Dispersion of white light by a prism. White light is a mixture of all the wavelengths of the visible spectrum. Waves of different wavelengths undergo different amounts of refraction.

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 (section 30.3 and problem 2). As we’ll see in section 35.2, 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.
31.5 ★ 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 un-bent 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 9, p. 908).

31.6 ★ Case study: the eye of the jumping spider

Figure 5 shows an exceptionally cute jumping spider. The jumping spider does not build a web. It stalks its prey like a cat, so it needs excellent eyesight. In some ways, its visual system is more sophisticated and more functional than that of a human, illustrating how evolution does not progress systematically toward “higher” forms of life.
One way in which the spider outdoes us is that it has eight eyes to our two. (Each eye is simple, not compound like that of a fly.) The reason this works well has to do with the trade-off between magnification and field of view. The elongated principal eyes at the front of the head have a large value of $d_i$, resulting in a large magnification $M = d_i/d_o$. This high magnification is used for sophisticated visual tasks like distinguishing prey from a potential mate. (The pretty stripes on the legs in the photo are probably evolved to aid in making this distinction, which is a crucial one on a Saturday night.) As always with a high magnification, this results in a reduction in the field of view: making the image bigger means reducing the amount of the potential image that can actually fit on the retina. The animal has tunnel vision in these forward eyes. To allow it to glimpse prey from other angles, it has the additional eyes on the sides of its head. These are not elongated, and the smaller $d_i$ gives a smaller magnification but a larger field of view. When the spider sees something moving in these eyes, it turns its body so that it can take a look with the front eyes. The tiniest pair of eyes are too small to be useful. These vestigial organs, like the maladaptive human appendix, are an example of the tendency of evolution to produce unfortunate accidents due to the lack of intelligent design. The use of multiple eyes for these multiple purposes is far superior to the two-eye arrangement found in humans, octopuses, etc., especially because of its compactness. If the spider had only two spherical eyes, they would have to have the same front-to-back dimension in order to produce the same acuity, but then the eyes would take up nearly all of the front of the head.

Another beautiful feature of these eyes is that they will never need bifocals. A human eye uses muscles to adjust for seeing near and far, varying $f$ in order to achieve a fixed $d_i$ for differing values of $d_o$. On older models of $H. sap.$, this poorly engineered feature is usually one of the first things to break down. The spider’s front eyes have muscles, like a human’s, that rotate the tube, but none that vary $f$, which is fixed. However, the retina consists of four separate layers at slightly different values of $d_i$. The figure only shows the detailed cellular structure of the rearmost layer, which is the most acute. Depending on $d_o$, the image may lie closest to any one of the four layers, and the spider can then use that layer to get a well-focused view. The layering is also believed to help eliminate problems caused by the variation of the index of refraction with wavelength (cf. problem 2, p. 906).

Although the spider’s eye is different in many ways from a human’s or an octopus’s, it shares the same fundamental construction, being essentially a lens that forms a real image on a screen inside a darkened chamber. From this perspective, the main difference is simply the scale, which is miniaturized by about a factor of $10^2$ in the linear dimensions. How far down can this scaling go? Does an
amoeba or a white blood cell lack an eye merely because it doesn’t have a nervous system that could make sense of the signals? In fact there is an optical limit on the miniaturization of any eye or camera. The spider’s eye is already so small that on the scale of the bottom panel in figure s, one wavelength of visible light would be easily distinguishable — about the length of the comma in this sentence. Chapter 32 is about optical effects that occur when the wave nature of light is important, and problem 14 on p. 932 specifically addresses the effect on this spider’s vision.
Summary

Selected vocabulary

refraction . . . . . . . the change in direction that occurs when a wave encounters the interface between two media.

index of refraction . . . . . . . an optical property of matter; the speed of light in a vacuum divided by the speed of light in the substance in question.

Notation

\( n \) . . . . . . . . . . . the index of refraction

Summary

Refraction is a change in direction that occurs when a wave encounters the interface between two media. Together, refraction and reflection account for the basic principles behind nearly all optical devices.

Snell discovered the equation for refraction,

\[ n_1 \sin \theta_1 = n_2 \sin \theta_2, \]

[angles measured with respect to the normal]

through experiments with light rays, long before light was proven to be a wave. Snell’s law can be proven based on the geometrical behavior of waves. Here \( n \) is the index of refraction. Snell invented this quantity to describe the refractive properties of various substances, but it was later found to be related to the speed of light in the substance,

\[ n = \frac{c}{v}, \]

where \( c \) is the speed of light in a vacuum. In general a material’s index of refraction is different for different wavelengths of light.

As discussed in chapter 20, any wave is partially transmitted and partially reflected at the boundary between two media in which its speeds are different. It is not particularly important to know the equation that tells what fraction is transmitted (and thus refracted), but important technologies such as fiber optics are based on the fact that this fraction becomes zero for sufficiently oblique angles. This phenomenon is referred to as total internal reflection. It occurs when there is no angle that satisfies Snell’s law.
Problems

Key
✓ A computerized answer check is available online.
∫ A problem that requires calculus.
★ A difficult problem.

1 Suppose a converging lens is constructed of a type of plastic whose index of refraction is less than that of water. How will the lens’s behavior be different if it is placed underwater?

▷ Solution, p. 1038

2 There are two main types of telescopes, refracting (using a lens) and reflecting (using a mirror as in figure i on p. 862). (Some telescopes use a mixture of the two types of elements: the light first encounters a large curved mirror, and then goes through an eyepiece that is a lens. To keep things simple, assume no eyepiece is used.) What implications would the color-dependence of focal length have for the relative merits of the two types of telescopes? Describe the case where an image is formed of a white star. You may find it helpful to draw a ray diagram.

3 Based on Snell’s law, explain why rays of light passing through the edges of a converging lens are bent more than rays passing through parts closer to the center. It might seem like it should be the other way around, since the rays at the edge pass through less glass — shouldn’t they be affected less? In your answer:

• Include a ray diagram showing a huge, full-page, close-up view of the relevant part of the lens.

• Make use of the fact that the front and back surfaces aren’t always parallel; a lens in which the front and back surfaces are always parallel doesn’t focus light at all, so if your explanation doesn’t make use of this fact, your argument must be incorrect.

