vacuum and light undergoing reflection, we will also see in a later chapter that it works for the bending of light when it passes from one medium into another.

Although it is beautiful that the entire ray model of light can be reduced to one simple rule, the principle of least time, it may seem a little spooky to speak as if the ray of light is intelligent, and has carefully planned ahead to find the shortest route to its destination. How does it know in advance where it’s going? What if we moved the mirror while the light was en route, so conditions along its planned path were not what it “expected”? The answer is that the principle of least time is really a shortcut for finding certain results of the wave model of light, which is the topic of the last chapter of this book.

There are a couple of subtle points about the principle of least time. First, the path does not have to be the quickest of all possible paths; it only needs to be quicker than any path that differs infinitesimally from it. In figure p, for instance, light could get from A to B either by the reflected path AQB or simply by going straight from A to B. Although AQB is not the shortest possible path, it cannot be shortened by changing it infinitesimally, e.g., by moving Q a little to the right or left. On the other hand, path APB is physically impossible, because it is possible to improve on it by moving point P infinitesimally to the right.

It’s not quite right to call this the principle of *least* time. In figure q, for example, the four physically possible paths by which a ray can return to the center consist of two shortest-time paths and two longest-time paths. Strictly speaking, we should refer to the *principle of least or greatest time*, but most physicists omit the niceties, and assume that other physicists understand that both maxima and minima are possible.
12.2 Images by Reflection

Infants are always fascinated by the antics of the Baby in the Mirror. Now if you want to know something about mirror images that most people don’t understand, try this. First bring this page closer and closer to your eyes, until you can no longer focus on it without straining. Then go in the bathroom and see how close you can get your face to the surface of the mirror before you can no longer easily focus on the image of your own eyes. You will find that the shortest comfortable eye-mirror distance is much less than the shortest comfortable eye-paper distance. This demonstrates that the image of your face in the mirror acts as if it had depth and existed in the space behind the mirror. If the image was like a flat picture in a book, then you wouldn’t be able to focus on it from such a short distance.

In this chapter we will study the images formed by flat and curved mirrors on a qualitative, conceptual basis. Although this type of image is not as commonly encountered in everyday life as images formed by lenses, images formed by reflection are simpler to understand, so we discuss them first. In section 12.3 we will turn to a more mathematical treatment of images made by reflection. Surprisingly, the same equations can also be applied to lenses, which are the topic of section 12.4.

12.2.1 A virtual image

We can understand a mirror image using a ray diagram. Figure a shows several light rays, 1, that originated by diffuse reflection at the person’s nose. They bounce off the mirror, producing new rays, 2. To anyone whose eye is in the right position to get one of these rays, they appear to have come from a behind the mirror, 3, where they would have originated from a single point. This point is where the tip of the image-person’s nose appears to be. A similar analysis applies to every other point on the person’s face, so it looks as though there was an entire face behind the mirror. The customary way of describing the situation requires some explanation:

**Customary description in physics:** There is an image of the face behind the mirror.

**Translation:** The pattern of rays coming from the mirror is exactly the same as it would be if there were a face behind the mirror. Nothing is really behind the mirror.

This is referred to as a virtual image, because the rays do not actually cross at the point behind the mirror. They only appear to have originated there.
Imagine that the person in figure a moves his face down quite a bit — a couple of feet in real life, or a few inches on this scale drawing. The mirror stays where it is. Draw a new ray diagram. Will there still be an image? If so, where is it visible from? > Answer, p. 932

The geometry of specular reflection tells us that rays 1 and 2 are at equal angles to the normal (the imaginary perpendicular line piercing the mirror at the point of reflection). This means that ray 2’s imaginary continuation, 3, forms the same angle with the mirror as ray 1. Since each ray of type 3 forms the same angles with the mirror as its partner of type 1, we see that the distance of the image from the mirror is the same as that of the actual face from the mirror, and it lies directly across from it. The image therefore appears to be the same size as the actual face.

<b>Example 2.</b>

An eye exam example 2
Figure b shows a typical setup in an optometrist’s examination room. The patient’s vision is supposed to be tested at a distance of 6 meters (20 feet in the U.S.), but this distance is larger than the amount of space available in the room. Therefore a mirror is used to create an image of the eye chart behind the wall.

The Praxinoscope example 3
Figure c shows an old-fashioned device called a praxinoscope, which displays an animated picture when spun. The removable strip of paper with the pictures printed on it has twice the radius of the inner circle made of flat mirrors, so each picture’s virtual image is at the center. As the wheel spins, each picture’s image is replaced by the next.
Discussion Question

A  The figure shows an object that is off to one side of a mirror. Draw a ray diagram. Is an image formed? If so, where is it, and from which directions would it be visible?
12.2.2 Curved mirrors

An image in a flat mirror is a pretechnological example: even animals can look at their reflections in a calm pond. We now pass to our first nontrivial example of the manipulation of an image by technology: an image in a curved mirror. Before we dive in, let’s consider why this is an important example. If it was just a question of memorizing a bunch of facts about curved mirrors, then you would rightly rebel against an effort to spoil the beauty of your liberally educated brain by force-feeding you technological trivia. The reason this is an important example is not that curved mirrors are so important in and of themselves, but that the results we derive for curved bowl-shaped mirrors turn out to be true for a large class of other optical devices, including mirrors that bulge outward rather than inward, and lenses as well. A microscope or a telescope is simply a combination of lenses or mirrors or both. What you’re really learning about here is the basic building block of all optical devices from movie projectors to octopus eyes.

Because the mirror in figure d is curved, it bends the rays back closer together than a flat mirror would: we describe it as converging. Note that the term refers to what it does to the light rays, not to the physical shape of the mirror’s surface. (The surface itself would be described as concave. The term is not all that hard to remember, because the hollowed-out interior of the mirror is like a cave.) It is surprising but true that all the rays like 3 really do converge on a point, forming a good image. We will not prove this fact, but it is true for any mirror whose curvature is gentle enough and that is symmetric with respect to rotation about the perpendicular line passing through its center (not asymmetric like a potato chip). The old-fashioned method of making mirrors and lenses is by grinding them in grit by hand, and this automatically tends to produce an almost perfect spherical surface.

Bending a ray like 2 inward implies bending its imaginary continuation 3 outward, in the same way that raising one end of a seesaw causes the other end to go down. The image therefore forms deeper behind the mirror. This doesn’t just show that there is extra distance between the image-nose and the mirror; it also implies that the image itself is bigger from front to back. It has been magnified in the front-to-back direction.

It is easy to prove that the same magnification also applies to the image’s other dimensions. Consider a point like E in figure e. The trick is that out of all the rays diffusely reflected by E, we pick the one that happens to head for the mirror’s center, C. The equal-angle property of specular reflection plus a little straightforward geometry easily leads us to the conclusion that triangles ABC and CDE are the same shape, with ABC being simply a scaled-up version of CDE. The magnification of depth equals the ratio BC/CD, and the up-
down magnification is AB/DE. A repetition of the same proof shows that the magnification in the third dimension (out of the page) is also the same. This means that the image-head is simply a larger version of the real one, without any distortion. The scaling factor is called the magnification, $M$. The image in the figure is magnified by a factor $M = 1.9$.

Note that we did not explicitly specify whether the mirror was a sphere, a paraboloid, or some other shape. However, we assumed that a focused image would be formed, which would not necessarily be true, for instance, for a mirror that was asymmetric or very deeply curved.

12.2.3 A real image

If we start by placing an object very close to the mirror, $g/1$, and then move it farther and farther away, the image at first behaves as we would expect from our everyday experience with flat mirrors, receding deeper and deeper behind the mirror. At a certain point, however, a dramatic change occurs. When the object is more than a certain distance from the mirror, $g/2$, the image appears upside-down and in front of the mirror.

![Diagram](g1.png) 1. A virtual image. 2. A real image. As you’ll verify in homework problem 12, the image is upside-down.

Here’s what’s happened. The mirror bends light rays inward, but when the object is very close to it, as in $g/1$, the rays coming from a
given point on the object are too strongly diverging (spreading) for
the mirror to bring them back together. On reflection, the rays are
still diverging, just not as strongly diverging. But when the object
is sufficiently far away, $g/2$, the mirror is only intercepting the rays
that came out in a narrow cone, and it is able to bend these enough
so that they will reconverge.

Note that the rays shown in the figure, which both originated at
the same point on the object, reunite when they cross. The point
where they cross is the image of the point on the original object.
This type of image is called a \textit{real image}, in contradistinction to the
virtual images we’ve studied before.

\textbf{Definition:} A real image is one where rays actually cross. A virtual
image is a point from which rays only appear to have come.

The use of the word “real” is perhaps unfortunate. It sounds
as though we are saying the image was an actual material object,
which of course it is not.

The distinction between a real image and a virtual image is an
important one, because a real image can be projected onto a screen
or photographic film. If a piece of paper is inserted in figure $g/2$
at the location of the image, the image will be visible on the paper
(provided the object is bright and the room is dark). Your eye uses
a lens to make a real image on the retina.

\textit{self-check C}

Sketch another copy of the face in figure $g/1$, even farther from the
mirror, and draw a ray diagram. What has happened to the location of
the image? $\triangleright$ Answer, p. 932

\section*{12.2.4 Images of images}

If you are wearing glasses right now, then the light rays from the
page are being manipulated first by your glasses and then by the lens
of your eye. You might think that it would be extremely difficult
to analyze this, but in fact it is quite easy. In any series of optical
elements (mirrors or lenses or both), each element works on the rays
furnished by the previous element in exactly the same manner as if
the image formed by the previous element was an actual object.

Figure h shows an example involving only mirrors. The Newto-
nian telescope, invented by Isaac Newton, consists of a large curved
mirror, plus a second, flat mirror that brings the light out of the
tube. (In very large telescopes, there may be enough room to put
a camera or even a person inside the tube, in which case the sec-
ond mirror is not needed.) The tube of the telescope is not vital; it
is mainly a structural element, although it can also be helpful for
blocking out stray light. The lens has been removed from the front
of the camera body, and is not needed for this setup. Note that the
two sample rays have been drawn parallel, because an astronomical telescope is used for viewing objects that are extremely far away. These two “parallel” lines actually meet at a certain point, say a crater on the moon, so they can’t actually be perfectly parallel, but they are parallel for all practical purposes since we would have to follow them upward for a quarter of a million miles to get to the point where they intersect.

