtains a much reduced—but valid—Fraunhofer diffraction pattern located nearer the aperture. She chooses to use the latter setup, with a positive lens of focal length 10 cm, enabling her to arrange the laser, pinhole, and screen, all on a convenient 2-meter optical bench.

---

2. Several typical Fraunhofer diffraction patterns. In successive order, we show the far-field diffraction pattern for a *single slit* (Figure 4-19), a *circular aperture* (Figure 4-20), and a *rectangular aperture* (Figure 4-21). Equations that describe the locations of the bright and dark fringes in the patterns accompany each figure.

Single Slit



Half-angle beam spread to first minimum, $\theta_{1/2}$, is:

$$\theta_{1/2} = \frac{\lambda}{d} \tag{4-23}$$

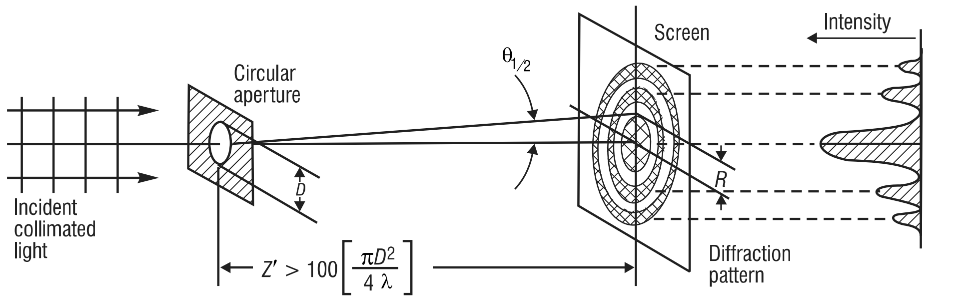
Half-width of bright central fringe, y_1, is:

$$y_1 = \frac{Z'\lambda}{d} \tag{4-24}$$

where λ = wavelength of light,
d = slit width, and
Z' = slit-to-screen distance

Figure 4-19 *Fraunhofer diffraction pattern for a single slit*

Circular Aperture



Half-angle beam spread to first minimum, $\theta_{1/2}$, is:

$$\theta_{1/2} = \frac{1.22\lambda}{D} \tag{4-25}$$

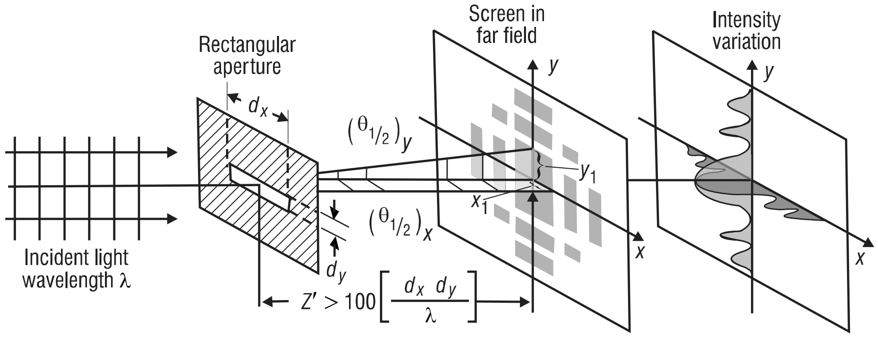
Radius of central bright disk (airy disk), R, is:

$$R = \frac{1.22\,\lambda\,Z'}{D} \tag{4-26}$$

where λ = wavelength of light,
D = diameter of pinhole, and
Z' = aperture-to-screen distance

Figure 4-20 *Fraunhofer diffraction pattern for a circular aperture*

Rectangular aperture



Half-angle beam divergences to first minimum in x and y directions:

$$\left(\theta_{1/2}\right)_x = \frac{\lambda}{d_x} \text{ and } \left(\theta_{1/2}\right)_y = \frac{\lambda}{d_y} \tag{4-27}$$

Half-widths of central bright fringe in x and y directions:

$$x_1 = \frac{Z'\lambda}{d_x} \text{ and } y_1 = \frac{Z'\lambda}{d_y} \tag{4-28}$$

Figure 4-21 *Fraunhofer diffraction pattern for a rectangular aperture*

C. Diffraction Grating

If we prepare an aperture with thousands of adjacent slits, we have a so-called *transmission-diffraction grating*. The width of a single slit—the opening—is given by d, and the distance between slit centers is given by ℓ (see Figure 4-22). For clarity, only a few of the thousands of slits normally present in a grating are shown. Note that the spreading of light occurs always in a direction perpendicular to the direction of the long edge of the slit opening—that is, since the long edge of the slit opening is *vertical* in Figure 4-22, the spreading is in the *horizontal* direction—along the screen.