• Make sure your argument still works even if the rays don’t come in parallel to the axis or from a point on the axis.
4 When you take pictures with a camera, the distance between the lens and the film or chip has to be adjusted, depending on the distance at which you want to focus. This is done by moving the lens. If you want to change your focus so that you can take a picture of something farther away, which way do you have to move the lens? Explain using ray diagrams. [Based on a problem by Eric Mazur.]

5 (a) Light is being reflected diffusely from an object 1.000 m underwater. The light that comes up to the surface is refracted at the water-air interface. If the refracted rays all appear to come from the same point, then there will be a virtual image of the object in the water, above the object’s actual position, which will be visible to an observer above the water. Consider three rays, A, B and C, whose angles in the water with respect to the normal are $\theta_i = 0.000^\circ$, $1.000^\circ$ and $20.000^\circ$ respectively. Find the depth of the point at which the refracted parts of A and B appear to have intersected, and do the same for A and C. Show that the intersections are at nearly the same depth, but not quite. [Check: The difference in depth should be about 4 cm.]

(b) Since all the refracted rays do not quite appear to have come from the same point, this is technically not a virtual image. In practical terms, what effect would this have on what you see? 

(c) In the case where the angles are all small, use algebra and trig to show that the refracted rays do appear to come from the same point, and find an equation for the depth of the virtual image. Do not put in any numerical values for the angles or for the indices of refraction — just keep them as symbols. You will need the approximation $\sin \theta \approx \tan \theta \approx \theta$, which is valid for small angles measured in radians.

6 The drawing shows the anatomy of the human eye, at twice life size. Find the radius of curvature of the outer surface of the cornea by measurements on the figure, and then derive the focal length of the air-cornea interface, where almost all the focusing of light occurs. You will need to use physical reasoning to modify the lensmaker’s equation for the case where there is only a single refracting surface. Assume that the index of refraction of the cornea is essentially that of water.

7 When swimming underwater, why is your vision made much clearer by wearing goggles with flat pieces of glass that trap air behind them? [Hint: You can simplify your reasoning by considering the special case where you are looking at an object far away, and along the optic axis of the eye.]
Problem 8. The figure shows four lenses. Lens 1 has two spherical surfaces. Lens 2 is the same as lens 1 but turned around. Lens 3 is made by cutting through lens 1 and turning the bottom around. Lens 4 is made by cutting a central circle out of lens 1 and recessing it.

(a) A parallel beam of light enters lens 1 from the left, parallel to its axis. Reasoning based on Snell’s law, will the beam emerging from the lens be bent inward, or outward, or will it remain parallel to the axis? Explain your reasoning. As part of your answer, make a huge drawing of one small part of the lens, and apply Snell’s law at both interfaces. Recall that rays are bent more if they come to the interface at a larger angle with respect to the normal.

(b) What will happen with lenses 2, 3, and 4? Explain. Drawings are not necessary.

Solution, p. 1040

9 Prove that the principle of least time leads to Snell’s law.

10 An object is more than one focal length from a converging lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in chapter 30, determine the positive and negative signs in the equation \(1/f = \pm 1/d_i \pm 1/d_o\). (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 80 cm from the rose, locate the image.

11 An object is less than one focal length from a converging lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in chapter 30, determine the positive and negative signs in the equation \(1/f = \pm 1/d_i \pm 1/d_o\). (c) The images of the rose in section 4.2 were made using a lens with a focal length of 23 cm. If the lens is placed 10 cm from the rose, locate the image.

12 Nearsighted people wear glasses whose lenses are diverging. (a) Draw a ray diagram. For simplicity pretend that there is no eye behind the glasses. (b) Using reasoning like that developed in chapter 30, determine the positive and negative signs in the equation \(1/f = \pm 1/d_i \pm 1/d_o\). (c) If the focal length of the lens is 50.0 cm, and the person is looking at an object at a distance of 80.0 cm, locate the image.
Problem 13

Two standard focal lengths for camera lenses are 50 mm (standard) and 28 mm (wide-angle). To see how the focal lengths relate to the angular size of the field of view, it is helpful to visualize things as represented in the figure. Instead of showing many rays coming from the same point on the same object, as we normally do, the figure shows two rays from two different objects. Although the lens will intercept infinitely many rays from each of these points, we have shown only the ones that pass through the center of the lens, so that they suffer no angular deflection. (Any angular deflection at the front surface of the lens is canceled by an opposite deflection at the back, since the front and back surfaces are parallel at the lens’s center.) What is special about these two rays is that they are aimed at the edges of one 35-mm-wide frame of film; that is, they show the limits of the field of view. Throughout this problem, we assume that $d_o$ is much greater than $d_i$. (a) Compute the angular width of the camera’s field of view when these two lenses are used. (b) Use small-angle approximations to find a simplified equation for the angular width of the field of view, $\theta$, in terms of the focal length, $f$, and the width of the film, $w$. Your equation should not have any trig functions in it. Compare the results of this approximation with your answers from part a. (c) Suppose that we are holding constant the aperture (amount of surface area of the lens being used to collect light). When switching from a 50-mm lens to a 28-mm lens, how many times longer or shorter must the exposure be in order to make a properly developed picture, i.e., one that is not under- or overexposed? [Based on a problem by Arnold Arons.]
> Solution, p. 1040

Problem 13.
15 Diamond has an index of refraction of 2.42, and part of the reason diamonds sparkle is that this encourages a light ray to undergo many total internal reflections before it emerges. (a) Calculate the critical angle at which total internal reflection occurs in diamond. (b) Explain the interpretation of your result: Is it measured from the normal, or from the surface? Is it a minimum angle for total internal reflection, or is it a maximum? How would the critical angle have been different for a substance such as glass or plastic, with a lower index of refraction?

16 Fred’s eyes are able to focus on things as close as 5.0 cm. Fred holds a magnifying glass with a focal length of 3.0 cm at a height of 2.0 cm above a flatworm. (a) Locate the image, and find the magnification. (b) Without the magnifying glass, from what distance would Fred want to view the flatworm to see its details as well as possible? With the magnifying glass? (c) Compute the angular magnification.

17 This problem has been deleted.

18 It would be annoying if your eyeglasses produced a magnified or reduced image. Prove that when the eye is very close to a lens, and the lens produces a virtual image, the angular magnification is always approximately equal to 1 (regardless of whether the lens is diverging or converging).

19 A typical mirror consists of a pane of glass of thickness $t$ and index of refraction $n$, “silvered” on the back with a reflective coating. Let $d_o$ and $d_i$ be measured from the back of the mirror. Show that $d_i = d_o - 2(1 - 1/n)t$. Use the result of, and make the approximation employed in, problem 5c. As a check on your result, consider separately the special values of $n$ and $t$ that would recover the case without any glass.

