

Optics

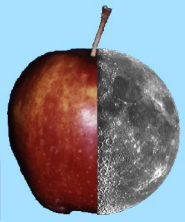
Benjamin Crowell



Optics

The **Light and Matter** series of introductory physics textbooks:


- 1 Newtonian Physics
- 2 Conservation Laws
- 3 Vibrations and Waves
- 4 Electricity and Magnetism
- 5 Optics
- 6 The Modern Revolution in Physics



Optics

Benjamin Crowell

www.lightandmatter.com

 Light and Matter
Fullerton, California
www.lightandmatter.com

© 1999-2001 by Benjamin Crowell
All rights reserved.

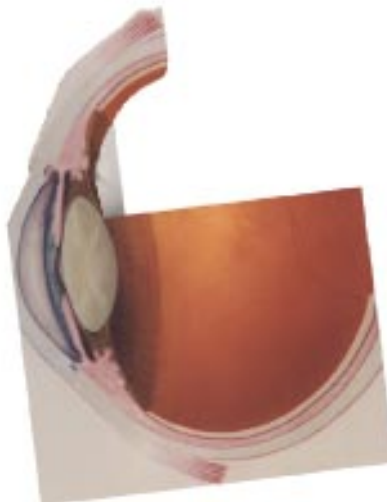
Edition 2.1
rev. 2002-09-27

ISBN 0-9704670-5-2



Brief Contents

1	The Ray Model of Light	11
2	Images by Reflection, Part I	25
3	Images by Reflection, Part II	33
4	Refraction and Images	45
5	Wave Optics	59



Contents

1 The Ray Model of Light 11

1.1 The Nature of Light	12
1.2 Interaction of Light with Matter	15
1.3 The Ray Model of Light	17
1.4 Geometry of Specular Reflection	20
1.5* The Principle of Least Time for Reflection	22
Summary	23
Homework Problems	24

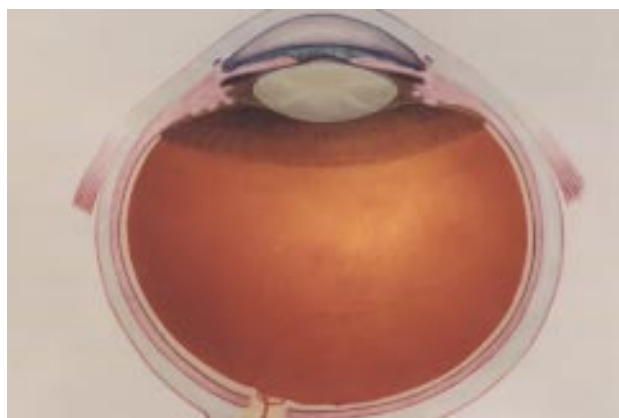


3 Images by Reflection, Part II 33

3.1 A Real Image Formed by an Inbending Mirror	34
3.2 Other Cases With Curved Mirrors	37
3.3* Aberrations	41
Summary	43
Homework Problems	44

2 Images by Reflection, Part I 25

2.1 A Virtual Image	26
2.2 Curved Mirrors	27
2.3 A Real Image	28
2.4 Images of Images	29
Summary	31
Homework Problems	32



4	Refraction and Images	45
4.1	Refraction	46
4.2	Lenses	52
4.3*	The Lensmaker's Equation	54
4.4*	Refraction and the Principle of Least Time	54
	Summary	55
	Homework Problems	56

Exercises	77
Solutions to Selected Problems	85
Glossary	87
Index	89



5	Wave Optics	59
5.1	Diffraction	60
5.2	Scaling of Diffraction	61
5.3	The Correspondence Principle	62
5.4	Huygens' Principle	63
5.5	Double-Slit Diffraction	64
5.6	Repetition	68
5.7	Single-Slit Diffraction	69
5.8]*	The Principle of Least Time	72
	Summary	73
	Homework Problems	74



1 The Ray Model of Light

Ads for the latest Macintosh computer brag that it can do an arithmetic calculation in less time than it takes for the light to get from the screen to your eye. We find this impressive because of the contrast between the speed of light and the speeds at which we interact with physical objects in our environment. Perhaps it shouldn't surprise us, then, that Newton succeeded so well in explaining the motion of objects, but was far less successful with the study of light.

This textbook series is billed as the Light and Matter series, but only now, in the fifth of the six volumes, are we ready to focus on light. If you are reading the series in order, then you know that the climax of our study of electricity and magnetism was discovery that light is an electromagnetic wave. Knowing this, however, is not the same as knowing everything about eyes and telescopes. In fact, the full description of light as a wave can be rather cumbersome. We will instead spend most of this book making use of a simpler model of light, the ray model, which does a fine job in most practical situations. Not only that, but we will even backtrack a little and start with a discussion of basic ideas about light and vision that predated the discovery of electromagnetic waves. Research in physics education has shown conclusively that the mere assertion that light is an electromagnetic wave is woefully insufficient to allow a student to interpret ordinary phenomena involving light.

1.1 The Nature of Light

The cause and effect relationship in vision

Despite its title, this chapter is far from your first look at light. That familiarity might seem like an advantage, but most people have never thought carefully about light and vision. Even smart people who have thought hard about vision have come up with incorrect ideas. The ancient Greeks, Arabs and Chinese had theories of light and vision, all of which were mostly wrong, and all of which were accepted for thousands of years.

One thing the ancients did get right is that there is a distinction between objects that emit light and objects that don't. When you see a leaf in the forest, it's because three different objects are doing their jobs: the leaf, the eye, and the sun. But luminous objects like the sun, a flame, or the filament of a light bulb can be seen by the eye without the presence of a third object. Emission of light is often, but not always, associated with heat. In modern times, we are familiar with a variety of objects that glow without being heated, including fluorescent lights and glow-in-the-dark toys.

How do we see luminous objects? The Greek philosophers Pythagoras (b. ca. 560 BC) and Empedocles of Acragas (b. ca. 492 BC), who unfortunately were very influential, claimed that when you looked at a candle flame, the flame and your eye were both sending out some kind of mysterious stuff, and when your eye's stuff collided with the candle's stuff, the candle would become evident to your sense of sight.

Bizarre as the Greek "collision of stuff theory" might seem, it had a couple of good features. It explained why both the candle and your eye had to be present for your sense of sight to function. The theory could also easily be expanded to explain how we see nonluminous objects. If a leaf, for instance, happened to be present at the site of the collision between your eye's stuff and the candle's stuff, then the leaf would be stimulated to express its green nature, allowing you to perceive it as green.

Modern people might feel uneasy about this theory, since it suggests that greenness exists only for our seeing convenience, implying a human precedence over natural phenomena. Nowadays, people would expect the cause and effect relationship in vision to be the other way around, with the leaf doing something to our eye rather than our eye doing something to the leaf. But how can you tell? The most common way of distinguishing cause from effect is to determine which happened first, but the process of seeing seems to occur too quickly to determine the order in which things happened. Certainly there is no obvious time lag between the moment when you move your head and the moment when your reflection in the mirror moves.

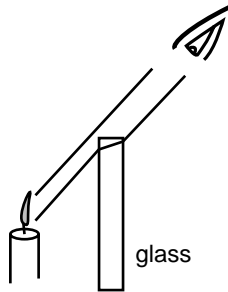
Today, photography provides the simplest experimental evidence that nothing has to be emitted from your eye and hit the leaf in order to make it "greenify." A camera can take a picture of a leaf even if there are no eyes anywhere nearby. Since the leaf appears green regardless of whether it is being sensed by a camera, your eye, or an insect's eye, it seems to make more sense to say that the leaf's greenness is the cause, and something happening in the camera or eye is the effect.

Light is a thing, and it travels from one point to another.

Another issue that few people have considered is whether a candle's flame simply affects your eye directly, or whether it sends out light which then gets into your eye. Again, the rapidity of the effect makes it difficult to tell what's happening. If someone throws a rock at you, you can see the rock on its way to your body, and you can tell that the person affected you by sending a material substance your way, rather than just harming you directly with an arm motion, which would be known as "action at a distance." It is not easy to do a similar observation to see whether there is some "stuff" that travels from the candle to your eye, or whether it is a case of action at a distance.

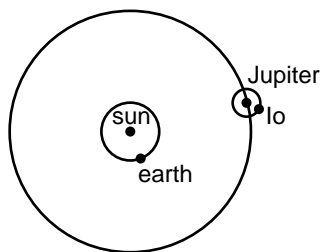
Newtonian physics includes both action at a distance (e.g. the earth's gravitational force on a falling object) and contact forces such as the normal force, which only allow distant objects to exert forces on each other by shooting some substance across the space between them (e.g. a garden hose spraying out water that exerts a force on a bush).

One piece of evidence that the candle sends out stuff that travels to your eye is that intervening transparent substances can make the candle appear to be in the wrong location, suggesting that light is a thing that can be bumped off course. Many people would dismiss this kind of observation as an optical illusion, however. (Some optical illusions are purely neurological or psychological effects, although some others, including this one, turn out to be caused by the behavior of light itself.)



Light from a candle is bumped off course by a piece of glass. Inserting the glass causes the apparent location of the candle to shift. The same effect can be produced by taking off your eyeglasses and looking at what you see in the lens, but a flat piece of glass works just as well as a lens for this purpose.

A more convincing way to decide in which category light belongs is to find out if it takes time to get from the candle to your eye; in Newtonian physics, action at a distance is supposed to be instantaneous. The fact that we speak casually today of "the speed of light" implies that at some point in history, somebody succeeded in showing that light did not travel infinitely fast. Galileo tried, and failed, to detect a finite speed for light, by arranging with a person in a distant tower to signal back and forth with lanterns. Galileo uncovered his lantern, and when the other person saw the light, he uncovered his lantern. Galileo was unable to measure any time lag that was significant compared to the limitations of human reflexes.



The first person to prove that light's speed was finite, and to determine it numerically, was Ole Roemer, in a series of measurements around the year 1675. Roemer observed Io, one of Jupiter's moons, over a period of several years. Since Io presumably took the same amount of time to complete each orbit of Jupiter, it could be thought of as a very distant, very accurate clock. A practical and accurate pendulum clock had recently been invented, so Roemer could check whether the ratio of the two clocks' cycles, about 42.5 hours to 1 orbit, stayed exactly constant or changed a little. If the process of seeing the distant moon was instantaneous, there would be no reason for the two to get out of step. Even if the speed of light was finite, you might expect that the result would be only to offset one cycle relative to the other. The earth does not, however, stay at a constant distance from Jupiter and its moons. Since the distance is changing gradually due to the two planets' orbit motions, a finite speed of light would make the "Io clock" appear to run faster as the planets drew near each other, and more slowly as their separation increased. Roemer did find a variation in the apparent speed of Io's orbits, which caused Io's eclipses by Jupiter (the moments when Io passed in front of or behind Jupiter) to occur about 7 minutes early when the earth was closest to Jupiter, and 7 minutes late when it was farthest. Based on these measurements, Roemer estimated the speed of light to be approximately 2×10^8 m/s, which is in the right ballpark compared to modern measurements of 3×10^8 m/s. (I'm not sure whether the fairly large experimental error was mainly due to imprecise knowledge of the radius of the earth's orbit or limitations in the reliability of pendulum clocks.)

Light can travel through a vacuum.

Many people are confused by the relationship between sound and light. Although we use different organs to sense them, there are some similarities. For instance, both light and sound are typically emitted in all directions by their sources. Musicians even use visual metaphors like "tone color," or "a bright timbre" to describe sound. One way to see that they are clearly different phenomena is to note their very different velocities. Sure, both are pretty fast compared to a flying arrow or a galloping horse, but as we have seen, the speed of light is so great as to appear instantaneous in most situations. The speed of sound, however, can easily be observed just by watching a group of schoolchildren a hundred feet away as they clap their hands to a song. There is an obvious delay between when you see their palms come together and when you hear the clap.

The fundamental distinction between sound and light is that sound is an oscillation in air pressure, so it requires air (or some other medium such as water) in which to travel. Today, we know that outer space is a vacuum, so the fact that we get light from the sun, moon and stars clearly shows that air is not necessary for the propagation of light. Also, a light bulb has a near vacuum inside, but that does not prevent the light from getting out. (The reason why the air is pumped out of light bulbs is to keep the oxygen from reacting violently with the hot filament and destroying it.)

Discussion Questions



- A. If you observe thunder and lightning, you can tell how far away the storm is. Do you need to know the speed of sound, of light, or of both?
- B. When phenomena like X-rays and cosmic rays were first discovered, suggest a way one could have tested whether they were forms of light.
- C. Why did Roemer only need to know the radius of the earth's orbit, not Jupiter's, in order to find the speed of light?

1.2 Interaction of Light with Matter

Absorption of light

The reason why the sun feels warm on your skin is that the sunlight is being absorbed, and the light energy is being transformed into heat energy. The same happens with artificial light, so the net result of leaving a light turned on is to heat the room. It doesn't matter whether the source of the light is hot, like the sun, a flame, or an incandescent light bulb, or cool, like a fluorescent bulb. (If your house has electric heat, then there is absolutely no point in fastidiously turning off lights in the winter; the lights will help to heat the house at the same dollar rate as the electric heater.)

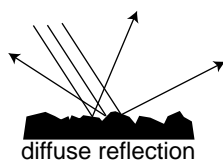
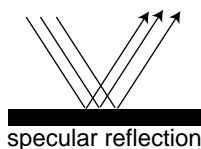
This process of heating by absorption is entirely different from heating by thermal conduction, as when an electric stove heats spaghetti sauce through a pan. Heat can only be conducted through matter, but there is vacuum between us and the sun, or between us and the filament of an incandescent bulb. Also, heat conduction can only transfer heat energy from a hotter object to a colder one, but a cool fluorescent bulb is perfectly capable of heating something that had already started out being warmer than the bulb itself.

How we see nonluminous objects

Not all the light energy that hits an object is transformed into heat. Some is reflected, and this leads us to the question of how we see nonluminous objects. If you ask the average person how we see a light bulb, the most likely answer is "The light bulb makes light, which hits our eyes." But if you ask how we see a book, they are likely to say "The bulb lights up the room, and that lets me see the book." All mention of light actually entering our eyes has mysteriously disappeared.

Most people would disagree if you told them that light was reflected from the book to the eye, because they think of reflection as something that mirrors do, not something that a book does. They associate reflection with the formation of a reflected image, which does not seem to appear in a piece of paper.

Imagine that you are looking at your reflection in a nice smooth piece of aluminum foil, fresh off the roll. You perceive a face, not a piece of metal. Perhaps you also see the bright reflection of a lamp over your shoulder behind you. Now imagine that the foil is just a little bit less smooth. The different parts of the image are now a little bit out of alignment with each other. Your brain can still recognize a face and a lamp, but it's a little scrambled, like a Picasso painting. Now suppose you use a piece of aluminum foil that has been crumpled up and then flattened out again. The parts of the image are so scrambled that you cannot recognize an image. Instead, your brain tells you you're looking at a rough, silvery surface.



(a) Diffuse and specular reflection.

Mirrorlike reflection at a specific angle is known as specular reflection, and random reflection in many directions is called diffuse reflection. Diffuse reflection is how we see nonluminous objects. Specular reflection only allows us to see images of objects other than the one doing the reflecting. In top part of figure (a), imagine that the rays of light are coming from the sun. If you are looking down at the reflecting surface, there is no way for your eye-brain system to tell that the rays are not really coming from a sun down below you.

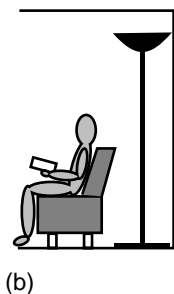


Figure (b) shows another example of how we can't avoid the conclusion that light bounces off of things other than mirrors. The lamp is one I have in my house. It has a bright bulb, housed in a completely opaque bowl-shaped metal shade. The only way light can get out of the lamp is by going up out of the top of the bowl. The fact that I can read a book in the position shown in the figure means that light must be bouncing off of the ceiling, then bouncing off of the book, then finally getting to my eye.

This is where the shortcomings of the Greek theory of vision become glaringly obvious. In the Greek theory, the light from the bulb and my mysterious "eye rays" are both supposed to go to the book, where they collide, allowing me to see the book. But we now have a total of four objects: lamp, eye, book, and ceiling. Where does the ceiling come in? Does it also send out its own mysterious "ceiling rays," contributing to a three-way collision at the book? That would just be too bizarre to believe!

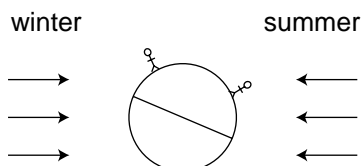
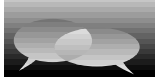
The differences among white, black, and the various shades of gray in between is a matter of what percentage of the light they absorb and what percentage they reflect. That's why light-colored clothing is more comfortable in the summer, and light-colored upholstery in a car stays cooler than dark upholstery.

Numerical measurement of the brightness of light

We have already seen that the physiological sensation of loudness relates to the sound's intensity (power per unit area), but is not directly proportional to it. If sound A has an intensity of 1 nW/m^2 , sound B is 10 nW/m^2 , and sound C is 100 nW/m^2 , then the increase in loudness from C to B is perceived to be the same as the increase from A to B, not ten times greater. That is, the sensation of loudness is logarithmic.

The same is true for the brightness of light. Brightness is related to power per unit area, but the psychological relationship is a logarithmic one rather than a proportionality. For doing physics, it's the power per unit area that we're interested in. The relevant SI unit is W/m^2 , although various non-SI units are still used in photography, for example. One way to determine the brightness of light is to measure the increase in temperature of a black object exposed to the light. The light energy is being converted to heat energy, and the amount of heat energy absorbed in a given amount of time can be related to the power absorbed, using the known heat capacity of the object. More practical devices for measuring light intensity, such as the light meters built into some cameras, are based on the conversion of light into electrical energy, but these meters have to be calibrated somehow against heat measurements.