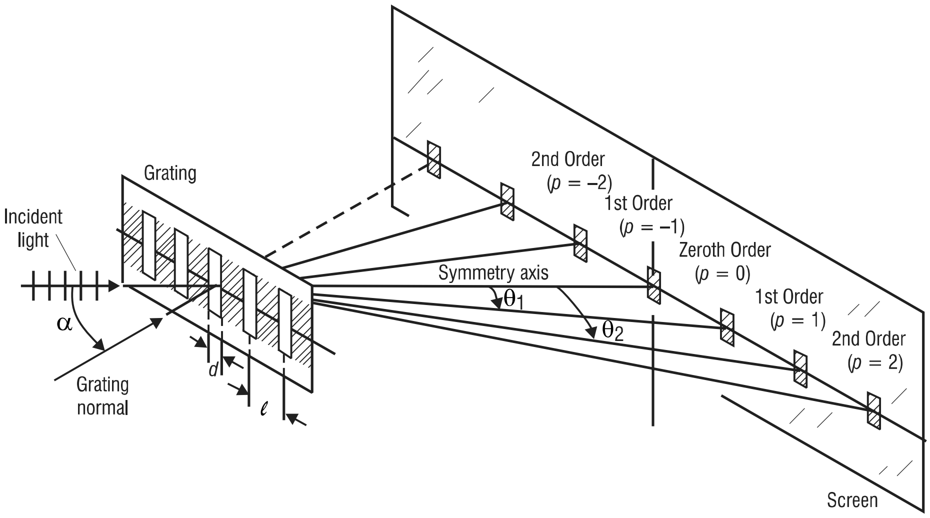


Figure 4-22 *Diffraction of light through a grating under Fraunhofer conditions*

The resulting diffraction pattern is a series of sharply defined, widely spaced fringes, as shown. The central fringe, on the symmetry axis, is called the *zeroth-order* fringe. The successive fringes on either side are called *lst order*, *2nd order*, etc., respectively. They are numbered according to their positions relative to the central fringe, as denoted by the letter p.

The intensity pattern on the screen is a *superposition* of the *diffraction effects from each slit* as well as the *interference effects of the light from all the adjacent slits*. The combined effect is to cause overall cancellation of light over most of the screen with marked enhancement over only limited regions, as shown in Figure 4-22. The location of the bright fringes is given by the following expression, called the *grating equation*, assuming that Fraunhofer conditions hold.

$$\boxed{\ell\,(\sin\alpha + \sin\theta_p) = p\lambda \text{ where } p = 0, \pm1, \pm2, \ldots} \tag{4-29}$$

where ℓ = distance between slit centers

α = angle of incidence of light measured with respect to the normal to the grating surface

θ_p = angle locating the pth-order fringe

p = an integer taking on values of 0, ±1, ±2, etc.

λ = wavelength of light

Note that, if the light is incident on the grating along the grating normal ($\alpha = 0$), the grating equation, Equation 4-29, reduces to the more common form shown in Equation 4-30.

$$\boxed{\ell\,(\sin\theta_p) = p\lambda} \tag{4-30}$$

If, for example, you shine a HeNe laser beam perpendicularly onto the surface of a transmission grating, you will see a series of brilliant red dots, spread out as shown in Figure 4-22. A complete calculation would show that less light falls on each successively distant red dot or fringe, the $p = 0$ or central fringe being always the brightest. Nevertheless, the location of each bright spot, or fringe, is given accurately by Equation 4-29 for either normal incidence ($\alpha = 0$) or oblique incidence ($\alpha \neq 0$). If light containing a mixture of wavelengths (white light, for example) is directed onto the transmission grating, Equation 4-29 holds for *each* component color or wavelength. So each color will be spread out on the screen according to Equation 4-29, with the longer wavelengths (red) spreading out farther than the shorter wavelengths (blue). In any case, the central fringe ($p = 0$) always remains the same color as the incident beam, since all wavelengths in the $p = 0$ fringe have $\theta_p = 0$, hence all overlap to re-form the "original" beam and therefore the original "color." Example 9 shows calculations for a typical diffraction grating under Fraunhofer conditions.

Example 9

Michael has been handed a transmission grating by his supervisor who wants to know how widely the red light and blue light fringes—in *second* order—are separated on a screen one meter from the grating. Michael is told that the separation distance between the red and blue colors is a critical piece of information needed for an experiment with a grating spectrometer. The transmission grating is to be illuminated at *normal incidence* with red light at $\lambda = 632.8$ nm and blue light at $\lambda = 420.0$ nm. Printed on the frame surrounding the ruled grating, Michael sees that there are 5000 slits (lines) per centimeter on this grating. Michael decides he must, in turn:

(a) Determine the distance ℓ between the slit centers.

(b) Determine the angular deviation θ_p in 2nd order for both the red and the blue light.

(c) Determine the separation distance on the screen between the red and blue fringes.

Solution:

(a) Since there are 5000 slits or grooves per centimeter, Michael knows that the distance ℓ between the slits, center to center, must be $\ell = \dfrac{1\text{ cm}}{5000} = 2 \times 10^{-4}\text{ cm}$.