The large curved mirror by itself would form an image I, but the small flat mirror creates an image of the image, I’. The relationship between I and I’ is exactly the same as it would be if I was an actual object rather than an image: I and I’ are at equal distances from the plane of the mirror, and the line between them is perpendicular to the plane of the mirror.

One surprising wrinkle is that whereas a flat mirror used by itself forms a virtual image of an object that is real, here the mirror is forming a real image of virtual image I. This shows how pointless it would be to try to memorize lists of facts about what kinds of images are formed by various optical elements under various circumstances. You are better off simply drawing a ray diagram.

Although the main point here was to give an example of an image of an image, figure i also shows an interesting case where we need to make the distinction between magnification and angular magnification. If you are looking at the moon through this telescope, then the images I and I’ are much smaller than the actual moon. Otherwise, for example, image I would not fit inside the telescope! However, these images are very close to your eye compared to the actual moon. The small size of the image has been more than compensated for by the shorter distance. The important thing here is the amount of angle within your field of view that the image covers, and it is this angle that has been increased. The factor by which it is increased is called the angular magnification, $M_a$. 

---

i / A Newtonian telescope being used for visual rather than photographic observing. In real life, an eyepiece lens is normally used for additional magnification, but this simpler setup will also work.

j / The angular size of the flower depends on its distance from the eye.
The person uses a mirror to get a view of both sides of the ladybug. Although the flat mirror has $M = 1$, it doesn't give an angular magnification of 1. The image is farther from the eye than the object, so the angular magnification $M_a = \alpha_i/\alpha_o$ is less than one.

**Discussion Questions**

A Locate the images of you that will be formed if you stand between two parallel mirrors.
B Locate the images formed by two perpendicular mirrors, as in the figure. What happens if the mirrors are not perfectly perpendicular?
Locate the images formed by the periscope.
12.3 Images, Quantitatively

It sounds a bit odd when a scientist refers to a theory as “beautiful,” but to those in the know it makes perfect sense. One mark of a beautiful theory is that it surprises us by being simple. The mathematical theory of lenses and curved mirrors gives us just such a surprise. We expect the subject to be complex because there are so many cases: a converging mirror forming a real image, a diverging lens that makes a virtual image, and so on for a total of six possibilities. If we want to predict the location of the images in all these situations, we might expect to need six different equations, and six more for predicting magnifications. Instead, it turns out that we can use just one equation for the location of the image and one equation for its magnification, and these two equations work in all the different cases with no changes except for plus and minus signs. This is the kind of thing the physicist Eugene Wigner referred to as “the unreasonable effectiveness of mathematics.” Sometimes we can find a deeper reason for this kind of unexpected simplicity, but sometimes it almost seems as if God went out of Her way to make the secrets of universe susceptible to attack by the human thought-tool called math.

12.3.1 A real image formed by a converging mirror

*Location of the image*

We will now derive the equation for the location of a real image formed by a converging mirror. We assume for simplicity that the mirror is spherical, but actually this isn’t a restrictive assumption, because any shallow, symmetric curve can be approximated by a sphere. The shape of the mirror can be specified by giving the location of its center, C. A deeply curved mirror is a sphere with a small radius, so C is close to it, while a weakly curved mirror has C farther away. Given the point O where the object is, we wish to find the point I where the image will be formed.

To locate an image, we need to track a minimum of two rays coming from the same point. Since we have proved in the previous chapter that this type of image is not distorted, we can use an on-axis point, O, on the object, as in figure a/1. The results we derive will also hold for off-axis points, since otherwise the image would have to be distorted, which we know is not true. We let one of the rays be the one that is emitted along the axis; this ray is especially easy to trace, because it bounces straight back along the axis again. As our second ray, we choose one that strikes the mirror at a distance of 1 from the axis. “One what?” asks the astute reader. The answer is that it doesn’t really matter. When a mirror has shallow curvature, all the reflected rays hit the same point, so 1 could be expressed in any units you like. It could, for instance, be 1 cm, unless your mirror is smaller than 1 cm!
The only way to find out anything mathematical about the rays is to use the sole mathematical fact we possess concerning specular reflection: the incident and reflected rays form equal angles with respect to the normal, which is shown as a dashed line. Therefore the two angles shown in figure a/2 are the same, and skipping some straightforward geometry, this leads to the visually reasonable result that the two angles in figure a/3 are related as follows:

\[ \theta_i + \theta_o = \text{constant} \]

(Note that \( \theta_i \) and \( \theta_o \), which are measured from the image and the object, not from the eye like the angles we referred to in discussing angular magnification on page 758.) For example, move O farther from the mirror. The top angle in figure a/2 is increased, so the bottom angle must increase by the same amount, causing the image point, I, to move closer to the mirror. In terms of the angles shown in figure a/3, the more distant object has resulted in a smaller angle \( \theta_o \), while the closer image corresponds to a larger \( \theta_i \); one angle increases by the same amount that the other decreases, so their sum remains constant. These changes are summarized in figure a/4.

The sum \( \theta_i + \theta_o \) is a constant. What does this constant represent? Geometrically, we interpret it as double the angle made by the dashed radius line. Optically, it is a measure of the strength of the mirror, i.e., how strongly the mirror focuses light, and so we call it the focal angle, \( \theta_f \),

\[ \theta_i + \theta_o = \theta_f \]

Suppose, for example, that we wish to use a quick and dirty optical test to determine how strong a particular mirror is. We can lay it on the floor as shown in figure c, and use it to make an image of a lamp mounted on the ceiling overhead, which we assume is very far away compared to the radius of curvature of the mirror, so that the mirror intercepts only a very narrow cone of rays from the lamp. This cone is so narrow that its rays are nearly parallel, and \( \theta_o \) is nearly zero. The real image can be observed on a piece of paper. By moving the paper nearer and farther, we can bring the image into focus, at which point we know the paper is located at the image point. Since \( \theta_o \approx 0 \), we have \( \theta_i \approx \theta_f \), and we can then determine this mirror’s focal angle either by measuring \( \theta_i \) directly with a protractor, or indirectly via trigonometry. A strong mirror will bring the rays together to form an image close to the mirror, and these rays will form a blunt-angled cone with a large \( \theta_i \) and \( \theta_f \).
The object and image angles are the same; the angle labeled \( \theta \) in the figure equals both of them. We therefore have \( \theta_i + \theta_o = \theta = \theta_f \). Comparing figures b and c, it is indeed plausible that the angles are related by a factor of two.

At this point, we could consider our work to be done. Typically, we know the strength of the mirror, and we want to find the image location for a given object location. Given the mirror’s focal angle and the object location, we can determine \( \theta_o \) by trigonometry, subtract to find \( \theta_i = \theta_f - \theta_o \), and then do more trig to find the image location.

There is, however, a shortcut that can save us from doing so much work. Figure a/3 shows two right triangles whose legs of length 1 coincide and whose acute angles are \( \theta_o \) and \( \theta_i \). These can be related by trigonometry to the object and image distances shown in figure d:

\[
\tan \theta_o = \frac{1}{d_o} \quad \tan \theta_i = \frac{1}{d_i}
\]

Ever since chapter 2, we’ve been assuming small angles. For small angles, we can use the small-angle approximation \( \tan x \approx x \) (for \( x \) in radians), giving simply

\[
\theta_o = \frac{1}{d_o} \quad \theta_i = \frac{1}{d_i}
\]

We likewise define a distance called the focal length, \( f \) according to \( \theta_f = \frac{1}{f} \). In figure b, \( f \) is the distance from the mirror to the place where the rays cross. We can now reexpress the equation relating the object and image positions as

\[
\frac{1}{f} = \frac{1}{d_i} + \frac{1}{d_o}
\]

Figure e summarizes the interpretation of the focal length and focal angle.\(^1\)

Which form is better, \( \theta_f = \theta_i + \theta_o \) or \( \frac{1}{f} = \frac{1}{d_i} + \frac{1}{d_o} \)? The angular form has in its favor its simplicity and its straightforward visual interpretation, but there are two reasons why we might prefer the second version. First, the numerical values of the angles depend on what we mean by “one unit” for the distance shown as 1 in

\(^1\)There is a standard piece of terminology which is that the “focal point” is the point lying on the optical axis at a distance from the mirror equal to the focal length. This term isn’t particularly helpful, because it names a location where nothing normally happens. In particular, it is \textit{not} normally the place where the rays come to a focus! — that would be the \textit{image} point. In other words, we don’t normally have \( d_i = f \), unless perhaps \( d_o = \infty \). A recent online discussion among some physics teachers (https://carnot.physics.buffalo.edu/archives, Feb. 2006) showed that many disliked the terminology, felt it was misleading, or didn’t know it and would have misinterpreted it if they had come across it. That is, it appears to be what grammarians call a “skunked term” — a word that bothers half the population when it’s used incorrectly, and the other half when it’s used correctly.
Second, it is usually easier to measure distances rather than angles, so the distance form is more convenient for number crunching. Neither form is superior overall, and we will often need to use both to solve any given problem.2

\[ \text{A searchlight example 5} \]

Suppose we need to create a parallel beam of light, as in a searchlight. Where should we place the lightbulb? A parallel beam has zero angle between its rays, so \( \theta_i = 0 \). To place the lightbulb correctly, however, we need to know a distance, not an angle: the distance \( d_o \) between the bulb and the mirror. The problem involves a mixture of distances and angles, so we need to get everything in terms of one or the other in order to solve it. Since the goal is to find a distance, let’s figure out the image distance corresponding to the given angle \( \theta_i = 0 \). These are related by \( d_i = 1/\theta_i \), so we have \( d_i = \infty \). (Yes, dividing by zero gives infinity. Don’t be afraid of infinity. Infinity is a useful problem-solving device.) Solving the distance equation for \( d_o \), we have

\[
\begin{align*}
  d_o &= (1/f - 1/d_i)^{-1} \\
  &= (1/f - 0)^{-1} \\
  &= f
\end{align*}
\]

The bulb has to be placed at a distance from the mirror equal to its focal point.

\[ \text{Diopters example 6} \]

An equation like \( d_i = 1/\theta_i \) really doesn’t make sense in terms of units. Angles are unitless, since radians aren’t really units, so the right-hand side is unitless. We can’t have a left-hand side with units of distance if the right-hand side of the same equation is unitless. This is an artifact of my cavalier statement that the conical bundles of rays spread out to a distance of 1 from the axis where they strike the mirror, without specifying the units used to measure this 1. In real life, optometrists define the thing we’re calling \( \theta_i = 1/d_i \) as the “dioptic strength” of a lens or mirror, and measure it in units of inverse meters (m\(^{-1}\)), also known as diopters (1 D=1 m\(^{-1}\)).

