The Fundamentals of

PHYSICS

Volume II

Using Models

Robert A. Cohen
Physics Department
East Stroudsburg University

January 4, 2024
Version 18.1

Acknowledgments

Several students and faculty have made informal comments on a study guide that I wrote which has since evolved into this textbook. Unfortunately, I don't recall the exact comments nor who has made them. I thank them nonetheless.

One student, Richard Cowell, reviewed a very early version and provided specific suggestions for improvement. Another student, Lisa Suabedissen, corrected several errors in a later version. Several errors in the previous drafts were pointed out by Michael Petrullo, Christopher Briggs and Blanche Healy. The supportive comments of several students (including David Rebar, Christopher Briggs, Aliraza Somji and Brian Napert) provided the motivation to carry this through to completion.

I have learned a great deal by lurking on the phys-L discussion, particularly the thought-provoking posts by John Denker, who reviewed an early version of the book and provided detailed comments and edits. Discussions with Professors David Larrabee, John Elwood, David Buckley, Jerry Ross and Maria Cohen were also helpful. Most importantly, perhaps, was the support of my wife, Maria, even though my free time (and many vacations) was often spent working on revisions.

Contents

Part A

Forces

1. Mass and the Gravitational Force

Puzzle #1: If I release a light wood ball and a heavy iron ball, from the same height and at the same time, they fall and hit the ground at the same time. Why?

Introduction

Each chapter in this book starts with a puzzle, usually a question. You may not know the answer to the puzzle right away, but that's okay. As you read the chapter, think about the puzzle and how the information can be used to answer the puzzle. You should be able to answer the puzzle by the end of chapter.

In this case, the puzzle asks a situation involving gravity, and thus leads us to discuss the properties of gravity, which we will do in this chapter. In part A, we examine the four fundamental forces. and the gravitational force is one of those four fundamental forces. The other fundamental forces are electric, magnetic and nuclear.

It is assumed that you have already learned about the gravitational force (from the first semester of physics; see volume I), but we'll review its properties, not only to refresh your memory but also so that you can compare and contrast the gravitational force with the other fundamental forces that will be introduced in future chapters. Along the way, we'll also review some of the other main ideas from the first semester.

1.1 Why?

Before getting started with the gravitational force, let's first address a very important question that you may have.

Why do we have to take another semester of physics if we've already learned about physics?

The first semester focused on scientific laws and definitions as a context for how we approach problems in physics. Basically, we try to identify a few general ideas that can be applied to lots and lots of situations.[i] This makes them very powerful, which is why they were used over and over again in Volume I, and why we will continue to use them in Volume II. However, due to their generality there is a need for rigorous clarity so that we can apply them appropriately.

Applying a few general ideas to every situation is much better than learning lots and lots of solutions, one solution for each particular situation, which would be like learning how to navigate through a *particular* maze and then learning how to navigate through *another* particular maze and so on, never bothering to figure out if there was a general approach that might work for all mazes. While it may be trickier to learn how to navigate mazes in general, it ultimately requires less memorization and is more powerful.

In volume I, the context for learning this approach involved the law of force and motion[ii] and the definition of average velocity to predict an object's motion in various situations. In this sense, it served its purpose. In general, however, the situations examined in volume I were pretty idealized and not terribly interesting.[iii]

In comparison, this semester we will extend our physics toolbox to incorporate conceptual models of how the world works. This will allow us to tackle more complicated and more interesting phenomena.

Some phenomena you may already be familiar with. For example, you probably already know about electrons and protons. However, have you ever really thought about them? Extremely tiny particles that you can't see or weigh yet they somehow attract each other with a force that explains how we are able to run and why we get hot when we exercise?

---

[i] We assumed that the laws are universal, meaning that they apply in all situations. Indeed, they have been shown to be extraordinarily accurate when applied to a wide range of ordinary situations, from bacteria and galaxies and lots of things in between. There are limits, however. For example, the law of force and motion cannot be directly applied to quantum mechanics, special relativity, general relativity, and non-inertial frames.

[ii] The law of force and motion is commonly referred to as *Newton's second law.*

[iii] The first semester was a lot like training for a sport. A lot of time spent in the weight room, strengthening muscles and improving coordination.

Such explanations, like the ones utilizing electrons and protons, are called **models** (or scientific theories). They aren't *physical* models, like a model airplane, but rather *conceptual* models – ones that we "see" in our minds. While we can build a physical model to represent certain features of a conceptual model, we must be careful not to confuse the two. For example, we can use balls connected by sticks to represent molecules but the balls and sticks don't share any of the physical properties of the molecules they represent. In a similar way, a model airplane may *look* like a real airplane but it doesn't *perform* like a real airplane. As such, it doesn't give us any insight into how a real airplane works or flies. It just acts to serve as a visual representative sample (much like a person who sits as a subject for a painting or who models what clothes look like when worn). A good conceptual model captures the salient features of the phenomenon so that we can make reliable predictions involving that phenomenon.

✍ Compared to the first semester of physics, there may be a bit more to memorize because we'll be exploring a variety of different phenomena and so we'll need a few different conceptual models. In addition, for a long time scientists didn't realize they were all related and so different sets of definitions and terms were developed. Still, the goal is to take a few different conceptual models to explain a whole variety of phenomena rather than trying to memorize *everything*.

✓ *Check Point 1.1:* [iv] *This book examines conceptual models. Which of the following represent(s) the type of conceptual model we are talking about?*
(a) Visualizing atoms as billiard balls.
(b) Using the icing on a cake to illustrate the cell membrane.
(c) Using a cloud chart to identify the names of clouds.

[iv] Checkpoints are problems provided throughout the text for *you* to continually evaluate your understanding of the readings. They are called "checkpoints" instead of "homework" or "practice" in order to stress the importance of active learning (sometimes called "active thinking") and self-assessment. In keeping with this self-assessment philosophy, an answer key to the checkpoints is available for you to check your understanding. Please use these checkpoints in the way they were intended. In other words, do not use the checkpoint as a *starting* point and then search through the text for the answer. Instead, read through the text and, as you read, compare what you are reading with what you already know. Then, use the checkpoint as a check.

1.2 Properties of the fundamental forces

The first third of volume II looks at the four fundamental[v] interactions (gravitational, electric, magnetic and nuclear) and three ways we deal with such interactions (forces, fields, and energy).

In part A, we'll examine interactions in terms of forces. You are already familiar with using this approach, as we used forces throughout volume I.

In volume I, each interaction was imagined to consist of invisible springs that, depending on the type of interaction, are either compressed when the two objects move together (pushing the objects apart) or are stretched when the two objects move apart (pulling the objects together). At the time, the only non-contact force we considered was the gravitational force. However, it turns out that all four fundamental interactions act even when the two objects aren't touching in the sense that we typically consider to be touching.

For example, suppose you brought a magnet close to a paper clip. At some point, the paper clip would get attracted to the magnet. In such a case, the two interacting objects are the magnet and the paper clip and the force is the magnetic force, which is exerted on both of the interacting objects.

What is important to recognize is that the paper clip experiences a magnetic force before it touches the magnet. In addition, each object does not "have" a magnetic force. Instead, each object experiences a magnetic force, due to the invisible magnet springs that we can imagine as the interaction between the two objects.

As another example, consider the bond between the water molecules that make up water. They are attracted to one another, which is why liquid water does not spread out like a gas. The interacting objects are the individual molecules and the force is the electric force. The individual hydrogen and oxygen atoms that make up each molecule are also attracted to one another, and the force associated with those interacting objects (i.e., the individual atoms) is again the electric force.

In both cases, each object does not "have" an electric force. Instead, we say that each object experiences an electric force, due to the invisible electric springs that we can imagine as the interaction between the two objects.

[v] Scientists now view some of these as just different aspects of the same interaction. However, I will approach them as separate interactions, since the language we use to describe them are different.

Now let's consider a rock on the ground.

What prevents the rock from just rising up off the ground? Apparently, there is a force that attracts the rock to **Earth**[vi], just as the paper clip is attracted to the magnet and the individual water molecules are attracted to each other.

That force is the **gravitational force**, and the two interacting objects are the rock and Earth.

As with the other cases, we say that the rock experiences a force, due to its interaction with Earth. In this case, that force is the gravitational force. We do not say that the rock "has" a gravitational force, since the force is not just associated with the rock. The force is associated with the invisible gravitational springs that we can imagine as the interaction between the rock and Earth.

We'll examine the gravitational force first (simply because you are probably most familiar with that one from volume I).

✓ *Check Point 1.2: What is the name of the force associated with the following interactions?*
(a) a north pole of a magnet being attracted to a south pole of another magnet
(b) a rock being attracted to Earth

As mentioned before, the gravitational force (like all forces) is associated with the interaction of two objects. It is not associated with a single object.

Unfortunately, saying "the gravitational force on the rock due to its interaction with Earth" is quite a mouthful. Instead, I'll tend to write "the gravitational force on the rock (due to Earth)." Hopefully that shortened phrase still conveys the idea that the gravitational force is due to the interaction between the rock and Earth, rather than Earth alone.