(b) At normal incidence ($\alpha = 0$), Equation 4-29 reduces to Equation 4-30, so, for 2nd order ($p = 2$), Michael writes the following two equations and solves them for the deviation angles θ_2^{red} and θ_2^{blue}:

$$\sin\theta_2^{red} = \frac{p\lambda_{red}}{\ell} = \frac{(2)\left(632.8\times10^{-9}\text{ m}\right)}{2\times10^{-6}\text{ m}} = 0.6328$$

$$\therefore \theta_2^{red} = \sin^{-1}(0.6328) = 39.3°$$

$$\sin\theta_2^{blue} = \frac{p\lambda_{blue}}{\ell} = \frac{(2)\left(420\times10^{-9}\text{ m}\right)}{2\times10^{-6}\text{ m}} = 0.4200$$

$$\therefore \theta_2^{blue} = \sin^{-1}(0.4200) = 24.8°$$

(c) From the geometry shown in Figure 4-22, Michael sees that the screen distances y_2^{red} and y_2^{blue} to the red and blue fringes in 2nd order respectively, and the grating-to-screen distance Z' are related to deviation angles by the equation

$$\tan\theta_2 = \frac{y_2}{Z'}, \text{ where here, } Z' = 1 \text{ meter.}$$

Thus

$$\Delta y = y_2^{red} - y_2^{blue} = \left(Z' \tan \theta_2^{red}\right) - \left(Z' \tan \theta_2^{blue}\right)$$

which becomes

$\Delta y = (1\text{ m})\ (\tan 39.3° - \tan 24.8°)$

$\Delta y = (100\text{ cm})\ (0.8185 - 0.4621)$

$\Delta y = 35.6\text{ cm}$

Michael reports his finding of $\Delta y = 35.6$ cm to his supervisor, who decides that this grating will work in the proposed experiment.

D. Diffraction-Limited Optics

A lens of diameter *D* is in effect a large circular aperture through which light passes. Suppose a lens is used to focus plane waves (light from a distant source) to form a "spot" in the focal plane of the lens, much as is done in *geometrical optics*. Is the focused spot truly a *point*? Reference to Figure 4-20 indicates that the focused spot is actually a tiny diffraction pattern—with a bright disk at the center (the so-called *airy disk)* surrounded by dark and bright rings, as pictured earlier in Figure 4-13a.

In Figure 4-23, we see collimated light incident on a lens of focal length *f*. The lens serves as both a circular aperture of diameter *D* to intercept the plane waves and a lens to focus the light on the screen, as shown in Figure 4-18b. Since the setup in Figure 4-23 matches the conditions shown in Figure 4-18b, we are assured that a Fraunhofer diffraction pattern will form at the "focal spot" of the lens.

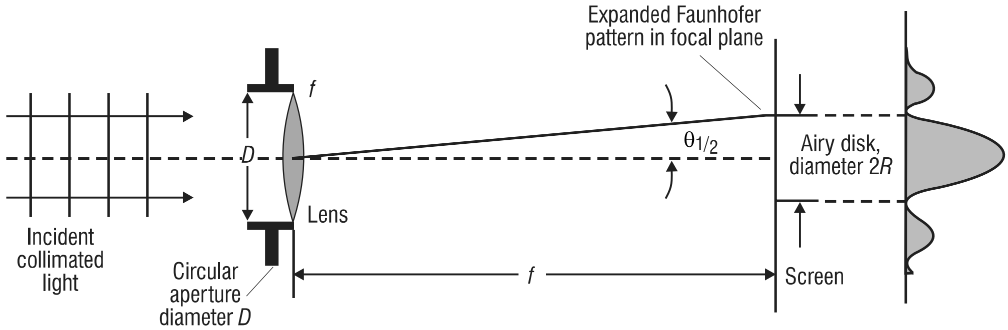


Figure 4-23 *Fraunhofer diffraction pattern formed in the focal plane of a lens of focal length* f (Drawing is not to scale.)

The diffraction pattern is, in truth then, an array of alternate bright and dark rings, with a bright spot at the center, even though the array is very small and hardly observable to the human eye. From the equations given with Figure 4-20, we see that the diameter of the central bright spot—inside the surrounding rings—is itself of size 2*R*, where, from Equation 4-26,

$$\boxed{2R = 2\left(\frac{1.22\lambda Z'}{D}\right)} \qquad (4\text{-}31)$$

where $Z' = f$

While indeed small, the diffraction pattern overall is greater than $2R$, demonstrating clearly that a lens focuses collimated light to a small diffraction pattern of rings and not to a point. However, when the lens is inches in size, we do justifiably refer to the focal plane pattern as a "point," ignoring all structure within the "point." Example 10 provides us with a "feel" for the *size of the structure* in the focused spot, when a lens of nominal size becomes the circular aperture that gives rise to the airy disk diffraction pattern.

Example 10

Determine the size of the airy disk at the center of the diffraction pattern formed by a lens such as that shown in Figure 4-23, if the lens is 4 cm in diameter and its focal length is 15 cm. Assume a wavelength of 550 nm incident on the lens.

Solution: Using Equation 4-31 with $Z' = f$, the diameter of the airy disk is

$$2R = \frac{2.44\lambda f}{D} = \frac{(2.44)\left(550 \times 10^{-9}\ \text{m}\right)(0.15\ \text{m})}{0.04\ \text{m}}$$

$$2R = 5.03 \times 10^{-6}\ \text{m}$$

Thus, the central bright spot (airy disk) in the diffraction pattern is **only** 5 micrometers in diameter. So, even though the focused spot is not a true point, it is small enough to be considered so in the world of large lenses, i.e., in the world of geometrical optics.