20 The figure shows a lens with surfaces that are curved, but whose thickness is constant along any horizontal line. Use the lensmaker’s equation to prove that this “lens” is not really a lens at all.

21 Estimate the radii of curvature of the two optical surfaces in the eye of the jumping spider in in figure s on p. 902. Use physical reasoning to modify the lensmaker’s equation for a case like this one, in which there are three indices of refraction $n_1$ (air), $n_2$ (lens), and $n_3$ (the material behind the lens), and $n_1 \neq n_3$. Show that the interface between $n_2$ and $n_3$ contributes negligibly to focusing, and verify that the image is produced at approximately the right place in the eye when the object is far away. As a check on your result, direct optical measurements by M.F. Land in 1969 gave $f = 512 \mu m$. 

\[\sqrt{\text{Solution, p. 1041}}\]
Zahra likes to play practical jokes on the friends she goes hiking with. One night, by a blazing camp fire, she stealthily uses a lens of focal length $f$ to gather light from the fire and make a hot spot on Becky’s neck. (a) Using the method of section 30.2, p. 877, draw a ray diagram and set up the equation for the image location, inferring the correct plus and minus signs from the diagram. (b) Let $A$ be the distance from the lens to the campfire, and $B$ the distance from the lens to Becky’s neck. Consider the following nine possibilities:

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt; f$</td>
<td>$&lt; f$</td>
</tr>
<tr>
<td>$= f$</td>
<td></td>
</tr>
<tr>
<td>$&gt; f$</td>
<td></td>
</tr>
</tbody>
</table>

By reasoning about your equation from part a, determine which of these are possible and which are not. Solution, p. 1041
Exercise 31: How strong are your glasses?

This exercise was created by Dan MacIsaac.

Equipment:

- eyeglasses
- diverging lenses for students who don’t wear glasses, or who use converging glasses
- rulers and metersticks
- scratch paper
- marking pens

Most people who wear glasses have glasses whose lenses are diverging, which allows them to focus on objects far away. Such a lens cannot form a real image, so its focal length cannot be measured as easily as that of an converging lens. In this exercise you will determine the focal length of your own glasses by taking them off, holding them at a distance from your face, and looking through them at a set of parallel lines on a piece of paper. The lines will be reduced (the lens’s magnification is less than one), and by adjusting the distance between the lens and the paper, you can make the magnification equal 1/2 exactly, so that two spaces between lines as seen through the lens fit into one space as seen simultaneously to the side of the lens. This object distance can be used in order to find the focal length of the lens.

1. Does this technique really measure magnification or does it measure angular magnification? What can you do in your experiment in order to make these two quantities nearly the same, so the math is simpler?

2. Before taking any numerical data, use algebra to find the focal length of the lens in terms of $d_o$, the object distance that results in a magnification of 1/2.

3. Use a marker to draw three evenly spaced parallel lines on the paper. (A spacing of a few cm works well.) Measure the object distance that results in a magnification of 1/2, and determine the focal length of your lens.
Chapter 32
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.

32.1 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 light-bulb 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?

32.2 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 section 1.2 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 chapter 20 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 $\lambda/d$, 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 $\lambda/d$ will be unchanged.

32.3 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 theory 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 $\lambda/d$ 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 $\lambda/d$ 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 too tiny to be visible.)
32.4 Huygens’ principle

Returning to the example of double-slit diffraction, note the strong visual impression of two overlapping sets of concentric semicircles. 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, which then spread out as circular ripples, 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. 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. 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 superposition, 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.

### 32.5 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 is 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 B**

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 object that diffracts it, so the triangle is long and skinny. Most real-world examples with diffraction of light, in fact, would have triangles with even skinnier proportions than this one. The two long sides are therefore very nearly parallel, and we are justified in drawing the right triangle shown in figure \( o \), labeling one leg of the right triangle as the difference in path length, \( L - L' \), and labeling the acute angle as \( \theta \). (In reality this angle is a tiny bit greater than the one labeled \( \theta \) in figure \( n \).)

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,

\[ L - L' = m\lambda. \quad \text{[condition for a maximum; } m \text{ is an integer]} \]

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 \( m\lambda/d = \sin \theta \), or

\[ \frac{\lambda}{d} = \frac{\sin \theta}{m}. \quad \text{[condition for a maximum; } m \text{ is an integer]} \]

Similarly, the condition for a minimum is

\[ \frac{\lambda}{d} = \frac{\sin \theta}{m + \frac{1}{2}}. \quad \text{[condition for a minimum; } m \text{ is an integer plus} \ 1/2] \]

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.

---

**Example 1**

**Double-slit diffraction of blue and red light**

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

---

**Example 2**

**The correspondence principle**

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.)

Spacial of the fringes at small angles
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 can still be 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.
32.6 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 paths 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 concentrated in narrower diffraction fringes, each fringe is brighter and easier to
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.

### 32.7 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 C**

How does the wavelength of the waves compare with the width of the slit in figure v?  
▷ Answer, p. 1046

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 D**

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

Answer, p. 1046

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 spacing corresponding to the large distance $d$, and a long spacing that relates to the small dimension $w$.

**Discussion question**

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

### 32.8 The principle of least time

In section 28.5 and 31.5, 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
aa / 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.

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 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 as 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 section 28.5, 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.
Summary

Selected vocabulary

- **diffraction** . . . . 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.

- **coherent** . . . . a light wave whose parts are all in phase with each other

Other terminology and notation

- **wavelets** . . . . the ripples in Huygens’ principle

Summary

Wave optics is a more general theory of light than ray optics. When light interacts with material objects that are much larger than one wavelength of the light, the ray model of light is approximately correct, but in other cases the wave model is required.

Huygens’ principle states that, given a wavefront at one moment in time, the future behavior of the wave can be found by breaking the wavefront up into a large number of small, side-by-side wave peaks, each of which then creates a pattern of circular or spherical ripples. As these sets of ripples add together, the wave evolves and moves through space. Since Huygens’ principle is a purely geometrical construction, diffraction effects obey a simple scaling rule: the behavior is unchanged if the wavelength and the dimensions of the diffracting objects are both scaled up or down by the same factor. If we wish to predict the angles at which various features of the diffraction pattern radiate out, scaling requires that these angles depend only on the unitless ratio $\lambda/d$, where $d$ is the size of some feature of the diffracting object.