✍ Try to avoid shortening the phrase further, like "the gravitational force of Earth on the rock," as that can lead people to think that the force belongs to Earth when, in fact, it belongs to the *interaction* between the rock and Earth. Without *both* objects, the force (and the interaction) wouldn't exist.

[vi] In keeping with the NASA Style Guide, I will capitalize the word "Earth" in order to emphasize that I am referring to the entire planet, and not just using the word "earth" to refer to the dirt. Mostly I will use "Earth", rather than "the Earth", but in both cases I am referring to the entire planet called Earth.

✓ *Check Point 1.3: Why is the phrase "the gravitational force on the rock" preferable to "the gravitational force of the rock"?*

1.3 The law of interactions

As mentioned in the previous section, each interaction can be imagined as invisible springs that act to pull two objects together (for attractive forces) or push them apart (for repulsive forces). Associated with each interaction, then, are two forces, one on each object participating in the interaction. The **law of interactions**[vii] describes the interaction and has a couple of parts.

First, the law of interactions states that when two objects interact, each object experiences a force due to that interaction. Forces are not something that objects "carry around" with them.[viii] For example, when two objects interact gravitationally (due to their mass), there is a gravitational force exerted on both objects. The gravitational force is not something an object "has."

As another example, if I touch you then there will be a contact force on *both* of us. I can't touch you (such that there is a force on you) without experiencing a force on me (due to that same interaction). This does *not* mean that if I hit you, you will hit me back (though that may indeed happen). Rather, it says that *while* I am hitting you, with you experiencing a force due to that interaction, I am simultaneously[ix] experiencing a force due to that same interaction. The two forces occur simultaneously, as they are associated with the same interaction.

Second, if you consider the two objects that make up the interacting pair, the law of interactions states that the force on each (due to that interaction)

[vii] The law of interactions, commonly known as **Newton's third law**, is called a law because it describes a relationship. In this case, it describes the relationship between the two forces associated with a single interaction. For more information on what a scientific law is, see the supplemental readings.

[viii] Quantities like mass, momentum and inertia, which are associated with the object itself, are not considered forces.

[ix] This means the two forces occur at the same time.

are opposite in direction. For example, if I am pushing on a car, trying to get it to move eastward, then at the same time there is an eastward force on the car (due to its interaction with me), there is a westward force on me (due to that same interaction).

⁉ The law of interactions states that each object of the interacting pair experiences an equal and opposite force.

Third, if you consider the two objects that make up the interacting pair, the law of interactions states that the forces due to that interaction on each of those two objects are equal in magnitude. So, when I touch you, you experience a force on you due to that interaction that is equal in magnitude to the force I experience due to that interaction.

Whether it hurts you more than it hurts me is irrelevant. If a large train runs into a small car, for example, the impact on the car is significantly greater than the impact on the truck. The law of interactions only says that the *forces* are equal.

In addition, this does *not* mean that the *net* force on each object is equal in magnitude. Each interaction is separate. For example, suppose there are *three* objects that interact. Let's call them A, B and C. There would be three separate interactions: an interaction between A and B, an interaction between A and C, and an interaction between B and C. If we consider just one of those objects, say object A, then there would be two forces on object A, one due to its interaction with B and one due to its interaction with C. Those two forces do *not* need to be equal in magnitude and opposite in direction, since they are due to two *different* interactions.

✓ *Check Point 1.4: A table tennis ball bounces off of a bowling ball. During the collision, the table tennis ball experiences an average force of 1 N toward the east due to the bowling ball. What was the average force exerted on the bowling ball due to the table tennis ball?*

1.4 Mass

As discussed earlier, when two objects interact, each interacting object experiences a force, as described by the *law of interactions*.

So, when a gravitational force is exerted on a rock due to its interaction with

Earth, there also happens to be a gravitational force exerted on *Earth* due to the same interaction. The two forces must occur simultaneously.

Not only that but the forces have to have the *same* magnitude (with opposite directions). This means that for the rock interacting with Earth, the force exerted on the rock (due to Earth) must be equal (in magnitude) to the force exerted on Earth (due to the rock).

For example, if there is a downward gravitational force of 690 N on me (due to my gravitational interaction with Earth) then there must simultaneously be an upward gravitational force of 690 N on Earth (due to its gravitational interaction with me). It is like there is an invisible spring connecting me with Earth and pulling us *both* together.[x]

IF THE FORCES ARE THE SAME, WHEN YOU JUMP OFF A CHAIR WHY DO YOU FALL DOWN TO EARTH RATHER THAN EARTH MOVING UPWARD TO MEET YOU?

Although the forces are the same, the effects are quite different. The difference has to do with Earth being so much more massive than I am.

Unlike force, **mass** is something that is inherent with an object (i.e., it "belongs" to the object). Earth's mass is very large. My mass is a bit more modest.

In fact, Earth's mass is about 10^{23} times bigger than mine, which means that the effect on Earth will be much, much less than the effect on me. Indeed, the effect on Earth is imperceptible because Earth's mass is so huge. So, a force is still exerted on Earth (due to me) – it is just not as obvious (because the effect is so tiny). That is why it looks like there is only a force on me and not on Earth, even though the forces actually have the same magnitude.

To be consistent with the law of interactions, it is important to say the force is associated with the interaction, not an individual object. This is also consistent with the forces being equal in magnitude (i.e., same interaction, so same magnitude of force).

One potential area of confusion is that for objects in contact with the ground there is also a *contact* force in addition to the *gravitational* force. The contact

[x]This conjures up images of Harry Potter and using wands to exert forces on objects without touching them. In a sense, such actions at a distance actually exist – the gravitational force being one example – it is just that the gravitational force (due to you) is too small to have any effect on other objects.

force is due to an interaction between the object and just the *surface* of Earth whereas the gravitational force is due to an interaction between the object and the *entire* mass of Earth. Those are two separate interactions and though there may be situations when they are equal in magnitude and opposite in direction there is no law that states that those two *must* be equal in magnitude or opposite in direction.

✓ *Check Point 1.5: At this very moment, I am standing on Earth. Which is larger in magnitude: the gravitational force exerted on me by Earth or the gravitational force exerted on Earth by me, are they equal in magnitude, or does it depend on the situation? Why?*

1.5 The universal law of gravitation

As mentioned before, the gravitational force is due to an interaction between the entire mass of each object. It should make sense, then, that the magnitude of the gravitational force depends upon the masses of the two objects.

The SI unit of mass is a **kilogram**. For example, my mass about 70 kg, where I have abbreviated[xi] the unit "kilogram" as "kg." In comparison, Earth's mass is about 6×10^{24} kg.

If you are unfamiliar with how much mass a kilogram represents, it might be useful to identify the masses of everyday objects in kilograms so that you can get a sense of what is reasonable. For reference, a one-liter bottle of water has a mass of 1 kilogram.[xii]

✓ *Check Point 1.6: An eight-fluid-ounce bottle or glass of water is about 240 ml (i.e., 240 milliliters). There are 1000 ml in a liter. What is its mass?*

If the gravitational force depends upon the mass, does the gravitational force exist between any two objects?

[xi] All unit abbreviations are listed in the supplemental readings.

[xii] Soda is similar to water, so a two-liter bottle of soda has a mass of two kilograms. Only materials with densities similar to water will have this proportion. A two-liter bottle of mercury, for example, would have a mass of 27 kilograms!

Yes. However, unless one or both of the objects are very massive, the gravitational force is typically too small to notice. Around here, the only hugely massive object is Earth. That is why we don't notice the force between two rocks but we do notice the force between a rock and Earth.

⁉ The gravitational force is easily evident only when one or both of the objects are very massive.

DOES THE GRAVITATIONAL FORCE HAVE ANYTHING TO DO WITH THE AIR?

No.

⁉ The gravitational force acts even if the two objects are not in contact.

Not only is air unnecessary, but the two objects involved do not have to be touching at all in order to attract gravitationally.

This recognition, that the two objects do not have to be touching in order to interact gravitationally, is traditionally attributed to **Isaac Newton**, an English scientist who lived in the seventeenth century. This was one of Isaac Newton's two great insights about gravity. The other insight was that the gravitational force exists between *any* two objects.

The first insight may not appear that great, seeing how two magnets can attract or repel without touching. However, the second insight really seems amazing when you think about it. After all, what evidence is there that two objects interact gravitationally even if one of the objects is not Earth?

Yes, we are all attracted down to Earth. That is why we don't float away. Still, we don't attract each other. We aren't even attracted to something big, like a building. Why would Newton conclude that every object is attracted to every other object?

Newton was able to see that the orbits of the planets around the Sun (and the orbits of moons around planets) could be explained by recognizing that there is a gravitational force between all of the planets and moons (the physics of orbits is explored in volume I).

He then concluded that the gravitational force must only be significant when at least one of objects is really, really big (like a planet or star), and that the gravitational force must be larger when the objects are separated by a smaller distance. He used this idea to construct his *universal law of gravitation* (so called because it works universally, or between any two objects).