The previous discussion and example indicate that the size of the focal spot—structure and all—is limited by diffraction. No matter what we do, we can never make the airy disk smaller than that given by $2R = \dfrac{2.44 f \lambda}{D}$. That is the limit *set by diffraction*. So all optical systems are *limited* by diffraction in their ability to form true point images of point objects. We recognize this when we speak of *diffraction-limited optics*. An ideal optical system therefore can do no better than that permitted by diffraction theory. In fact, a real optical system—which contains imperfections in the optical lenses, variations in the index of refraction of optical components, scattering centers, and the existence of temperature gradients in the intervening atmosphere—will <u>not</u> achieve the quality limit permitted by diffraction theory. Real optical systems are therefore poorer than those limited by diffraction only. We often refer to real systems as *many-times diffraction limited* and sometimes attach a numerical figure such as "five-times diffraction-limited" to indicate the deviation in quality expected from the given system compared with an ideal "diffraction-limited" system.

V. POLARIZATION

We continue our discussion of the main concepts in *physical optics* with a brief look at *polarization*. Before we describe the polarization of light waves, let's take a look at a simplistic—but helpful—analogy of "polarization" with *rope waves*.

A. Polarization—a simple analogy

Imagine a "magic" rope that you can whip up and down at one end, thereby sending a *transverse* "whipped pulse" (vibration) out along the rope. See Figure 4-24a. Imagine further that you can change the direction of the "whipped shape," quickly and randomly at your end, so that a person looking back along the rope toward you, sees the "vibration" occurring in all directions—up and down, left to right, northeast to southwest, and so on, as shown in Figure 4-24b.

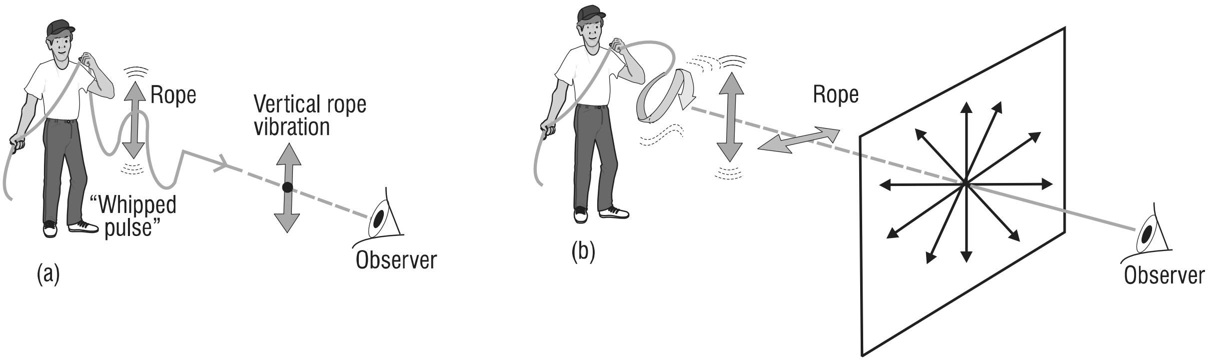


Figure 4-24 *Rope waves and polarization*

In Figure 4-24a, the rope wave is *linearly polarized*, that is, the rope vibrates in only one transverse direction—vertically in the sketch shown. In Figure 4-24b, the rope vibrations are in all transverse directions, so that the rope waves are said to be *unpolarized*.

Now imagine that the waves on the rope—representing all possible directions of vibration as shown in Figure 4-24b—are passed through a *picket fence*. Since the vertical slots of the fence pass only vertical vibrations, the many randomly oriented transverse vibrations incident on the picket fence emerge as only vertical vibrations, as depicted in Figure 4-25. In this example of transverse waves moving out along a rope, we see how we can—with the help of a polarizing device, the picket fence in this case—change unpolarized rope waves into polarized rope waves.

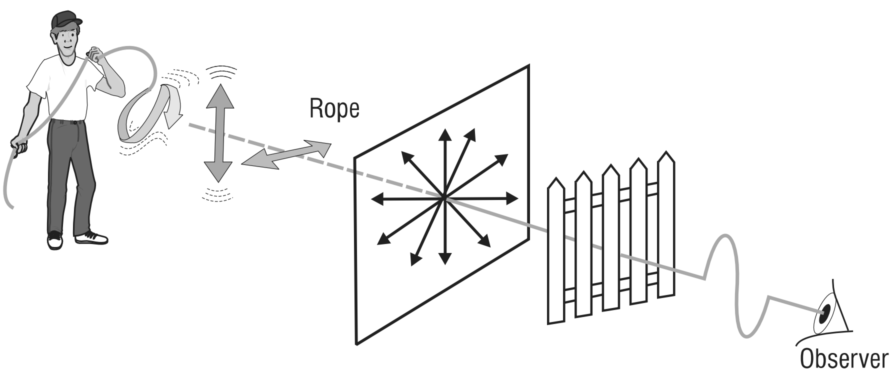


Figure 4-25 *Polarization of rope waves by a picket fence*

B. Polarization of light waves

The polarization of light waves refers to the *transverse* direction of vibration of the electric field vector of electromagnetic waves. (Refer back to Figure 4-3.) As described earlier, *transverse* means E-field vibrations *perpendicular* to the direction of wave propagation. If the electric field vector remains in a given direction in the transverse *x-y* plane—as shown in Figure 4-26—the light is said to be *linearly polarized*. (The "vibration" of the electric field referred to here is not the same as a *physical* displacement or movement in a rope. Rather, the vibration here refers to an increase and decrease of the electric field strength occurring in a particular transverse direction—at all given points along the propagation of the wave.) Figure 4-26 shows linearly polarized light propagating along the *z*-direction toward an observer at the left. The electric field *E* increases and decreases in strength, reversing itself as shown, always along a direction making an angle θ with the *y*-axis in the transverse plane. The E-field components $E_x = E \sin \theta$ and $E_y = E \cos \theta$ are shown also in the figure.