Discussion questions



Discussion question C.

A. The curtains in a room are drawn, but a small gap lets light through, illuminating a spot on the floor. It may or may not also be possible to see the beam of sunshine crossing the room, depending on the conditions. What's going on?

B. Laser beams are made of light. In science fiction movies, laser beams are often shown as bright lines shooting out of a laser gun on a spaceship. Why is this scientifically incorrect?

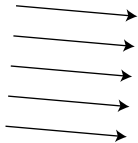
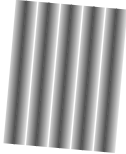
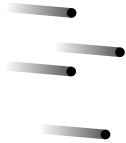
C. A documentary filmmaker went to Harvard's 1987 graduation ceremony and asked the graduates, on camera, to explain the cause of the seasons. Only two out of 23 were able to give a correct explanation, but you now have all the information needed to figure it out for yourself, assuming you didn't already know. The figure shows the earth in its winter and summer positions relative to

the sun. Hint: Consider the units used to measure the brightness of light, and recall that the sun is lower in the sky in winter, so its rays are coming in at a shallower angle.

1.3 The Ray Model of Light

Models of light

Note how I've been casually diagramming the motion of light with pictures showing light rays as lines on the page. More formally, this is known as the ray model of light. The ray model of light seems natural once we convince ourselves that light travels through space, and observe phenomena like sunbeams coming through holes in clouds. Having already been introduced to the concept of light as an electromagnetic wave, you know that the ray model is not the ultimate truth about light, but the ray model is simpler, and in any case science always deals with models of reality, not the ultimate nature of reality. The following table summarizes three models of light.

ray model		<i>Advantage:</i> Simplicity.
wave model		<i>Advantage:</i> Color is described naturally in terms of wavelength. <i>Required</i> in order to explain the interaction of light with material objects with sizes comparable to a wavelength of light or smaller.
particle model		<i>Required</i> in order to explain the interaction of light with individual atoms. At the atomic level, it becomes apparent that a beam of light has a certain graininess to it.

The ray model is essentially a generic one. By using it we can discuss the path taken by the light, without committing ourselves to any specific description of what it is that is moving along that path. We will use the nice simple ray model for most of this book, and with it we can analyze a great many devices and phenomena. Not until the last chapter will we concern ourselves specifically with wave optics, although in the intervening chapters I will sometimes analyze the same phenomenon using both the ray model and the wave model.

Note that the statements about the applicability of the various models are only rough guides. For instance, wave interference effects are often detectable, if small, when light passes around an obstacle that is quite a bit

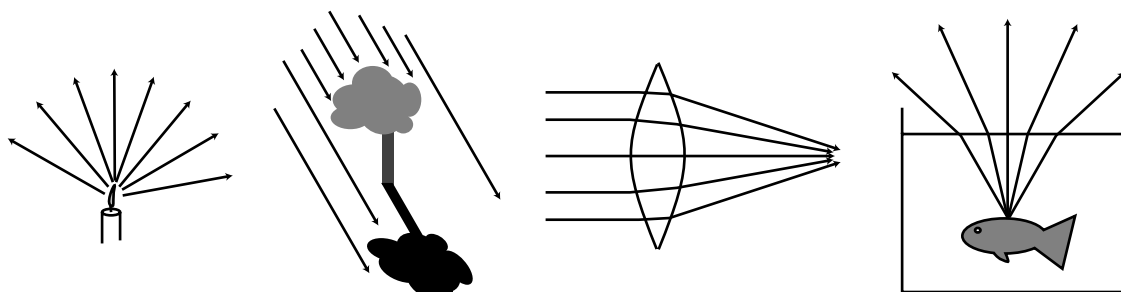
bigger than a wavelength. Also, the criterion for when we need the particle model really has more to do with energy scales than distance scales, although the two turn out to be related.

The alert reader may have noticed that the wave model is required at scales smaller than a wavelength of light (on the order of a micrometer for visible light), and the particle model is demanded on the atomic scale or lower (a typical atom being a nanometer or so in size). This implies that at the smallest scales we need *both* the wave model and the particle model. They appear incompatible, so how can we simultaneously use both? The answer is that they are not as incompatible as they seem. Light is both a wave and a particle, but a full understanding of this apparently nonsensical statement is a topic for the following book in this series.

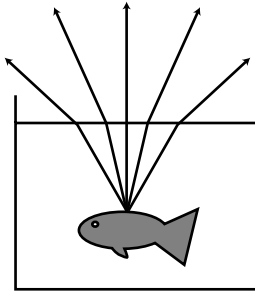
Ray diagrams

Without even knowing how to use the ray model to calculate anything numerically, we can learn a great deal by drawing ray diagrams. For instance, if you want to understand how eyeglasses help you to see in focus, a ray diagram is the right place to start. Many students underutilize ray diagrams in optics and instead rely on rote memorization or plugging into formulas. The trouble with memorization and plug-ins is that they can obscure what's really going on, and it is easy to get them wrong. Often the best plan is to do a ray diagram first, then do a numerical calculation, then check that your numerical results are in reasonable agreement with what you expected from the ray diagram.

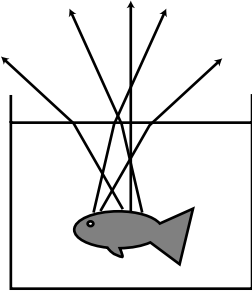
Examples (a) through (c) show some guidelines for using ray diagrams effectively. The light rays bend when they pass out through the surface of the water (a phenomenon that we'll discuss in more detail later). The rays appear to have come from a point above the goldfish's actual location, an



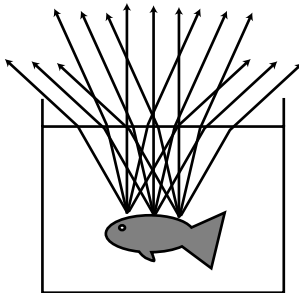
Examples of ray diagrams.



(a) correct



(b) Incorrect: implies that diffuse reflection only gives one ray from each reflecting point.



(c) Correct, but unnecessarily complicated.

effect that is familiar to people who have tried spearfishing.

- A stream of light is not really confined to a finite number of narrow lines. We just draw it that way. In (a), it has been necessary to choose a finite number of rays to draw (five), rather than the theoretically infinite number of rays that will diverge from that point.
- There is a tendency to conceptualize rays incorrectly as objects. In his Optics, Newton goes out of his way to caution the reader against this, saying that some people “consider ... the refraction of ... rays to be the bending or breaking of them in their passing out of one medium into another.” But a ray is a record of the path traveled by light, not a physical thing that can be bent or broken.
- In theory, rays may continue infinitely far into the past and future, but we need to draw lines of finite length. In (a), a judicious choice has been made as to where to begin and end the rays. There is no point in continuing the rays any farther than shown, because nothing new and exciting is going to happen to them. There is also no good reason to start them earlier, before being reflected by the fish, because the direction of the diffusely reflected rays is random anyway, and unrelated to the direction of the original, incoming ray.
- When representing diffuse reflection in a ray diagram, many students have a mental block against drawing many rays fanning out from the same point. Often, as in example (b), the problem is the misconception that light can only be reflected in one direction from one point.
- Another difficulty associated with diffuse reflection, example (c), is the tendency to think that in addition to drawing many rays coming out of one point, we should also be drawing many rays coming from many points. In (a), drawing many rays coming out of one point gives useful information, telling us, for instance, that the fish can be seen from any angle. Drawing many sets of rays, as in (c), does not give us any more useful information, and just clutters up the picture in this example. The only reason to draw sets of rays fanning out from more than one point would be if different things were happening to the different sets.

Discussion Question



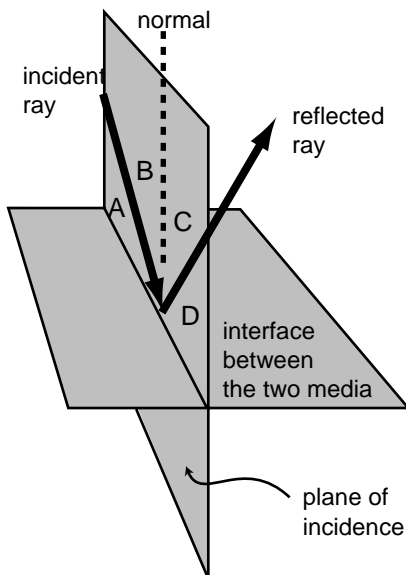
Suppose an intelligent tool-using fish is spearhunting for humans. Draw a ray diagram to show how the fish has to correct its aim. Note that although the rays are now passing from the air to the water, the same rules apply: the rays are closer to being perpendicular to the surface when they are in the water, and rays that hit the air-water interface at a shallow angle are bent the most.

1.4 Geometry of Specular Reflection

To change the motion of a material object, we use a force. Is there any way to exert a force on a beam of light? Experiments show that electric and magnetic fields do not deflect light beams, so apparently light has no electric charge. Light also has no mass, so until the twentieth century it was believed to be immune to gravity as well. Einstein predicted that light beams would be very slightly deflected by strong gravitational fields, and he was proven correct by observations of rays of starlight that came close to the sun, but obviously that's not what makes mirrors and lenses work!

If we investigate how light is reflected by a mirror, we will find that the process is horrifically complex, but the final result is surprisingly simple. What actually happens is that the light is made of electric and magnetic fields, and these fields accelerate the electrons in the mirror. Energy from the light beam is momentarily transformed into extra kinetic energy of the electrons, but because the electrons are accelerating they reradiate more light, converting their kinetic energy back into light energy. We might expect this to result in a very chaotic situation, but amazingly enough, the electrons move together to produce a new, reflected beam of light, which obeys two simple rules:

- The angle of the reflected ray is the same as that of the incident ray.
- The reflected ray lies in the plane containing the incident ray and the normal (perpendicular) line. This plane is known as the plane of incidence.



The two angles can be defined either with respect to the normal, like angles B and C in the figure, or with respect to the reflecting surface, like angles A and D. There is a convention of several hundred years' standing that one measures the angles with respect to the normal, but the rule about equal angles can logically be stated either as $B=C$ or as $A=D$.

The phenomenon of reflection occurs only at the boundary between two media, just like the change in the speed of light that passes from one medium to another. As we have seen in book 3 of this series, this is the way all waves behave.

Most people are surprised by the fact that light can be reflected back into a less dense medium. For instance, if you are diving and you look up at the surface of the water, you will see a reflection of yourself.

Reversibility of light rays

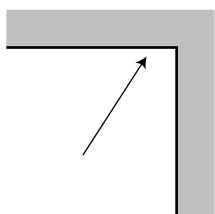
The fact that specular reflection displays equal angles of incidence and reflection means that there is a symmetry: if the ray had come in from the right instead of the left in the figure above, the angles would have looked exactly the same. This is not just a pointless detail about specular reflection. It's a manifestation of a very deep and important fact about nature, which is that the laws of physics do not distinguish between past and future. Cannonballs and planets have trajectories that are equally natural in reverse, and

so do light rays. This type of symmetry is called time-reversal symmetry.

Typically, time-reversal symmetry is a characteristic of any process that does not involve heat. For instance, the planets do not experience any friction as they travel through empty space, so there is no frictional heating. We should thus expect the time-reversed versions of their orbits to obey the laws of physics, which they do. In contrast, a book sliding across a table does generate heat from friction as it slows down, and it is therefore not surprising that this type of motion does not appear to obey time-reversal symmetry. A book lying still on a flat table is never observed to spontaneously start sliding, sucking up heat energy and transforming it into kinetic energy.

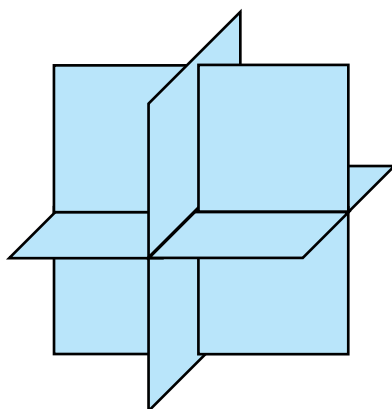
Similarly, the only situation we've observed so far where light does not obey time-reversal symmetry is absorption, which involves heat. Your skin absorbs visible light from the sun and heats up, but we never observe people's skin to glow, converting heat energy into visible light. People's skin does glow in infrared light, but that doesn't mean the situation is symmetric. Even if you absorb infrared, you don't emit visible light, because your skin isn't hot enough to glow in the visible spectrum.

Discussion Questions



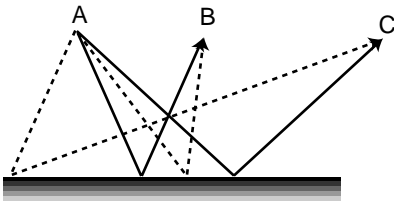
Discussion question B.

- A.** If a light ray has a velocity vector with components c_x and c_y , what will happen when it is reflected from a surface that lies along the y axis? Make sure your answer does not imply a change in the ray's speed.
- B.** Generalizing your reasoning from discussion question B, what will happen to the velocity components of a light ray that hits a corner, as shown in the figure, and undergoes two reflections?
- C.** Three pieces of sheet metal arranged perpendicularly as shown in the figure form what is known as a radar corner. Let's assume that the radar corner is large compared to the wavelength of the radar waves, so that the ray model makes sense. If the radar corner is bathed in radar rays, at least some of them will undergo three reflections. Making a further generalization of your reasoning from the two preceding discussion questions, what will happen to the three velocity components of such a ray? What would the radar corner be useful for?

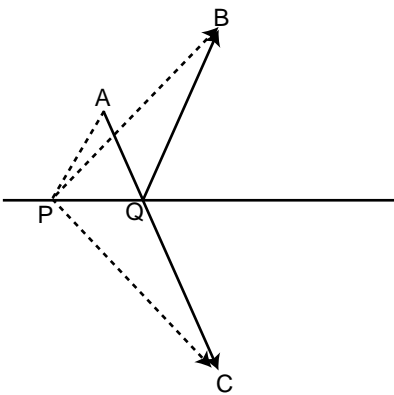


Discussion question C.

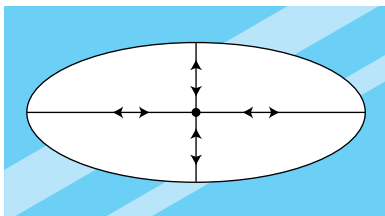
1.5* The Principle of Least Time for Reflection



(a) The solid lines are physically possible paths for light rays traveling from A to B and from A to C. They obey the principle of least time. The dashed lines do not obey the principle of least time, and are not physically possible.



(b) Paths AQB and APB are two conceivable paths that a ray could travel to get from A to B with one reflection, but only AQB is physically possible. We wish to prove that path AQB, with equal angles of incidence and reflection, is shorter than any other path, such as APB. The trick is to construct a third point, C, lying as far below the surface as B lies above it. Then path AQC is a straight line whose length is the same as AQB, and path APC has the same length as path APB. Since AQC is straight, it must be shorter than any other path such as APC that connects A and C, and therefore AQB must be shorter than any path such as APB.



(c) Light is emitted at the center of an elliptical mirror. There are four physically possible paths by which a ray can be reflected and return to the center.

We had to choose between an unwieldy explanation of reflection at the atomic level and a simpler geometric description that was not as fundamental. There is a third approach to describing the interaction of light and matter which is very deep and beautiful. Emphasized by the twentieth-century physicist Richard Feynman, it is called the principle of least time, or Fermat's principle.

Let's start with the motion of light that is not interacting with matter at all. In a vacuum, a light ray moves in a straight line. This can be rephrased as follows: of all the conceivable paths light could follow from P to Q, the only one that is physically possible is the path that takes the least time.

What about reflection? If light is going to go from one point to another, being reflected on the way, the quickest path is indeed the one with equal angles of incidence and reflection. If the starting and ending points are equally far from the reflecting surface, (a), it's not hard to convince yourself that this is true, just based on symmetry. There is also a tricky and simple proof, shown in the bottom panel of the figure, for the more general case where the points are at different distances from the surface.

Not only does the principle of least time work for light in a vacuum and light undergoing reflection, we will also see in a later chapter that it works for the bending of light when it passes from one medium into another.

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

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

It should also be noted that it's a misnomer to call this the principle of *least* time. In figure (c), for example, the four physically possible paths by which a ray can return to the center consist of two shortest-time paths and two longest-time paths. Strictly speaking, we should refer to the *principle of least or greatest time*, but most physicists omit the niceties, and assume that other physicists understand that both maxima and minima are possible.

Summary

Selected Vocabulary

- absorption what happens when light hits matter and gives up some of its energy
- reflection what happens when light hits matter and bounces off, retaining at least some of its energy
- specular reflection reflection from a smooth surface, in which the light ray leaves at the same angle at which it came in
- diffuse reflection reflection from a rough surface, in which a single ray of light is divided up into many weaker reflected rays going in many directions
- normal the line perpendicular to a surface at a given point

Notation

- c the speed of light

Summary

We can understand many phenomena involving light without having to use sophisticated models such as the wave model or the particle model. Instead, we simply describe light according to the path it takes, which we call a ray. The ray model of light is useful when light is interacting with material objects that are much larger than a wavelength of light. Since a wavelength of visible light is so short compared to the human scale of existence, the ray model is useful in many practical cases.