Double-slit diffraction is easily analyzed using Huygens’ principle if the slits are narrower than one wavelength. We need only construct two sets of ripples, one spreading out from each slit. The angles of the maxima (brightest points in the bright fringes) and minima (darkest points in the dark fringes) are given by the equation

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

where $d$ is the center-to-center spacing of the slits, and $m$ is an integer at a maximum or an integer plus 1/2 at a minimum.

If some feature of a diffracting object is repeated, the diffraction fringes remain in the same places, but become narrower with each repetition. By repeating a double-slit pattern hundreds or thousands of times, we obtain a diffraction grating.

A single slit can produce diffraction fringes if it is larger than one
wavelength. Many practical instances of diffraction can be interpreted as single-slit diffraction, e.g., diffraction in telescopes. The main thing to realize about single-slit diffraction is that it exhibits the same kind of relationship between $\lambda$, $d$, and angles of fringes as in any other type of diffraction.
Problems

Key
✓ A computerized answer check is available online.
∫ A problem that requires calculus.
* A difficult problem.

1 Why would blue or violet light be the best for microscopy?
   Solution, p. 1042

2 Match gratings A-C with the diffraction patterns 1-3 that they produce. Explain.

3 The beam of a laser passes through a diffraction grating, fans out, and illuminates a wall that is perpendicular to the original beam, lying at a distance of 2.0 m from the grating. The beam is produced by a helium-neon laser, and has a wavelength of 694.3 nm. The grating has 2000 lines per centimeter. (a) What is the distance on the wall between the central maximum and the maxima immediately to its right and left? (b) How much does your answer change when you use the small-angle approximations \( \theta \approx \sin \theta \approx \tan \theta \)?

4 When white light passes through a diffraction grating, what is the smallest value of \( m \) for which the visible spectrum of order \( m \) overlaps the next one, of order \( m + 1 \)? (The visible spectrum runs from about 400 nm to about 700 nm.)
5. Ultrasound, i.e., sound waves with frequencies too high to be audible, can be used for imaging fetuses in the womb or for breaking up kidney stones so that they can be eliminated by the body. Consider the latter application. Lenses can be built to focus sound waves, but because the wavelength of the sound is not all that small compared to the diameter of the lens, the sound will not be concentrated exactly at the geometrical focal point. Instead, a diffraction pattern will be created with an intense central spot surrounded by fainter rings. About 85% of the power is concentrated within the central spot. The angle of the first minimum (surrounding the central spot) is given by \( \sin \theta = \lambda/b \), where \( b \) is the diameter of the lens. This is similar to the corresponding equation for a single slit, but with a factor of 1.22 in front which arises from the circular shape of the aperture. Let the distance from the lens to the patient’s kidney stone be \( L = 20 \text{ cm} \). You will want \( f > 20 \text{ kHz} \), so that the sound is inaudible. Find values of \( b \) and \( f \) that would result in a usable design, where the central spot is small enough to lie within a kidney stone 1 cm in diameter.

6. For star images such as the ones in figure y, estimate the angular width of the diffraction spot due to diffraction at the mouth of the telescope. Assume a telescope with a diameter of 10 meters (the largest currently in existence), and light with a wavelength in the middle of the visible range. Compare with the actual angular size of a star of diameter \( 10^9 \text{ m} \) seen from a distance of \( 10^{17} \text{ m} \). What does this tell you?  
   Solution, p. 1042

7. Under what circumstances could one get a mathematically undefined result by solving the double-slit diffraction equation for \( \theta \)? Give a physical interpretation of what would actually be observed.  
   Solution, p. 1042

8. When ultrasound is used for medical imaging, the frequency may be as high as 5-20 MHz. Another medical application of ultrasound is for therapeutic heating of tissues inside the body; here, the frequency is typically 1-3 MHz. What fundamental physical reasons could you suggest for the use of higher frequencies for imaging?
9 The figure below shows two diffraction patterns, both made with the same wavelength of red light. (a) What type of slits made the patterns? Is it a single slit, double slits, or something else? Explain. (b) Compare the dimensions of the slits used to make the top and bottom pattern. Give a numerical ratio, and state which way the ratio is, i.e., which slit pattern was the larger one. Explain.

10 The figure below shows two diffraction patterns. The top one was made with yellow light, and the bottom one with red. Could the slits used to make the two patterns have been the same?
11. The figure below shows three diffraction patterns. All were made under identical conditions, except that a different set of double slits was used for each one. The slits used to make the top pattern had a center-to-center separation \( d = 0.50 \text{ mm} \), and each slit was \( w = 0.04 \text{ mm} \) wide. (a) Determine \( d \) and \( w \) for the slits used to make the pattern in the middle. (b) Do the same for the slits used to make the bottom pattern.

![Diffraction Patterns](image)

12. The figure shows a diffraction pattern made by a double slit, along with an image of a meter stick to show the scale. The slits were 146 cm away from the screen on which the diffraction pattern was projected. The spacing of the slits was 0.050 mm. What was the wavelength of the light?

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

13. The figure shows a diffraction pattern made by a double slit, along with an image of a meter stick to show the scale. Sketch the diffraction pattern from the figure on your paper. Now consider the four variables in the equation \( \lambda/d = \sin \theta/m \). Which of these are the same for all five fringes, and which are different for each fringe? Which variable would you naturally use in order to label which fringe was which? Label the fringes on your sketch using the values of that variable.

14. Figure s on p. 902 shows the anatomy of a jumping spider’s principal eye. The smallest feature the spider can distinguish is limited by the size of the receptor cells in its retina. (a) By making measurements on the diagram, estimate this limiting angular size in units of minutes of arc (60 minutes = 1 degree). (b) Show that this is greater than, but roughly in the same ballpark as, the limit imposed by diffraction for visible light.