**Magnification**

We have already discussed in the previous chapter how to find the magnification of a virtual image made by a curved mirror. The result is the same for a real image, and we omit the proof, which is very similar. In our new notation, the result is \( M = d_i/d_o \). A numerical example is given in subsection 12.3.2.

---

2I would like to thank Fouad Ajami for pointing out the pedagogical advantages of using both equations side by side.
12.3.2 Other cases with curved mirrors

The equation \( d_i = (1/f - 1/d_o)^{-1} \) can easily produce a negative result, but we have been thinking of \( d_i \) as a distance, and distances can’t be negative. A similar problem occurs with \( \theta_i = \theta_f - \theta_o \) for \( \theta_o > \theta_f \). What’s going on here?

The interpretation of the angular equation is straightforward. As we bring the object closer and closer to the image, \( \theta_o \) gets bigger and bigger, and eventually we reach a point where \( \theta_o = \theta_f \) and \( \theta_i = 0 \). This large object angle represents a bundle of rays forming a cone that is very broad, so broad that the mirror can no longer bend them back so that they reconverge on the axis. The image angle \( \theta_i = 0 \) represents an outgoing bundle of rays that are parallel. The outgoing rays never cross, so this is not a real image, unless we want to be charitable and say that the rays cross at infinity. If we go on bringing the object even closer, we get a virtual image.

To analyze the distance equation, let’s look at a graph of \( d_i \) as a function of \( d_o \). The branch on the upper right corresponds to the case of a real image. Strictly speaking, this is the only part of the graph that we’ve proven corresponds to reality, since we never did any geometry for other cases, such as virtual images. As discussed in the previous section, making \( d_o \) bigger causes \( d_i \) to become smaller, and vice-versa.
Letting $d_o$ be less than $f$ is equivalent to $\theta_o > \theta_f$: a virtual image is produced on the far side of the mirror. This is the first example of Wigner’s “unreasonable effectiveness of mathematics” that we have encountered in optics. Even though our proof depended on the assumption that the image was real, the equation we derived turns out to be applicable to virtual images, provided that we either interpret the positive and negative signs in a certain way, or else modify the equation to have different positive and negative signs.

**self-check D**
Interpret the three places where, in physically realistic parts of the graph, the graph approaches one of the dashed lines. [This will come more naturally if you have learned the concept of limits in a math class.]

Answer, p. 932

*A flat mirror example 7*
We can even apply the equation to a flat mirror. As a sphere gets bigger and bigger, its surface is more and more gently curved. The planet Earth is so large, for example, that we cannot even perceive the curvature of its surface. To represent a flat mirror, we let the mirror's radius of curvature, and its focal length, become infinite. Dividing by infinity gives zero, so we have

$$1/d_o = -1/d_i$$

or

$$d_o = -d_i$$

If we interpret the minus sign as indicating a virtual image on the far side of the mirror from the object, this makes sense.

It turns out that for any of the six possible combinations of real or virtual images formed by converging or diverging lenses or mirrors, we can apply equations of the form

$$\theta_f = \theta_i + \theta_o$$

and

$$\frac{1}{f} = \frac{1}{d_i} + \frac{1}{d_o}$$

with only a modification of plus or minus signs. There are two possible approaches here. The approach we have been using so far is the more popular approach in American textbooks: leave the equation the same, but attach interpretations to the resulting negative or positive values of the variables. The trouble with this approach is that one is then forced to memorize tables of sign conventions, e.g., that the value of $d_i$ should be negative when the image is a virtual image formed by a converging mirror. Positive and negative
signs also have to be memorized for focal lengths. Ugh! It’s highly unlikely that any student has ever retained these lengthy tables in his or her mind for more than five minutes after handing in the final exam in a physics course. Of course one can always look such things up when they are needed, but the effect is to turn the whole thing into an exercise in blindly plugging numbers into formulas.

As you have gathered by now, there is another method which I think is better, and which I’ll use throughout the rest of this book. In this method, all distances and angles are positive by definition, and we put in positive and negative signs in the equations depending on the situation. (I thought I was the first to invent this method, but I’ve been told that this is known as the European sign convention, and that it’s fairly common in Europe.) Rather than memorizing these signs, we start with the generic equations

\[
\theta_f = \pm \theta_i \pm \theta_o
\]

\[
\frac{1}{f} = \pm \frac{1}{d_i} \pm \frac{1}{d_o}
\]

and then determine the signs by a two-step method that depends on ray diagrams. There are really only two signs to determine, not four; the signs in the two equations match up in the way you’d expect. The method is as follows:

1. Use ray diagrams to decide whether \(\theta_o\) and \(\theta_i\) vary in the same way or in opposite ways. (In other words, decide whether making \(\theta_o\) greater results in a greater value of \(\theta_i\) or a smaller one.) Based on this, decide whether the two signs in the angle equation are the same or opposite. If the signs are opposite, go on to step 2 to determine which is positive and which is negative.

2. If the signs are opposite, we need to decide which is the positive one and which is the negative. Since the focal angle is never negative, the smaller angle must be the one with a minus sign.

In step 1, many students have trouble drawing the ray diagram correctly. For simplicity, you should always do your diagram for a point on the object that is on the axis of the mirror, and let one of your rays be the one that is emitted along the axis and reflected straight back on itself, as in the figures in subsection 12.3.1. As shown in figure a/4 in subsection 12.3.1, there are four angles involved: two at the mirror, one at the object (\(\theta_o\)), and one at the image (\(\theta_i\)). Make sure to draw in the normal to the mirror so that you can see the two angles at the mirror. These two angles are equal, so as you change the object position, they fan out or fan in, like opening or closing a book. Once you’ve drawn this effect, you should easily be able to tell whether \(\theta_o\) and \(\theta_i\) change in the same way or in opposite ways.

Although focal lengths are always positive in the method used in this book, you should be aware that diverging mirrors and lenses
are assigned negative focal lengths in the other method, so if you see a lens labeled \( f = -30 \text{ cm} \), you’ll know what it means.

**An anti-shoplifting mirror**

Convenience stores often install a diverging mirror so that the clerk has a view of the whole store and can catch shoplifters. Use a ray diagram to show that the image is reduced, bringing more into the clerk’s field of view. If the focal length of the mirror is 3.0 m, and the mirror is 7.0 m from the farthest wall, how deep is the image of the store?

As shown in ray diagram g/1, \( d_i \) is less than \( d_o \). The magnification, \( M = d_i / d_o \), will be less than one, i.e., the image is actually reduced rather than magnified.

Apply the method outlined above for determining the plus and minus signs. Step 1: The object is the point on the opposite wall. As an experiment, g/2, move the object closer. I did these drawings using illustration software, but if you were doing them by hand, you’d want to make the scale much larger for greater accuracy. Also, although I split figure g into two separate drawings in order to make them easier to understand, you’re less likely to make a mistake if you do them on top of each other.

The two angles at the mirror fan out from the normal. Increasing \( \theta_o \) has clearly made \( \theta_i \) larger as well. (All four angles got bigger.) There must be a cancellation of the effects of changing the two terms on the right in the same way, and the only way to get such a cancellation is if the two terms in the angle equation have opposite signs:

\[
\theta_f = + \theta_i - \theta_o
\]

or

\[
\theta_f = - \theta_i + \theta_o
\]

Step 2: Now which is the positive term and which is negative? Since the image angle is bigger than the object angle, the angle equation must be

\[
\theta_f = \theta_i - \theta_o
\]

in order to give a positive result for the focal angle. The signs of the distance equation behave the same way:

\[
\frac{1}{f} = \frac{1}{d_i} - \frac{1}{d_o}
\]

Solving for \( d_i \), we find

\[
d_i = \left( \frac{1}{f} + \frac{1}{d_o} \right)^{-1} = 2.1 \text{ m}
\]

The image of the store is reduced by a factor of \( 2.1/7.0 = 0.3 \), i.e., it is smaller by 70%.
h / A diverging mirror in the shape of a sphere. The image is reduced \((M < 1)\). This is similar to example 8, but here the image is distorted because the mirror’s curve is not shallow.

A shortcut for real images

In the case of a real image, there is a shortcut for step 1, the determination of the signs. In a real image, the rays cross at both the object and the image. We can therefore time-reverse the ray diagram, so that all the rays are coming from the image and reconverging at the object. Object and image swap roles. Due to this time-reversal symmetry, the object and image cannot be treated differently in any of the equations, and they must therefore have the same signs. They are both positive, since they must add up to a positive result.

12.3.3 ⋆ Aberrations

An imperfection or distortion in an image is called an aberration. An aberration can be produced by a flaw in a lens or mirror, but even with a perfect optical surface some degree of aberration is unavoidable. To see why, consider the mathematical approximation we’ve been making, which is that the depth of the mirror’s curve is small compared to \(d_o\) and \(d_i\). Since only a flat mirror can satisfy this shallow-mirror condition perfectly, any curved mirror will deviate somewhat from the mathematical behavior we derived by assuming that condition. There are two main types of aberration in curved mirrors, and these also occur with lenses.

(1) An object on the axis of the lens or mirror may be imaged correctly, but off-axis objects may be out of focus or distorted. In a camera, this type of aberration would show up as a fuzziness or warping near the sides of the picture when the center was perfectly focused. An example of this is shown in figure i, and in that particular example, the aberration is not a sign that the equipment was of low quality or wasn’t right for the job but rather an inevitable result of trying to flatten a panoramic view; in the limit of a 360-
degree panorama, the problem would be similar to the problem of representing the Earth’s surface on a flat map, which can’t be accomplished without distortion.