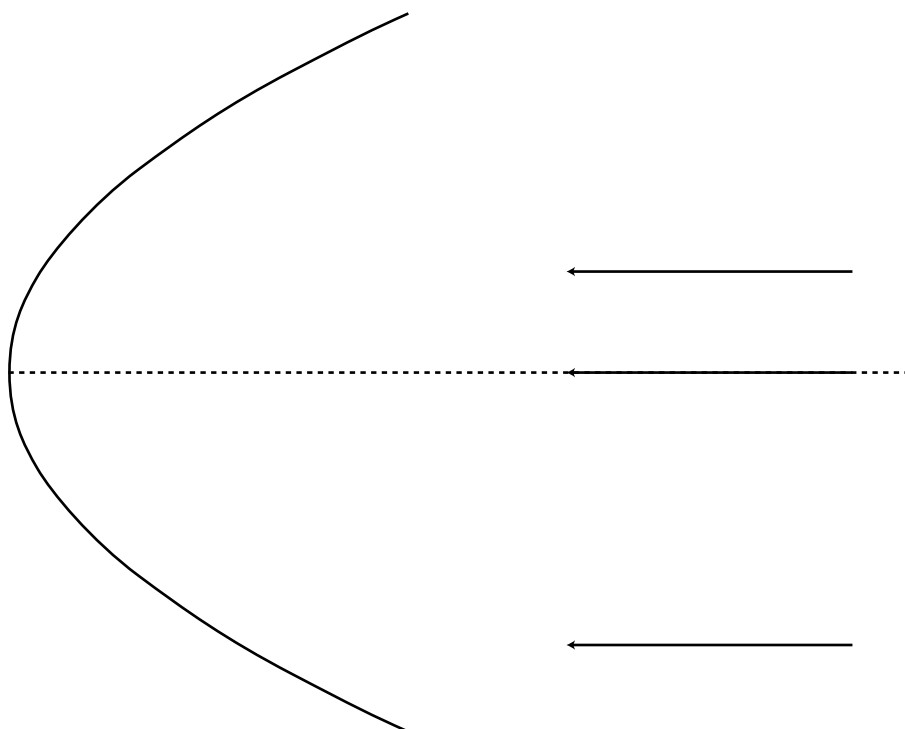
We see things because light comes from them to our eyes. Objects that glow may send light directly to our eyes, but we see an object that doesn't glow via light from another source that has been reflected by the object.

Many of the interactions of light and matter can be understood by considering what happens when light reaches the boundary between two different substances. In this situation, part of the light is reflected (bounces back) and part passes on into the new medium. This is not surprising — it is typical behavior for a wave, and light is a wave. Light energy can also be absorbed by matter, i.e. converted into heat.

A smooth surface produces specular reflection, in which the reflected ray exits at the same angle with respect to the normal as that of the incoming ray. A rough surface gives diffuse reflection, where a single ray of light is divided up into many weaker reflected rays going in many directions.

Homework Problems

1. Draw a ray diagram showing why a small light source (a candle, say) produces sharper shadows than a large one (e.g. a long fluorescent bulb).
2. A Global Positioning System (GPS) receiver is a device that lets you figure out where you are by exchanging radio signals with satellites. It works by measuring the round-trip time for the signals, which is related to the distance between you and the satellite. By finding the ranges to several different satellites in this way, it can pin down your location in three dimensions to within a few meters. How accurate does the measurement of the time delay have to be to determine your position to this accuracy?
3. Estimate the frequency of an electromagnetic wave whose wavelength is similar in size to an atom (about a nm). Referring back to book 4, in what part of the electromagnetic spectrum would such a wave lie (infrared, gamma-rays,...)?
4. The Stealth bomber is designed with flat, smooth surfaces. Why would this make it difficult to detect via radar?
5. The large figure shows a curved (parabolic) mirror, with three parallel light rays coming toward it. One ray is approaching along the mirror's center line. (a) Trace the drawing accurately, and continue the light rays until they are about to undergo their second reflection. To determine the angles accurately, you'll want to draw in the normal at the point where the ray hits the mirror. What do you notice? (b) Make up an example of a practical use for this device. (c) How could you use this mirror with a small lightbulb to produce a parallel beam of light rays going off to the right?



Problem 5.

S A solution is given in the back of the book.

✓ A computerized answer check is available.

★ A difficult problem.

∫ A problem that requires calculus.

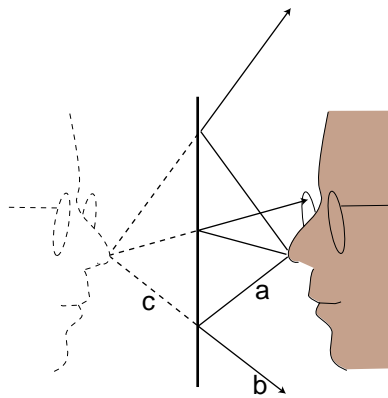


2 Images by Reflection, Part I

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

In this chapter we will study the images formed by flat and curved mirrors on a qualitative, conceptual basis. Although this type of image is not as commonly encountered in everyday life as images formed by lenses, images formed by reflection are simpler to understand, so we discuss them first. In chapter 3 we will turn to a more mathematical treatment of images made by reflection. Surprisingly, the same equations can also be applied to lenses, which are the topic of chapter 4.

2.1 A Virtual Image



(a) An image formed by a mirror.

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

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

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

This is referred to as a *virtual* image, because the rays do not actually cross at the point behind the mirror. They only appear to have originated there.

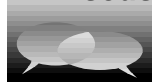
Self-Check



Imagine that the person in figure (a) moves his face down quite a bit — a couple of feet in real life, or a few inches on this scale drawing. Draw a new ray diagram. Will there still be an image? If so, where is it visible from?

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

Discussion Question



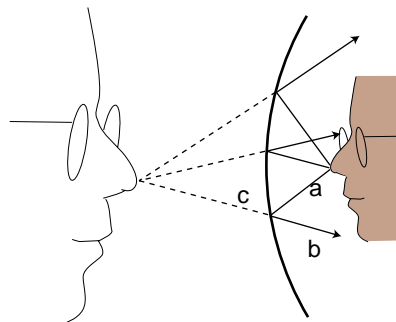
The figure shows an object that is off to one side of a mirror. Draw a ray diagram. Is an image formed? If so, where is it, and from which directions would it be visible?

○



You should have found from your ray diagram that an image is still formed, and it has simply moved down the same distance as the real face. However, this new image would only be visible from high up, and the person can no longer see his own image. If you couldn't draw a ray diagram that seemed to result in an image, the problem was probably that you didn't choose any rays that happened to go away from the face in the right direction to hit the mirror.

2.2 Curved Mirrors

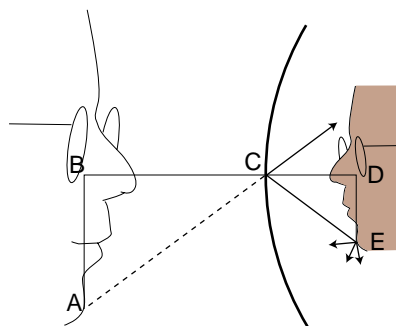


(b) An image formed by a curved mirror.

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

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

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



(c) The image is magnified by the same factor in depth and in its other dimensions.

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

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

2.3 A Real Image

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

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

Note that the rays shown in the figure, which both originated at the same point on the object, reunite when they cross. The point where they cross is the image of the point on the original object. This type of image is called a *real image*, in contradistinction to the virtual images we've studied before. The use of the word "real" is perhaps unfortunate. It sounds as though we are saying the image was an actual material object, which of course it is not.

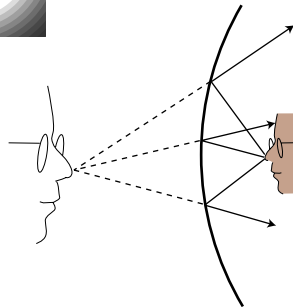
The distinction between a real image and a virtual image is an important one, because a real image can be projected onto a screen or photographic film. If a piece of paper is inserted in figure (e) at the location of the image, the image will be visible to someone looking at the paper from the left. Your eye uses a lens to make a real image on the retina.

Self-Check

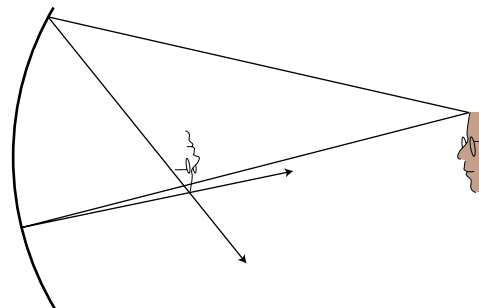


Sketch another copy of the face in figure (e), even farther from the mirror, and draw a ray diagram. What has happened to the location of the image?

(d) A virtual image.

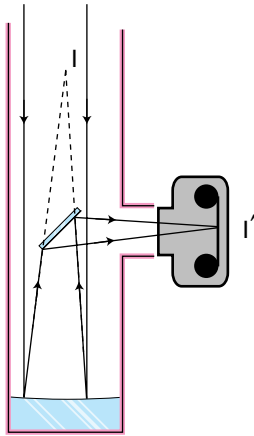


(e) A real image.



Increasing the distance from the face to the mirror has decreased the distance from the image to the mirror. This is the opposite of what happened with the virtual image.

2.4 Images of Images

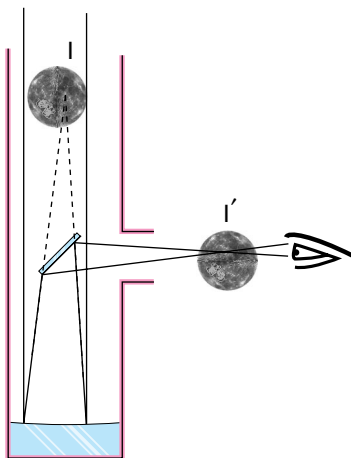


(f) A Newtonian telescope being used with a camera.

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

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

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



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

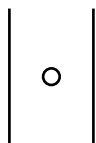
One surprising wrinkle is that whereas a flat mirror used by itself forms a virtual image of an object that is real, here the mirror is forming a real image of virtual image I . This shows how pointless it would be to try to memorize lists of facts about what kinds of images are formed by various optical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

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

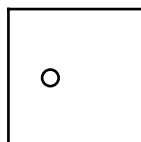


Discussion Questions

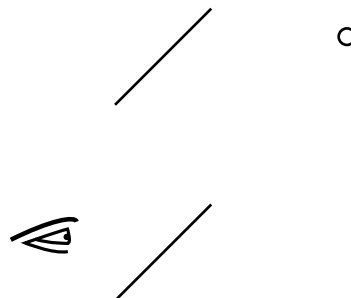
- A.** Describe the images that will be formed of you if you stand between two parallel mirrors.
- B.** Locate the images formed by two perpendicular mirrors, as in the figure. What happens if the mirrors are not perfectly perpendicular?
- C.** Locate the images formed by the periscope.



Discussion question A.



Discussion question B.



Discussion question C.

Summary

Selected Vocabulary

- real image a place where an object appears to be, because the rays diffusely reflected from any given point on the object have been bent so that they come back together and then spread out again from the new point
- virtual image like a real image, but the rays don't actually cross again; they only appear to have come from the point on the image
- inbending describes an optical device that brings light rays closer to the optical axis
- outbending bends light rays farther from the optical axis
- magnification the factor by which an image's linear size is increased (or decreased)
- angular magnification the factor by which an image's apparent angular size is increased (or decreased)

Vocabulary Used in Other Books

- concave describes a surface that is hollowed out like a cave
- convex describes a surface that bulges outward

Notation

- M the magnification of an image
- M_a the angular magnification of an image

Summary

A large class of optical devices, including lenses and flat and curved mirrors, operates by bending light rays to form an image. A real image is one for which the rays actually cross at each point of the image. A virtual image, such as the one formed behind a flat mirror, is one for which the rays only appear to have crossed at a point on the image. A real image can be projected onto a screen; a virtual one cannot.

Mirrors and lenses will generally make an image that is either smaller than or larger than the original object. The scaling factor is called the magnification. In many situations, the angular magnification is more important than the actual magnification.

Homework Problems

- 1 ✓. A man is walking at 1.0 m/s directly towards a flat mirror. At what speed is his separation from his image reducing?
2. If a mirror on a wall is only big enough for you to see yourself from your head down to your waist, can you see your entire body by backing up? Test this experimentally and come up with an explanation for your observations. Note that it is easy to confuse yourself if the mirror is even a tiny bit off of vertical; check whether you are able to see more of yourself both above *and* below.
3. In this chapter we've only done examples of mirrors with hollowed-out shapes (called concave mirrors). Now draw a ray diagram for a curved mirror that has a bulging outward shape (called a convex mirror). (a) How does the image's distance from the mirror compare with the actual object's distance from the mirror? From this comparison, determine whether the magnification is greater than or less than one. (b) Is the image real or virtual? Could this mirror ever make the other type of image?
4. As discussed in question 3, there are two types of curved mirrors, concave and convex. Make a list of all the possible combinations of types of images (virtual or real) with types of mirrors (concave and convex). (Not all of the four combinations are physically possible.) Now for each one, use ray diagrams to determine whether increasing the distance of the object from the mirror leads to an increase or a decrease in the distance of the image from the mirror.
5. If the user of an astronomical telescope moves her head closer to or farther away from the image she is looking at, does the magnification change? Does the angular magnification change? Explain.



Breakfast Table, by Willem Claesz. de Heda, 17th century. A variety of images occur in the painting, some distorted, as a result of both reflection and refraction (ch. 4).

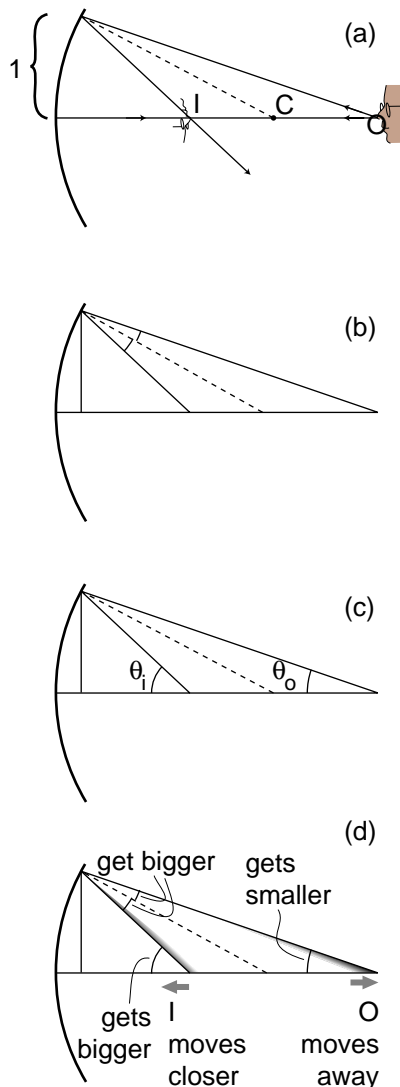
3 Images by Reflection, Part II

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

3.1 A Real Image Formed by an Inbending Mirror

Location of the image

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



To locate an image, we need to track a minimum of two rays coming from the same point. Since we have proved in the previous chapter that this type of image is not distorted, we can use an on-axis point, O , on the object, as in figure (a). The results we derive will also hold for off-axis points, since otherwise the image would have to be distorted, which we know is not true. We let one of the rays be the one that is emitted along the axis; this ray is especially easy to trace, because it bounces straight back along the axis again. As our second ray, we choose one that strikes the mirror at a distance of 1 from the axis. “One what?” asks the astute reader. The answer is that it doesn't really matter. When a mirror has shallow curvature, all the reflected rays hit the same point, so 1 could be expressed in any units you like. It could, for instance, be 1 cm, unless your mirror is smaller than 1 cm!

The only way to find out anything mathematical about the rays is to use the sole mathematical fact we possess concerning specular reflection: the incident and reflected rays form equal angles with respect to the normal, which is shown as a dashed line. Therefore the two angles shown in figure (b) are the same, and skipping some straightforward geometry, this leads to the visually reasonable result that the two angles in figure (c) are related as follows:

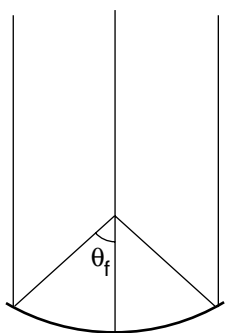
$$\theta_i + \theta_o = \text{constant}$$

Suppose, for example, that we move O farther from the mirror. The top angle in figure (b) is increased, so the bottom angle must increase by the same amount, causing the image point, I , to move closer to the mirror. In terms of the angles shown in figure (c), the more distant object has resulted in a smaller angle θ_o , while the closer image corresponds to a larger θ_i ; One angle increases by the same amount that the other decreases, so their sum remains constant. These changes are summarized in figure (d).

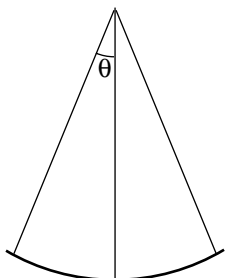
The sum $\theta_i + \theta_o$ is a constant. What does this constant represent? Geometrically, we interpret it as double the angle made by the dashed radius line. Optically, it is a measure of the strength of the mirror, i.e., how strongly the mirror focuses light, and so we call it the focal angle,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

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

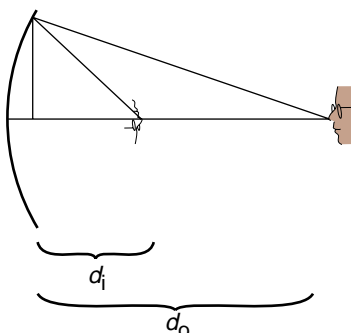
Suppose, for example, that we wish to use a quick and dirty optical test to determine how strong a particular mirror is. We can lay it on the floor as shown in figure (e), and use it to make an image of a lamp mounted on the ceiling overhead, which we assume is very far away compared to the radius of curvature of the mirror, so that the mirror intercepts only a very narrow



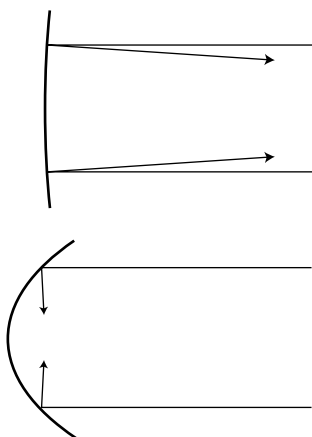
(e) The geometric interpretation of the focal angle.