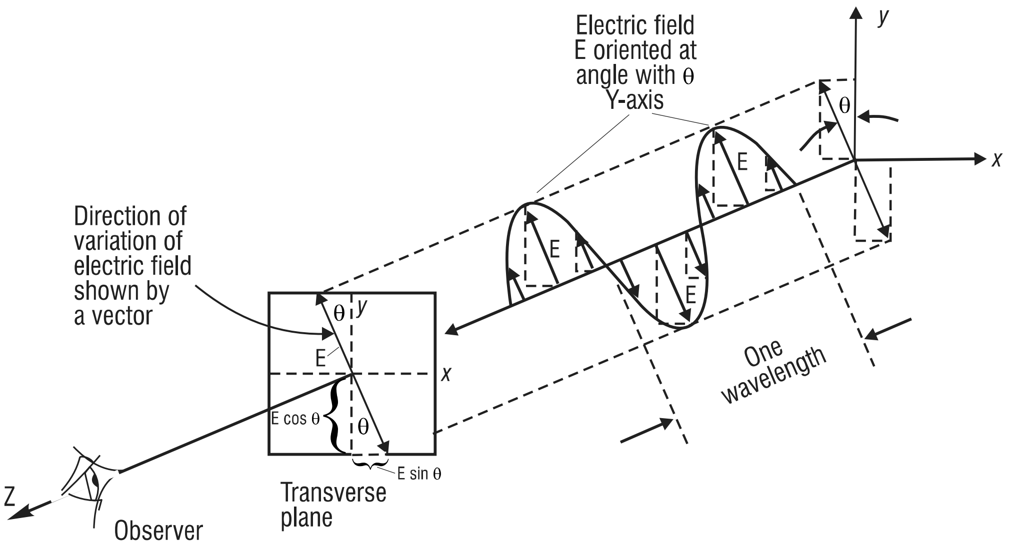
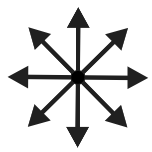
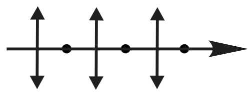
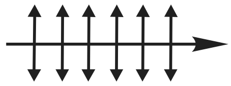


Figure 4-26 *Linearly polarized light with transverse electric field E propagating along the* z*-axis*

Table 1 lists the *symbols* used generally to indicate *unpolarized* light (*E*-vector vibrating randomly in all directions), *vertically polarized* light (*E*-vector vibrating in the vertical direction only), and *horizontally polarized* light (*E*-vector vibrating in the horizontal direction only). With reference to Figure 4-26, the vertical direction is along the *y*-axis, the horizontal direction along the *x*-axis.

Table 4-1 Standard Symbols for Polarized Light

Viewing Position	Unpolarized	Vertically Polarized	Horizontally Polarized
Viewed head-on; beam coming toward viewer			
Viewed from the side; beam moving from left to right			

Like the action of the picket fence described in Figure 4-25, a special optical filter—called either a *polarizer* or an *analyzer* depending on how it's used—transmits only the light wave vibrations of the E-vector that are lined up with the filter's *transmission axis*—like the slats in the picket fence. The combined action of a polarizer and an analyzer are shown in Figure 4-27. Unpolarized light, represented by the multiple arrows, is incident on a "polarizer" whose transmission axis (TA) is vertical. As a result, only *vertically polarized* light emerges from the polarizer. The vertically polarized light is then incident on an "analyzer" whose transmission axis is horizontal, at 90° to the direction of the vertically polarized light. As a result, **no** light is transmitted.