(2) The image may be sharp when the object is at certain distances and blurry when it is at other distances. The blurriness occurs because the rays do not all cross at exactly the same point. If we know in advance the distance of the objects with which the mirror or lens will be used, then we can optimize the shape of the optical surface to make in-focus images in that situation. For instance, a spherical mirror will produce a perfect image of an object that is at the center of the sphere, because each ray is reflected directly onto the radius along which it was emitted. For objects at greater distances, however, the focus will be somewhat blurry. In astronomy the objects being used are always at infinity, so a spherical mirror is a poor choice for a telescope. A different shape (a parabola) is better specialized for astronomy.

One way of decreasing aberration is to use a small-diameter mirror or lens, or block most of the light with an opaque screen with a hole in it, so that only light that comes in close to the axis can get

i / This photo was taken using a “fish-eye lens,” which gives an extremely large field of view.
Spherical mirrors are the cheapest to make, but parabolic mirrors are better for making images of objects at infinity. A sphere has equal curvature everywhere, but a parabola has tighter curvature at its center and gentler curvature at the sides. Through. Either way, we are using a smaller portion of the lens or mirror whose curvature will be more shallow, thereby making the shallow-mirror (or thin-lens) approximation more accurate. Your eye does this by narrowing down the pupil to a smaller hole. In a camera, there is either an automatic or manual adjustment, and narrowing the opening is called “stopping down.” The disadvantage of stopping down is that light is wasted, so the image will be dimmer or a longer exposure must be used.

Even though the spherical mirror (solid line) is not well adapted for viewing an object at infinity, we can improve its performance greatly by stopping it down. Now the only part of the mirror being used is the central portion, where its shape is virtually indistinguishable from a parabola (dashed line).

What I would suggest you take away from this discussion for the sake of your general scientific education is simply an understanding of what an aberration is, why it occurs, and how it can be reduced, not detailed facts about specific types of aberrations.
12.4 Refraction

Economists normally consider free markets to be the natural way of judging the monetary value of something, but social scientists also use questionnaires to gauge the relative value of privileges, disadvantages, or possessions that cannot be bought or sold. They ask people to imagine that they could trade one thing for another and ask which they would choose. One interesting result is that the average light-skinned person in the U.S. would rather lose an arm than suffer the racist treatment routinely endured by African-Americans. Even more impressive is the value of sight. Many prospective parents can imagine without too much fear having a deaf child, but would have a far more difficult time coping with raising a blind one.

So great is the value attached to sight that some have imbued it with mystical aspects. Joan of Arc saw visions, and my college has a “vision statement.” Christian fundamentalists who perceive a conflict between evolution and their religion have claimed that the eye is such a perfect device that it could never have arisen through a process as helter-skelter as evolution, or that it could not have evolved because half of an eye would be useless. In fact, the structure of an eye is fundamentally dictated by physics, and it has arisen separately by evolution somewhere between eight and 40 times, depending on which biologist you ask. We humans have a version of the eye that can be traced back to the evolution of a light-sensitive “eye spot” on the head of an ancient invertebrate. A sunken pit then developed so that the eye would only receive light from one direction, allowing the organism to tell where the light was coming from. (Modern flatworms have this type of eye.) The top of the pit then became partially covered, leaving a hole, for even greater directionality (as in the nautilus). At some point the cavity became filled with jelly, and this jelly finally became a lens, resulting in the...
12.4.1 Refraction

Refraction

The fundamental physical phenomenon at work in the eye is that when light crosses a boundary between two media (such as air and the eye’s jelly), part of its energy is reflected, but part passes into the new medium. In the ray model of light, we describe the original ray as splitting into a reflected ray and a transmitted one (the one that gets through the boundary). Of course the reflected ray goes in a direction that is different from that of the original one, according to the rules of reflection we have already studied. More surprisingly — and this is the crucial point for making your eye focus light — the transmitted ray is bent somewhat as well. This bending phenomenon is called refraction. The origin of the word is the same as that of the word “fracture,” i.e., the ray is bent or “broken.” (Keep in mind, however, that light rays are not physical objects that can really be “broken.”) Refraction occurs with all waves, not just light waves.

The actual anatomy of the eye, b, is quite complex, but in essence it is very much like every other optical device based on refraction. The rays are bent when they pass through the front surface of the eye, c. Rays that enter farther from the central axis are bent more, with the result that an image is formed on the retina. There is only one slightly novel aspect of the situation. In most human-built optical devices, such as a movie projector, the light is bent as it passes into a lens, bent again as it reemerges, and then reaches a focus beyond the lens. In the eye, however, the “screen” is inside the eye, so the rays are only refracted once, on entering the jelly, and never emerge again.

A common misconception is that the “lens” of the eye is what does the focusing. All the transparent parts of the eye are made of fairly similar stuff, so the dramatic change in medium is when a ray crosses from the air into the eye (at the outside surface of the cornea). This is where nearly all the refraction takes place. The lens medium differs only slightly in its optical properties from the rest of the eye, so very little refraction occurs as light enters and exits the lens. The lens, whose shape is adjusted by muscles attached to it, is only meant for fine-tuning the focus to form images of near or far objects.
Refraction has time-reversal symmetry. Regardless of whether the light is going into or out of the water, the relationship between the two angles is the same, and the ray is closer to the normal while in the water.

Snell’s law

The numerical rule governing refraction was discovered by Snell, who must have collected experimental data something like what is
shown on this graph and then attempted by trial and error to find the right equation. The equation he came up with was
\[ \frac{\sin \theta_1}{\sin \theta_2} = \text{constant} \]

The value of the constant would depend on the combination of media used. For instance, any one of the data points in the graph would have sufficed to show that the constant was 1.3 for an air-water interface (taking air to be substance 1 and water to be substance 2).

Snell further found that if media A and B gave a constant \( K_{AB} \) and media B and C gave a constant \( K_{BC} \), then refraction at an interface between A and C would be described by a constant equal to the product, \( K_{AC} = K_{AB}K_{BC} \). This is exactly what one would expect if the constant depended on the ratio of some number characterizing one medium to the number characteristic of the second medium. This number is called the \textit{index of refraction} of the medium, written as \( n \) in equations. Since measuring the angles would only allow him to determine the \textit{ratio} of the indices of refraction of two media, Snell had to pick some medium and define it as having \( n = 1 \). He chose to define vacuum as having \( n = 1 \). (The index of refraction of air at normal atmospheric pressure is 1.0003, so for most purposes it is a good approximation to assume that air has \( n = 1 \).) He also had to decide which way to define the ratio, and he chose to define it so that media with their rays closer to the normal would have larger indices of refraction. This had the advantage that denser media would typically have higher indices of refraction, and for this reason the index of refraction is also referred to as the optical density. Written in terms of indices of refraction, Snell’s equation becomes
\[ \frac{\sin \theta_1}{\sin \theta_2} = \frac{n_2}{n_1} \]

but rewriting it in the form
\[ n_1 \sin \theta_1 = n_2 \sin \theta_2 \]

[relationship between angles of rays at the interface between media with indices of refraction \( n_1 \) and \( n_2 \); angles are defined with respect to the normal]

makes us less likely to get the 1’s and 2’s mixed up, so this the way most people remember Snell’s law. A few indices of refraction are given in the back of the book.

**self-check E**

(1) What would the graph look like for two substances with the same index of refraction?

(2) Based on the graph, when does refraction at an air-water interface change the direction of a ray most strongly? 

\( \triangleright \) Answer, p. 932
Example 10.

A submarine shines its searchlight up toward the surface of the water. What is the angle $\alpha$ shown in the figure?

The tricky part is that Snell’s law refers to the angles with respect to the normal. Forgetting this is a very common mistake. The beam is at an angle of 30° with respect to the normal in the water. Let’s refer to the air as medium 1 and the water as 2. Solving Snell’s law for $\theta_1$, we find

$$\theta_1 = \sin^{-1} \left( \frac{n_2}{n_1} \sin \theta_2 \right) .$$

As mentioned above, air has an index of refraction very close to 1, and water’s is about 1.3, so we find $\theta_1 = 40^\circ$. The angle $\alpha$ is therefore 50°.

The index of refraction is related to the speed of light.

What neither Snell nor Newton knew was that there is a very simple interpretation of the index of refraction. This may come as a relief to the reader who is taken aback by the complex reasoning involving proportionalities that led to its definition. Later experiments showed that the index of refraction of a medium was inversely proportional to the speed of light in that medium. Since $c$ is defined as the speed of light in vacuum, and $n = 1$ is defined as the index of refraction of vacuum, we have

$$n = \frac{c}{v} .$$

[$n =$ medium’s index of refraction, $v =$ speed of light in that medium, $c =$ speed of light in a vacuum$]$.

Many textbooks start with this as the definition of the index of refraction, although that approach makes the quantity’s name somewhat of a mystery, and leaves students wondering why $c/v$ was used rather than $v/c$. It should also be noted that measuring angles of refraction is a far more practical method for determining $n$ than direct measurement of the speed of light in the substance of interest.

A mechanical model of Snell’s law

Why should refraction be related to the speed of light? The mechanical model shown in the figure may help to make this more plausible. Suppose medium 2 is thick, sticky mud, which slows down the car. The car’s right wheel hits the mud first, causing the right side of the car to slow down. This will cause the car to turn to the right until it moves far enough forward for the left wheel to cross into the mud. After that, the two sides of the car will once again be moving at the same speed, and the car will go straight.

Of course, light isn’t a car. Why should a beam of light have anything resembling a “left wheel” and “right wheel?” After all,
the mechanical model would predict that a motorcycle would go straight, and a motorcycle seems like a better approximation to a ray of light than a car. The whole thing is just a model, not a description of physical reality.

A derivation of Snell’s law

However intuitively appealing the mechanical model may be, light is a wave, and we should be using wave models to describe refraction. In fact Snell’s law can be derived quite simply from wave concepts. Figure \( j \) shows the refraction of a water wave. The water in the upper left part of the tank is shallower, so the speed of the waves is slower there, and their wavelengths is shorter. The reflected part of the wave is also very faintly visible.