(f) An alternative test for finding the focal angle. The mirror is the same as in figure (d),



(g) The object and image distances.



(h) The top mirror has a shallower curvature, a longer focal length, and a smaller focal angle. It reflects rays at angles not much different from those that would be produced with a flat mirror.

cone of rays from the lamp. This cone is so narrow that its rays are nearly parallel, and θ_o is nearly zero. The real image can be observed on a piece of paper. By moving the paper nearer and farther, we can bring the image into focus, at which point we know the paper is located at the image point. Since $\theta_o \approx 0$, we have $\theta_i \approx \theta_f$ and we can then determine this mirror's focal angle either by measuring θ_i directly with a protractor, or indirectly via trigonometry. A strong mirror will bring the rays to a focal point close to the mirror, and these rays will form a blunt-angled cone with a large θ_i and θ_f .

Example: An alternative optical test

Question: Figure (f) shows an alternative optical test. Rather than placing the object at infinity as in figure (e), we adjust it so that the image is right on top of the object. Points O and I coincide, and the rays are reflected right back on top of themselves. If we measure the angle θ shown in figure (f), how can we find the focal angle?

Solution: The object and image angles are the same; the angle labeled θ in the figure equals both of them. We therefore have $\theta_i + \theta_o = 2\theta = \theta_f$. Comparing figures (e) and (f), it is indeed plausible that the angles are related by a factor of two.

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

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

$$\tan \theta_o = 1/d_o \quad \tan \theta_i = 1/d_i$$

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

$$\theta_o = 1/d_o \quad \theta_i = 1/d_i$$

We likewise define a distance called the focal length, f according to $\theta_f = 1/f$. In figure (e), f is the distance from the mirror to the focal point. We can now reexpress the equation relating the object and image positions as

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

Figure (h) summarizes the interpretation of the focal length and focal angle.

Which form is better, $\theta_f = \theta_i + \theta_o$ or $1/f = 1/d_i + 1/d_o$? The angular form has in its favor its simplicity and its straightforward visual interpretation, but there are two reasons why we might prefer the second version. First, the numerical values of the angles depend on what we mean by "one unit" for the distance shown as 1 in figure (a). Second, it is usually easier to measure distances rather than angles, so the distance form is more convenient for number crunching. Neither form is superior overall, and we will often need to use both to solve any given problem. (I would like to thank Fouad Ajami

for pointing out the pedagogical advantages of using both equations side by side.)

Example: A searchlight

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

$$\begin{aligned} d_o &= (1/f - 1/d_i)^{-1} \\ &= (1/f - 0)^{-1} \\ &= (1/f)^{-1} \\ &= f \end{aligned}$$

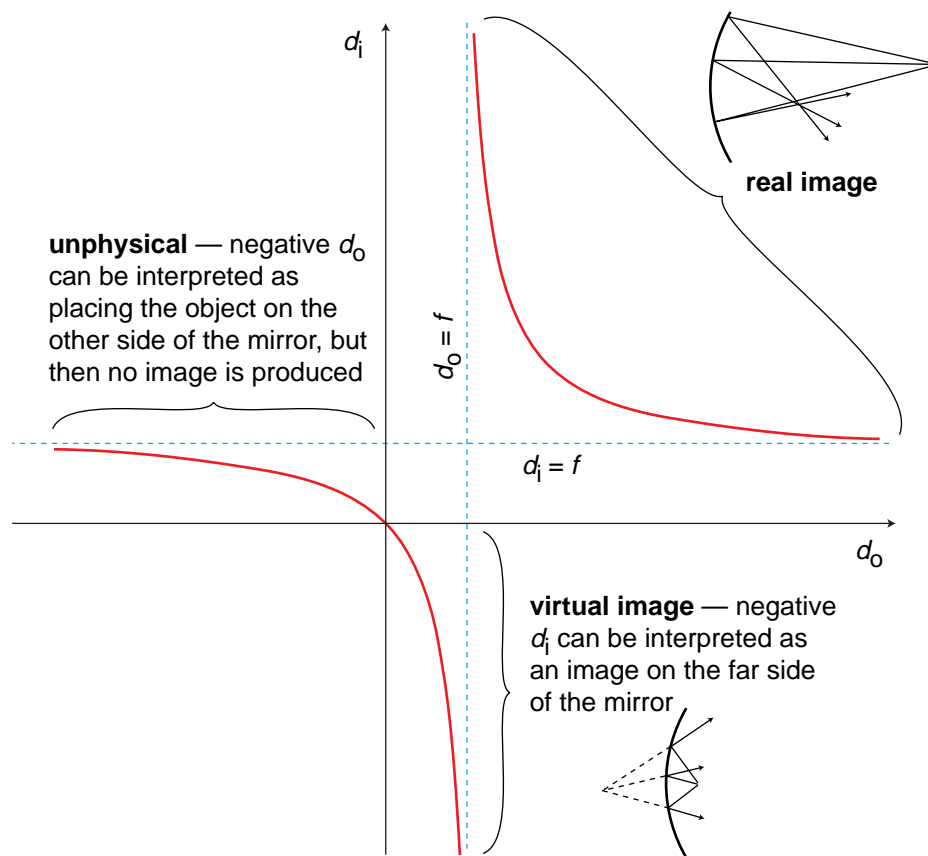
The bulb has to be placed at a distance from the mirror equal to its focal point.

Example: Diopters

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

Magnification

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



3.2 Other Cases With Curved Mirrors

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

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

To analyze the distance equation, let's look at a graph of d_i as a function of d_o . The branch on the upper right corresponds to the case of a real image. Strictly speaking, this is the only part of the graph that we've proven corresponds to reality, since we never did any geometry for other cases, such as virtual images. As discussed in the previous section, making d_o bigger causes d_i to become smaller, and vice-versa.

Letting d_o be less than f is equivalent to $\theta_o > \theta_f$: a virtual image is produced on the far side of the mirror. This is the first example of Wigner's "unreasonable effectiveness of mathematics" that we have encountered in

optics. Even though our proof depended on the assumption that the image was real, the equation we derived turns out to be applicable to virtual images, provided that we either interpret the positive and negative signs in a certain way, or else modify the equation to have different positive and negative signs.

Self-Check

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

Example: A flat mirror

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

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

or

$$d_o = -d_i .$$

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

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

$$\theta_i = \theta_1 + \theta_o$$

and

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

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

At the top of the graph, d_i approaches infinity when d_o approaches f ; interpretation: the rays just barely converge to the right of the mirror. On the far right, d_i approaches f as d_o approaches infinity; this is the definition of the focal length. At the bottom, d_i approaches negative infinity when d_o approaches f from the other side; interpretation: the rays don't quite converge on the right side of the mirror, so they appear to have come from a virtual image point very far to the left of the mirror.

As you have gathered by now, I have a method I think is better, and which I'll use throughout the rest of this book. In this method, all distances and angles are *positive by definition*, and we put in positive and negative signs in the *equations* depending on the situation. Rather than memorizing these signs, we start with the generic equations

$$\theta_f = \pm \theta_i \pm \theta_o$$

$$\frac{1}{f} = \pm \frac{1}{d_i} \pm \frac{1}{d_o} ,$$

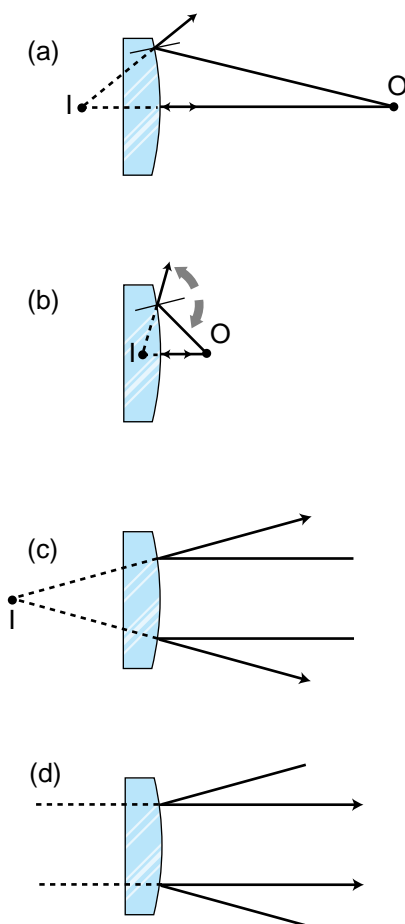
and then determine the signs by a two-step method that depends on ray diagrams. There are really only two signs to determine, not four; the signs in the two equations match up in the way you'd expect. The method is as follows:

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

2. It is normally only physically possible for either θ_o or θ_i to be zero, not both. This corresponds to either an object at infinity or an image at infinity. (Of course it is always possible to put an object at infinity, but that might for instance result in the formation of a real image, when you are interested in the case of a virtual image.) If we imagine the case where that angle is zero, then the only term on the right side of the angle equation would be the one that has the other variable in it. Since the left-hand side of the equation is positive by definition, the term on the right that we didn't eliminate must be the one that has a plus sign.

In step 1, many students have trouble drawing the ray diagram correctly. For simplicity, you should always do your diagram for a point on the object that is on the axis of the mirror, and let one of your rays be the one that is emitted along the axis and reflect straight back on itself, as in the figures in section 3.1. As shown in figure (d) in section 3.1, there are four angles involved: two at the mirror, one at the object (θ_o), and one at the image (θ_i). Make sure to draw in the normal to the mirror so that you can see the two angles at the mirror. These two angles are equal, so as you change the object position, they fan out or fan in, like opening or closing a book. Once you've drawn this effect, you should easily be able to tell whether θ_o and θ_i change in the same way or in opposite ways.

Although focal lengths are always positive in the method used in this book, you should be aware that out-bending mirrors and lenses are assigned negative focal lengths in the other method, so if you see a lens labeled $f = -30$ cm, you'll know what it means.



Example: An anti-shoplifting mirror

Question: Convenience stores often install an out-bending mirror so that the clerk has a view of the whole store and can catch shoplifters. Use a ray diagram to show that the image is reduced, bringing more into the clerk's field of view. If the focal length of the mirror is 3.0 m, and the mirror is 7.0 m from the farthest wall, how deep is the image of the store? (Note that in the other method of handling the signs, the focal length would have been given as -3.0 m.)

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

We now apply the method outlined above for determining the plus and minus signs. Step 1: The object is the point on the opposite wall. As an experiment, (b), we try making the object closer — *much much* closer, so that even if our drawing isn't perfectly accurate we'll still get the right result for the change in the image's location. (I did these drawings using illustration software, but if you were doing them by hand, you'd also want to make much larger ones for greater accuracy.) The two angles at the mirror fan out from the normal. Increasing θ_o has clearly made θ_i larger as well. (All four angles got bigger.) There must be a cancellation of the effects of changing the two terms on the right in the same way, and the only way to get such a cancellation is if the two terms in the angle equation have *opposite signs*:

$$\theta_i = +\theta_i - \theta_o$$

or

$$\theta_i = -\theta_i + \theta_o$$

Step 2: Now which is the positive term and which is negative? Figure (c) shows a perfectly reasonable ray diagram of an image formed of an object at infinity, the moon, for example. We have $\theta_o = 0$, since there is no angle between the rays arriving at the mirror from the object. Figure (d) shows an attempt to make an image at infinity. To get an image at infinity, we would have had to start with a converging set of rays, which is not physically possible, since diffuse reflection creates diverging rays. If θ_o can be zero, then the sign of the θ_i term must be positive.



An outbending mirror in the shape of a sphere. The image is reduced (minified), and is also distorted because the mirror's curve is not shallow.

We have now determined that the form of the angle equation must be

$$\theta_i = \theta_o - \theta_i,$$

and the signs of the distance equation must behave the same way:

$$\frac{1}{f} = \frac{1}{d_i} - \frac{1}{d_o}.$$

Solving for d_i , we find

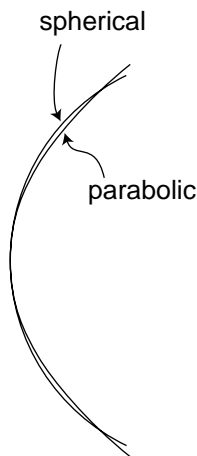
$$d_i = \left(\frac{1}{f} + \frac{1}{d_o} \right)^{-1} = 2.1 \text{ m}.$$

The image of the store is reduced by a factor of $2.1/7.0=0.3$, i.e. it is smaller by 70%.

Example: A shortcut for real images

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

3.3* Aberrations

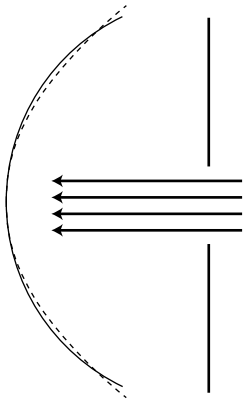


Spherical mirrors are the cheapest to make, but parabolic mirrors are better for making images of objects at infinity. A sphere has equal curvature everywhere, but a parabola has tighter curvature at its center and gentler curvature at the sides.

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

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

(2) An object on the axis of the lens or mirror may be imaged correctly, but off-axis objects may be out of focus. In a camera, this type of aberration would show up as a fuzziness near the sides of the picture when the center was perfectly focused.



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

One way of decreasing aberration is to use a small-diameter mirror or lens, or block most of the light with an opaque screen with a hole in it, so that only light that comes in close to the axis can get through. Either way, we are using a smaller portion of the lens or mirror whose curvature will be more shallow, thereby making the shallow-mirror (or thin-lens) approximation more accurate. Your eye does this by narrowing down the pupil to a smaller hole. In a camera, there is either an automatic or manual adjustment, and narrowing the opening is called “stopping down.” The disadvantage of stopping down is that light is wasted, so the image will be dimmer or a longer exposure must be used.

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

Summary

Selected Vocabulary

focal length..... a property of a lens or mirror, equal to the distance from the lens or mirror to the image it forms of an object that is infinitely far away

Notation

f the focal length
 d_o the distance of the object from the mirror (technically from the plane tangent to the center of the mirror, although this seldom matters much for a mirror whose curve is shallow)
 d_i the distance of the image from the mirror
 θ_f the focal angle, defined as $1/f$
 θ_o the object angle, defined as $1/d_o$
 θ_i the image angle, defined as $1/d_i$

Notation Used in Other Books

$f > 0$ describes an inbending lens or mirror; in this book, all focal lengths are positive, so there is no such implication
 $f < 0$ describes an out-bending lens or mirror; in this book, all focal lengths are positive
 $M < 0$ indicates an inverted image

Summary

Every lens or mirror has a property called the focal length, which is defined as the distance from the lens or mirror to the image it forms of an object that is infinitely far away. A stronger lens or mirror has a shorter focal length.

The relationship between the locations of an object and its image formed by a lens or mirror can always be expressed by equations of the form

$$\theta_i = \pm \theta_i \pm \theta_o$$
$$\frac{1}{f} = \pm \frac{1}{d_i} \pm \frac{1}{d_o} ,$$

The choice of plus and minus signs depends on whether we are dealing with a lens or a mirror, whether the lens or mirror is inbending or outbending, and whether the image is real or virtual. A method for determining the plus and minus signs is as follows:

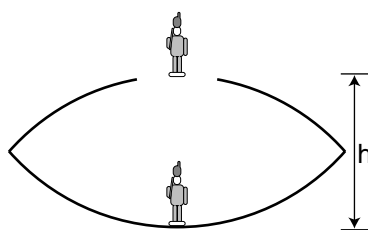
1. Use ray diagrams to decide whether θ_i and θ_o vary in the same way or in opposite ways. Based on this, decide whether the two signs in the equation are the same or opposite. If the signs are opposite, go on to step 2 to determine which is positive and which is negative.
2. It is normally only physically possible for either θ_i or θ_o to be zero, not both. Imagine the case where that variable is zero. Since the left-hand side of the equation is positive by definition, the term on the right that we didn't eliminate must be the one that has a plus sign.

Once the correct form of the equation has been determined, the magnification can be found via the equation

$$M = \frac{d_i}{d_o} .$$

Homework Problems

1. Apply the equation $M=d_i/d_o$ to the case of a flat mirror.
- 2 S. Use the method described in the text to derive the equation relating object distance to image distance for the case of a virtual image produced by an inbending mirror.
3. (a) Make up a numerical example of a virtual image formed by an inbending mirror with a certain focal length, and determine the magnification. (You will need the result of the previous problem.) Now change the location of the object *a little bit* and redetermine the magnification, showing that it changes. At my local department store, the cosmetics department sells mirrors advertised as giving a magnification of 5 times. How would you interpret this?
(b ★) Suppose a Newtonian telescope is being used for astronomical observing. Assume for simplicity that no eyepiece is used, and assume a value for the focal length of the mirror that would be reasonable for an amateur instrument that is to fit in a closet. Is the angular magnification different for the moon than for a distant galaxy?
4. (a) Find a case where the magnification of a curved mirror is infinite. Is the *angular* magnification infinite from any realistic viewing position? (b) Explain why an arbitrarily large magnification can't be achieved by having a sufficiently small value of d_o .



Problem 5.