The **universal law of gravitation** states that there exists a gravitational force between any two objects and that the magnitude is proportional to the masses of both objects (which explains why we are attracted to Earth but not each other) and inversely proportional to the square of their center to

center distance (which explains why we are attracted to Earth more than to the Jupiter).

Mathematically, we can express the universal law of gravitation in terms of a universal gravity equation:

$$F_{\text{g}} = G\frac{m_1 m_2}{r^2} \tag{1.1}$$

where F_{g} is the magnitude of the gravitational force, m_1 and m_2 are the masses of the two interacting objects, r is the distance from the center of one object to the center of the other, and G is equal to 6.67408×10^{-11} N m^2/kg^2.[xiii]

✍ Remember from Volume I that mathematical equations *represent* physical ideas – they don't *replace* the physical ideas. In other words, we use equations to guide our application of the physical ideas – we don't use equations *instead* of the physical ideas.

Why does G have such strange units?

The somewhat strange-looking units for G has to do with the units we use for the force (N; newtons), mass (kg; kilograms) and distance (m; meters). These units are part of the SI system of units.

As long as we use SI units for all quantities, we know the units will work out. For this reason, it is easiest to convert all values into SI units when using the universal gravity equation (or any equation we'll be using). For example, if the mass value is in grams, not kilograms, first convert the mass into kilograms before using the value in the universal gravity equation to ensure that the value we obtain for the force is in newtons.

✍ A newton is equivalent to a $\text{kg}\cdot\text{m}/\text{s}^2$.

✓ *Check Point 1.7: Suppose two 1-kg balls are one foot apart (center to center). Since each number is one, does that mean multiplying by G gives a gravitational force equal to* 6.67408×10^{-11} *N? Why or why not?*

How does the universal gravity equation reflect what we know about the universal law of gravitation?

[xiii] All variable abbreviations are listed in the supplemental readings. Subscripts are used to describe the variable.

To interpret an equation that has more than one or two quantities, like the universal gravity equation, it helps to consider one quantity at a time. To do this, we can compare two situations where everything on the right side of the equation is the same except for that one quantity.

For example, let's suppose we compare the magnitude of the gravitational force (F_g) when two objects are close rather than far apart. In the two situations, m_1 and m_2 are unchanged since we are not replacing any of the objects and G always has the same value. Since r^2 is in the denominator on the right side of the equation while F_g is in the numerator on the left side of the equation, they have an inverse relationship – namely that F_g is smaller when r^2 is larger, and visa-versa F_g is larger when r^2 is smaller. This is consistent with how the gravitational is larger when the two objects are closer.

In the same way, we can see that m_2 is in the numerator on the right side of the equation while F_g is in the numerator on the left side of the equation. This means that they are proportional – and F_g is larger when m_2 is larger, and visa-versa F_g is smaller when m_2 is smaller. This is consistent with how, for the same distance, the gravitational is larger when an object's mass is larger.

WHICH QUANTITIES ARE LIKELY TO BE DIFFERENT IN DIFFERENT SITUATIONS?

The answer is that it usually depends on the situation, but let's consider each of the letters in the universal gravity equation.

- G The gravitational constant G always has the same value, whether we are considering a rock interacting gravitationally with Earth, or Jupiter interacting gravitationally with the Sun.
- m The mass values m_1 and m_2 depend on which two objects we are considering. However, for the same two objects, the masses won't change. For example, as a rock falls to Earth, the rock's mass and Earth's mass do not change as the rock falls because it is the same rock and same Earth as the rock falls.
- r The distance between the centers of the two objects r will change as the two objects move together (as with the rock falling to Earth) or apart (as when you throw an object upward). However, it may not change significantly enough to have any effect on the force value F_g. For example, if the rock and Earth start out six million meters apart

and then the rock moves two meters closer, that change of two meters is insignificant compared to the value of r (six million meters). On the other hand, if the rock falls from twelve million meters out and falls six million meters from there, then the change in r would be significant compared to the value of r, and we'd expect the force value to likely change significantly.

✍ If we have two rocks that are three meters apart and then we move them together so that they are two meters closer, the change of two meters would be significant compared to their distance, so that there would be a significant change to the force as well. However, the masses are so small that the gravitational would be small either way.

✓ *Check Point 1.8: Suppose a 1-kg ball is thrown off a three-story building. While the ball is falling, is it reasonable to treat the gravitational force as being constant, even as it gets closer to the center of Earth? Explain.*

1.6 G vs. g

Regarding the last point in the previous section about r in the universal gravity equation, we can see that the gravitational force on a falling object may or may not change, depending on where it is and how far it falls. For most of the situations examined in volume I, the objects were always within several kilometers of Earth's surface, so the r value (distance from object's center to Earth's center) didn't change much *when compared to* r (which would be more than six thousand kilometers).

When faced with situations where one of the objects, and the distance to that object, is the same for all of the situations, it is easier to use the *simplified* gravity equation:

$$F_{\text{g}} = mg \tag{1.2}$$

where m is the mass of the object upon which the gravitational force acts (like the rock) and g represents the gravitational field strength of the *other* object (like Earth), with which the first object is interacting. Near or on the surface of Earth, the Earth's gravitational field strength is 9.8 N/kg. We can

then use that value as g in equation 1.2 to determine the magnitude of the gravitational force on any object (near or on Earth's surface) due to Earth.

WHAT IS THE DIFFERENCE BETWEEN G AND g?

There are several differences.

⁉ Whereas the value of G is the same for all cases, the value of g depends on the situation.

First, notice that the units of g differ from the units of G, as they represent different things. Second, the value of G is the same for all cases, whereas the value of g depends on the situation. The value of 9.8 N/kg only works for the gravitational field strength associated with Earth, and only when near or on the surface of Earth.

✍ If you know the gravitational field strength then it is easier to use the simplified gravity equation than to use the universal gravity equation.

✓ *Check Point 1.9: (a) Does g ever equal 9.8 N/kg? If so, when? (b) Does G ever equal 9.8 N/kg? If so, when?*

1.7 The law of force and motion

IF ALL OBJECTS INTERACT GRAVITATIONALLY, WHY DO WE ONLY SEE THINGS ATTRACTED TO EARTH, AND NOT TO EACH OTHER?

As you can see in the universal gravity equation, G has a very small value (6.67408×10^{-11} N m^2/kg^2), which means that the gravitational force will be very small unless one or both of the masses is really large. The gravitational attraction between two regular objects is much smaller than the gravitational attraction between an object and Earth because Earth's mass is much greater.

BUT AREN'T WE CLOSER TO OTHER OBJECTS THAN EARTH'S CENTER?

Even though the separation distance (center to center) between us and, say, our neighbor is much smaller than the separation distance between us and Earth (center to center), the distance is not small enough to counter the effect of the smaller mass.[xiv]

[xiv] In other words, you cannot get close enough to a small object (like another person) for the gravitational force on you due to the other object to compare to the gravitational force on you due to Earth, even though the center of the Earth is so far away.

Why are we attracted to Earth, yet Earth is not attracted to us?

Actually, for the same interaction, like between me and Earth, the gravitational force pulls equally on both me and Earth, like an invisible spring that connects me and Earth. This is what the law of interactions describes.

However, an object's change in motion not only depends upon the force exerted on it but also its mass. In particular, an object's change in motion is inversely proportional to its mass. For the same force exerted on it, a lower-mass object will experience a greater change in motion. Consequently, due to my interaction with Earth, I experience a greater change in motion than Earth does.

For that reason, we will assume that everyday objects (like you, me or ball) don't exert a measurable force on each other unless they are able to interact via another force (i.e., electric, magnetic or nuclear).

✍ Particles like electrons and protons are so light that we ignore the gravitational force between them, regardless of their separation distance.

✓ *Check Point 1.10: Suppose a 1-kg ball is sitting on a table two meters away from me. When predicting the motion of the ball, is it reasonable to ignore the gravitational force exerted by me on the ball? Explain.*

The idea that an object's change in motion is proportional to the force imbalance and inversely proportional to the mass is known as the **law of force and motion** (also known as **Newton's second law**).

⁝ The law of force and motion describes what happens to an object when forces exerted upon it.

Notice that I say force *imbalance* because there oould be multiple forces acting on the object (see section 1.8). Another term for force imbalance is the **net force**.[xv]

Mathematically, we can express the law of force and motion in terms of a force and motion equation:

$$\vec{a} = \frac{\vec{F}_{\text{net}}}{m} \tag{1.3}$$

[xv] The word "net" in "net force" is being used in the same sense as one would use the word in business (as in "net" vs. "gross").

where $\vec{a}$ is the rate at which an object's motion changes while the force imbalance $\vec{F}_{\text{net}}$ is acting on the object (of mass m).[xvi]

We will typically measure force, mass and acceleration in SI units of **newtons** (N), kilograms (kg), and meters per second squared (m/s^2), respectively.

Notice that in equation 1.3 the mass is in the denominator. For the same force exerted on it, a less massive object (like a table tennis ball) experiences a greater change in velocity than does a more massive object (like a bowling ball).