In the close-up view on the right, the dashed lines are normals to the interface. The two marked angles on the right side are both equal to \( \theta_1 \), and the two on the left to \( \theta_2 \).

Trigonometry gives

\[
\sin \theta_1 = \frac{\lambda_1}{h} \quad \text{and} \quad \sin \theta_2 = \frac{\lambda_2}{h}.
\]

Eliminating \( h \) by dividing the equations, we find

\[
\frac{\sin \theta_1}{\sin \theta_2} = \frac{\lambda_1}{\lambda_2}.
\]

The frequencies of the two waves must be equal or else they would get out of step, so by \( v = f\lambda \) we know that their wavelengths are
proportional to their velocities. Combining $\lambda \propto v$ with $v \propto 1/n$ gives $\lambda \propto 1/n$, so we find

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{n_2}{n_1},$$

which is one form of Snell’s law.

Ocean waves near and far from shore example 11
Ocean waves are formed by winds, typically on the open sea, and the wavefronts are perpendicular to the direction of the wind that formed them. At the beach, however, you have undoubtedly observed that waves tend come in with their wavefronts very nearly (but not exactly) parallel to the shoreline. This is because the speed of water waves in shallow water depends on depth: the shallower the water, the slower the wave. Although the change from the fast-wave region to the slow-wave region is gradual rather than abrupt, there is still refraction, and the wave motion is nearly perpendicular to the normal in the slow region.

Color and refraction
In general, the speed of light in a medium depends both on the medium and on the wavelength of the light. Another way of saying it is that a medium’s index of refraction varies with wavelength. This is why a prism can be used to split up a beam of white light into a rainbow. Each wavelength of light is refracted through a different angle.

How much light is reflected, and how much is transmitted?
In section 6.2 we developed an equation for the percentage of the wave energy that is transmitted and the percentage reflected at a boundary between media. This was only done in the case of waves in one dimension, however, and rather than discuss the full three-dimensional generalization it will be more useful to go into some qualitative observations about what happens. First, reflection happens only at the interface between two media, and two media with the same index of refraction act as if they were a single medium. Thus, at the interface between media with the same index of refraction, there is no reflection, and the ray keeps going straight. Continuing this line of thought, it is not surprising that we observe very little reflection at an interface between media with similar indices of refraction.

The next thing to note is that it is possible to have situations where no possible angle for the refracted ray can satisfy Snell’s law. Solving Snell’s law for $\theta_2$, we find

$$\theta_2 = \sin^{-1} \left( \frac{n_1}{n_2} \sin \theta_1 \right),$$

and if $n_1$ is greater than $n_2$, then there will be large values of $\theta_1$ for which the quantity $(n_1/n_2) \sin \theta$ is greater than one, meaning
that your calculator will flash an error message at you when you try to take the inverse sine. What can happen physically in such a situation? The answer is that all the light is reflected, so there is no refracted ray. This phenomenon is known as total internal reflection, and is used in the fiber-optic cables that nowadays carry almost all long-distance telephone calls. The electrical signals from your phone travel to a switching center, where they are converted from electricity into light. From there, the light is sent across the country in a thin transparent fiber. The light is aimed straight into the end of the fiber, and as long as the fiber never goes through any turns that are too sharp, the light will always encounter the edge of the fiber at an angle sufficiently oblique to give total internal reflection. If the fiber-optic cable is thick enough, one can see an image at one end of whatever the other end is pointed at.

Alternatively, a bundle of cables can be used, since a single thick cable is too hard to bend. This technique for seeing around corners is useful for making surgery less traumatic. Instead of cutting a person wide open, a surgeon can make a small “keyhole” incision and insert a bundle of fiber-optic cable (known as an endoscope) into the body.

Since rays at sufficiently large angles with respect to the normal may be completely reflected, it is not surprising that the relative amount of reflection changes depending on the angle of incidence, and is greatest for large angles of incidence.

**Discussion Questions**

**A** What index of refraction should a fish have in order to be invisible to other fish?

**B** Does a surgeon using an endoscope need a source of light inside the body cavity? If so, how could this be done without inserting a light bulb through the incision?

**C** A denser sample of a gas has a higher index of refraction than a less dense sample (i.e., a sample under lower pressure), but why would it not make sense for the index of refraction of a gas to be proportional to density?

**D** The earth’s atmosphere gets thinner and thinner as you go higher in altitude. If a ray of light comes from a star that is below the zenith, what will happen to it as it comes into the earth’s atmosphere?

**E** Does total internal reflection occur when light in a denser medium encounters a less dense medium, or the other way around? Or can it occur in either case?

### 12.4.2 Lenses

Figures n/1 and n/2 show examples of lenses forming images. There is essentially nothing for you to learn about imaging with lenses that is truly new. You already know how to construct and use ray diagrams, and you know about real and virtual images. The
The concept of the focal length of a lens is the same as for a curved mirror. The equations for locating images and determining magnifications are of the same form. It’s really just a question of flexing your mental muscles on a few examples. The following self-checks and discussion questions will get you started.

**self-check F**

1. In figures n/1 and n/2, classify the images as real or virtual.

2. Glass has an index of refraction that is greater than that of air. Consider the topmost ray in figure n/1. Explain why the ray makes a slight left turn upon entering the lens, and another left turn when it exits.

3. If the flame in figure n/2 was moved closer to the lens, what would happen to the location of the image?  

**Discussion Questions**

**A** In figures n/1 and n/2, the front and back surfaces are parallel to each other at the center of the lens. What will happen to a ray that enters near the center, but not necessarily along the axis of the lens? Draw a BIG ray diagram, and show a ray that comes from off axis.

   In discussion questions B-F, don’t draw ultra-detailed ray diagrams as in A.

**B** Suppose you wanted to change the setup in figure n/1 so that the location of the actual flame in the figure would instead be occupied by an image of a flame. Where would you have to move the candle to achieve this? What about in n/2?

**C** There are three qualitatively different types of image formation that can occur with lenses, of which figures n/1 and n/2 exhaust only two. Figure out what the third possibility is. Which of the three possibilities can result in a magnification greater than one? Cf. problem 10, p. 801.
D. Classify the examples shown in figure o according to the types of images delineated in discussion question C.

E. In figures n/1 and n/2, the only rays drawn were those that happened to enter the lenses. Discuss this in relation to figure o.

F. In the right-hand side of figure o, the image viewed through the lens is in focus, but the side of the rose that sticks out from behind the lens is not. Why?

o / Two images of a rose created by the same lens and recorded with the same camera.

12.4.3 *The lensmaker’s equation*

The focal length of a spherical mirror is simply $r/2$, but we cannot expect the focal length of a lens to be given by pure geometry, since it also depends on the index of refraction of the lens. Suppose we have a lens whose front and back surfaces are both spherical. (This is no great loss of generality, since any surface with a sufficiently shallow curvature can be approximated with a sphere.) Then if the lens is immersed in a medium with an index of refraction of 1, its focal length is given approximately by

$$ f = \left[ (n - 1) \left| \frac{1}{r_1} \pm \frac{1}{r_2} \right| \right]^{-1}, $$

where $n$ is the index of refraction and $r_1$ and $r_2$ are the radii of curvature of the two surfaces of the lens. This is known as the lensmaker’s equation. In my opinion it is not particularly worthy
of memorization. The positive sign is used when both surfaces are curved outward or both are curved inward; otherwise a negative sign applies. The proof of this equation is left as an exercise to those readers who are sufficiently brave and motivated.

12.4.4 Dispersion

For most materials, we observe that the index of refraction depends slightly on wavelength, being highest at the blue end of the visible spectrum and lowest at the red. For example, white light disperses into a rainbow when it passes through a prism, q. Even when the waves involved aren’t light waves, and even when refraction isn’t of interest, the dependence of wave speed on wavelength is referred to as dispersion. Dispersion inside spherical raindrops is responsible for the creation of rainbows in the sky, and in an optical instrument such as the eye or a camera it is responsible for a type of aberration called chromatic aberration (subsection 12.3.3 and problem 28). As we’ll see in subsection 13.3.2, dispersion causes a wave that is not a pure sine wave to have its shape distorted as it travels, and also causes the speed at which energy and information are transported by the wave to be different from what one might expect from a naive calculation. The microscopic reasons for dispersion of light in matter are discussed in optional subsection 12.4.6.

12.4.5 ★ The principle of least time for refraction

We have seen previously how the rules governing straight-line motion of light and reflection of light can be derived from the principle of least time. What about refraction? In the figure, it is indeed plausible that the bending of the ray serves to minimize the time required to get from a point A to point B. If the ray followed the unbent path shown with a dashed line, it would have to travel a longer distance in the medium in which its speed is slower. By bending the correct amount, it can reduce the distance it has to cover in the slower medium without going too far out of its way. It is true that Snell’s law gives exactly the set of angles that minimizes the time required for light to get from one point to another. The proof of this fact is left as an exercise (problem 38, p. 806).
12.4.6 *Microscopic description of refraction*

Given that the speed of light is different in different media, we’ve seen two different explanations (on p. 778 and in subsection 12.4.5 above) of why refraction must occur. What we haven’t yet explained is why the speed of light does depend on the medium.

![Graph showing the index of refraction of silica glass](https://www.seas.ucla.edu/~pilon/Publications/AO2007-1.pdf)

A good clue as to what’s going on comes from the figure s. The relatively minor variation of the index of refraction within the visible spectrum was misleading. At certain specific frequencies, $n$ exhibits wild swings in the positive and negative directions. After each such swing, we reach a new, lower plateau on the graph. These frequencies are resonances. For example, the visible part of the spectrum lies on the left-hand tail of a resonance at about $2 \times 10^{15}$ Hz, corresponding to the ultraviolet part of the spectrum. This resonance arises from the vibration of the electrons, which are bound to the nuclei as if by little springs. Because this resonance is narrow, the effect on visible-light frequencies is relatively small, but it is stronger at the blue end of the spectrum than at the red end. Near each resonance, not only does the index of refraction fluctuate wildly, but the glass becomes nearly opaque; this is because the vibration becomes very strong, causing energy to be dissipated as heat. The “staircase” effect is the same one visible in any resonance, e.g., figure k on p. 180: oscillators have a finite response for $f \ll f_0$, but the response approaches zero for $f \gg f_0$.