- 5 ★. The figure shows a device for constructing a realistic optical illusion. Two mirrors of equal focal length are put against each other with their silvered surfaces facing inward. A small object placed in the bottom of the cavity will have its image projected in the air above. The way it works is that the top mirror produces a virtual image, and the bottom mirror then creates a real image of the virtual image. (a) Show that if the image is to be positioned as shown, at the mouth of the cavity, then the focal length of the mirrors is related to the dimension h via the equation

$$\frac{1}{f} = \frac{1}{h} + \frac{1}{h + \left(\frac{1}{h} - \frac{1}{f}\right)^{-1}}.$$

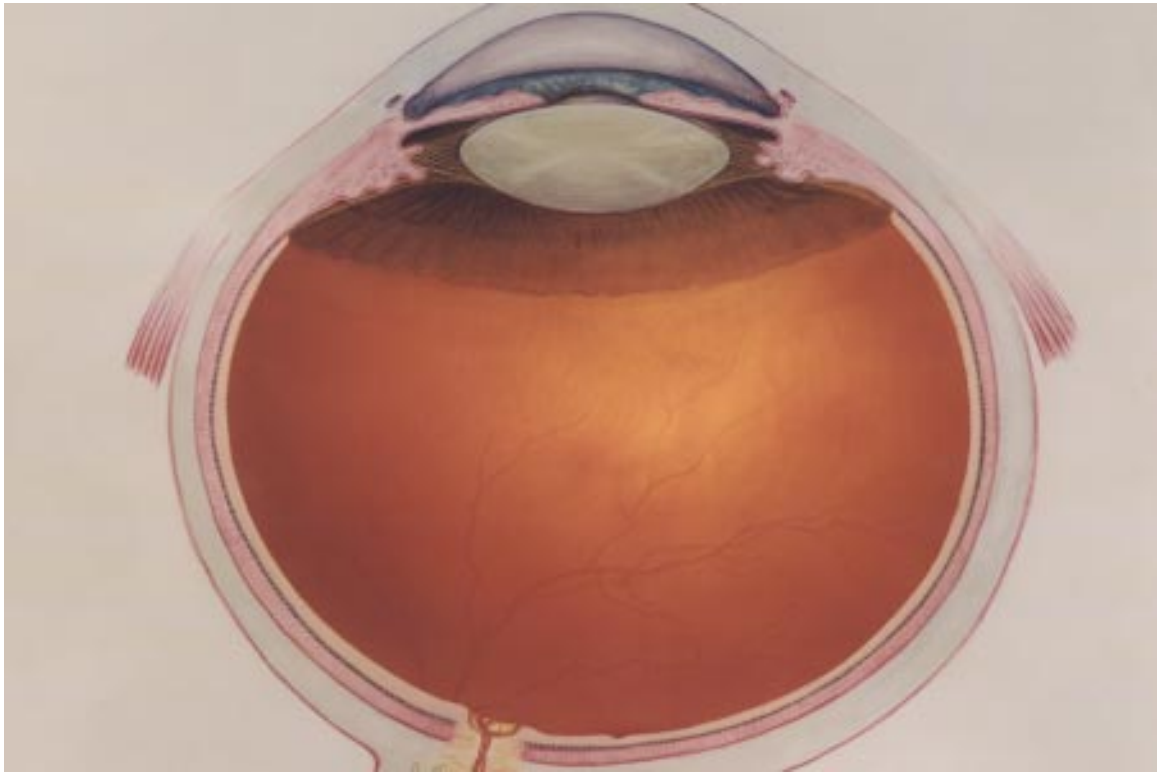
- (b) Restate the equation in terms of a single variable $x=h/f$, and show that there are two solutions for x . Which solution is physically consistent with the assumptions of the calculation?
6. A hollowed-out surface that reflects sound waves can act just like an inbending mirror. Suppose that, standing near such a surface, you are able to find point where you can place your head so that your own whispers are focused back on your head, so that they sound loud to you. Given your distance to the surface, what is the surface's focal length?

S A solution is given in the back of the book.

✓ A computerized answer check is available.

★ A difficult problem.

∫ A problem that requires calculus.



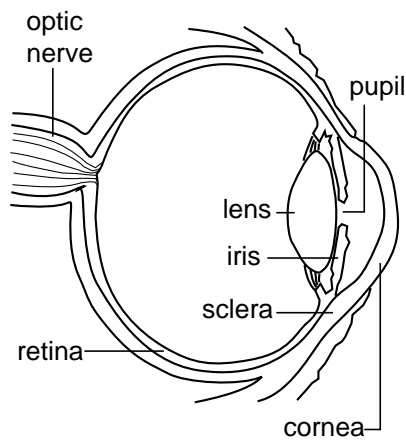
4 Refraction and Images

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

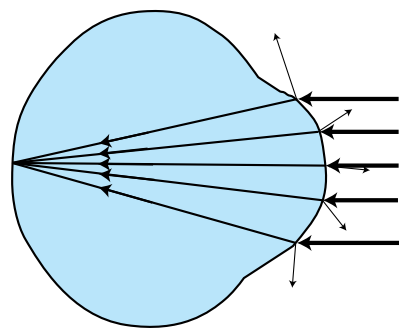
So great is the value attached to sight that some have imbued it with mystical aspects. Moses “had vision,” George Bush did not. Christian fundamentalists who perceive a conflict between evolution and their religion have claimed that the eye is such a perfect device that it could never have arisen through a process as helter-skelter as evolution, or that it could not have evolved because half of an eye would be useless. In fact, the structure of an eye is fundamentally dictated by physics, and it has arisen separately by evolution somewhere between eight and 40 times, depending

on which biologist you ask. We humans have a version of the eye that can be traced back to the evolution of a light-sensitive “eye spot” on the head of an ancient invertebrate. A sunken pit then developed so that the eye would only receive light from one direction, allowing the organism to tell where the light was coming from. (Modern flatworms have this type of eye.) The top of the pit then became partially covered, leaving a hole, for even greater directionality (as in the nautilus). At some point the cavity became filled with jelly, and this jelly finally became a lens, resulting in the general type of eye that we share with the bony fishes and other vertebrates. Far from being a perfect device, the vertebrate eye is marred by a serious design flaw due to the lack of planning or intelligent design in evolution: the nerve cells of the retina and the blood vessels that serve them are all in front of the light-sensitive cells, blocking part of the light. Squids and other molluscs, whose eyes evolved on a separate branch of the evolutionary tree, have a more sensible arrangement, with the light-sensitive cells out in front.

4.1 Refraction



(a) The anatomy of the human eye.
(After an uncopyrighted diagram by the National Eye Institute, NIH.)



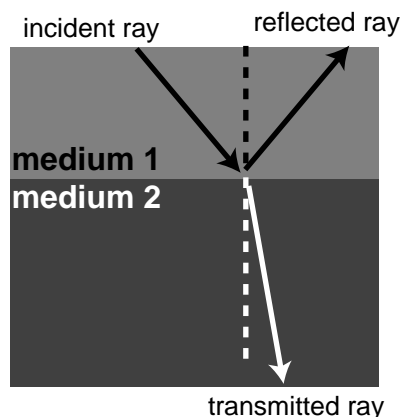
(b) A simplified optical diagram of the eye. Light rays are bent when they cross from the air into the eye.

Refraction

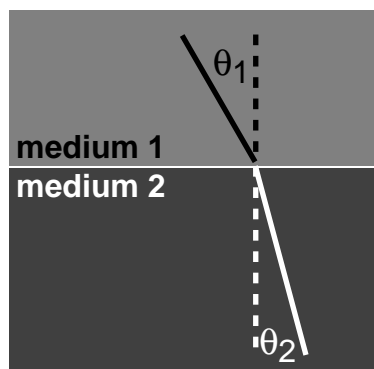
The fundamental physical phenomenon at work in the eye is that when light crosses a boundary between two media (such as air and the eye’s jelly), part of its energy is reflected, but part passes into the new medium. In the ray model of light, we describe the original ray as splitting into a reflected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

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

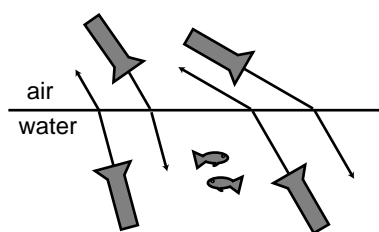
A common misconception is that the “lens” of the eye is what does the focusing. All the transparent parts of the eye are made of fairly similar stuff, so the dramatic change in medium is when a ray crosses from the air into the eye (at the outside surface of the cornea). This is where nearly all the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



(c) The incident, reflected, and refracted rays all lie in a plane that includes the normal (dashed line).



(d) The angles θ_1 and θ_2 are related to each other, and also depend on the properties of the two media. Because refraction is time-reversal symmetric, there is no need to label the rays with arrowheads.



(e) Refraction has time-reversal symmetry. Regardless of whether the light is going in or out of the water, the relationship between the two angles is the same, and the ray is closer to the normal while in the water.

Refractive properties of media

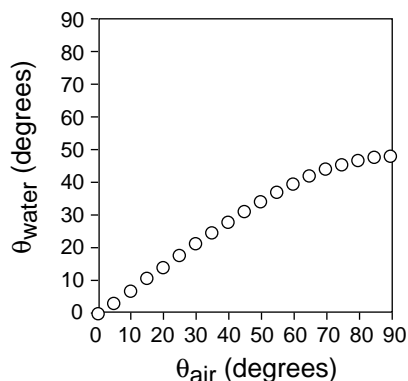
What are the rules governing refraction? The first thing to observe is that just as with reflection, the new, bent part of the ray lies in the same plane as the normal (perpendicular) and the incident ray, (c).

If you try shooting a beam of light at the boundary between two substances, say water and air, you'll find that regardless of the angle at which you send in the beam, the part of the beam in the water is always closer to the normal line, (d). It doesn't matter if the ray is entering the water or leaving, so refraction is symmetric with respect to time-reversal, (e).

If, instead of water and air, you try another combination of substances, say plastic and gasoline, again you'll find that the ray's angle with respect to the normal is consistently smaller in one and larger in the other. Also, we find that if substance A has rays closer to normal than in B, and B has rays closer to normal than in C, then A has rays closer to normal than C. This means that we can rank-order all materials according to their refractive properties. Isaac Newton did so, including in his list many amusing substances, such as "Danzig vitriol" and "a pseudo-topazius, being a natural, pellucid, brittle, hairy stone, of a yellow color." Several general rules can be inferred from such a list:

- Vacuum lies at one end of the list. In refraction across the interface between vacuum and any other medium, the other medium has rays closer to the normal.
- Among gases, the ray gets closer to the normal if you increase the density of the gas by pressurizing it more.
- The refractive properties of liquid mixtures and solutions vary in a smooth and systematic manner as the proportions of the mixture are changed.
- Denser substances usually, but not always, have rays closer to the normal.

The second and third rules provide us with a method for measuring the density of an unknown sample of gas, or the concentration of a solution. The latter technique is very commonly used, and the CRC Handbook of Physics and Chemistry, for instance, contains extensive tables of the refractive properties of sugar solutions, cat urine, and so on.



Snell's law

The numerical rule governing refraction was discovered by Snell, who must have collected experimental data something like what is shown on this graph and then attempted by trial and error to find the right equation. The equation he came up with was

$$\frac{\sin \theta_1}{\sin \theta_2} = \text{constant} .$$

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

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

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

but rewriting it in the form

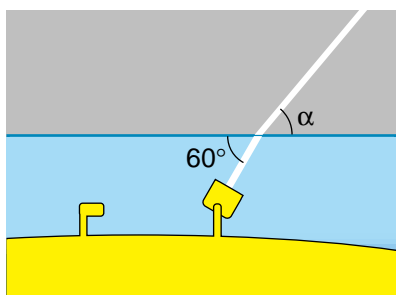
$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad \text{[relationship between angles of rays at the interface between media with indices of refraction } n_1 \text{ and } n_2; \text{ angles are defined with respect to the normal]}$$

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

Self-Check



- (1) What would the graph look like for two substances with the same index of refraction?
 - (2) Based on the graph, when does refraction at an air-water interface change the direction of a ray most strongly?
- [Answers on next page.]



Example

Question: A submarine shines its searchlight up toward the surface of the water. What is the angle θ shown in the figure?

Solution: The tricky part is that Snell's law refers to the angles with respect to the normal. Forgetting this is a very common mistake. The beam is at an angle of 30° with respect to the normal in the water. Let's call the air medium 1 and the water medium 2. Solving Snell's law for θ_1 , we find

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

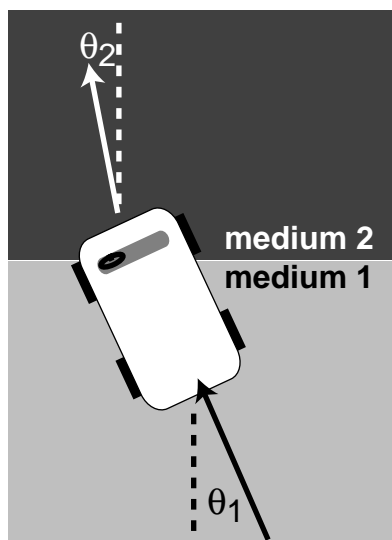
As mentioned above, air has an index of refraction very close to 1, and water's is about 1.3, so we find $\theta_1 = 40^\circ$. The angle α is therefore 50° .

The index of refraction is related to the speed of light

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

$$n = c/v \quad [n = \text{medium's index of refraction, } v = \text{speed of light in that medium, } c = \text{speed of light in a vacuum}]$$

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



A mechanical model of Snell's law

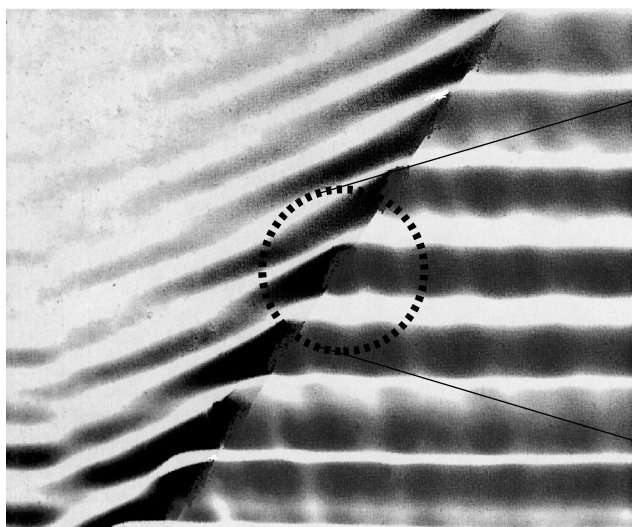
Why should refraction be related to the speed of light? The mechanical model shown in the figure may help to make this more plausible. Suppose medium 2 is thick, sticky mud, which slows down the car. The car's right wheel hits the mud first, causing the right side of the car to slow down. This will cause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

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

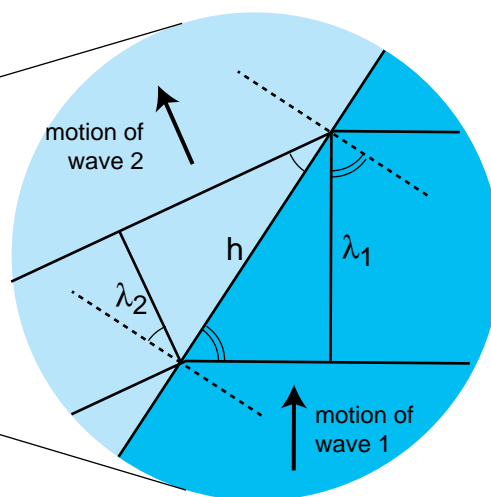


(1) If n_1 and n_2 are equal, Snell's law becomes $\sin\theta_1 = \sin\theta_2$, which implies $\theta_1 = \theta_2$. The graph would be a straight line along the diagonal of the graph.

(2) The graph is farthest from the diagonal when the angles are large, i.e. when the ray strikes the interface at an oblique or grazing angle.



(a) Refraction of a water wave. The water in the upper left part of the tank is shallower, so the speed of the waves is slower there, and their wavelength is shorter. The reflected part of the wave is also very faintly visible. Retouched from an uncopyrighted PSSC College Physics photograph.



(b) A close-up view of what happens at the interface between the deeper medium and the shallower medium. The dashed lines are normals to the interface. The two marked angles on the right side are both equal to θ_1 , and the two on the left equal θ_2 .

A derivation of Snell's law

However intuitively appealing the mechanical model may be, light is a wave, and we should be using wave models to describe refraction. In fact Snell's law can be derived quite simply from wave concepts. In figure (b), trigonometry gives

$$\sin \theta_1 = \lambda_1 / h \quad \text{and}$$

$$\sin \theta_2 = \lambda_2 / h \quad .$$

Eliminating h by dividing the equations, we find

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{\lambda_1}{\lambda_2} \quad .$$

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

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

which is one form of Snell's law.

Example: Ocean waves near and far from shore

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

Color and refraction

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

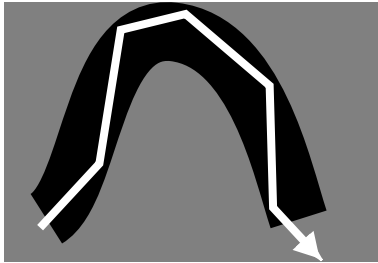
How much light is reflected, and how much is transmitted?

In book 3 we developed an equation for the percentage of the wave energy that is transmitted and the percentage reflected at a boundary between media. This was only done in the case of waves in one dimension, however, and rather than discuss the full three dimensional generalization it will be more useful to go into some qualitative observations about what happens. First, reflection happens only at the interface between two media, and two media with the same index of refraction act as if they were a single medium. Thus, at the interface between media with the same index of refraction, there is no reflection, and the ray keeps going straight. Continuing this line of thought, it is not surprising that we observe very little reflection at an interface between media with similar indices of refraction.