✓ *Check Point 1.11: If a* 10-*N force is exerted on a* 10-*kg ball, the ball accelerates at a rate of* 1 m/s^2*. If the same force is exerted on an electron, of mass* 9.1×10^{-31} kg*, it accelerates at* 1.1×10^{31} m/s^2*. Why are the two accelerations so different?*

It is important to recognize that the force and motion equation relates the net force with the acceleration, not the velocity. This is because, when a non-zero net force is acting on the object, its velocity is *changing*, meaning that the object is speeding up (when the net force is in the direction of the motion), slowing down (when the net force is opposite the direction of motion) and/or turning/changing directions (when the net force is perpendicular to the direction of motion).

In other words, there is no *single* velocity value that corresponds to a particular net force, yet there *is* a single acceleration value because acceleration is defined as the rate at which the velocity is changing:

$$\vec{a}_{\text{avg}} = \frac{\Delta\vec{v}}{\Delta t} \tag{1.4}$$

where $\Delta\vec{v}$ is the change in the object's **velocity**[xvii] and Δt is the length of time it takes for the velocity to change. If the velocity changes a great deal during a short time period, the acceleration is large. Conversely, if the velocity changes a small amount over a long time period, the acceleration is small.

[xvi] The arrow over the a is used to indicate that acceleration is a vector quantity, which means that it has a direction.

[xvii] The "Δ" is a Greek letter called "Delta". It is short-hand for "the change in...". Consequently, "$\Delta\vec{v}$" means "the change in velocity".

In the puzzle, a light wood ball and a heavy iron ball hit the ground at the same time when released from the same height and at the same time. Why?

From the law of gravitation, the gravitational force on the more massive iron ball is greater than the gravitational force on the less massive wood ball. However, from the law of force and motion, a greater gravitational force is required on the more massive iron ball for it to have the same acceleration (since the acceleration is inversely proportional to the mass).

As a result, with a greater force on the more massive object, each object accelerates at the same rate (speeding up as they fall).

✓ *Check Point 1.12: Since a light wood ball and a heavy iron ball hit the ground at the same time when released from the same height and at the same time, does that mean the force on each is the same? Why or why not?*

1.8 Multiple forces

Although we will generally only consider single interactions between pairs of objects, we know from volume I that, in most cases, there are multiple interactions occurring, which means multiple forces can be acting on an object.

As mentioned earlier, forces in the direction of the motion make an object speed up while forces opposite the direction of motion make an object slow down. If there are multiple forces acting on an object then whether it speeds up or slows down depends on which forces are greater – the ones acting to speed up the object or the ones acting to slow it down.

For example, drag acts against an object's motion and thus acts to slow down the object. However, if there is a propulsion force acting in the direction of the object's motion, the object will speed up if the propulsion force is greater in magnitude than the drag force.

What happens if the propulsion force is equal in magnitude to the drag force?

If the propulsion is equal in magnitude to the drag then the net force or force imbalance is zero. In that case, the object neither speeds up nor slows down. Its acceleration is zero.

Does that mean the object is stationary?

No. If the forces are balanced, that just means the object is neither speeding up nor slowing down. It could be stationary or it could be moving at a constant velocity.

How can the object be moving if the forces are balanced?

You do need a force imbalance to *get* the object moving from rest but, once moving, you don't need a force imbalance to *keep* the object moving.

⁉ The law of inertia describes what happens to an object when the forces exerted upon it are balanced.

This idea is called the **law of inertia**.[xviii] Certainly, an object at rest will remain at rest if the forces exerted on it are balanced, and the object will start moving when the forces become unbalanced. However, once started moving, if the forces again become balanced then, from that point on, the object will continue to move with the same speed and direction.

✍ Remember that the **net force** refers to the force imbalance exerted on the object. If the forces are balanced then the net force is zero.

✓ *Check Point 1.13: In which of the following situations is the net force exerted on the object zero?*
(a) A block sliding on a horizontal, frictionless surface such that it slides with a constant speed, without slowing down.
(b) A ball on a string being swung in horizontally-oriented circle, such that it moves about the circle at a constant speed.
(c) A block at rest on a table top.
(d) A ball in free fall (falling without air resistance).

Summary

This chapter examined how the gravitational force is associated with the interaction of two objects.

The main points of this chapter are as follows:

- The gravitational force is easily evident only when one or both of the objects are very massive.

[xviii] The law of inertia is commonly known as **Newton's first law**.

- The gravitational force acts even if the two objects are not in contact.
- Whereas the value of G is the same for all cases, the value of g depends on the situation.
- The law of inertia describes what happens to an object when the forces exerted upon it are balanced.
- The law of force and motion describes what happens to an object when forces exerted upon it.
- The law of interactions states that each object of the interacting pair experiences an equal and opposite force.

By now you should be able to recognize when there will be a significant gravitational force associated with the interaction of two objects and, if so, calculate the gravitational force on each object.

Frequently asked questions

If the moon is massive, why don't we notice a force between us and the moon?

The gravitational force not only depends on the masses of the two objects but also how far apart they are. The moon is so far away that its gravitational force on us is not as great as Earth's. We'd notice a force between us and the moon if we were on the moon instead of Earth.

But there is no air on the moon. Don't we need air to have gravity?

No. The air actually pushes *up* on us, since the pressure decreases as one goes up (i.e., the air pushing up on us from the bottom is greater than the air pushing down on us from above). However, that upward force (called the buoyancy force) on us is small compared to the downward gravitational force (due to Earth) so we can ignore it. We can't ignore the buoyancy force with very light objects (like a helium balloon), though.

What's the difference between a theory and a law?

For our purposes, a **law** is simply the observed relationship between two or more variables. A theory, on the other hand, is the explanation of why the variables are related in that way.[xix]

[xix] Not surprisingly, the line between theories and laws is not universally agreed upon.

For example, an English scientist named Robert Hooke[xx] found that there is a linear relationship between how much a spring stretches and how much weight is hung from it. We call this relationship Hooke's law because it describes this relationship. The law does not attempt to *explain why* the relationship is linear (as opposed to quadratic or whatever).

For more information on what a scientific law is, and how it differs from a scientific theory, see the supplemental readings.

Is a law necessarily proven correct?

No. We can never prove that a relationship is correct, although we can support it through repeated testing. There are laws that aren't correct all the time or are only correct under special circumstances (like the linear relationship for springs mentioned above).

If an object is moving at a constant speed and direction, like a ball rolling along the floor, is there a force keeping the ball moving, like a "force of the motion"?

No. While it may be common to think that there is something that seems to keep an object in motion once it is already set in motion, we cannot call that a "force" since, as pointed out in section 1.3, forces are due to *interactions between objects*.

Rather, we use the word **inertia** to describe the "tendency" of objects to continue moving in the direction they are moving. This distinction is important because the laws and such that we use are based on a precise meaning of force. If you confuse "force" with "inertia" or "effect", you are just asking for trouble. Only use the word "force" for the pushes and pulls that are applied via a source *external* to the object.

To keep an object moving, do you have to exert a force on it?

That depends on whether there are any other forces acting. If you are pushing a box across the floor, there is friction acting on the box, acting to slow it down. To prevent it from slowing down, you have to continually exert a force on it. However, if the floor was icy, with very little friction, you wouldn't need to push on the box to keep it moving (once it has already started moving).

Don't be surprised if you run across relationships in science that don't adhere to the strict definition of law and theory provided here.

[xx]Robert Hooke was born in 1635 and, in addition to his studies of springs, also wrote the first book describing observations made through a microscope and was the first person to use the word "cell" to identify microscopic structures.

Terminology introduced

Acceleration	Law	Newton's first law
Earth	Law of force and motion	Newton's second law
Gravitational force	Law of interactions	Newton's third law
Hypotheses	Mass	Newtons
Inertia	Law of inertia	Speed
Isaac Newton	Models	Theory
Kilogram	Net force	Universal law of gravitation

Abbreviations introduced

Quantity	**SI unit**
acceleration ($\vec{a}$)	meter per second squared (m/s^2)
distance (r)	meter (m)
force ($\vec{F}$)	newton (N)
gravitational field strength (g)	newton per kilogram (N/kg)
mass (m)	kilogram (kg)
time (t)	second (s)
velocity ($\vec{v}$)	meter per second (m/s)
volume (V)	meter cubed (m^3)[xxi]
gravitational constant (G)	newton meter squared per square kilogram ($N{\cdot}m^2/kg^2$)

Additional problems

Problem 1.1: Calculate the gravitational force on an electron that is 10^{-10} m away from a proton (typical radius of a hydrogen atom). Compare that to the gravitational force on the electron due to Earth, whose center is 6.37×10^6 m away (average radius of Earth). See supplemental readings for relevant mass values.

Problem 1.2: In your field of study, which of the four forces (gravitational, electric, magnetic or nuclear) do you think you'll need to understand most often, and which of the three descriptions (force, field, or energy) do you

[xxi] A non-SI unit for volume is the liter (l).

think you'll use most often? Provide an example from your field to support your answer. Ask your advisor if you have no idea.