So far, we have a qualitative explanation of the frequency-variation of the loosely defined “strength” of the glass’s effect on a light wave, but we haven’t explained why the effect is observed as a change in speed, or why each resonance is an up-down swing rather than a single positive peak. To understand these effects in more detail, we need to consider the phase response of the oscillator. As shown in the bottom panel of figure j on p. 181, the phase response reverses itself as we pass through a resonance.

Suppose that a plane wave is normally incident on the left side of a thin sheet of glass, $t/1$, at $f \ll f_0$. The light wave observed on the
right side consists of a superposition of the incident wave consisting of \( \mathbf{E}_0 \) and \( \mathbf{B}_0 \) with a secondary wave \( \mathbf{E}^* \) and \( \mathbf{B}^* \) generated by the oscillating charges in the glass. Since the frequency is far below resonance, the response \( qx \) of a vibrating charge \( q \) is in phase with the driving force \( \mathbf{E}_0 \). The current is the derivative of this quantity, and therefore 90 degrees ahead of it in phase. The magnetic field generated by a sheet of current has been analyzed in subsection 11.2.1, and the result, shown in figure e on p. 668, is just what we would expect from the right-hand rule. We find, t/1, that the secondary wave is 90 degrees ahead of the incident one in phase. The incident wave still exists on the right side of the sheet, but it is superposed with the secondary one. Their addition is shown in t/2 using the complex number representation introduced in subsection 10.5.7. The superposition of the two fields lags behind the incident wave, which is the effect we would expect if the wave had traveled more slowly through the glass.

In the case \( f \gg 0 \), the same analysis applies except that the phase of the secondary wave is reversed. The transmitted wave is advanced rather than retarded in phase. This explains the dip observed in figure s after each spike.

All of this is in accord with our understanding of relativity, ch. 7, in which we saw that the universal speed \( c \) was to be understood fundamentally as a conversion factor between the units used to measure time and space — not as the speed of light. Since \( c \) isn’t defined as the speed of light, it’s of no fundamental importance whether light has a different speed in matter than it does in vacuum. In fact, the picture we’ve built up here is one in which all of our electromagnetic waves travel at \( c \); propagation at some other speed is only what appears to happen because of the superposition of the \( (\mathbf{E}_0, \mathbf{B}_0) \) and \( (\mathbf{E}^*, \mathbf{B}^*) \) waves, both of which move at \( c \).

But it is worrisome that at the frequencies where \( n < 1 \), the speed of the wave is greater than \( c \). According to special relativity, information is never supposed to be transmitted at speeds greater than \( c \), since this would produce situations in which a signal could be received before it was transmitted! This difficulty is resolved in subsection 13.3.2, where we show that there are two different velocities that can be defined for a wave in a dispersive medium, the phase velocity and the group velocity. The group velocity is the velocity at which information is transmitted, and it is always less than \( c \).
12.5 Wave Optics

Electron microscopes can make images of individual atoms, but why will a visible-light microscope never be able to? Stereo speakers create the illusion of music that comes from a band arranged in your living room, but why doesn’t the stereo illusion work with bass notes? Why are computer chip manufacturers investing billions of dollars in equipment to etch chips with x-rays instead of visible light?

The answers to all of these questions have to do with the subject of wave optics. So far this book has discussed the interaction of light waves with matter, and its practical applications to optical devices like mirrors, but we have used the ray model of light almost exclusively. Hardly ever have we explicitly made use of the fact that light is an electromagnetic wave. We were able to get away with the simple ray model because the chunks of matter we were discussing, such as lenses and mirrors, were thousands of times larger than a wavelength of light. We now turn to phenomena and devices that can only be understood using the wave model of light.

12.5.1 Diffraction

Figure a shows a typical problem in wave optics, enacted with water waves. It may seem surprising that we don’t get a simple pattern like figure b, but the pattern would only be that simple if the wavelength was hundreds of times shorter than the distance between the gaps in the barrier and the widths of the gaps.

Wave optics is a broad subject, but this example will help us to pick out a reasonable set of restrictions to make things more manageable:

(1) We restrict ourselves to cases in which a wave travels through a uniform medium, encounters a certain area in which the medium has different properties, and then emerges on the other side into a second uniform region.

(2) We assume that the incoming wave is a nice tidy sine-wave pattern with wavefronts that are lines (or, in three dimensions, planes).

(3) In figure a we can see that the wave pattern immediately beyond the barrier is rather complex, but farther on it sorts itself out into a set of wedges separated by gaps in which the water is still. We will restrict ourselves to studying the simpler wave patterns that occur farther away, so that the main question of interest is how intense the outgoing wave is at a given angle.

The kind of phenomenon described by restriction (1) is called *diffraction*. Diffraction can be defined as the behavior of a wave when it encounters an obstacle or a nonuniformity in its medium. In general, diffraction causes a wave to bend around obstacles and
make patterns of strong and weak waves radiating out beyond the obstacle. Understanding diffraction is the central problem of wave optics. If you understand diffraction, even the subset of diffraction problems that fall within restrictions (2) and (3), the rest of wave optics is icing on the cake.

Diffraction can be used to find the structure of an unknown diffracting object: even if the object is too small to study with ordinary imaging, it may be possible to work backward from the diffraction pattern to learn about the object. The structure of a crystal, for example, can be determined from its x-ray diffraction pattern.

Diffraction can also be a bad thing. In a telescope, for example, light waves are diffracted by all the parts of the instrument. This will cause the image of a star to appear fuzzy even when the focus has been adjusted correctly. By understanding diffraction, one can learn how a telescope must be designed in order to reduce this problem — essentially, it should have the biggest possible diameter.

There are two ways in which restriction (2) might commonly be violated. First, the light might be a mixture of wavelengths. If we simply want to observe a diffraction pattern or to use diffraction as a technique for studying the object doing the diffracting (e.g., if the object is too small to see with a microscope), then we can pass the light through a colored filter before diffracting it.

A second issue is that light from sources such as the sun or a lightbulb does not consist of a nice neat plane wave, except over very small regions of space. Different parts of the wave are out of step with each other, and the wave is referred to as incoherent. One way of dealing with this is shown in figure c. After filtering to select a certain wavelength of red light, we pass the light through a small pinhole. The region of the light that is intercepted by the pinhole is so small that one part of it is not out of step with another. Beyond the pinhole, light spreads out in a spherical wave; this is analogous to what happens when you speak into one end of a paper towel roll and the sound waves spread out in all directions from the other end. By the time the spherical wave gets to the double slit it has spread out and reduced its curvature, so that we can now think of it as a simple plane wave.

If this seems laborious, you may be relieved to know that modern technology gives us an easier way to produce a single-wavelength, coherent beam of light: the laser.

The parts of the final image on the screen in c are called diffraction fringes. The center of each fringe is a point of maximum brightness, and halfway between two fringes is a minimum.

Discussion Question

A Why would x-rays rather than visible light be used to find the structure...
of a crystal? Sound waves are used to make images of fetuses in the womb. What would influence the choice of wavelength?

12.5.2 Scaling of diffraction

This chapter has “optics” in its title, so it is nominally about light, but we started out with an example involving water waves. Water waves are certainly easier to visualize, but is this a legitimate comparison? In fact the analogy works quite well, despite the fact that a light wave has a wavelength about a million times shorter. This is because diffraction effects scale uniformly. That is, if we enlarge or reduce the whole diffraction situation by the same factor, including both the wavelengths and the sizes of the obstacles the wave encounters, the result is still a valid solution.

This is unusually simple behavior! In subsection 0.2.2 we saw many examples of more complex scaling, such as the impossibility of bacteria the size of dogs, or the need for an elephant to eliminate heat through its ears because of its small surface-to-volume ratio, whereas a tiny shrew’s life-style centers around conserving its body heat.

Of course water waves and light waves differ in many ways, not just in scale, but the general facts you will learn about diffraction are applicable to all waves. In some ways it might have been more appropriate to insert this chapter after section 6.2 on bounded waves, but many of the important applications are to light waves, and you would probably have found these much more difficult without any background in optics.

Another way of stating the simple scaling behavior of diffraction is that the diffraction angles we get depend only on the unitless ratio \( \frac{\lambda}{d} \), where \( \lambda \) is the wavelength of the wave and \( d \) is some dimension of the diffracting objects, e.g., the center-to-center spacing between the slits in figure a. If, for instance, we scale up both \( \lambda \) and \( d \) by a factor of 37, the ratio \( \frac{\lambda}{d} \) will be unchanged.

12.5.3 The correspondence principle

The only reason we don’t usually notice diffraction of light in everyday life is that we don’t normally deal with objects that are comparable in size to a wavelength of visible light, which is about a millionth of a meter. Does this mean that wave optics contradicts ray optics, or that wave optics sometimes gives wrong results? No. If you hold three fingers out in the sunlight and cast a shadow with them, either wave optics or ray optics can be used to predict the straightforward result: a shadow pattern with two bright lines where the light has gone through the gaps between your fingers. Wave optics is a more general theory than ray optics, so in any case where ray optics is valid, the two theories will agree. This is an example of a general idea enunciated by the physicist Niels Bohr, called the *correspondence principle*: when flaws in a physical theory...
lead to the creation of a new and more general theory, the new
time must still agree with the old theory within its more restricted
area of applicability. After all, a theory is only created as a way of
describing experimental observations. If the original theory had not
worked in any cases at all, it would never have become accepted.

In the case of optics, the correspondence principle tells us that
when λ/d is small, both the ray and the wave model of light must
give approximately the same result. Suppose you spread your fingers
and cast a shadow with them using a coherent light source. The
quantity λ/d is about 10⁻⁴, so the two models will agree very closely.
(To be specific, the shadows of your fingers will be outlined by a
series of light and dark fringes, but the angle subtended by a fringe
will be on the order of 10⁻⁴ radians, so they will be invisible and
washed out by the natural fuzziness of the edges of sun-shadows,
caused by the finite size of the sun.)

self-check G

What kind of wavelength would an electromagnetic wave have to have
in order to diffract dramatically around your body? Does this contradict
the correspondence principle?