The next thing to note is that it is possible to have situations where no possible angle for the refracted ray can satisfy Snell's law. Solving Snell's law for θ_2 , we find

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

and if n_1 is greater than n_2 , then there will be large values of θ_1 for which the quantity $(n_1/n_2) \sin \theta$ is greater than one, meaning that your calculator will flash an error message at you when you try to take the inverse sine. What can happen physically in such a situation? The answer is that all the light is reflected, so there is no refracted ray. This phenomenon is known as *total internal reflection*, and is used in the fiber-optic cables that nowadays carry almost all long-distance telephone calls. The electrical signals from your phone travel to a switching center, where they are converted from electricity into light. From there, the light is sent across the country in a thin transparent fiber. The light is aimed straight into the end of the fiber, and as long as the fiber never goes through any turns that are too sharp, the light will always encounter the edge of the fiber at an angle sufficiently oblique to give total internal reflection. If the fiber-optic cable is thick enough, one can see an image at one end of whatever the other end is pointed at.



Total internal reflection in a fiber-optic cable.

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

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

Discussion questions



- A. What index of refraction should a fish have in order to be invisible?
- B. Does a surgeon using an endoscope need a source of light inside the body cavity? If so, how could this be done without inserting a light bulb through the incision?
- C. A denser sample of a gas has a higher index of refraction than a less dense sample (i.e. a sample under lower pressure), but why would it not make sense for the index of refraction of a gas to be proportional to density?
- D. The earth's atmosphere gets thinner and thinner as you go higher in altitude. If a ray of light comes from a star that is below the zenith, what will happen to it as it comes into the earth's atmosphere?
- E. Does total internal reflection occur when light in a denser medium encounters a less dense medium, or the other way around? Or can it occur in either case?

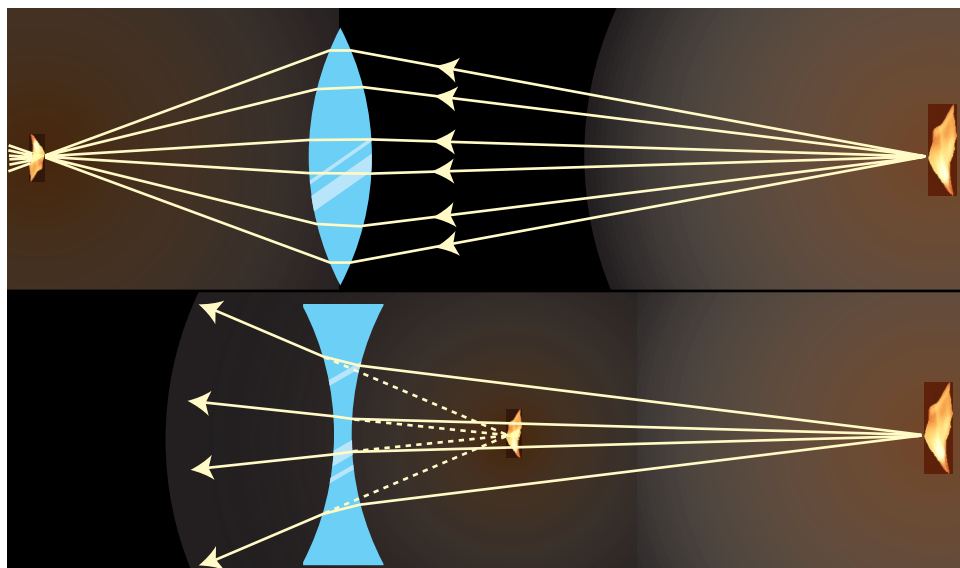
4.2 Lenses

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

Self-Checks



- (1) In figures (a) and (b), classify the images as real or virtual.
- (2) Glass has an index of refraction that is greater than that of air. Consider the topmost ray in figure (a). Explain why the ray makes a slight left turn upon entering the lens, and another left turn when it exits.
- (3) If the flame in figure (b) was moved closer to the lens, what would happen to the location of the image?



(a) An inbending lens is making an image of a candle flame.

(b) Now an outbending lens is making an image of the flame.



- (1) In (a), the rays cross at the image, so it is real. In (b), the rays only appear to have come from the image point, so the image is virtual.
- (2) A ray is always closer to the normal in the medium with the higher index of refraction. The first left turn makes the ray closer to the normal, as it should be in glass. The second left turn makes the ray farther from the normal, which is how it should be in air.
- (3) Take the topmost ray as an example. It will still take two right turns, but since it is entering the lens at a steeper angle, it will also leave at a steeper angle. Tracing backward to image, the steeper lines will meet closer to the lens.



(c) Two images of a rose created by the same lens and recorded with the same camera.

Discussion Questions



- A.** In figures (a) and (b), the front and back surfaces are parallel to each other at the center of the lens. What will happen to a ray that enters near the center, but not necessarily along the axis of the lens?
- B.** Suppose you wanted to change the setup in figure (a) so that the location of the actual flame in the figure would instead be occupied by an image of a flame. Where would you have to move the candle to achieve this? What about in (b)?
- C.** There are three qualitatively different types of image formation that can occur with lenses, of which figures (a) and (b) exhaust only two. Figure out what the third possibility is. Which of the three possibilities can result in a magnification greater than one?
- D.** Classify the examples shown in figure (c) according to the types of images delineated in the previous discussion question.
- E.** In figures (a) and (b), the only rays drawn were those that happened to enter the lenses. Discuss this in relation to figure (c).
- F.** In the right-hand side of figure (c), the image viewed through the lens is in focus, but the side of the rose that sticks out from behind the lens is not. Why?
- G.** In general, the index of refraction depends on the color of the light. What effect would this have on images formed by lenses?

4.3* The Lensmaker's Equation

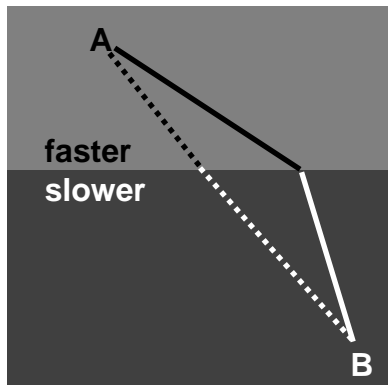
The focal length of a spherical mirror is simply $r/2$, but we cannot expect the focal length of a lens to be given by pure geometry, since it also depends on the index of refraction of the lens. Suppose we have a lens whose front and back surfaces are both spherical. (This is no great loss of generality,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

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

This is known as the lensmaker's equation. In my opinion it is not particularly worthy of memorization. The positive sign is used when both surfaces are curved outward or both are curved inward; otherwise a negative sign applies. The proof of this equation is left as an exercise to those readers who are sufficiently brave and motivated.

4.4* Refraction and the Principle of Least Time

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



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 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,

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad [\text{angles measured with respect to the normal}]$$

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,

$$n = c/v ,$$

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 the third book of this series, 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.

Homework Problems

1. Suppose an inbending 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?
2. There are two main types of telescopes, refracting (using lenses) and reflecting (using mirrors). (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.) What implications would the color-dependence of focal length have for the relative merits of the two types of telescopes? What would happen with white starlight, for example?
3. Based on Snell's law, explain why rays of light passing through the edges of an inbending 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?
4. By changing the separation distance between lens and film, a camera can focus on subjects at a variety of distances. Suppose the proper lens-film separation for taking an in-focus picture of a distant object such as the moon is x . To take an in-focus picture of a nearby object, will the proper lens-film separation be greater than, equal to, or less than x ? Explain using diagrams. [Based on a problem by Eugene Hecht.]
- 5 ★. (a) Light is being reflected diffusely from an object 1.000 m under water. 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° and 20.000° 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?

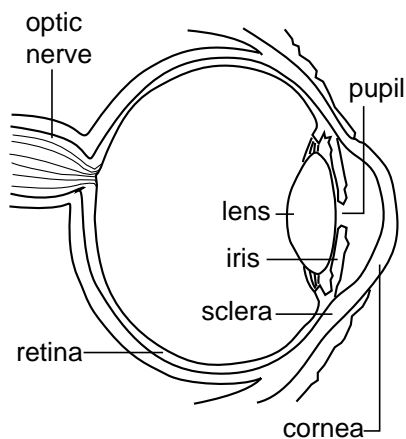
(c) In the case where the angles are all small, use algebra and trig to show that the refracted rays do appear to come from the same point, and find an equation for the depth of the virtual image. Do not put in any numerical values for the angles or for the indices of refraction — just keep them as symbols.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

S A solution is given in the back of the book.

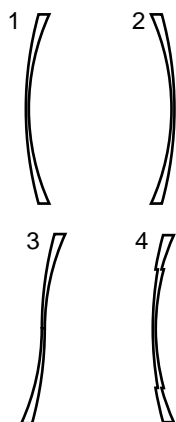
✓ A computerized answer check is available.

★ A difficult problem.

∫ A problem that requires calculus.



Problem 6.



Problem 8.

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.]

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. [Hint: It may be helpful to make an enlarged 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.

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

10 ✓. An object is more than one focal length from an inbending lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in the previous chapter, determine the positive and negative signs in the equation

$$\frac{1}{f} = \pm \frac{1}{d_i} \pm \frac{1}{d_o}.$$

(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 an inbending lens. (a) Draw a ray diagram. (b) Using reasoning like that developed in the previous chapter, determine the positive and negative signs in the equation

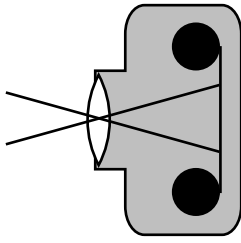
$$\frac{1}{f} = \pm \frac{1}{d_i} \pm \frac{1}{d_o}.$$

(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 outbending. (a) Draw a ray diagram. For simplicity pretend that there is no eye behind the glasses. (b) Using reasoning like that developed in the previous chapter,

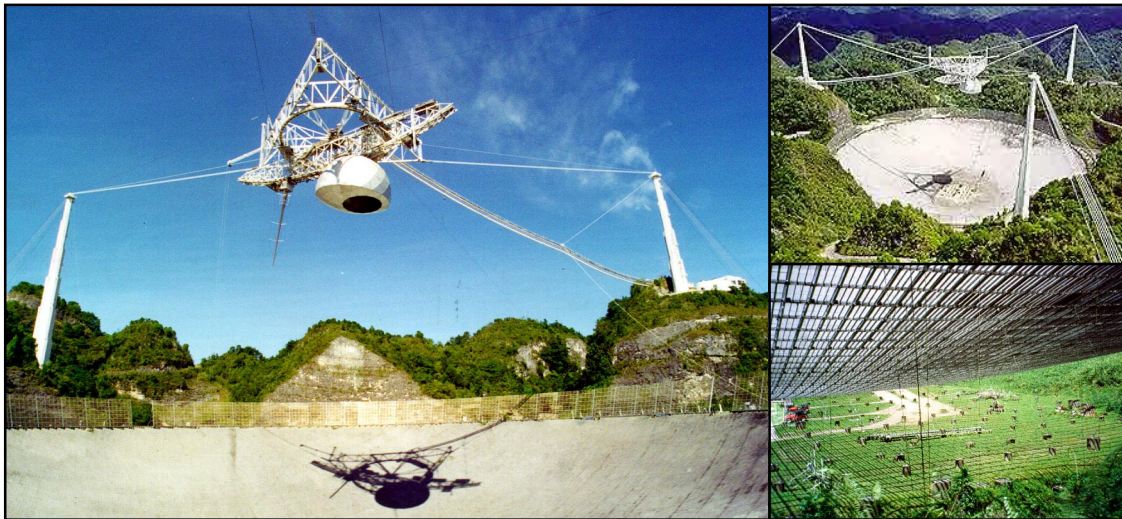
determine the positive and negative signs in the equation $\frac{1}{f} = \pm \frac{1}{d_i} \pm \frac{1}{d_o}$.

(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.

13 S. 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, θ , 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.]



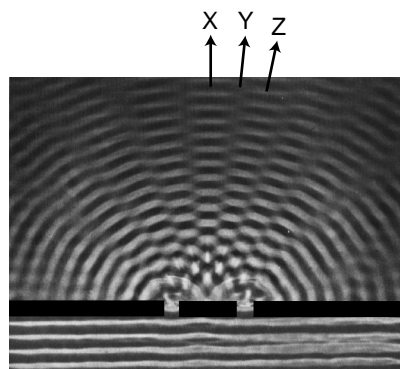
This huge radio dish nestles in a natural canyon at Arecibo, Puerto Rico. Its apparently solid surface is actually a scaffolding that hangs like a suspension bridge. It is a telescope that images the universe using radio waves rather than visible light, and it is also used to search for artificial radio signals from intelligent beings on other planets. Why does it have to be so huge? It's not mainly a question of picking up weak signals. The dish's sensitivity is overkill for most jobs, and for example it would have easily been able to pick up signals from a species on the other side of the galaxy that were no more intense than the ones humans themselves have transmitted out into space. (This is assuming that it was tuned to the right frequency at the right time, and that the signals came from within the limited field of view it sweeps out as the world spins.) No, the reason it has to be so big is a matter of wave optics. To make the antenna select signals from only a certain direction in space and reject those coming from other angles, it must be as large as possible compared to the wavelength of a radio wave.

5 Wave Optics

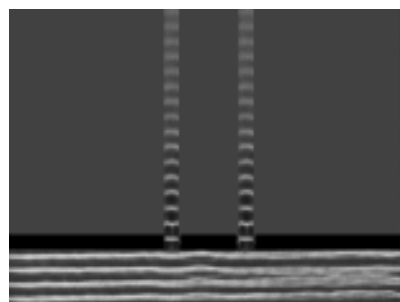
Electron microscopes can make images of individual atoms, but why will a visible-light microscope never be able to? Stereo speakers create the illusion of music that comes from a band arranged in your living room, but why doesn't the stereo illusion work with bass notes? Why are computer chip manufacturers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

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

5.1 Diffraction



(a) In this view from overhead, a straight, sinusoidal water wave encounters a barrier with two gaps in it. Strong wave vibration occurs at angles X and Z, but there is none at all at angle Y.



(b) This doesn't happen.

Figures (a) and (b) were constructed as collages of uncopyrighted photos from PSSC College Physics. Figure (a), although essentially correct, is a little unrealistic because the waves beyond the barrier would be much weaker.

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

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

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

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

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

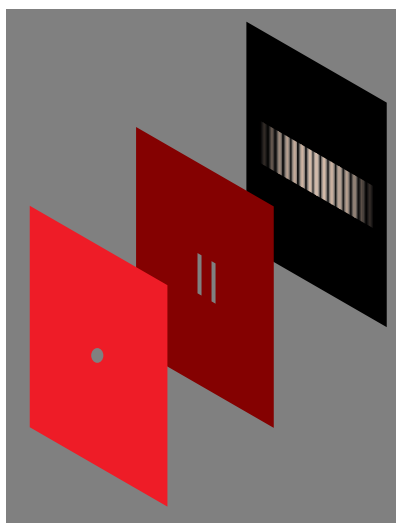
The kind of phenomenon described by restriction (1) is called *diffraction*. Diffraction can be defined as the behavior of a wave when it encounters an obstacle or a nonuniformity in its medium. In general, diffraction causes a wave to bend around obstacles and make patterns of strong and weak waves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

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

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

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

A second issue is that light from sources such as the sun or a lightbulb



(c) A practical setup for observing diffraction of light.

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



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?

5.2 Scaling of Diffraction

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

This is unusually simple behavior! In the first book of this series 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 at the end of book 3, Vibrations and 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 λ/d , where λ 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 λ and d by a factor of 37, the ratio λ/d will be unchanged.

5.3 The Correspondence Principle

The only reason we don't usually notice diffraction of light in everyday life is that we don't normally deal with objects that are comparable in size to a wavelength of visible light, which is about a millionth of a meter. Does this mean that wave optics contradicts ray optics, or that wave optics sometimes gives wrong results? No. If you hold three fingers out in the sunlight and cast a shadow with them, *either* wave optics or ray optics can be used to predict the straightforward result: a shadow pattern with two bright lines where the light has gone through the gaps between your fingers. Wave optics is a more general theory than ray optics, so in any case where ray optics is valid, the two theories will agree. This is an example of a general idea enunciated by the physicist Niels Bohr, called the *correspondence principle*: when flaws in a physical theory lead to the creation of a new and more general theory, the new 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 λ/d is small, both the ray and the wave model of light must give approximately the same result. Suppose you spread your fingers and cast a shadow with them using a coherent light source. The quantity λ/d is about 10^{-4} , so the two models will agree very closely. (To be specific, the shadows of your fingers will be outlined by a series of light and dark fringes, but the angle subtended by a fringe will be on the order of 10^{-4} radians, so they will be invisible and washed out by the natural fuzziness of the edges of sunshadows, caused by the finite size of the sun.)

Self-Check

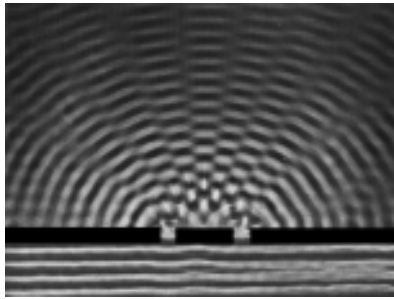


What kind of wavelength would an electromagnetic wave have to have in order to diffract dramatically around your body? Does this contradict the correspondence principle?

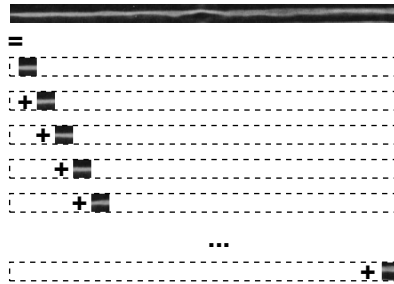


It would have to have a wavelength on the order of centimeters or meters, the same distance scale as that of your body. These would be microwaves or radio waves. (This effect can easily be noticed when a person affects a TV's reception by standing near the antenna.) None of this contradicts the correspondence principle, which only states that the wave model must agree with the ray model when the ray model is applicable. The ray model is not applicable here because λ/d is on the order of 1.