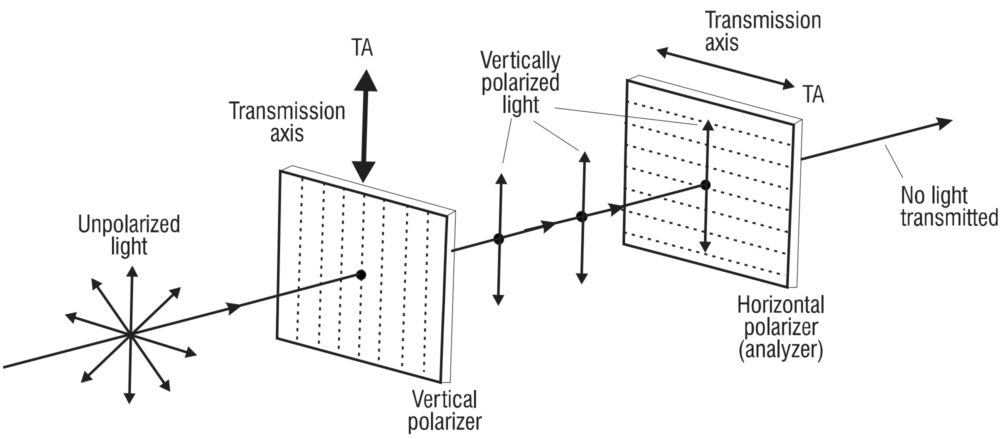


Figure 4-27 *Effect of polarizers on unpolarized light*

C. Law of Malus

When unpolarized light passes through a polarizer, the light intensity—proportional to the square of its electric field strength—is reduced, since only the E-field component along the transmission axis of the polarizer is passed. When linearly polarized light is directed through a polarizer and the direction of the E-field is at an angle θ to the transmission axis of the polarizer, the light intensity is likewise reduced. The reduction in intensity is expressed by the *law of Malus*, given in Equation 4-32.

$$\boxed{I = I_0 \cos^2 \theta} \tag{4-32}$$

where I = intensity of light that is passed through the polarizer

I_0 = intensity of light that is incident on the polarizer

θ = angle between the transmission axis of the polarizer and the direction of the E-field vibration

Application of the law of Malus is illustrated in Figure 4-28, where two polarizers are used to control the intensity of the transmitted light. The first polarizer changes the incident unpolarized light to linearly polarized light, represented by the vertical vector labeled E_0. The second polarizer, whose TA is at an angle θ with E_0, passes only the component $E_0 \cos \theta$, that is, the part of E_0 that lies along the direction of the transmission axis. Since the intensity goes as the square of the electric field, we see that I, the light intensity transmitted through polarizer 2, is equal to $(E_0 \cos)^2$, or $I = E_0^2 \cos^2$. Since E_0^2 is equal to I_0, we have demonstrated how the *law of Malus* ($I = I_0 \cos^2 \theta$) comes about.

We can see that, by rotating polarizer 2 to change θ, we can vary the amount of light passed. Thus, if $\theta = 90°$ (TA of polarizer 1 is 90° to TA of polarizer 2) no light is passed, since $\cos 90° = 0$. If $\theta = 0°$ (TA of polarizer 1 is parallel to TA of polarizer 2) all of the light is passed, since $\cos 0° = 1$. For any other θ between 0° and 90°, an amount $I_0 \cos^2 \theta$ is passed.

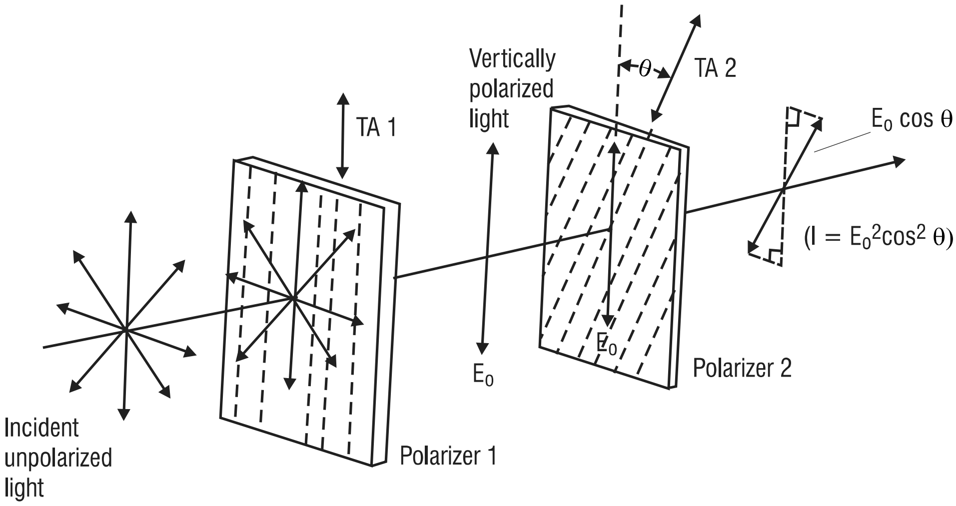


Figure 4-28 *Controlling light intensity with a pair of polarizers*

Example 11 shows how to use the law of Malus in a light-controlling experiment.

Example 11

Unpolarized light is incident on a pair of polarizers as shown in Figure 4-28.

(a) Determine the angle θ required—between the transmission axes of polarizers 1 and 2—that will reduce the intensity of light I_0 incident on polarizer 2 by 50%.

(b) For this same reduction, determine by how much the field E_0 incident on polarizer 2 has been reduced.

Solution:

(a) Based on the statement of the problem, we see that $I = 0.5\ I_0$. By applying the *law of Malus,* we have:

$$
\begin{aligned}
I &= I_0 \cos^2\theta \\
0.5\, I_0 &= I_0 \cos^2\theta \\
\cos\theta &= \sqrt{0.5} = 0.707 \\
\theta &= 45^\circ
\end{aligned}
$$

So the two TAs should be at an angle of 45° with each other.

(b) Knowing that the E-field passed by polarizer 2 is equal to $E_0 \cos\theta$, we have

$$
\begin{aligned}
E_2 &= E_0 \cos\theta \\
E_2 &= E_0 \cos 45^\circ \\
E_2 &= 0.707\, E_0 \cong 71\%\, E_0
\end{aligned}
$$

Thus, the E-field incident on polarizer 2 has been reduced by about 29% after passing through polarizer 2.