Answer, p. 932

12.5.4 Huygens' principle

Returning to the example of double-slit diffraction, note the
strong visual impression of two overlapping sets of concentric semi-
circles. This is an example of Huygens' principle, named after a
Dutch physicist and astronomer. (The first syllable rhymes with
"boy.") Huygens' principle states that any wavefront can be broken
down into many small side-by-side wave peaks, g, which then spread
out as circular ripples, h, and by the principle of superposition, the
result of adding up these sets of ripples must give the same result
as allowing the wave to propagate forward, i. In the case of sound
or light waves, which propagate in three dimensions, the "ripples"
are actually spherical rather than circular, but we can often imagine
things in two dimensions for simplicity.

In double-slit diffraction the application of Huygens' principle is
visually convincing: it is as though all the sets of ripples have been
blocked except for two. It is a rather surprising mathematical fact,
however, that Huygens' principle gives the right result in the case of
an unobstructed linear wave, h and i. A theoretically infinite number
of circular wave patterns somehow conspire to add together and
produce the simple linear wave motion with which we are familiar.

Since Huygens' principle is equivalent to the principle of super-
position, and superposition is a property of waves, what Huygens
had created was essentially the first wave theory of light. However,
he imagined light as a series of pulses, like hand claps, rather than
as a sinusoidal wave.

The history is interesting. Isaac Newton loved the atomic theory
of matter so much that he searched enthusiastically for evidence that light was also made of tiny particles. The paths of his light particles would correspond to rays in our description; the only significant difference between a ray model and a particle model of light would occur if one could isolate individual particles and show that light had a “graininess” to it. Newton never did this, so although he thought of his model as a particle model, it is more accurate to say he was one of the builders of the ray model.

Almost all that was known about reflection and refraction of light could be interpreted equally well in terms of a particle model or a wave model, but Newton had one reason for strongly opposing Huygens’ wave theory. Newton knew that waves exhibited diffraction, but diffraction of light is difficult to observe, so Newton believed that light did not exhibit diffraction, and therefore must not be a wave. Although Newton’s criticisms were fair enough, the debate also took on the overtones of a nationalistic dispute between England and continental Europe, fueled by English resentment over Leibniz’s supposed plagiarism of Newton’s calculus. Newton wrote a book on optics, and his prestige and political prominence tended to discourage questioning of his model.

Thomas Young (1773-1829) was the person who finally, a hundred years later, did a careful search for wave interference effects with light and analyzed the results correctly. He observed double-slit diffraction of light as well as a variety of other diffraction effects, all of which showed that light exhibited wave interference effects, and that the wavelengths of visible light waves were extremely short. The crowning achievement was the demonstration by the experimentalist Heinrich Hertz and the theorist James Clerk Maxwell that light was an electromagnetic wave. Maxwell is said to have related his discovery to his wife one starry evening and told her that she was the only other person in the world who knew what starlight was.

12.5.5 Double-slit diffraction

Let’s now analyze double-slit diffraction, k, using Huygens’ principle. The most interesting question is how to compute the angles such as X and Z where the wave intensity is at a maximum, and the in-between angles like Y where it is minimized. Let’s measure all our angles with respect to the vertical center line of the figure, which was the original direction of propagation of the wave.

If we assume that the width of the slits is small (on the order of the wavelength of the wave or less), then we can imagine only a single set of Huygens ripples spreading out from each one, l. White lines represent peaks, black ones troughs. The only dimension of the diffracting slits that has any effect on the geometric pattern of the overlapping ripples is then the center-to-center distance, d, between the slits.
We know from our discussion of the scaling of diffraction that there must be some equation that relates an angle like $\theta_Z$ to the ratio $\lambda/d$,

$$\frac{\lambda}{d} \leftrightarrow \theta_Z .$$

If the equation for $\theta_Z$ depended on some other expression such as $\lambda + d$ or $\lambda^2/d$, then it would change when we scaled $\lambda$ and $d$ by the same factor, which would violate what we know about the scaling of diffraction.

Along the central maximum line, X, we always have positive waves coinciding with positive ones and negative waves coinciding with negative ones. (I have arbitrarily chosen to take a snapshot of the pattern at a moment when the waves emerging from the slit are experiencing a positive peak.) The superposition of the two sets of ripples therefore results in a doubling of the wave amplitude along this line. There is constructive interference. This is easy to explain, because by symmetry, each wave has had to travel an equal number of wavelengths to get from its slit to the center line, m: Because both sets of ripples have ten wavelengths to cover in order to reach the point along direction X, they will be in step when they get there.

At the point along direction Y shown in the same figure, one wave has traveled ten wavelengths, and is therefore at a positive extreme, but the other has traveled only nine and a half wavelengths, so it at a negative extreme. There is perfect cancellation, so points along this line experience no wave motion.

But the distance traveled does not have to be equal in order to get constructive interference. At the point along direction Z, one wave has gone nine wavelengths and the other ten. They are both at a positive extreme.

**self-check H**

At a point half a wavelength below the point marked along direction X, carry out a similar analysis. 

To summarize, we will have perfect constructive interference at any point where the distance to one slit differs from the distance to the other slit by an integer number of wavelengths. Perfect destructive interference will occur when the number of wavelengths of path length difference equals an integer plus a half.

Now we are ready to find the equation that predicts the angles of the maxima and minima. The waves travel different distances to get to the same point in space, n. We need to find whether the waves are in phase (in step) or out of phase at this point in order to predict whether there will be constructive interference, destructive interference, or something in between.

One of our basic assumptions in this chapter is that we will only be dealing with the diffracted wave in regions very far away from the
Cutting \( d \) in half doubles the angles of the diffraction fringes.

Double-slit diffraction patterns of long-wavelength red light (top) and short-wavelength blue light (bottom).

The difference in path length is related to \( d \) and \( \theta \) by the equation

\[
\frac{L - L'}{d} = \sin \theta
\]

Constructive interference will result in a maximum at angles for which \( L - L' \) is an integer number of wavelengths,

\[
L - L' = m\lambda
\]

[condition for a maximum; \( m \) is an integer]

Here \( m \) equals 0 for the central maximum, \(-1\) for the first maximum to its left, \(+2\) for the second maximum on the right, etc. Putting all the ingredients together, we find

\[
\frac{\lambda}{d} = \frac{\sin \theta}{m}
\]

[condition for a maximum; \( m \) is an integer]

Similarly, the condition for a minimum is

\[
\frac{\lambda}{d} = \frac{\sin \theta}{m + 1/2}
\]

[condition for a minimum; \( m \) is an integer plus \( 1/2 \)]

That is, the minima are about halfway between the maxima.

As expected based on scaling, this equation relates angles to the unitless ratio \( \lambda/d \). Alternatively, we could say that we have proven the scaling property in the special case of double-slit diffraction. It was inevitable that the result would have these scaling properties, since the whole proof was geometric, and would have been equally valid when enlarged or reduced on a photocopying machine!

Counterintuitively, this means that a diffracting object with smaller dimensions produces a bigger diffraction pattern, \( p \).
Double-slit diffraction of blue and red light  

Example 12

Blue light has a shorter wavelength than red. For a given double-slit spacing $d$, the smaller value of $\lambda/d$ for blue light leads to smaller values of $\sin \theta$, and therefore to a more closely spaced set of diffraction fringes. (g)

The correspondence principle  

Example 13

Let's also consider how the equations for double-slit diffraction relate to the correspondence principle. When the ratio $\lambda/d$ is very small, we should recover the case of simple ray optics. Now if $\lambda/d$ is small, $\sin \theta$ must be small as well, and the spacing between the diffraction fringes will be small as well. Although we have not proven it, the central fringe is always the brightest, and the fringes get dimmer and dimmer as we go farther from it. For small values of $\lambda/d$, the part of the diffraction pattern that is bright enough to be detectable covers only a small range of angles. This is exactly what we would expect from ray optics: the rays passing through the two slits would remain parallel, and would continue moving in the $\theta = 0$ direction. (In fact there would be images of the two separate slits on the screen, but our analysis was all in terms of angles, so we should not expect it to address the issue of whether there is structure within a set of rays that are all traveling in the $\theta = 0$ direction.)

Spacing of the fringes at small angles  

Example 14

At small angles, we can use the approximation $\sin \theta \approx \theta$, which is valid if $\theta$ is measured in radians. The equation for double-slit diffraction becomes simply

$$\frac{\lambda}{d} = \frac{\theta}{m},$$

which can be solved for $\theta$ to give

$$\theta = \frac{m\lambda}{d}.$$

The difference in angle between successive fringes is the change in $\theta$ that results from changing $m$ by plus or minus one,

$$\Delta \theta = \frac{\lambda}{d}.$$

For example, if we write $\theta_7$ for the angle of the seventh bright fringe on one side of the central maximum and $\theta_8$ for the neighboring one, we have

$$\theta_8 - \theta_7 = \frac{8\lambda}{d} - \frac{7\lambda}{d} = \frac{\lambda}{d},$$

and similarly for any other neighboring pair of fringes.
Although the equation $\lambda/d = \sin \theta/m$ is only valid for a double slit, it is still a guide to our thinking even if we are observing diffraction of light by a virus or a flea’s leg: it is always true that

(1) large values of $\lambda/d$ lead to a broad diffraction pattern, and

(2) diffraction patterns are repetitive.

In many cases the equation looks just like $\lambda/d = \sin \theta/m$ but with an extra numerical factor thrown in, and with $d$ interpreted as some other dimension of the object, e.g., the diameter of a piece of wire.

12.5.6 Repetition

Suppose we replace a double slit with a triple slit, $s$. We can think of this as a third repetition of the structures that were present in the double slit. Will this device be an improvement over the double slit for any practical reasons?

The answer is yes, as can be shown using figure $u$. For ease of visualization, I have violated our usual rule of only considering points very far from the diffracting object. The scale of the drawing is such that a wavelength is one cm. In $u/1$, all three waves travel an integer number of wavelengths to reach the same point, so there is a bright central spot, as we would expect from our experience with the double slit. In figure $u/2$, we show the path lengths to a new point. This point is farther from slit $A$ by a quarter of a wavelength, and correspondingly closer to slit $C$. The distance from slit $B$ has hardly changed at all. Because the path lengths traveled from slits $A$ and $C$ differ by half a wavelength, there will be perfect destructive interference between these two waves. There is still some uncanceled wave intensity because of slit $B$, but the amplitude will be three times less than in figure $u/1$, resulting in a factor of 9 decrease in brightness. Thus, by moving off to the right a little, we have gone from the bright central maximum to a point that is quite dark.