Remark: Evolution is a scientific theory that makes testable predictions, and if observations contradict its predictions, the theory can be disproved. It would be maladaptive for the spider to have retinal receptor cells with sizes much less than the limit imposed by diffraction, since it would increase complexity without giving any improvement in visual acuity. The results of this problem confirm that, as predicted by Darwinian evolution, this is not the case. Work by M.F. Land in 1969 shows that in this spider’s eye, aberration is a somewhat bigger effect than diffraction, so that the size of the receptors is very nearly at

\[ \frac{\lambda}{d} \]
an evolutionary optimum. √ √
Exercise 32A: Double-source interference

1. Two sources separated by a distance $d = 2$ cm make circular ripples with a wavelength of $\lambda = 1$ cm. On a piece of paper, make a life-size drawing of the two sources in the default setup, and locate the following points:

   A. The point that is 10 wavelengths from source #1 and 10 wavelengths from source #2.
   B. The point that is 10.5 wavelengths from #1 and 10.5 from #2.
   C. The point that is 11 wavelengths from #1 and 11 from #2.
   D. The point that is 10 wavelengths from #1 and 10.5 from #2.
   E. The point that is 11 wavelengths from #1 and 11.5 from #2.
   F. The point that is 10 wavelengths from #1 and 11 from #2.
   G. The point that is 11 wavelengths from #1 and 12 from #2.

   You can do this either using a compass or by putting the next page under your paper and tracing. It is not necessary to trace all the arcs completely, and doing so is unnecessarily time-consuming; you can fairly easily estimate where these points would lie, and just trace arcs long enough to find the relevant intersections.

   What do these points correspond to in the real wave pattern?

2. Make a fresh copy of your drawing, showing only point F and the two sources, which form a long, skinny triangle. Now suppose you were to change the setup by doubling $d$, while leaving $\lambda$ the same. It’s easiest to understand what’s happening on the drawing if you move both sources outward, keeping the center fixed. Based on your drawing, what will happen to the position of point F when you double $d$? Measure its angle with a protractor.

3. What would happen if you doubled both $\lambda$ and $d$ compared to the standard setup?

4. Combining the ideas from parts 2 and 3, what do you think would happen to your angles if, starting from the standard setup, you doubled $\lambda$ while leaving $d$ the same?

5. Suppose $\lambda$ was a millionth of a centimeter, while $d$ was still as in the standard setup. What would happen to the angles? What does this tell you about observing diffraction of light?
Exercise 32B: Single-slit diffraction

Equipment:

- rulers
- computer with web browser

The following page is a diagram of a single slit and a screen onto which its diffraction pattern is projected. The class will make a numerical prediction of the intensity of the pattern at the different points on the screen. Each group will be responsible for calculating the intensity at one of the points. (Either 11 groups or six will work nicely – in the latter case, only points a, c, e, g, i, and k are used.) The idea is to break up the wavefront in the mouth of the slit into nine parts, each of which is assumed to radiate semicircular ripples as in Huygens’ principle. The wavelength of the wave is 1 cm, and we assume for simplicity that each set of ripples has an amplitude of 1 unit when it reaches the screen.

1. For simplicity, let’s imagine that we were only to use two sets of ripples rather than nine. You could measure the distance from each of the two points inside the slit to your point on the screen. Suppose the distances were both 25.0 cm. What would be the amplitude of the superimposed waves at this point on the screen?

Suppose one distance was 24.0 cm and the other was 25.0 cm. What would happen?

What if one was 24.0 cm and the other was 26.0 cm?

What if one was 24.5 cm and the other was 25.0 cm?

In general, what combinations of distances will lead to completely destructive and completely constructive interference?

Can you estimate the answer in the case where the distances are 24.7 and 25.0 cm?

2. Although it is possible to calculate mathematically the amplitude of the sine wave that results from superimposing two sine waves with an arbitrary phase difference between them, the algebra is rather laborious, and it become even more tedious when we have more than two waves to superimpose. Instead, one can simply use a computer spreadsheet or some other computer program to add up the sine waves numerically at a series of points covering one complete cycle. This is what we will actually do. You just need to enter the relevant data into the computer, then examine the results and pick off the amplitude from the resulting list of numbers. You can run the software through a web interface at http://lightandmatter.com/cgi-bin/diffraction1.cgi.

3. Measure all nine distances to your group’s point on the screen, and write them on the board - that way everyone can see everyone else’s data, and the class can try to make sense of why the results came out the way they did. Determine the amplitude of the combined wave, and write it on the board as well.

The class will discuss why the results came out the way they did.
Exercise 32C: Diffraction of light

Equipment:

slit patterns, lasers, straight-filament bulbs

station 1
You have a mask with a bunch of different double slits cut out of it. The values of w and d are as follows:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>w</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04 mm</td>
<td>.250 mm</td>
</tr>
<tr>
<td>B</td>
<td>0.04 mm</td>
<td>.500 mm</td>
</tr>
<tr>
<td>C</td>
<td>0.08 mm</td>
<td>.250 mm</td>
</tr>
<tr>
<td>D</td>
<td>0.08 mm</td>
<td>.500 mm</td>
</tr>
</tbody>
</table>

Predict how the patterns will look different, and test your prediction. The easiest way to get the laser to point at different sets of slits is to stick folded up pieces of paper in one side or the other of the holders.

station 2
This is just like station 1, but with single slits:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.02 mm</td>
</tr>
<tr>
<td>B</td>
<td>0.04 mm</td>
</tr>
<tr>
<td>C</td>
<td>0.08 mm</td>
</tr>
<tr>
<td>D</td>
<td>0.16 mm</td>
</tr>
</tbody>
</table>

Predict what will happen, and test your predictions. If you have time, check the actual numerical ratios of the w values against the ratios of the sizes of the diffraction patterns.

station 3
This is like station 1, but the only difference among the sets of slits is how many slits there are:

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>double slit</td>
</tr>
<tr>
<td>B</td>
<td>3 slits</td>
</tr>
<tr>
<td>C</td>
<td>4 slits</td>
</tr>
<tr>
<td>D</td>
<td>5 slits</td>
</tr>
</tbody>
</table>

station 4
Hold the diffraction grating up to your eye, and look through it at the straight-filament light bulb. If you orient the grating correctly, you should be able to see the $m = 1$ and $m = -1$ diffraction patterns off the left and right. If you have it oriented the wrong way, they’ll be above and below the bulb instead, which is inconvenient because the bulb’s filament is vertical. Where is the $m = 0$ fringe? Can you see $m = 2$, etc.?

*Station 5 has the same equipment as station 4. If you’re assigned to station 5 first, you should actually do activity 4 first, because it’s easier.*

station 5
Use the transformer to increase and decrease the voltage across the bulb. This allows you to control the filament’s temperature. Sketch graphs of intensity as a function of wavelength for various temperatures. The inability of the wave model of light to explain the mathematical shapes of these curves was historically one of the reasons for creating a new model, in which light is both a particle and a wave.
The Modern Revolution in Physics
The continental U.S. got its first taste of volcanism in recent memory with the eruption of Mount St. Helens in 1980.