D. Polarization by reflection and Brewster's angle

Unpolarized light—the light we normally see around us—can be polarized through several methods. The polarizers and analyzers we have introduced above polarize by *selective absorption*. That is, we can prepare materials—called *dichroic* polarizers—that selectively *absorb* components of E-field vibrations along a given direction and largely *transmit* the components of the E-field vibration perpendicular to the absorption direction. The perpendicular (transmitting) direction defines the TA of the material. This phenomena of selective absorption is what E. H. Land discovered in 1938 when he produced such a material—and called it *Polaroid.*

Polarization is produced also by the phenomenon of *scattering*. If light is incident on a collection of particles, as in a gas, the electrons in the particles absorb and reradiate the light. The light radiated in a direction perpendicular to the direction of propagation is partially polarized. For example, if you look into the north sky at dusk through a polarizer, the light being scattered toward the south—toward you—is partially polarized. You will see variations in the intensity of the light as you rotate the polarizer, confirming the state of partial polarization of the light coming toward you.

Another method of producing polarized light is by *reflection*. Figure 4-29 shows the *complete* polarization of the *reflected light* at a *particular angle of incidence* B, called the *Brewster angle*.

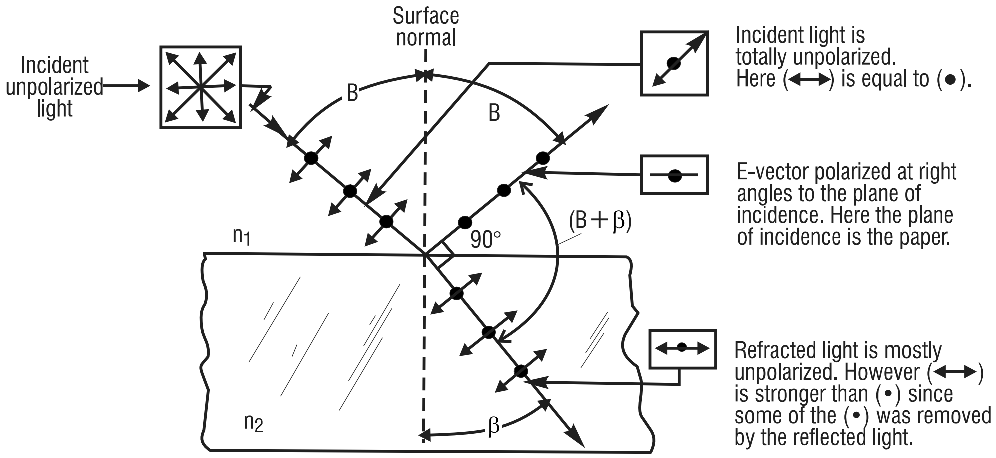


Figure 4-29 *Polarization by reflection at Brewster's angle*

The refracted light on the other hand becomes only partially polarized. Note that the symbols introduced in Table 4-1 are used to keep track of the different components of polarization. One of these is the dot (•) which indicates E-field vibrations perpendicular to both the light ray and the plane of incidence, that is, in and out of the paper. The other is an arrow (↔) indicating E-field vibrations **in** the plane of incidence **and perpendicular** to the ray of light. The reflected E-field coming off at Brewster's angle is totally polarized in a direction in and out of the paper, perpendicular to the reflected ray. This happens only at Brewster's angle, that particular angle of incidence for which the angle between the reflected and refracted rays, $B + \beta$, is exactly 90°. At the angle of incidence B, the E-field component (↔) cannot exist, for if it did it would be **along** the reflected ray, violating the requirement that E-field vibrations must always be transverse—that is, perpendicular to the direction of propagation. Thus, only the *E*-field component perpendicular to the plane of incidence (•) is reflected.

Referring to Figure 4-29 and Snell's law at the Brewster angle of incidence, we can write:

$$n_1 \sin B = n_2 \sin \beta$$

Since $\beta + B = 90°$, $\beta = 90 - B$, which then allows us to write

$$n_1 \sin B = n_2 \sin (90 - B) = n_2 \cos B$$

or

$$\frac{\sin B}{\cos B} = \frac{n_2}{n_1}$$

and finally

$$\boxed{\tan B = \frac{n_2}{n_1}} \qquad (4\text{-}33)$$

Equation 4-33 is an expression for Brewster's law. Knowing n_1 (the refractive index of the *incident* medium) and n_2 (the refractive index of the *refractive* medium), we can calculate the Brewster angle B. Shining light on a reflecting surface at this angle ensures complete polarization of the reflected ray. We make use of Equation 4-33 in Example 12.

Example 12

In one instance, unpolarized light in air is to be reflected off a glass surface (n = 1.5). In another instance, internal unpolarized light in a glass prism is to be reflected at the glass-air interface, where n for the prism is also 1.5. Determine the Brewster angle for each instance.

Solution:

(a) Light going from air to glass. In this case, $n_1 = 1$ and $n_2 = 1.5$.