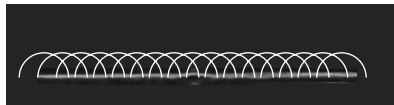
5.4 Huygens' Principle



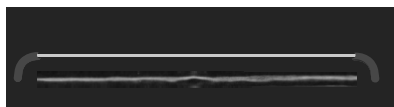
(a) Double-slit diffraction.



(b) A wavefront can be analyzed by the principle of superposition, breaking it down into many small parts.



(c) If it was by itself, each of the parts would spread out as a circular ripple.



(d) Adding up the ripples produces a new wavefront.

Returning to the example of double-slit diffraction, (a), note the strong visual impression of two overlapping sets of concentric semicircles. This is an example of *Huygens' principle*, named after a Dutch physicist. (The first syllable rhymes with “boy.”) Huygens' principle states that any wavefront can be broken down into many small side-by-side wave peaks, (b), which then spread out as circular ripples, (c), 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, (d). 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, (c) and (d). 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 Leibnitz'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 person in the world who knew what starlight was.

5.5 Double-Slit Diffraction

Let's now analyze double-slit diffraction, (a), 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 us 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, (b). 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 θ_Z to the ratio λ/d ,

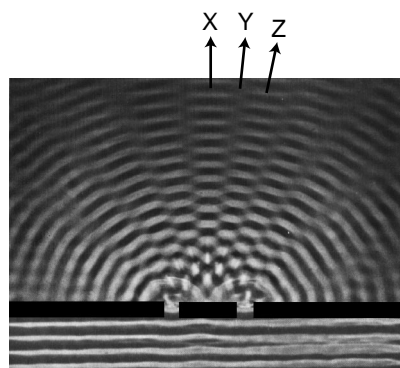
$$\lambda / d \leftrightarrow \theta_Z .$$

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

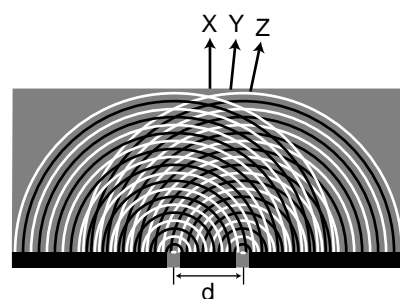
Along the central maximum line, X, we always have positive waves coinciding with positive ones and negative waves coinciding with negative ones. (I have arbitrarily chosen to take a snapshot of the pattern at a moment when the waves emerging from the slit are experiencing a positive peak.) The superposition 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, (c).

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.

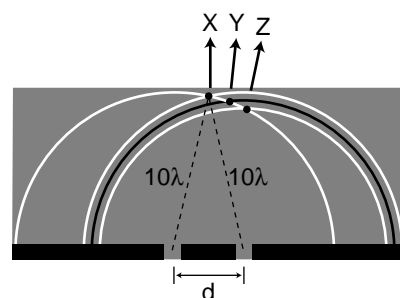
But the distance traveled does not have to be equal in order to get constructive interference. At the point along direction Z, one wave has gone nine wavelengths and the other ten. They are both at a positive extreme.



(a) Double-slit diffraction.



(b) Application of Huygens' principle. White lines represent peaks, black lines represent troughs, which we can refer to as positive and negative.



(c) 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.

Self-Check



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, (d). 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 skinner 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 (e), labeling one leg of the right triangle as the difference in path length, $L - L'$, and labeling the acute angle as θ . (In reality this angle is a tiny bit greater than the one labeled θ in the previous figure.)

The difference in path length is related to d and θ by the equation

$$\frac{L - L'}{d} = \sin \theta \quad .$$

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

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

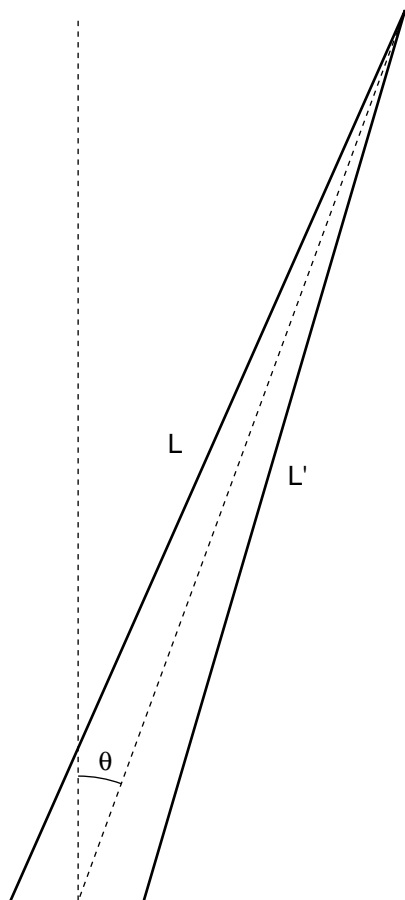
Here m equals 0 for the central maximum, -1 for the first maximum to its left, $+2$ for the second maximum on the right, etc. Putting all the ingredients together, we find $m\lambda/d = \sin \theta$, or

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

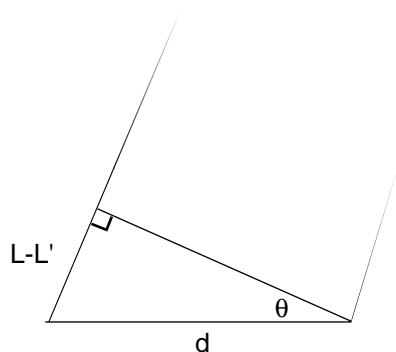
Similarly, the condition for a minimum is

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

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



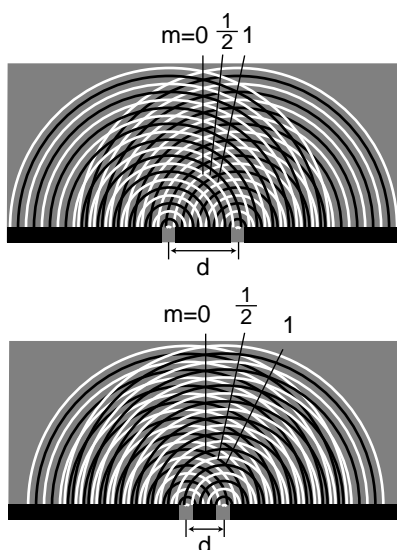
(d) The waves travel distances L_1 and L_2 from the two slits to get to the same point in space, at an angle θ from the center line.



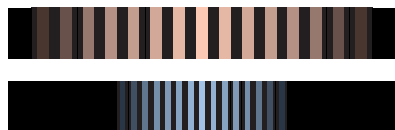
(e) A closeup of the previous figure, showing how the path length difference $L - L'$ is related to d and to the angle θ .



At this point, both waves would have traveled nine and a half wavelengths. They would both be at a negative extreme, so there would be constructive interference.



(f) Cutting d in half doubles the angles of the diffraction fringes.



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

As expected based on scaling, this equation relates angles to the unitless ratio λ/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, (f).

Example: 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 λ/d for leads to smaller values of $\sin \theta$, and therefore to a more closely spaced set of diffraction fringes, (g)

Example: The correspondence principle

Let's also consider how the equations for double-slit diffraction relate to the correspondence principle. When the ratio λ/d is very small, we should recover the case of simple ray optics. Now if λ/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 λ/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.)

Example: Spacing of the fringes at small angles

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

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

which can be solved for θ to give

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

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

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

For example, if we write θ_7 for the angle of the seventh bright fringe on one side of the central maximum and θ_8 for the neighboring one, we have

$$\begin{aligned}\theta_8 - \theta_7 &= \frac{8\lambda}{d} - \frac{7\lambda}{d} \\ &= \frac{\lambda}{d},\end{aligned}$$

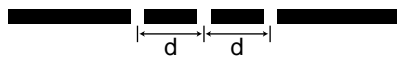
and similarly for any other neighboring pair of fringes.

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 λ/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.

5.6 Repetition



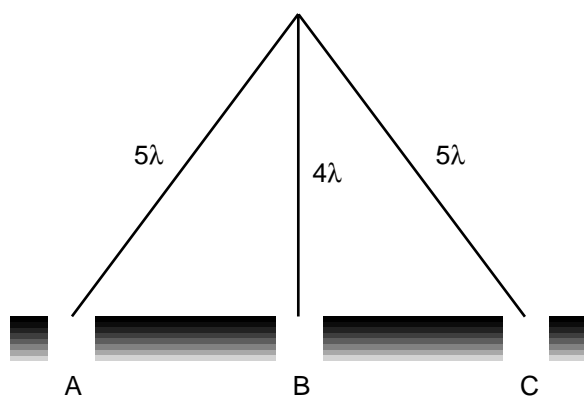
(a) A triple slit.

Suppose we replace a double slit with a triple slit, (a). 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?

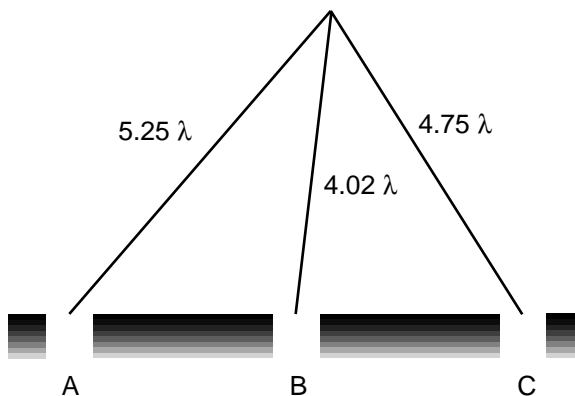
The answer is yes, as can be shown using figures (b) and (c). For ease of visualization, I have violated our usual rule of only considering points very far from the diffracting object. The scale of the drawing is such that a wavelength is one cm. In (b), all three waves travel an integer number of wavelengths 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 (c), 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 from 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 (b), 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 (c) would have been quite brightly lit up.

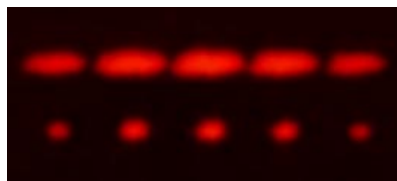
To summarize, we have found that adding a third slit narrows down the central fringe dramatically. The same is true for all the other fringes as well,



(b) There is a bright central maximum.



(c) At this point just off the central maximum, the path lengths traveled by the three waves have changed.



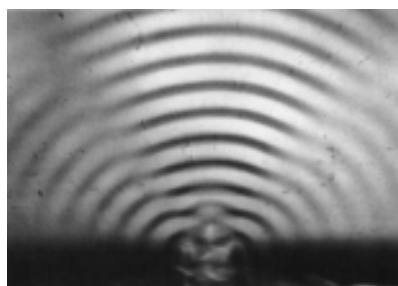
(d) A double-slit diffraction pattern (top), and a pattern made by five slits (bottom).. (Photo by the author.)

and since the same amount of energy is concentrated in narrower diffraction fringes, each fringe is brighter and easier to see, (d).

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.

5.7 Single-Slit Diffraction



(a) Single-slit diffraction of water waves. (PSSC Physics.)

If we use only a single slit, is there diffraction? If the slit is very narrow 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 (a) shows an example with water waves, and figure (b) with light.

Self-Check



How does the wavelength of the waves compare with the width of the slit in figure (a)?

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 λ/a , so all the familiar qualitative facts are true. For instance, shorter-wavelength light will produce a more closely spaced diffraction pattern.



(b) Single-slit diffraction of red light. Note the double width of the central maximum. (Photo by the author.)



(c) A pretty good simulation of the single-slit pattern of figure (a), made by using three motors to produce overlapping ripples from three neighboring points in the water. (PSSC Physics)



Judging by the distance from one bright wave crest to the next, the wavelength appears to be about 2/3 or 3/4 as great as the width of the slit.



An important scientific example of single-slit diffraction is in telescopes. Images of individual stars, as in the figure above, 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 slit, but the idea is the same.) Since the angular sizes of the diffracted images depend on λ/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



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

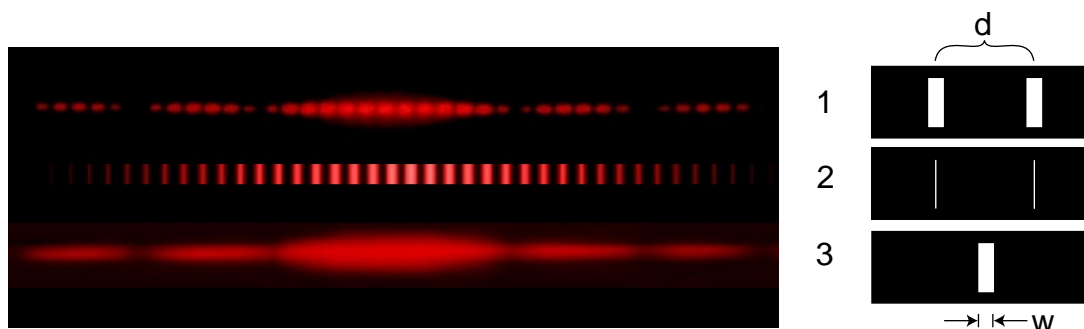
Discussion Question



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



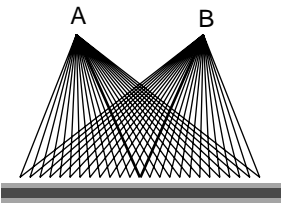
Since the wavelengths of radio waves are thousands of times longer, diffraction causes the resolution of a radio telescope to be thousands of times worse, all other things being equal.



1. A diffraction pattern formed by a real double slit. The width of each slit is not negligible 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 narrow compared 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.
- (Photos by the author.)

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 pattern 1 in the figure above, rather than the simpler one, 2, you were expecting. This is because the slits are not narrower than 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 .

5.8]* The Principle of Least Time



(a) Light could take many different paths from A to B.

12.28
11.93
11.61
11.31
11.04
10.80
10.59
10.41
10.26
10.15
10.07
10.02
10.00
10.02
10.07
10.15
10.26
10.41
10.59
10.80
11.04
11.31
11.61
11.93
12.28

(b) The distance traveled along the 25 possible paths ranges from 10.00 wavelengths to 12.28.

In sections 1.5 and 4.4, 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, (a). Starting at point A, Huygens' principle for waves tells us that we can think of the wave as spreading out in all directions. Suppose we imagine all the possible ways that a ray could travel from A to B. We show this by drawing 25 possible paths, of which the central one is the shortest. Since the principle of least time connects the wave model to the ray model, we should expect to get the most accurate results when the wavelength is much shorter than the distances involved — for the sake of this numerical example, let's say that a wavelength is 1/10 of the shortest reflected path from A to B. The table, (b), 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 25, what happens if we think of the angle, θ , 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, (b), turned out to have a group of rays that all traveled nearly the same distance.

As discussed in section 1.5, the principle of least time is really a principle of least *or greatest* time. This makes perfect sense, since $dL/d\theta=0$ can in general describe either a minimum or a maximum

The principle of least time is very general. It does not apply just to refraction and reflection — it can even be used to prove that light rays travel in a straight line through empty space, without taking detours! This general approach to wave motion was used by Richard Feynman, one of the pioneers who in the 1950's reconciled quantum mechanics with relativity (book 6). 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

Terminology Used in Other Books

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 λ/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

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

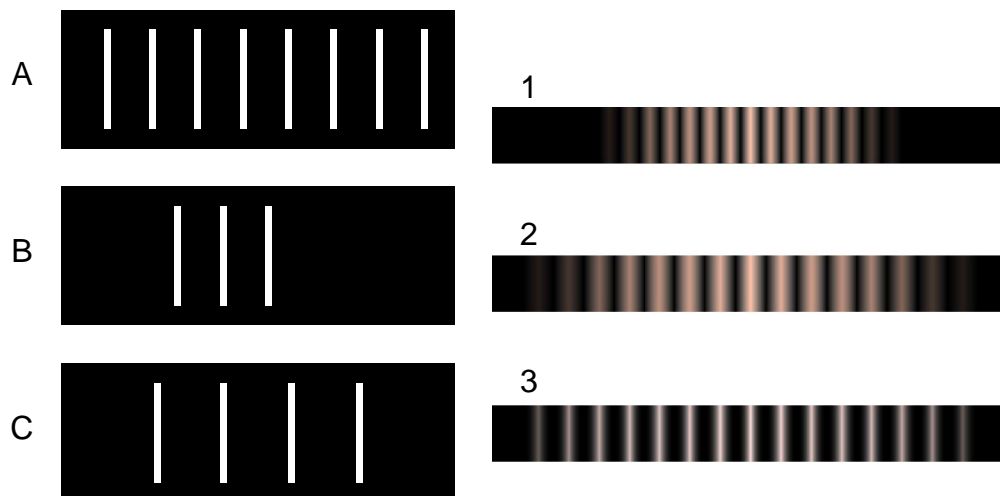
where d is the center-to-center spacing of the slits, and m is an integer at a maximum or an integer plus 1/2 at a minimum.

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 λ , d , and angles of fringes as in any other type of diffraction.

Homework Problems

1. Why would blue or violet light be the best for microscopy?
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 approximation $\sin \theta \approx \theta$?

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.)

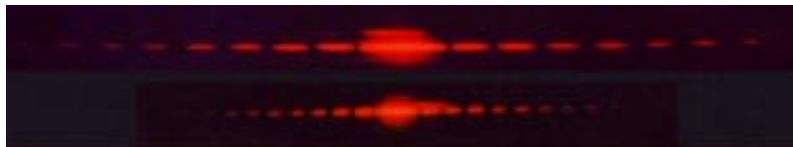
S A solution is given in the back of the book.

✓ A computerized answer check is available.