Chapter 33
Rules of Randomness

Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective positions of the things which compose it...nothing would be uncertain, and the future as the past would be laid out before its eyes.  

*Pierre Simon de Laplace, 1776*

The Quantum Mechanics is very imposing. But an inner voice tells me that it is still not the final truth. The theory yields much, but it hardly brings us nearer to the secret of the Old
One. In any case, I am convinced that He does not play dice.

*Albert Einstein*

However radical Newton’s clockwork universe seemed to his contemporaries, by the early twentieth century it had become a sort of smugly accepted dogma. Luckily for us, this deterministic picture of the universe breaks down at the atomic level. The clearest demonstration that the laws of physics contain elements of randomness is in the behavior of radioactive atoms. Pick two identical atoms of a radioactive isotope, say the naturally occurring uranium 238, and watch them carefully. They will decay at different times, even though there was no difference in their initial behavior.

We would be in big trouble if these atoms’ behavior was as predictable as expected in the Newtonian world-view, because radioactivity is an important source of heat for our planet. In reality, each atom chooses a random moment at which to release its energy, resulting in a nice steady heating effect. The earth would be a much colder planet if only sunlight heated it and not radioactivity. Probably there would be no volcanoes, and the oceans would never have been liquid. The deep-sea geothermal vents in which life first evolved would never have existed. But there would be an even worse consequence if radioactivity was deterministic: after a few billion years of peace, all the uranium 238 atoms in our planet would presumably pick the same moment to decay. The huge amount of stored nuclear energy, instead of being spread out over eons, would all be released at one instant, blowing our whole planet to Kingdom Come.¹

The new version of physics, incorporating certain kinds of randomness, is called quantum physics (for reasons that will become clear later). It represented such a dramatic break with the previous, deterministic tradition that everything that came before is considered “classical,” even the theory of relativity. The remainder of this book is a basic introduction to quantum physics.

**Discussion question**

A. I said “Pick two identical atoms of a radioactive isotope.” Are two atoms really identical? If their electrons are orbiting the nucleus, can we distinguish each atom by the particular arrangement of its electrons at some instant in time?

¹This is under the assumption that all the radioactive heating comes from uranium atoms, and that all the atoms were created at the same time. In reality, both uranium and thorium atoms contribute, and they may not have all been created at the same time. We have only a general idea of the processes that created these heavy elements in the gas cloud from which our solar system condensed. Some portion of them may have come from nuclear reactions in supernova explosions in that particular nebula, but some may have come from previous supernova explosions throughout our galaxy, or from exotic events like collisions of white dwarf stars.
33.1 Randomness isn’t random

Einstein’s distaste for randomness, and his association of determinism with divinity, goes back to the Enlightenment conception of the universe as a gigantic piece of clockwork that only had to be set in motion initially by the Builder. Many of the founders of quantum mechanics were interested in possible links between physics and Eastern and Western religious and philosophical thought, but every educated person has a different concept of religion and philosophy. Bertrand Russell remarked, “Sir Arthur Eddington deduces religion from the fact that atoms do not obey the laws of mathematics. Sir James Jeans deduces it from the fact that they do.”

Russell’s witticism, which implies incorrectly that mathematics cannot describe randomness, reminds us how important it is not to oversimplify this question of randomness. You should not simply surmise, “Well, it’s all random, anything can happen.” For one thing, certain things simply cannot happen, either in classical physics or quantum physics. The conservation laws of mass, energy, momentum, and angular momentum are still valid, so for instance processes that create energy out of nothing are not just unlikely according to quantum physics, they are impossible.

A useful analogy can be made with the role of randomness in evolution. Darwin was not the first biologist to suggest that species changed over long periods of time. His two new fundamental ideas were that (1) the changes arose through random genetic variation, and (2) changes that enhanced the organism’s ability to survive and reproduce would be preserved, while maladaptive changes would be eliminated by natural selection. Doubters of evolution often consider only the first point, about the randomness of natural variation, but not the second point, about the systematic action of natural selection. They make statements such as, “the development of a complex organism like Homo sapiens via random chance would be like a whirlwind blowing through a junkyard and spontaneously assembling a jumbo jet out of the scrap metal.” The flaw in this type of reasoning is that it ignores the deterministic constraints on the results of random processes. For an atom to violate conservation of energy is no more likely than the conquest of the world by chimpanzees next year.

Discussion question

A Economists often behave like wannabe physicists, probably because it seems prestigious to make numerical calculations instead of talking about human relationships and organizations like other social scientists. Their striving to make economics work like Newtonian physics extends to a parallel use of mechanical metaphors, as in the concept of a market’s supply and demand acting like a self-adjusting machine, and the idealization of people as economic automatons who consistently strive to maximize their own wealth. What evidence is there for randomness rather than mechanical determinism in economics?
33.2 Calculating randomness

You should also realize that even if something is random, we can still understand it, and we can still calculate probabilities numerically. In other words, physicists are good bookmakers. A good bookie can calculate the odds that a horse will win a race much more accurately than an inexperienced one, but nevertheless cannot predict what will happen in any particular race.

Statistical independence

As an illustration of a general technique for calculating odds, suppose you are playing a 25-cent slot machine. Each of the three wheels has one chance in ten of coming up with a cherry. If all three wheels come up cherries, you win $100. Even though the results of any particular trial are random, you can make certain quantitative predictions. First, you can calculate that your odds of winning on any given trial are \( \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} = \frac{1}{1000} = 0.001 \). Here, I am representing the probabilities as numbers from 0 to 1, which is clearer than statements like “The odds are 999 to 1,” and makes the calculations easier. A probability of 0 represents something impossible, and a probability of 1 represents something that will definitely happen.

Also, you can say that any given trial is equally likely to result in a win, and it doesn’t matter whether you have won or lost in prior games. Mathematically, we say that each trial is statistically independent, or that separate games are uncorrelated. Most gamblers are mistakenly convinced that, to the contrary, games of chance are correlated. If they have been playing a slot machine all day, they are convinced that it is “getting ready to pay,” and they do not want anyone else playing the machine and “using up” the jackpot that they “have coming.” In other words, they are claiming that a series of trials at the slot machine is negatively correlated, that losing now makes you more likely to win later. Craps players claim that you should go to a table where the person rolling the dice is “hot,” because she is likely to keep on rolling good numbers. Craps players, then, believe that rolls of the dice are positively correlated, that winning now makes you more likely to win later.