Problem 1.3: (a) Describe a situation where an object has a non-zero acceleration and a zero velocity. (b) Describe a situation where an object has a non-zero velocity and a zero acceleration.

Problem 1.4: (a) If possible, describe a situation where an object has a non-zero net force exerted on it and a non-zero velocity. If not possible, explain why not.
(b) If possible, describe a situation where an object has a zero net force exerted on it, and a non-zero velocity. If not possible, explain why not.

2. Charge and the Electric Force

Puzzle #2: Blow up a balloon so the rubber is taut. If you bring the balloon near some small pieces of paper, nothing happens. The paper just sits there. However, if you first rub the balloon with hair or fabric and then bring the balloon near the paper, you should notice that the paper is attracted to the balloon.[i] Why are the pieces of paper attracted to the balloon after rubbing the balloon with hair/fabric but not before? What does the rubbing do and how do you know?

Introduction

In the last chapter, we looked at the gravitational force. However, most of the interactions we experience on a day-to-day basis are not gravitational in nature but electrical.

This may sound strange because saying "electric force" sounds like someone getting electrocuted. However, it turns out every physical contact you have with another object (like sitting on a chair, holding a pencil or typing on a computer) is actually an interaction involving the **electric force**.

In this chapter, we look at the characteristics of the electric force and, in particular, the idea of **charge**.

2.1 The charge model

The demonstration discussed in the puzzle can be explained in terms of a **charge model**. As discussed in section 1.1, a model is a way of looking at a

[i] This may work better when the air is dry, for reasons discussed in this chapter.

phenomenon in terms of something that is familiar to us and shares certain properties with the phenomenon. An example of a model is the way we imagined the gravitational force to be like invisible springs between objects, pulling the objects together.

Chances are that you are already familiar with the charge model.[ii] It is the idea that objects are made up of two types of incredibly tiny particles, called electrons and protons, that repel or attract depending on whether the two particles are the same (like two electrons) or different (i.e., one proton and electron).

The concept of charge can be kind of abstract but the charge model, like the other models we'll consider in this volume, can also be very powerful. It not only tells us about the electric force but also tells us something about the basic structure of matter.

Most people believe such particles exist because that is what they were taught. However, you can see these particles so what evidence do you have they exist?

I want you to have a better reason to accept the charge model than just because that is what the book says. To help you develop a strong rationale for the model, I am first going to go through why we need the model in the first place.

Basically, I'll start with considering why gravity alone cannot explain the puzzle. Then, I'll try successively better, but still weak, explanations.

We know that the "correct" explanation consists of electrons and protons. By going through the "incorrect" models you'll not only get a better sense of how we know electrons and protons exist but will hopefully also get a better sense of what it means to develop and test scientific hypotheses.

Model 1 - gravity only

As mentioned above, before considering the charge model, let's first consider why gravitation is insufficient to explain the model. After all, we know that gravitation is like invisible springs that pull objects together, so why can't the gravitational force be responsible for pulling the paper and balloon together?

[ii] The charge model is often considered to be part of the atomic theory. For information on what is a theory and how it differs from laws, see the supplemental readings.

One way to show it isn't gravitational is to use the universal gravity equation (on page 13) to determine the force on the balloon and paper.[iii] However, you don't need to use any math to show that it can't be gravitational.

Basically, we see that there is not an attraction initially. It is only by rubbing the balloon with the hair/fabric that there is an attraction. Consequently, the question becomes: could some mass have been transferred to the balloon by rubbing with the hair/fabric? And, if so, could the increase in mass and the corresponding increase in gravitational force be enough to make the paper attract to the balloon?

We don't see anything being transferred. Is that sufficient to invalidate this model?

Just because we don't see anything being transferred doesn't mean that something wasn't. After all, even a little bit of mass would change the gravitational attraction.

But would it be noticeable?

Perhaps not, but it isn't sufficient to simply state that it is *probably* not noticeable. To really test the model, we use the model to make a prediction. If the prediction is correct, that lends support for the model. If not, the model is falsified.

In this case, one prediction we can make is that, if the model is correct, we should see the same effect regardless of *how* we add mass to the balloon. For example, rather than rubbing the balloon by hair/fabric, we could simply add some tape to it. If adding tape makes the balloon attract pieces of paper, that would support our model. If not, our model would not be supported.

So what happens?

It turns out that just adding weight to the balloon doesn't make it attract pieces of paper. This serves to invalidate this particular model.

[iii] For example, in problem 2.1, you are asked to calculate the gravitational force on a 1-gram piece of paper due to a 1-gram balloon that happens to be 3 cm away. It turns out that the gravitational force on the paper due to Earth is almost 10^{11} times stronger, despite the fact that the center of Earth is much farther away.

Model 2 - the electric force

Let's consider a second hypothesis. This second hypothesis is still not the charge model but it is *similar* to it. It still has a weakness, which I will get to, but I am introducing it so you can better understand the charge model (when I get to it). Don't skip this section. As you read through it, see if you can pick up on what makes it inadequate to explain our observations.

We'll start with the idea that *something* was transferred by the rubbing but that something must be made of particles so incredibly small and practically massless that there was no perceptible difference in the balloon.

Why would anything be transferred by rubbing?

In chapter 3 we'll see why rubbing does what it does. For now, we are just trying to explain why the balloon and paper attract, and to explain that attraction we'll hypothesize that a *different* force, not the gravitational force, is responsible and that other force acts with a magnitude much *greater* than gravity. We'll call this force the **electric** force.

An advantage of this model is that it not only explains the balloon and paper demonstration but it also explains why a table stays together. In particular, if the particles in the table were attracted to each other solely by gravity, they'd fall to the floor because the gravitational force attracting the surrounding particles to each other would be much smaller than the gravitational force on them due to Earth. It is the *electric* force that allows the particles to stick together and allows the table to remain as a table.

This model is still not the charge model, since it only specifies one type of particle, rather than two (i.e., electrons and protons). Still, how can we test this model in a way that suggests the model is invalid?

The answer, again, is to use the model to make a prediction. If the prediction is correct, that lends support for the model. If not, the model is falsified.

In this case, it is relatively easy to show that this model is incorrect: these particles must be present in the paper (since they are attracted to the balloon) and so the model predicts that the pieces of paper should attract each other.

They don't. Thus, the model is not supported by our observations and so the model is falsified.

Model 3 - two particle types

Just because a model has been shown to be invalid does not mean we must throw it out entirely. In this case, let's keep the idea that particles were transferred, particles so incredibly small and practically massless that we didn't notice it. We'll also keep the idea that these practically massless particles attract one another, not with gravity, but with a *second* type of force that we call the electric force.

We'll improve upon the model, however, by utilizing the idea that the particles can either attract or *repel*. In other words, the electric force can be either attractive or repulsive. The gravitational force, in comparison, is only an attractive force.

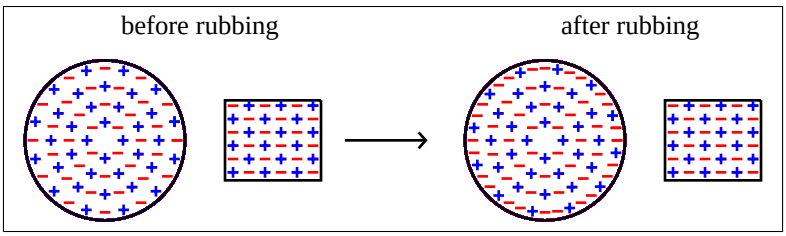
Like the previous model, this new model still has a weakness, which I will get to. And, as before, it is important that you don't skip this section. As you read through it, see if you can pick up on what makes the new model inadequate to explain our observations.

To account for both attraction and repulsion, the model distinguishes between *two* types of particles, which we can call **positive** particles and **negative** particles. The new model proposes that **opposite** particles attract (i.e., the *positive* particles are attracted to the *negative* particles) and **like** particles repel (e.g., the *positive* particles are repelled by other *positive* particles).

Like particles repel and opposite particles attract.

✍ In this model, the words "positive" and "negative" do not refer to opposite directions, as it did in volume I when dealing with vectors. Instead, it just refers to opposite types of particles.

We can use this revised model to explain why objects don't normally attract one another. Basically, we treat objects as initially consisting of equal numbers of both positive and negative particles, as illustrated on the left side of the diagram below, where the balloon is indicated by a circle and the paper is indicated by a square. Positive (+) and negative (−) signs indicate positive and negative particles, respectively.



When an object contains just as many positive particles as negative particles, we say that the object is **neutral**. According to the model so far, neutral objects should neither attract nor repel because they attract one another with the same force that they repel one another, leading to no net attraction or repulsion on each other, consistent with our observations.

⁝ By utilizing the idea of positive and negative charge, we can explain both attraction and lack of an attraction.