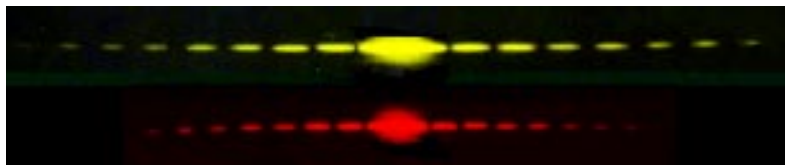
★ A difficult problem.

∫ A problem that requires calculus.

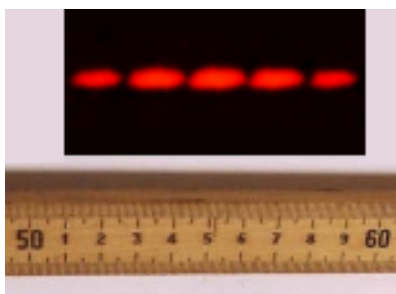
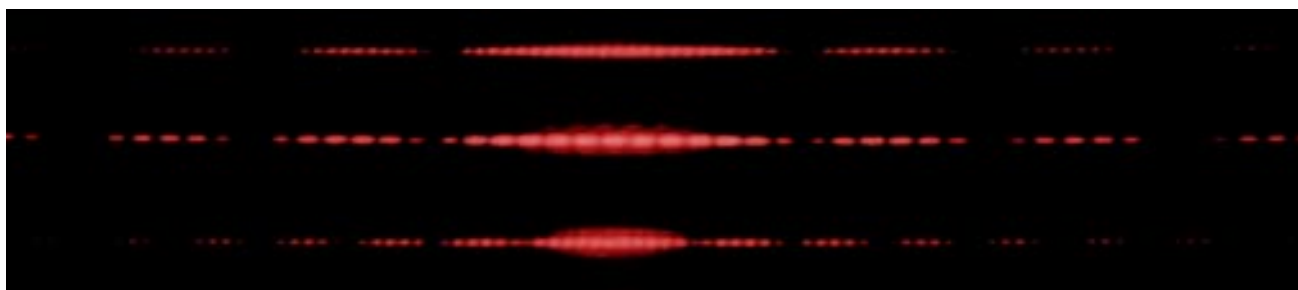
5. Ultrasound, i.e. sound waves with frequencies too high to be audible, can be used for imaging fetuses in the womb or for breaking up kidney stones so that they can be eliminated by the body. Consider the latter application. Lenses can be built to focus sound waves, but because the wavelength of the 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 $\sin \theta = 1.22 \lambda/b$, 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 $L=20$ cm. You will want $f>20$ 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 the photo in section 5.6, 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 m seen from a distance of 10^{17} m. What does this tell you?
7. Under what circumstances could one get a mathematically undefined result by solving the double-slit diffraction equation for θ ? Give a physical interpretation of what would actually be observed.
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 $d=0.50$ mm, and each slit was $w=0.04$ 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.



Problems 12 and 13.

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?

13. 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.

Exercises

Exercise 2A: Exploring Images With a Curved Mirror

Equipment:

curved mirrors like the ones described in this chapter
curved mirrors that bulge outward (for part 6 only)

1. Obtain a curved mirror from your instructor. If it is silvered on both sides, make sure you're working with the hollowed-out side, which bends light rays inward. Look at your own face in the mirror. Now change the distance between your face and the mirror, and see what happens. How do you explain your observations?

2. With the mirror held far away from you, observe the image of something behind you, over your shoulder. Now bring your eye closer and closer to the mirror. Can you see the image with your eye very close to the mirror? Explain what's happening.

3. Now imagine the following new situation, but *don't actually do it yet*. Suppose you lay the mirror face-up on a piece of tissue paper, put your finger 5 or 10 cm or so above the mirror, and look at the image of your finger. As in part 2, you can bring your eye closer and closer to the mirror.

Write down a prediction of what will happen. Will you be able to see the image with your eye very close to the mirror?

Prediction: _____

Now test your prediction. If your prediction was incorrect, can you explain your results?

4. Lay the mirror on the tissue paper, and use it to create an image of the overhead lights on a piece of paper above it and a little off to the side. What do you have to do in order to make the image clear? Can you explain this observation?

5. Now imagine the following experiment, but *don't do it yet*. What will happen to the image on the paper if you cover half of the mirror with your hand?

Prediction: _____

Test your prediction. If your prediction was incorrect, can you explain what happened?

6. Now imagine forming an image with a curved mirror that bulges outward, and that therefore bends light rays away from the central axis. Draw a typical ray diagram. Is the image real or virtual? Will there be more than one type of image?

Prediction: _____

Test your prediction with the new type of mirror.

Exercise 3A: Object and Image Distances

Equipment:

optical benches
inbending mirrors
illuminated objects

1. Set up the optical bench with the mirror at zero on the centimeter scale. Set up the illuminated object on the bench as well.
2. Each group will locate the image for their own value of the object distance, by finding where a piece of paper has to be placed in order to see the image on it. (The instructor will do one point as well.) Note that you will have to tilt the mirror a little so that the paper on which you project the image doesn't block the light from the illuminated object.

Is the image real or virtual? How do you know? Is it inverted or uninverted?

Draw a ray diagram.

3. Measure the image distance and write your result in the table on the board. Do the same for the magnification.
4. What do you notice about the trend of the data on the board? Draw a second ray diagram with a different object distance, and show why this makes sense. Some tips for doing this correctly: (1) For simplicity, use the point on the object that is on the mirror's axis. (2) You need to trace two rays to locate the image. To save work, don't just do two rays at random angles. You can either use the on-axis ray as one ray, or do two rays that come off at the same angle, one above and one below the axis. (3) Where each ray hits the mirror, draw the normal line, and make sure the ray is at equal angles on both sides of the normal.
5. We will find the mirror's focal length from the instructor's data-point. Then, using this focal length, calculate a theoretical prediction of the image distance, and write it on the board next to the experimentally determined image distance.

Exercise 4A: How strong are your glasses?

This exercise was created by Dan MacIsaac.

Equipment:

- eyeglasses
- outbending lenses for students who don't wear glasses, or who use inbending glasses
- rulers and metersticks
- scratch paper
- marking pens

Most people who wear glasses have glasses whose lenses are outbending, 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 inbending 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. Use a marker to draw three evenly spaced parallel lines on the paper. (A spacing of a few cm works well.)
2. 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?
3. 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 $1/2$.
4. Measure the object distance that results in a magnification of $1/2$, and determine the focal length of your lens.

Exercise 5A: Double-Source Interference

Equipment:

ripple tank
ruler, protractor, and compass

1. Observe the wave pattern formed by a single source. Try adjusting the frequency at which the motor runs. What do you have to do to the frequency in order to increase the wavelength, and what do you have to do to decrease it?
2. Observe the interference pattern formed by two sources. For convenience, try to get your wavelength as close as possible to 1 cm. We'll call this setup, with $\lambda = 1$ cm and $d = 2.5$ cm, the default setup.
3. Imagine that you were to double the wavelength and double the distance between the sources. How would a snapshot of this wave pattern compare with a snapshot of the pattern made by the default setup? Based on this, how do you predict the angles of the maxima and minima will compare?

Test your predictions.

4. 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 11 wavelengths from #1 and 11 from #2.
 - C. The point that is 10 wavelengths from #1 and 10.5 from #2.
 - D. The point that is 11 wavelengths from #1 and 11.5 from #2.
 - E. The point that is 10 wavelengths from #1 and 11 from #2.
 - F. The point that is 11 wavelengths from #1 and 12 from #2.

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

5. Make a fresh copy of your drawing, showing only point E and the two sources, which form a long, skinny triangle. Now suppose you were to change the default setup by doubling d , while leaving λ the same. Realistically this involves moving one peg over one hole, while leaving the other peg in the same place, but it's easier 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 E when you double d ? How has the angle of point E changed? _____

Test your prediction.

6. In the previous part of the exercise, you saw the effect of doubling d while leaving λ the same. Now what do you think would happen to your angles if, starting from the standard setup, you doubled λ while leaving d the same? _____

Try it.

7. Suppose λ 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 5B: Single-slit diffraction

Equipment:

rulers

computer spreadsheet or computer program for adding sine waves

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 becomes 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.

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.

4. Why do you think the intensity at the center came out the way it did? Would it have mattered if we had used 900 sets of ripples rather than 9?

5. Looking at the raw data for the point that had the least intensity, can you see why it came out that way?

6. What do you notice about the width of the central maximum compared to the width of the first side maximum? How is this different from double-slit interference? Compare with figure (b) in section 5.7.

7. Although the pattern goes up and down, the general trend is that the farther away we get from the center, the weaker it gets. Why does it make sense that the intensity at some random angle far from the center would tend to be small?

8. Single-slit diffraction can actually be calculated using equations in closed form rather than doing it numerically, and one result is that the intensity of the second maximum is always smaller than the intensity of the central maximum by a factor of $4/9\pi^2$. Note that the intensity (in units of watts per unit area) is proportional to the square of the wave's amplitude. Compare our results with the exact result.



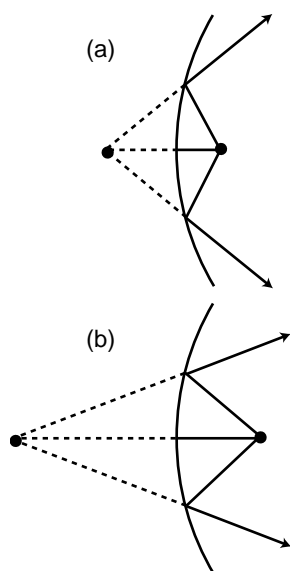
a b c d e f g h i j k



Solutions to Selected Problems

Chapter 3

2. See the ray diagrams below. Increasing d_i increases d_o , so the equation $1/f = \pm 1/d_i \pm 1/d_o$ must have opposite signs on the right. Physically, we can have a virtual image with $d_i = \infty$, but not with $d_o = \infty$, so the positive sign has to be the one in front of d_o , giving $1/f = -1/d_i + 1/d_o$.



Problem 2.

Chapter 4

13. Since d_o is much greater than d_i , the lens-film distance d_i is essentially the same as f . (a) Splitting the triangle inside the camera into two right triangles, straightforward trigonometry gives

$$\theta = 2 \arctan(w/2f)$$

for the field of view. This comes out to be 39° and 64° for the two lenses. (b) For small angles, the tangent is approximately the same as the angle itself, provided we measure everything in radians. The equation above then simplifies to

$$\theta = w/f$$

The results for the two lenses are $.70 \text{ rad} = 40^\circ$, and

$1.25 \text{ rad} = 72^\circ$. This is a decent approximation.

(c) With the 28-mm lens, which is closer to the film, the entire field of view we had with the 50-mm lens is now confined to a small part of the film. Using our small-angle approximation $\theta = w/f$, the amount of light contained within the same angular width θ is now striking a piece of the film whose linear dimensions are smaller by the ratio $28/50$. Area depends on the square of the linear dimensions, so all other things being equal, the film would now be overexposed by a factor of $(50/28)^2 = 3.2$. To compensate, we need to shorten the exposure by a factor of 3.2.

Glossary

Absorption. What happens when light hits matter and gives up some of its energy.

Angular magnification. The factor by which an image's apparent angular size is increased (or decreased). Cf. magnification.

Coherent. A light wave whose parts are all in phase with each other.

Concave. Describes a surface that is hollowed out like a cave.

Convex. Describes a surface that bulges outward.

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.

Diffuse reflection. Reflection from a rough surface, in which a single ray of light is divided up into many weaker reflected rays going in many directions.

Focal length. A property of a lens or mirror, equal to the distance from the lens or mirror to the image it forms of an object that is infinitely far away.

Image. A place where an object appears to be, because the rays diffusely reflected from any given point on the object have been bent so that they come back together and then spread out again from the image point, or spread apart as if they had originated from the image.

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.

Magnification. The factor by which an image's linear size is increased (or decreased). Cf. angular magnification.

Real image. A place where an object appears to be, because the rays diffusely reflected from any given point on the object have been bent so that they come back together and then spread out again from the new point. Cf. virtual image.

Reflection. What happens when light hits matter and bounces off, retaining at least some of its energy.

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

Specular reflection. Reflection from a smooth surface, in which the light ray leaves at the same angle at which it came in.

Virtual image. Like a real image, but the rays don't actually cross again; they only appear to have come from the point on the image. Cf. real image.

Index

A

aberration 41
absorption 15
angular magnification 29

B

Bohr
 Niels 62
brightness of light 16
Bush, George 45

C

color 51
concave
 defined 31
convex
 defined 31

D

diffraction
 defined 60
 double-slit 64
 fringe 61
 scaling of 61
 single-slit 69
diffraction grating 69
diffuse reflection 15
diopter 36
double-slit diffraction 64

E

Empedocles of Acragas 12
evolution 45
eye
 evolution of 45
 human 46

F

Fermat's principle 22
flatworm 46
focal angle 34
focal length 35
 negative 43
fringe
 diffraction 61

G

Galileo 13

H

Hertz, Heinrich
 Heinrich 63
Huygens' principle 63

I

images
 formed by curved mirrors 27
 formed by plane mirrors 26
 location of 34
 of images 29
 real 28
 virtual 26
inbending
 defined 27
incoherent light 61
index of refraction
 defined 48
 related to speed of light 49
Io 14

J

Jupiter 14

L

least time
 principle of
 for reflection 22
 for refraction 54
lensmaker's equation 54
light
 absorption of 15
 brightness of 16
 particle model of 17
 ray model of 17
 speed of 13
 wave model of 17

M

magnification
 angular 29
 by an inbending mirror 27
 negative 43
Maxwell, James Clerk 64
mirror
 inbending 34
mollusc 46
Moses 45

N

nautilus 46
Newton, Isaac 29, 63

O

optical density. *See* index of refraction: defined
orrespondence principle 62
outbending
 defined 31

P

particle model of light 17, 63
Pythagoras 12

R

ray diagrams 18
ray model of light 17, 63
reflection
 diffuse 15
 specular 20
refraction
 and color 51
 defined 46
repetition of diffracting objects 68
retina 28
reversibility 20
Roemer 14

S

single-slit
 diffraction 69
Snell's law 48
 derivation of 50
 mechanical model of 49
Squid 46

T

telescope 70
time reversal 20
total internal reflection 51

V

vision 12

W

wave model of light 17, 63
Wigner, Eugene 33

Y

Young, Thomas 63

Photo Credits

All photographs are by Benjamin Crowell, except as noted below or in the captions of the photos. As noted here and in the captions, many of the photos in this chapter are from PSSC Physics, and are used under a license provided in PSSC College Physics.

Cover

Cross-section of eye: National Eye Institute, National Institutes of Health.

Chapter 4

Cross-section of eye: National Eye Institute, National Institutes of Health.

Diffraction of water wave: PSSC Physics (retouched).

Line drawing of eye: National Eye Institute, National Institutes of Health.

Chapter 5

As noted in the captions, many images in this chapter are from PSSC Physics. All the diffraction patterns in the homework problems are by the author.

Arecibo: Photos from above by Tony Acevedo, from below by David Parker. Courtesy of NAIC - Arecibo Observatory, a facility of the NSF.

Star field: Space Telescope Science Institute/Digitized Sky Survey.

Useful Data

Metric Prefixes

M-	mega-	10^6
k-	kilo-	10^3
m-	milli-	10^{-3}
μ - (Greek mu)	micro-	10^{-6}
n-	nano-	10^{-9}
p-	pico-	10^{-12}
f-	femto-	10^{-15}

(Centi-, 10^{-2} , is used only in the centimeter.)

Notation and Units

quantity	unit	symbol
distance	meter, m	$x, \Delta x$
time	second, s	$t, \Delta t$
mass	kilogram, kg	m
density	kg/m^3	ρ
force	newton, 1 N=1 $\text{kg}\cdot\text{m}/\text{s}^2$	F
velocity	m/s	v
acceleration	m/s^2	a
energy	joule, J	E
momentum	$\text{kg}\cdot\text{m}/\text{s}$	p
angular momentum	$\text{kg}\cdot\text{m}^2/\text{s}$	L
period	s	T
wavelength	m	λ
frequency	s^{-1} or Hz	f
focal length	m	f
magnification	unitless	M
index of refraction	unitless	n

Fundamental Constants

gravitational constant	$G=6.67\times 10^{-11} \text{ N}\cdot\text{m}^2/\text{kg}^2$
Coulomb constant	$k=8.99\times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2$
quantum of charge	$e=1.60\times 10^{-19} \text{ C}$
speed of light	$c=3.00\times 10^8 \text{ m/s}$

Conversions

Conversions between SI and other units:

1 inch	=	2.54 cm (exactly)
1 mile	=	1.61 km
1 pound	=	4.45 N
(1 kg):g	=	2.2 lb
1 gallon	=	$3.78\times 10^3 \text{ cm}^3$
1 horsepower	=	746 W
1 kcal*	=	$4.18\times 10^3 \text{ J}$

*When speaking of food energy, the word "Calorie" is used to mean 1 kcal, i.e. 1000 calories. In writing, the capital C may be used to indicate

1 Calorie=1000 calories.

Conversions between U.S. units:

1 foot	=	12 inches
1 yard	=	3 feet
1 mile	=	5280 ft

Some Indices of Refraction

substance	index of refraction
vacuum	1 by definition
air	1.0003
water	1.3
glass	1.5 to 1.9
diamond	2.4

Note that all indices of refraction depend on wavelength. These values are about right for the middle of the visible spectrum (yellow).

Subatomic Particles

particle	mass (kg)	charge	radius (fm)
electron	9.109×10^{-31}	$-e$	<0.01
proton	1.673×10^{-27}	$+e$	~ 1.1
neutron	1.675×10^{-27}	0	~ 1.1
neutrino	$\sim 10^{-39} \text{ kg?}$	0	?

The radii of protons and neutrons can only be given approximately, since they have fuzzy surfaces. For comparison, a typical atom is about a million fm in radius.