My method of calculating the probability of winning on the slot machine was an example of the following important rule for calculations based on independent probabilities:

the law of independent probabilities

If the probability of one event happening is \( P_A \), and the probability of a second statistically independent event happening is \( P_B \), then the probability that they will both occur is the product of the probabilities, \( P_A P_B \). If there are more than two events involved, you simply keep on multiplying.
This can be taken as the definition of statistical independence.

Note that this only applies to independent probabilities. For instance, if you have a nickel and a dime in your pocket, and you randomly pull one out, there is a probability of 0.5 that it will be the nickel. If you then replace the coin and again pull one out randomly, there is again a probability of 0.5 of coming up with the nickel, because the probabilities are independent. Thus, there is a probability of 0.25 that you will get the nickel both times.

Suppose instead that you do not replace the first coin before pulling out the second one. Then you are bound to pull out the other coin the second time, and there is no way you could pull the nickel out twice. In this situation, the two trials are not independent, because the result of the first trial has an effect on the second trial. The law of independent probabilities does not apply, and the probability of getting the nickel twice is zero, not 0.25.

Experiments have shown that in the case of radioactive decay, the probability that any nucleus will decay during a given time interval is unaffected by what is happening to the other nuclei, and is also unrelated to how long it has gone without decaying. The first observation makes sense, because nuclei are isolated from each other at the centers of their respective atoms, and therefore have no physical way of influencing each other. The second fact is also reasonable, since all atoms are identical. Suppose we wanted to believe that certain atoms were “extra tough,” as demonstrated by their history of going an unusually long time without decaying. Those atoms would have to be different in some physical way, but nobody has ever succeeded in detecting differences among atoms. There is no way for an atom to be changed by the experiences it has in its lifetime.

**Addition of probabilities**

The law of independent probabilities tells us to use multiplication to calculate the probability that both A and B will happen, assuming the probabilities are independent. What about the probability of an “or” rather than an “and?” If two events A and B are mutually exclusive, then the probability of one or the other occurring is the sum $P_A + P_B$. For instance, a bowler might have a 30% chance of getting a strike (knocking down all ten pins) and a 20% chance of knocking down nine of them. The bowler’s chance of knocking down either nine pins or ten pins is therefore 50%.

It does not make sense to add probabilities of things that are not mutually exclusive, i.e., that could both happen. Say I have a 90% chance of eating lunch on any given day, and a 90% chance of eating dinner. The probability that I will eat either lunch or dinner is not 180%.
Normalization

If I spin a globe and randomly pick a point on it, I have about a 70% chance of picking a point that’s in an ocean and a 30% chance of picking a point on land. The probability of picking either water or land is 70% + 30% = 100%. Water and land are mutually exclusive, and there are no other possibilities, so the probabilities had to add up to 100%. It works the same if there are more than two possibilities — if you can classify all possible outcomes into a list of mutually exclusive results, then all the probabilities have to add up to 1, or 100%. This property of probabilities is known as normalization.

Averages

Another way of dealing with randomness is to take averages. The casino knows that in the long run, the number of times you win will approximately equal the number of times you play multiplied by the probability of winning. In the game mentioned above, where the probability of winning is 0.001, if you spend a week playing, and pay $2500 to play 10,000 times, you are likely to win about 10 times (10,000 × 0.001 = 10), and collect $1000. On the average, the casino will make a profit of $1500 from you. This is an example of the following rule.

rule for calculating averages

If you conduct $N$ identical, statistically independent trials, and the probability of success in each trial is $P$, then on the average, the total number of successful trials will be $NP$. If $N$ is large enough, the relative error in this estimate will become small.

The statement that the rule for calculating averages gets more and more accurate for larger and larger $N$ (known popularly as the “law of averages”) often provides a correspondence principle that connects classical and quantum physics. For instance, the amount of power produced by a nuclear power plant is not random at any detectable level, because the number of atoms in the reactor is so large. In general, random behavior at the atomic level tends to average out when we consider large numbers of atoms, which is why physics seemed deterministic before physicists learned techniques for studying atoms individually.

We can achieve great precision with averages in quantum physics because we can use identical atoms to reproduce exactly the same situation many times. If we were betting on horses or dice, we would be much more limited in our precision. After a thousand races, the horse would be ready to retire. After a million rolls, the dice would be worn out.
self-check A
Which of the following things must be independent, which could be independent, and which definitely are not independent?

(1) the probability of successfully making two free-throws in a row in basketball
(2) the probability that it will rain in London tomorrow and the probability that it will rain on the same day in a certain city in a distant galaxy
(3) your probability of dying today and of dying tomorrow

▷ Answer, p. 1046

Discussion questions

A Newtonian physics is an essentially perfect approximation for describing the motion of a pair of dice. If Newtonian physics is deterministic, why do we consider the result of rolling dice to be random?

B Why isn’t it valid to define randomness by saying that randomness is when all the outcomes are equally likely?

C The sequence of digits 121212121212121212 seems clearly nonrandom, and 41592653589793 seems random. The latter sequence, however, is the decimal form of pi, starting with the third digit. There is a story about the Indian mathematician Ramanujan, a self-taught prodigy, that a friend came to visit him in a cab, and remarked that the number of the cab, 1729, seemed relatively uninteresting. Ramanujan replied that on the contrary, it was very interesting because it was the smallest number that could be represented in two different ways as the sum of two cubes. The Argentine author Jorge Luis Borges wrote a short story called “The Library of Babel,” in which he imagined a library containing every book that could possibly be written using the letters of the alphabet. It would include a book containing only the repeated letter “a;” all the ancient Greek tragedies known today, all the lost Greek tragedies, and millions of Greek tragedies that were never actually written; your own life story, and various incorrect versions of your own life story; and countless anthologies containing a short story called “The Library of Babel.” Of course, if you picked a book from the shelves of the library, it would almost certainly look like a nonsensical sequence of letters and punctuation, but it’s always possible that the seemingly meaningless book would be a science-fiction screenplay written in the language of a Neanderthal tribe, or the lyrics to a set of incomparably beautiful love songs written in a language that never existed. In view of these examples, what does it really mean to say that something is random?
33.3 Probability distributions

So far we’ve discussed random processes having only two possible outcomes: yes or no, win or lose, on or off. More generally, a random process could have a result that is a number. Some processes yield integers, as when you roll a die and get a result from one to six, but some are not restricted to whole numbers, for example the number of seconds that a uranium-238 atom will exist before undergoing radioactive decay.