Now let’s compare with what would have happened if slit $C$ had been covered, creating a plain old double slit. The waves coming from slits $A$ and $B$ would have been out of phase by 0.23 wavelengths, but this would not have caused very severe interference. The point in figure $u/2$ would have been quite brightly lit up.

To summarize, we have found that adding a third slit narrows down the central fringe dramatically. The same is true for all the other fringes as well, and since the same amount of energy is con-
Single-slit diffraction of water waves. Note the double width of the central maximum.

A pretty good simulation of the single-slit pattern of figure v, made by using three motors to produce overlapping ripples from three neighboring points in the water.

u / 1. There is a bright central maximum. 2. At this point just off the central maximum, the path lengths traveled by the three waves have changed.

centrated in narrower diffraction fringes, each fringe is brighter and easier to see, t.

This is an example of a more general fact about diffraction: if some feature of the diffracting object is repeated, the locations of the maxima and minima are unchanged, but they become narrower.

Taking this reasoning to its logical conclusion, a diffracting object with thousands of slits would produce extremely narrow fringes. Such an object is called a diffraction grating.

12.5.7 Single-slit diffraction

If we use only a single slit, is there diffraction? If the slit is not wide compared to a wavelength of light, then we can approximate its behavior by using only a single set of Huygens ripples. There are no other sets of ripples to add to it, so there are no constructive or destructive interference effects, and no maxima or minima. The result will be a uniform spherical wave of light spreading out in all directions, like what we would expect from a tiny lightbulb. We could call this a diffraction pattern, but it is a completely featureless one, and it could not be used, for instance, to determine the wavelength of the light, as other diffraction patterns could.

All of this, however, assumes that the slit is narrow compared to a wavelength of light. If, on the other hand, the slit is broader, there will indeed be interference among the sets of ripples spreading out from various points along the opening. Figure v shows an example with water waves, and figure w with light.

self-check I

How does the wavelength of the waves compare with the width of the slit in figure v?  

We will not go into the details of the analysis of single-slit diffraction, but let us see how its properties can be related to the general
things we’ve learned about diffraction. We know based on scaling arguments that the angular sizes of features in the diffraction pattern must be related to the wavelength and the width, $a$, of the slit by some relationship of the form

$$\frac{\lambda}{a} \leftrightarrow \theta.$$  

This is indeed true, and for instance the angle between the maximum of the central fringe and the maximum of the next fringe on one side equals $1.5\lambda/a$. Scaling arguments will never produce factors such as the 1.5, but they tell us that the answer must involve $\lambda/a$, so all the familiar qualitative facts are true. For instance, shorter-wavelength light will produce a more closely spaced diffraction pattern.

An important scientific example of single-slit diffraction is in telescopes. Images of individual stars, as in figure y, are a good way to examine diffraction effects, because all stars except the sun are so far away that no telescope, even at the highest magnification, can image their disks or surface features. Thus any features of a star’s image must be due purely to optical effects such as diffraction. A prominent cross appears around the brightest star, and dimmer ones surround the dimmer stars. Something like this is seen in most telescope photos, and indicates that inside the tube of the telescope there were two perpendicular struts or supports. Light diffracted around these struts. You might think that diffraction could be eliminated entirely by getting rid of all obstructions in the tube, but the circles around the stars are diffraction effects arising from single-slit diffraction at the mouth of the telescope’s tube! (Actually we have not even talked about diffraction through a circular opening, but the idea is the same.) Since the angular sizes of the diffracted images depend on $\lambda/a$, the only way to improve the resolution of the images is to increase the diameter, $a$, of the tube. This is one of the main reasons (in addition to light-gathering power) why the best telescopes must be very large in diameter.

**Self-check J**

What would this imply about radio telescopes as compared with visible-light telescopes?

Answer, p. 932

Double-slit diffraction is easier to understand conceptually than single-slit diffraction, but if you do a double-slit diffraction experiment in real life, you are likely to encounter a complicated pattern like figure aa/1, rather than the simpler one, 2, you were expecting. This is because the slits are fairly big compared to the wavelength of the light being used. We really have two different distances in our pair of slits: $d$, the distance between the slits, and $w$, the width of each slit. Remember that smaller distances on the object the light diffracts around correspond to larger features of the diffraction pattern. The pattern 1 thus has two spacings in it: a short spac-
ing corresponding to the large distance $d$, and a long spacing that relates to the small dimension $w$.

1. A diffraction pattern formed by a real double slit. The width of each slit is fairly big compared to the wavelength of the light. This is a real photo. 2. This idealized pattern is not likely to occur in real life. To get it, you would need each slit to be so narrow that its width was comparable to the wavelength of the light, but that’s not usually possible. This is not a real photo. 3. A real photo of a single-slit diffraction pattern caused by a slit whose width is the same as the widths of the slits used to make the top pattern.

**Discussion Question**

A. Why is it optically impossible for bacteria to evolve eyes that use visible light to form images?

12.5.8  **The principle of least time**

In subsection 12.1.5 and 12.4.5, we saw how in the ray model of light, both refraction and reflection can be described in an elegant and beautiful way by a single principle, the principle of least time. We can now justify the principle of least time based on the wave model of light. Consider an example involving reflection, ab. Starting at point A, Huygens' principle for waves tells us that we can think of the wave as spreading out in all directions. Suppose we imagine all the possible ways that a ray could travel from A to B. We show this by drawing 25 possible paths, of which the central one is the shortest. Since the principle of least time connects the wave model to the ray model, we should expect to get the most accurate results when the wavelength is much shorter than the distances involved — for the sake of this numerical example, let’s say that a wavelength is $1/10$ of the shortest reflected path from A to B. The table, 2, shows the distances traveled by the 25 rays.

Note how similar are the distances traveled by the group of 7 rays, indicated with a bracket, that come closest to obeying the principle of least time. If we think of each one as a wave, then all 7 are again nearly in phase at point B. However, the rays that are farther from satisfying the principle of least time show more
Light could take many different paths from A to B.

rapidly changing distances; on reuniting at point B, their phases are a random jumble, and they will very nearly cancel each other out. Thus, almost none of the wave energy delivered to point B goes by these longer paths. Physically we find, for instance, that a wave pulse emitted at A is observed at B after a time interval corresponding very nearly to the shortest possible path, and the pulse is not very “smeared out” when it gets there. The shorter the wavelength compared to the dimensions of the figure, the more accurate these approximate statements become.

Instead of drawing a finite number of rays, such 25, what happens if we think of the angle, \( \theta \), of emission of the ray as a continuously varying variable? Minimizing the distance \( L \) requires

\[
\frac{dL}{d\theta} = 0.
\]

Because \( L \) is changing slowly in the vicinity of the angle that satisfies the principle of least time, all the rays that come out close to this angle have very nearly the same \( L \), and remain very nearly in phase when they reach B. This is the basic reason why the discrete table, ab/2, turned out to have a group of rays that all traveled nearly the same distance.

As discussed in subsection 12.1.5, the principle of least time is really a principle of least or greatest time. This makes perfect sense, since \( dL/d\theta = 0 \) can in general describe either a minimum or a maximum.

The principle of least time is very general. It does not apply just to refraction and reflection — it can even be used to prove that light rays travel in a straight line through empty space, without taking detours! This general approach to wave motion was used by Richard Feynman, one of the pioneers who in the 1950’s reconciled quantum mechanics with relativity. A very readable explanation is given in a book Feynman wrote for laypeople, QED: The Strange Theory of Light and Matter.
Problems

The symbols $\sqrt{}$, $\equiv$, etc. are explained on page 814.

1. Draw a ray diagram showing why a small light source (a candle, say) produces sharper shadows than a large one (e.g., a long fluorescent bulb).

2. A Global Positioning System (GPS) receiver is a device that lets you figure out where you are by receiving timed radio signals from satellites. It works by measuring the travel time for the signals, which is related to the distance between you and the satellite. By finding the ranges to several different satellites in this way, it can pin down your location in three dimensions to within a few meters. How accurate does the measurement of the time delay have to be to determine your position to this accuracy?

3. Estimate the frequency of an electromagnetic wave whose wavelength is similar in size to an atom (about a nm). Referring back to figure o on p. 707, in what part of the electromagnetic spectrum would such a wave lie (infrared, gamma-rays, . . .)?

4. The Stealth bomber is designed with flat, smooth surfaces. Why would this make it difficult to detect via radar?

5. The natives of planet Wumpus play pool using light rays on an eleven-sided table with mirrors for bumpers, shown in the figure on the next page. Trace this shot accurately with a ruler to reveal the hidden message. To get good enough accuracy, you’ll need to photocopy the page (or download the book and print the page) and construct each reflection using a protractor.

Problem 5.
6. The figure on the next page shows a curved (parabolic) mirror, with three parallel light rays coming toward it. One ray is approaching along the mirror’s center line. (a) Trace the drawing accurately, and continue the light rays until they are about to undergo their second reflection. To get good enough accuracy, you’ll need to photocopy the page (or download the book and print the page) and draw in the normal at each place where a ray is reflected. What do you notice? (b) Make up an example of a practical use for this device. (c) How could you use this mirror with a small lightbulb to produce a parallel beam of light rays going off to the right?

Problem 6.

7. A man is walking at 1.0 m/s directly towards a flat mirror. At what speed is his separation from his image decreasing?

8. If a mirror on a wall is only big enough for you to see yourself from your head down to your waist, can you see your entire body by backing up? Test this experimentally and come up with an explanation for your observations, including a ray diagram.

Note that when you do the experiment, it’s easy to confuse yourself if the mirror is even a tiny bit off of vertical. One way to check yourself is to artificially lower the top of the mirror by putting a piece of tape or a post-it note where it blocks your view of the top of your head. You can then check whether you are able to see more of yourself both above and below by backing up.