To see an attraction (or repulsion), there has to be *more* particles of one kind than another. According to the model, this is accomplished by rubbing the balloon with hair or fabric. The right side of the diagram illustrates how the balloon ends up with an imbalance of particles (note that the outermost layer of the balloon now has more negative particles than before; the arrangement in the paper remains the same as before). With an imbalance, the balloon is no longer neutral and there can be an attraction or repulsion.

To test this model, we could take two balloons. According to the model, two rubbed balloons (each with, say, an excess of negative particles) would repel one another. This is indeed what happens.

Does that mean this model is correct?

Not quite. While the model explains why the paper is not initially attracted to the balloon (because the balloon had equal amounts of positive and negative particles), it does not explain the attraction after the rubbing.

Why not?

Rubbing the balloon changes the balloon, not the paper. The paper has not been rubbed so it should remain the same as before: neutral. According to the model, the only way there can be an attraction is if both the balloon *and the paper* have an imbalance of particles.

Why do both objects have to have an imbalance?

If the paper has equal amounts of positive and negative particles, the net force on the paper would be zero. One type of particle would be attracted to the balloon and the other type would be repelled by the balloon. The end result would be no net force on the paper.

So, the model predicts that the paper must have an imbalance of particle type (i.e., more positive or more negative) but, at the same time, if the paper had an imbalance, the pieces of paper would repel one another (each having the same imbalance).

Such a prediction is not supported by our observations. The paper pieces did not attract or repel one another, which means that they are indeed neutral

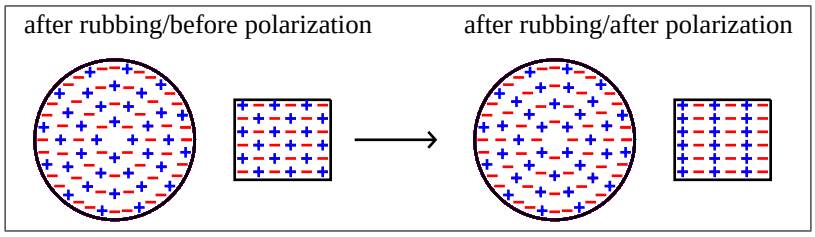
(i.e., equal amounts of positive and negative particles). Thus, there is still something wrong with our model.

Model 4 - dependence on distance

As before, just because a model has been shown to be invalid does not mean we must throw it out entirely. Model 3 is actually pretty close to being valid. The missing piece is to add the idea that the electric force on the particles (due to each other) is greater the closer they are to each other. In other words, there is an *inverse* relationship between the electric force on the two particles and their separation distance.[iv]

How does this help?

Let's suppose that the balloon has more negative particles than positive particles (as a result of the rubbing). This is illustrated in the left side of the figure below (repeated from the right side of the figure on page 29).



Since the balloon has an excess of negative particles, it repels the negative particles in the paper and attracts the positive particles in the paper. Consequently, when the balloon is close enough, the negative particles in the paper move slightly away from it and the positive particles move slightly closer. The end result is illustrated on the right side of the figure. We say that the paper is now **polarized**, meaning that it is still neutral but one side is more negative and the other is more positive.

⁉ A force inversely related to distance explains how a charged object can attract a neutral object.

Since the positive particles are now closer, they affect the force more than the negative particles. The end result is a net attraction.

[iv] The gravitational force is similar in that it also decreases as the two objects are separated. However, the gravitational force is only attractive, whereas the electric force can be either attractive or repulsive.

The model also predicts that the paper would be attracted to *any* object that has an excess of one type: positive *or* negative. This is what happens. This simple activity supports model 4, which we call the charge model.

In summary, then, our model of the electric force is that there are small particles transferred between the hair/fabric and the balloon when the balloon is rubbed. These particles, though they are practically massless (and thus don't seem to add or subtract any weight from the balloon), have a very strange property: they repel (or attract) one another with what we've been calling the electric force, a force that depends upon the separation distance.

These small particles are not only responsible for the balloon trick mentioned above, but they also let you listen to the radio, allow you to see this page and even prevent you from falling through the floor, just to name a few!

✓ *Check Point 2.1: The pieces of paper are attracted to the balloon. This shows how the positive and negative particles attract each other. Describe a simple way to demonstrate the* repulsive *nature of the particles (i.e., how positive particles are repelled by other positive particles, or how negative particles are repelled by other negative particles).*

2.2 Electric dipoles

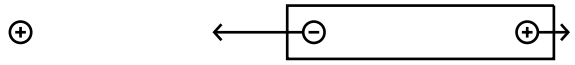
We say an object is electrically **charged**[v] when it has an imbalance of positive vs. negative particles (more of one than the other), whereas it is electrically **neutral** when there are equal numbers of positive and negative particles.

In the previous section, it was pointed out that there can be an attraction between a neutral object and a charged object if the neutral object is polarized, as with the paper when a charged balloon is nearby. However, there are some objects that are naturally polarized in that they always have a positive

[v] When we "charge" a battery we are providing energy to the battery but we are *not* making it electrically charged. Batteries are neutral and remain neutral when charged – they do not develop an imbalance of positive vs. negative particles. Notice the somewhat different use of the term "charged" in the two contexts (a charged battery vs. a charged object). Batteries are explored in chapter 11.

side and a negative side. Many atoms and molecules are naturally polarized. We call such polarized objects electric **dipoles**.

As explained in the previous section, a charged object and a polarized object attract because, according to the charge model, the electric force is stronger the closer the interacting particles. The same is true for a charged object and an electric dipole. The attraction is illustrated below.



In the figure, a dipole is indicated by a box with one negative particle on the left and one positive particle on the right. A dipole need not be a single negative particle and a single positive particle, but illustrating it in this way allows us to more easily describe what is going on.

The charged object on the left is represented as a single positive particle. Due to the electric interaction between the charged object and the dipole, there is a force on each end of the dipole. Since the negative part of the dipole is attracted to the charged object, I've drawn an arrow on the negative side of the dipole directed toward the left (toward the charged object). Conversely, since the positive part of the dipole is repelled by the charged object, I've drawn an arrow on the positive part of the dipole directed toward the right (away from the charged object).

✍ The arrows in the figure do not indicate that the two ends of the dipole repel. The two ends attract, since they are opposite types. The arrows are just showing the influence of the charged object (single positive particle on left side of figure) on each part of the dipole.

Notice how the force on the negative part of the dipole has a *greater* magnitude than the force on the positive part of the dipole. This means the net force is toward the left, and the dipole is attracted to the charged object, consistent with what was described in the previous section.

WOULD THERE STILL BE AN ATTRACTION IF WE HAD A NEGATIVE OBJECT INSTEAD OF A POSITIVE OBJECT?

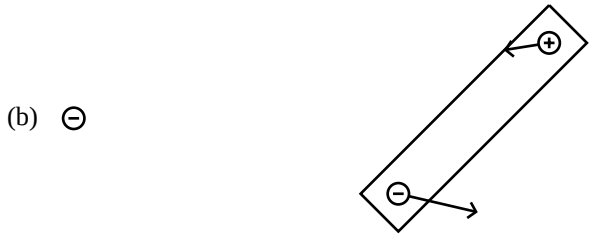
Yes, but not with the dipole orientation that is shown. However, that particular orientation is unlikely, as the forces on the dipole will tend to make the dipole flip to an orientation where there *is* an attraction.

First, I'll show you why there isn't an attraction with this particular orientation, and then I'll show you why the dipole will likely flip.

The following illustration shows the same dipole as before, with the same orientation as before (negative on left, positive on right), but with a particle to the left that is *negative* rather than positive.

Again, the arrows on the dipole indicate the influence of the left object (negative particle) on the dipole.[vi] Notice, as before, that the force on the left end has a greater magnitude. However, since that end is *repelled* by the charged object (negative this time), the net force on the dipole is away from the charged object (to the right), not toward it.

However, it is very hard for a dipole to maintain this orientation. Just due to random variations in orientation (and bumping into other molecules and such), the dipole will at some point shift its orientation. As it does so, the dipole will rotate toward an orientation where it will be attracted, not repelled. To see why, consider the illustration below, where I've drawn the dipole at a small angle from how it was before.



✍ Due to random fluctuations, it is just as likely to turn the opposite way but, regardless of which way it initially turns, the key point is that it will then continue to turn in that direction until it is aligned.

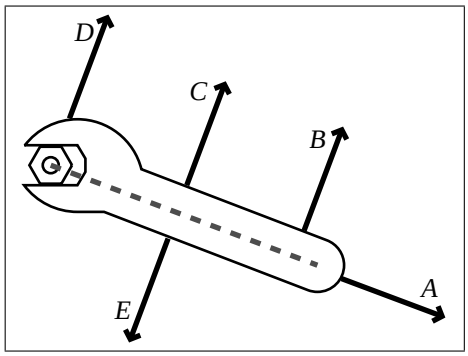
[vi] Admittedly, it looks like the arrows are indicating how the two ends of the dipole attract one another. While it is true that the two ends attract, since they are opposite, that is *not* what the arrows represent. If they did represent that, the two arrows would have to be the same size, since the forces associated with a single interaction must be equal in magnitude (as per the law of interactions).