Consider a throw of a die. If the die is “honest,” then we expect all six values to be equally likely. Since all six probabilities must add up to 1, then probability of any particular value coming up must be 1/6. We can summarize this in a graph, d. Areas under the curve can be interpreted as total probabilities. For instance, the area under the curve from 1 to 3 is $1/6 + 1/6 + 1/6 = 1/2$, so the probability of getting a result from 1 to 3 is 1/2. The function shown on the graph is called the probability distribution.

Figure e shows the probabilities of various results obtained by rolling two dice and adding them together, as in the game of craps. The probabilities are not all the same. There is a small probability of getting a two, for example, because there is only one way to do it, by rolling a one and then another one. The probability of rolling a seven is high because there are six different ways to do it: 1+6, 2+5, etc.

If the number of possible outcomes is large but finite, for example the number of hairs on a dog, the graph would start to look like a smooth curve rather than a ziggurat.

What about probability distributions for random numbers that are not integers? We can no longer make a graph with probability on the $y$ axis, because the probability of getting a given exact number is typically zero. For instance, there is zero probability that a radioactive atom will last for exactly 3 seconds, since there are infinitely many possible results that are close to 3 but not exactly three, for example $2.9999999999999996876587658465436$. It doesn’t usually make sense, therefore, to talk about the probability of a single numerical result, but it does make sense to talk about the probability of a certain range of results. For instance, the probability that an atom will last more than 3 and less than 4 seconds is a perfectly reasonable thing to discuss. We can still summarize the probability information on a graph, and we can still interpret areas under the curve as probabilities.

But the $y$ axis can no longer be a unitless probability scale. In radioactive decay, for example, we want the $x$ axis to have units of time, and we want areas under the curve to be unitless probabilities.
The area of a single square on the graph paper is then

\[
\frac{\text{(unitless area of a square)}}{} = (\text{width of square with time units}) \times (\text{height of square}).
\]

If the units are to cancel out, then the height of the square must evidently be a quantity with units of inverse time. In other words, the \( y \) axis of the graph is to be interpreted as probability per unit time, not probability.

Figure \( f \) shows another example, a probability distribution for people’s height. This kind of bell-shaped curve is quite common.

**self-check B**

Compare the number of people with heights in the range of 130-135 cm to the number in the range 135-140.  

▶ Answer, p. 1046

**Looking for tall basketball players example 1**

▶ A certain country with a large population wants to find very tall people to be on its Olympic basketball team and strike a blow against western imperialism. Out of a pool of \( 10^8 \) people who are the right age and gender, how many are they likely to find who are over 225 cm (7 feet 4 inches) in height? Figure \( g \) gives a close-up of the “tail” of the distribution shown previously in figure \( f \).

▶ The shaded area under the curve represents the probability that a given person is tall enough. Each rectangle represents a probability of \( 0.2 \times 10^{-7} \) cm\(^{-1} \) × 1 cm = \( 2 \times 10^{-8} \). There are about 35 rectangles covered by the shaded area, so the probability of having a height greater than 225 cm is \( 7 \times 10^{-7} \), or just under one in a million. Using the rule for calculating averages, the average, or expected number of people this tall is \( (10^8) \times (7 \times 10^{-7}) = 70 \).

**Average and width of a probability distribution**

If the next Martian you meet asks you, “How tall is an adult human?,” you will probably reply with a statement about the average human height, such as “Oh, about 5 feet 6 inches.” If you wanted to explain a little more, you could say, “But that’s only an average. Most people are somewhere between 5 feet and 6 feet tall.” Without bothering to draw the relevant bell curve for your new extraterrestrial acquaintance, you’ve summarized the relevant information by giving an average and a typical range of variation.

The average of a probability distribution can be defined geometrically as the horizontal position at which it could be balanced if it was constructed out of cardboard, \( h \). A convenient numerical measure of the amount of variation about the average, or amount of uncertainty, is the full width at half maximum, or FWHM, defined in figure \( g \). (The FWHM was introduced on p. 481.)
A great deal more could be said about this topic, and indeed an introductory statistics course could spend months on ways of defining the center and width of a distribution. Rather than force-feeding you on mathematical detail or techniques for calculating these things, it is perhaps more relevant to point out simply that there are various ways of defining them, and to inoculate you against the misuse of certain definitions.

The average is not the only possible way to say what is a typical value for a quantity that can vary randomly; another possible definition is the median, defined as the value that is exceeded with 50% probability. When discussing incomes of people living in a certain town, the average could be very misleading, since it can be affected massively if a single resident of the town is Bill Gates. Nor is the FWHM the only possible way of stating the amount of random variation; another possible way of measuring it is the standard deviation (defined as the square root of the average squared deviation from the average value).

33.4 Exponential decay and half-life

Most people know that radioactivity “lasts a certain amount of time,” but that simple statement leaves out a lot. As an example, consider the following medical procedure used to diagnose thyroid function. A very small quantity of the isotope $^{131}$I, produced in a nuclear reactor, is fed to or injected into the patient. The body’s biochemical systems treat this artificial, radioactive isotope exactly the same as $^{127}$I, which is the only naturally occurring type. (Nutritionally, iodine is a necessary trace element. Iodine taken into the body is partly excreted, but the rest becomes concentrated in the thyroid gland. Iodized salt has had iodine added to it to prevent the nutritional deficiency known as goiters, in which the iodine-starved thyroid becomes swollen.) As the $^{131}$I undergoes beta decay, it emits electrons, neutrinos, and gamma rays. The gamma rays can be measured by a detector passed over the patient’s body. As the radioactive iodine becomes concentrated in the thyroid, the amount of gamma radiation coming from the thyroid becomes greater, and that emitted by the rest of the body is reduced. The rate at which the iodine concentrates in the thyroid tells the doctor about the health of the thyroid.

If you ever undergo this procedure, someone will presumably explain a little about radioactivity to you, to allay your fears that you will turn into the Incredible Hulk, or that your next child will have an unusual number of limbs. Since iodine stays in your thyroid for a long time once it gets there, one thing you’ll want to know is whether your thyroid is going to become radioactive forever. They may just tell you that the radioactivity “only lasts a certain amount of time,” but we can now carry out a quantitative derivation of how