Using Equation 4-33

$$\tan B = \frac{n_2}{n_1} = \frac{1.5}{1}$$

$$B = \tan^{-1} 1.5 = 56.3°$$

The Brewster angle is 56.3°.

(b) Light going from glass to air: In this case, $n_1 = 1.5$ and $n_2 = 1.0$.

Then, $$\tan B = \frac{n_2}{n_1} = \frac{1}{1.5} = 0.667$$

$$B = \tan^{-1}(0.667) = 33.7°$$

The Brewster angle is 33.7°.

E. Brewster windows in a laser cavity

Brewster windows are used in laser cavities to ensure that the laser light—after bouncing back and forth between the cavity mirrors—emerges as linearly polarized light. Figure 4-30 shows the general arrangement of the windows—thin slabs of glass with parallel sides—mounted on the opposite edges of the gas laser tube—in this case a helium-neon gas laser.

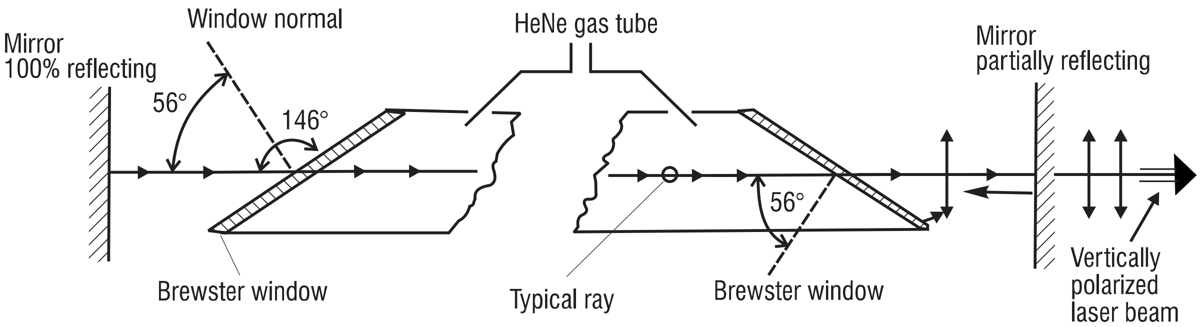


Figure 4-30 *Brewster windows in a HeNe gas laser*

As you can see, the light emerging is linearly polarized in a vertical direction. Why this is so is shown in detail in Figure 4-31. Based on Figure 4-29 and Example 12, Figure 4-31 shows that it is the refracted light—and not the reflected light—that is eventually linearly polarized.

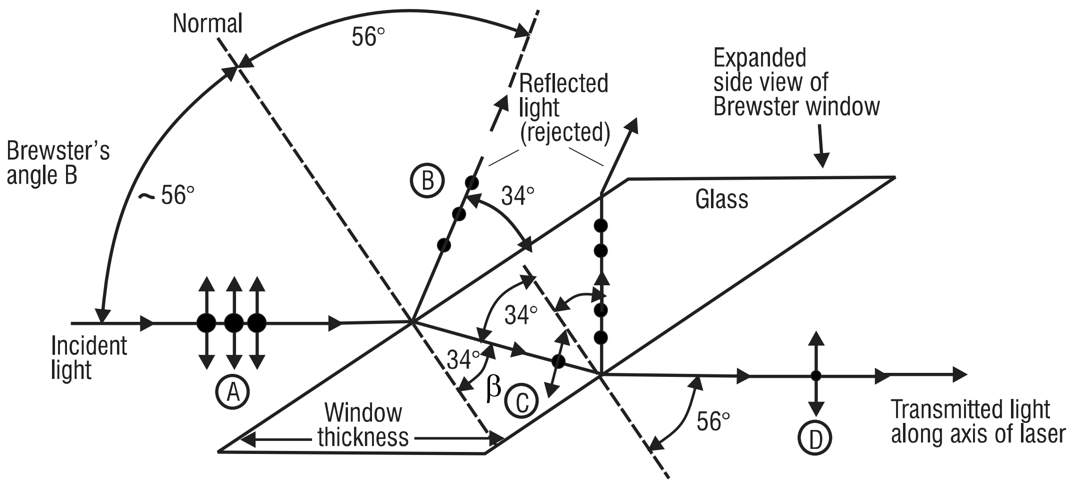


Figure 4-31 *Unpolarized light passing through both faces at a Brewster angle*

The unpolarized light at A is incident on the left face of the window—from air to glass—defining, as in Example 12, a Brewster angle of 56.3°. The reflected light at B is totally polarized and is *rejected*. The refracted (transmitted) light at C is now partially polarized since the reflected light has carried away part of the vibration perpendicular to the paper (shown by the dots). At the right face, the ray is incident again at a Brewster angle (34°) for a glass-to-air interface—as was shown in Example 12. Here again, the reflected light, totally polarized, is *rejected*. The light transmitted through the window, shown at D, now has even less of the vibration perpendicular to the paper. After hundreds of such passes back and forth through the Brewster windows, as the laser light bounces between the cavity mirrors, the transmitted light is left with only the vertical polarization, as shown exiting the laser in Figure 4-30. And since all of the reflected light is removed (50% of the initial incident light) we see that 50% of the initial incident light *remains* in the refracted light, hence in the laser beam.