With this orientation, the net force is still away (and downward) since the force repelling the negative end is still stronger than the force attracting the positive end. However, there is also a **torque** on the dipole, making it rotate counter-clockwise.

What is torque?

You might remember from volume I that the torque[vii] on an object tells us whether the object will rotate and won't (or, more precisely, whether it will spin faster or slower and when it won't).

To see why rotation has to do with torque and not *force*, consider the figure to the right, which shows a wrench being used to loosen or tighten a nut (the hexagon) that sits on a bolt (the circle inside the hexagon). The arrows represent different forces that will be discussed shortly. All five forces have the same magnitude but not the same directions nor the same application point.



Suppose the nut is free to spin around the bolt (which, in turn, is fixed to some other object and not free to move). Which of the forces indicated would make the wrench unscrew the nut?

To unscrew the nut, we need to get the nut rotating in the counter-clockwise direction. To change the rotation rate in the counter-clockwise direction, we need to apply a force in the counter-clockwise direction.[viii]

For the situation shown in the figure, only the forces represented by arrows B and C would make the wrench rotate counter-clockwise. In comparison, the force represented by arrow E is acting in the opposite direction, clockwise around the rotation axis, and thus would act to tighten the nut.

What about the forces represented by arrows A and D?

Arrow A represents a force that is not in the correct direction. Since force A is directed away from the nut, it simply pulls the wrench off the nut, and

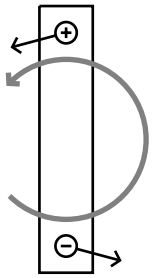
[vii] The word "torque" comes from Latin word *torquere*, which means "to twist". The words "torsion" and "torture" come from the same root word.

[viii] If friction is present (which makes the nut hard to loosen) then that friction applies a clockwise torque, opposing the counter-clockwise torque that is being applied to loosen it.

doesn't change its rotation rate. Pushing in a direction opposite A would also not work – it would just push the wrench onto the nut, not make it tighten or loosen the nut. And force D doesn't do anything because it is applied right at the nut. It would be like trying to open a door by pushing on the hinges. The further a force is applied from the axis of rotation, the greater the torque and the more effective it is at making the nut spin faster or slower.

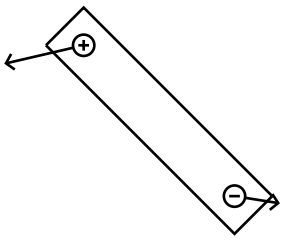
So, getting back to our dipole example, the dipole will rotate counter-clockwise due to the counter-clockwise torque on it associated with its interaction with the charged object. Let's now consider what happens as it rotates. At some point, it will rotate to the following orientation.

(c) ⊖



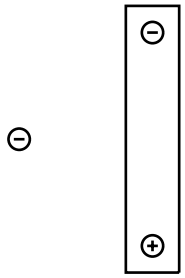
Again, notice the directions of the forces. I've added a gray circular arrow to indicate the direction of the torque, which is counter-clockwise. Due to this counter-clockwise torque, the dipole continues to rotate. At some point, it will look like the following.

(d) ⊖

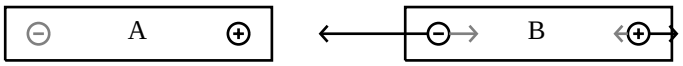


Notice how the positive end of the dipole is now closer to the negative particle, which results in a net attraction. There is still a torque, and the dipole will continue to rotate as it attracts, eventually ending up aligned with the positive end of the dipole toward the negative particle and the negative end of the dipole away.

✓ *Check Point 2.2: (a) For the negative particle and dipole depicted in the illustrate to the right, is the dipole attracted to the negative charge, repelled away from the negative particle, or neither?*
(b) Is there a net torque on the dipole? If so, will the torque rotate the dipole into a new orientation where your answer to (a) would be different?



Not only is there an attraction between a charged object and a neutral electric dipole, but two neutral dipoles will also attract each other. This is illustrated in the figure below.



In the figure, I've drawn two dipoles, A (on the left) and B (on the right). For each particle in dipole B, I've drawn two arrows, one for the force due to its interaction with A's positive particle (black) and one for the force due to its interaction with A's negative particle (gray).

Since the negative particle in A is farther away from dipole B, the forces due to that negative particle (gray arrows) have a smaller magnitude than the forces due to the positive particle in A (black arrows).

The net effect is to have a greater force of attraction than repulsion, and the two dipoles attract.

⁉ By utilizing the idea of an inverse relationship with distance, we explain how two dipoles attract.

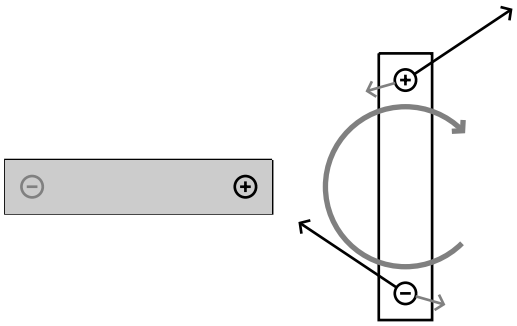
Indeed, the strongest force is between the positive particle in A and the negative particle in B, since they are nearest each other. Consequently, if we are only considering whether two dipoles attract or repel, we can just consider the ends that are closest to each other.

✍ The forces due to the negative particle in A (in gray) not only have a smaller magnitude than the forces due to the positive particle (in black), but the difference in magnitude between the two gray arrows is smaller than the difference in magnitude between the two black arrows. That is why there is a net force of attraction.[ix]

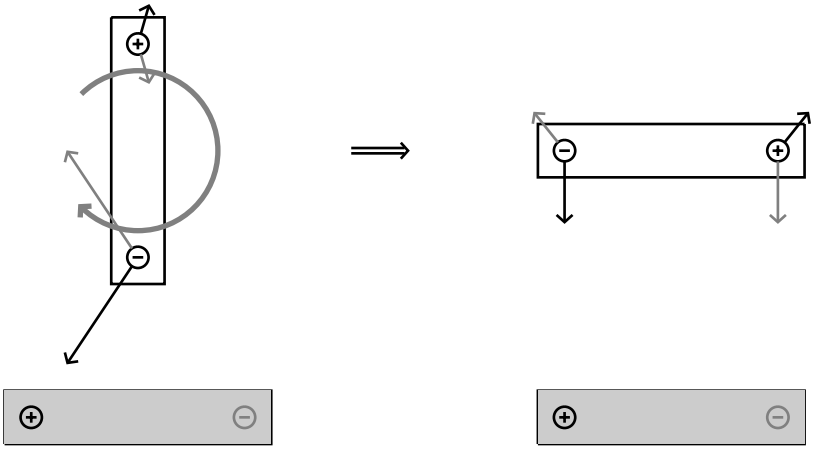
[ix] This is similar to how the gravitational force on you doesn't significantly change when you jump in the air, since the difference in distance is small compared to how far away the Earth's center is from you.

The two dipoles are aligned in this example but, regardless of their initial orientations, they'll still end up attracting because the torque eventually leads to a configuration where they attract. However, with dipoles there are two final configurations that are possible: aligned and opposite.

The first case occurs when the two dipoles are oriented as shown below. The black and gray arrows representing the forces on the right dipole (unshaded) due to the top (positive) and bottom (negative) parts, respectively, of the left dipole (shaded), which is fixed and unable to rotate. Due to the forces on the right dipole, there is a clockwise torque on it, rotating it toward the orientation illustrated in the previous figure.



The second case results when the two dipoles are oriented as shown on the left half of the figure below. Again, the left dipole (shaded) is fixed. As you can see, there is a clockwise torque on the right dipole (unshaded). This makes the dipole rotate clockwise, leading it to the orientation shown on the right half of the figure, where the two dipoles are oppositely aligned. Notice how the end result is still an attraction.



WHICH FINAL ORIENTATION IS MORE LIKELY?

It generally depends on whether the dipoles are rod-like (long and slender) or button-like (short and thick). Rods tend to end up side-by-side whereas buttons tend to end up aligned.

Keep in mind, however, that they only attract because they are dipoles. If not dipoles, neutral objects do not attract or repel.

✍ This is why atoms bond. For example, water molecules are naturally polarized. Consequently, if water molecules in the vapor state are moving slow enough (i.e., the temperature is not so great), they can attract one another and bond, forming liquid water. Being dipoles, they will also be attracted to a charged object (like the balloon in the puzzle). This can serve to "siphon" off some of he excess charge on the balloon, making it less charged and less likely to attract the neutral pieces of paper. This is why the balloon demonstration works better when the air is dry (as in the water) than when the air is humid (like in the summer).

✓ *Check Point 2.3: Why do two neutral people not attract one another electrically but two neutral water molecules do?*

2.3 Electrons and protons

⁝ Electrons carry negative charge and protons carry positive charge.

As you may have suspected, the positive and negative particles are called **protons** and **electrons**, respectively. Each particle has a property called **charge**. Protons have positive charge and electrons have negative charge.

All matter has lots of electrons and lots of protons. In fact, there are zillions and zillions of electrons inside you at this very moment!

IF ALL OF THESE ELECTRONS ARE INSIDE ME AND EACH ONE REPELS THE OTHERS, WHY DON'T WE JUST EXPLODE?

The reason why you don't explode is because you are electrically neutral – for every electron in your body, there is one proton. The protons and electrons **attract** each other. The protons keep the electrons from flying off of us. Likewise, the electrons keep the protons from flying off of us.

For objects that are not neutral, there is an imbalance of charge. We say that such an object has a charge, and that charge refers to the imbalance. For example, if an object has an excess of electrons (negatively-charged particles), we say that it has a negative charge. A neutral object, with equal numbers of electrons (negatively-charged particles) and protons (positively-charged particles), is said to have no charge.

Example 2.1: A hydrogen atom is made up of one proton and one electron. Is a hydrogen atom positively charged, negatively charged or neutral?

Answer 2.1: A hydrogen atom is neutral. The proton has positive charge and the electron has an equal but opposite negative charged.

As mentioned earlier, protons and electrons must be very tiny, since in our balloon demonstration some particles were transferred yet we were unable to see any of them. Indeed, the mass of a proton is only 1.673×10^{-27} kg and the mass of an electron is about a thousand times smaller at only 9.11×10^{-31} kg. Neither is big enough to see, even with the most powerful microscopes.

❣ Protons are about a thousand times more massive than an electron.

Mathematically, we abbreviate the masses of the proton and electron as m_p and m_e, respectively. Consequently, we can write the following:

$$\begin{aligned} m_p &= 1.673 \times 10^{-27} \text{ kg} \\ m_e &= 9.11 \times 10^{-31} \text{ kg} \end{aligned}$$

To get a sense of just how tiny these particles are, consider how many of these particles are needed to be equal to the mass of a typical person, who might have a mass of 75 kg (about 165 pounds).

Each proton is only 1.673×10^{-27} kg. Divide the person's mass by the mass of a proton to get a total of about 4.48×10^{28} protons.

Is that the number of protons in a typical person?

No, that would only be the number of protons if the person was made up solely of protons. In reality, a person's mass is roughly split between that due to protons and that due to neutrons.[x]

What is a neutron?

[x] According to the CRC Handbook (2008-9 89th edition 7-24), 88% of the human body

A **neutron** is a third type of particle. A neutron is approximately the same mass as a proton but does not have any charge. There are roughly equal amounts of protons and neutrons in your body. Consequently, only half our body is made up of protons. That is still a huge number (2.24×10^{28} protons)!

✍ Since neutrons have no charge, they are neither attracted to nor repelled by electrons and protons. For this reason, we don't need to concern ourselves with them when dealing with the electric force. Since they have mass, they still contribute to the overall mass of an object and the gravitational interactions, however. We'll examine them in more detail in the next chapter.

ISN'T SOME OF OUR MASS DUE TO ELECTRONS?

Since we are neutral, for every proton in our body there is an electron. However, the mass of each electron is about 1000 times smaller than the mass of each proton. Consequently, we can ignore the contribution of the electrons when considering the mass of the person.

✓ *Check Point 2.4: In our estimate of the number of protons in a typical person, we ignored electrons.*
(a) Is it reasonable to ignore the mass of the electrons? Why or why not?
(a) Since we are neutral, how electrons should there be in a typical person?

2.4 Law of electric force

Suppose we had two protons. Since they both carry positive charge, they repel. But what is the force? To make any numerical predictions, we need to quantify the force.

As you have seen in section 2.1, the electric force associated with the interaction of two particles depends upon two things:

---

is made up of oxygen (O), carbon (C), nitrogen (N) and calcium (Ca), and almost all atoms of O, C, N and Ca have equal numbers of protons and neutrons (actually it is 99.8% of O, 98.9% of C, 99.6% of N and 96.9% of Ca). It is for this reason that I say "roughly half" is made up of protons. Actually, there are probably more protons than neutrons, since 10% of the remaining 12% not covered by O, C, N and Ca is made up of hydrogen (H, which is mostly 1 proton with no neutron). I've ignored electrons since their mass is more than a thousand times less than that of the protons (even though they are equal in number).

1. The charge of each particle
2. how far apart the particles are

In equations, we'll represent the value of the charge via the letter q and we'll use the letter r for the distance between the two particles (actually, the distance from center to center, as we did for the gravitational force; see equation 1.1).

WHY DO WE USE THE LETTER q FOR CHARGE?

You can think of q as standing for the *quantity* of charge. The key point is that the magnitude of the electric force, which will be indicated as $|\vec{F}_e|$, depends upon the charges of the interacting objects, which will be indicated as q_1 and q_2, as well as their separation distance r.

WHAT IS THE RELATIONSHIP BETWEEN $|\vec{F}_e|$, q_1, q_2 AND r?

The **law of electric force** provides the relationship between the magnitude of the electric force $|\vec{F}_e|$, the charges of the interacting objects q_1 and q_2, and the separation distance r. Mathematically, it is written as follows:

⁝ The law of electric force describes how the electric force depends upon the charge on the interacting objects and their separation distance.

$$|\vec{F}_e| = k\frac{q_1 q_2}{r^2} \tag{2.1}$$

where k always has a value of 8.98755×10^9 N m^2/C^2.

✍ The law of electric force is often called **Coulomb's law**, after the French physicist Charles Augustin Coulomb (1736-1806), who published this relationship in 1784.

You may have noticed that the electric force equation is very similar in structure to the gravitational force equation, where the charge of the objects is analogous to the masses of the objects. In addition, like the gravitational force, the electric force decreases according to the square of the separation distance (r).

✍ Equation 2.1 not only works for individual particles like protons and electrons but also for objects that contain an excess (or deficiency) of one or the other. In such cases, we can simply treat the object as a single particle with a large charge.

✓ *Check Point 2.5: (a) What does the letter q represent in equation 2.1? (b) What does F_e represent in equation 2.1?*

2.5 Units of electric charge

What is the meaning of k in equation 2.1?

In equation (2.1), k represents a conversion factor between (a) the units of charge and distance on the right side of the equation and (b) the unit of force on the left side of the equation. We call k the **electric force constant** because it has the same value for all electric interactions.

I know that force is measured in units like newtons and that distance is measured in units like meters. What kind of units do we use for charge?

We quantify the charge via the **coulomb** (abbreviated as C), a unit named after Charles Augustin Coulomb, the same French physicist who first published the law of electric force. Note that we use Roman type for unit abbreviations (like C) and Italic type for variable abbreviations (like q).[xi]

⁇ Charge is measured in coulombs.

Is a Coulomb equal to the number of electrons or protons?

No. The number of excess electrons or protons is *related* to the charge in coulombs but the charge in coulombs is not equal to the number of excess electrons or protons.

✍ In chemistry, it may be common to indicate the amount of imbalance in terms of the excess *number* of one particle over the other (as with an **ion**, which is an atom that has an unequal number of protons and electrons). However, that number is *not* the charge in coulombs.[xii]

Instead, coulomb is a unit, like kilogram, except for charge, not mass. Just like you need lots of atoms to equal a mass of one kilogram, you need lots of electrons to equal a charge of one coulomb. Indeed, one coulomb is defined to be the charge of 6.2415×10^{18} protons. That is a lot of protons! It makes sense to use a unit that represents the charge of lots and lots of protons since most charged objects have lots and lots of excess electrons or protons.

Since it requires a whole bunch of protons to have a charge of one coulomb, that means that the charge of a single proton must be a tiny fraction of one

[xi] All unit abbreviations are listed in the supplemental readings.

[xii] In a similar way, one can refer to an atom's **valence number**, which is the *number* of electrons an atom's outermost shell needs to be "full" (which influences how the atom will bond with other atoms). The valence number is associated with a *neutral* atom (i.e., the atom has zero charge), however, with equal numbers of protons and electrons.

coulomb. Using q_e and q_p to represent the charge of a single electron and proton, respectively, one can show[xiii] that they are equal to the following:

$$\begin{aligned} q_e &= -1.60218 \times 10^{-19} \text{ C} \\ q_p &= +1.60218 \times 10^{-19} \text{ C} \end{aligned}$$

WHY IS ONE POSITIVE AND THE OTHER NEGATIVE?

By convention, we say that the electron has a negative charge and the proton has a positive charge. Consequently, 6.2415×10^{18} electrons would have a charge of negative one coulomb.

Since we are usually only concerned with the charge imbalance, not the total amount of negative or positive charge, it is sufficient to just consider how many more electrons we have than protons, or how many more protons we have than electrons. We can then use the "excess" amount to determine the charge imbalance.

For example, suppose we had 10 million protons and 11 million electrons. Since we have one million more electrons than protons and each electron has a charge of -1.60218×10^{-19} C, we can multiply that charge by one million to get a charge imbalance equal to -1.60218×10^{-13} C.[xiv]

✍ Just as the force imbalance is also called the net force, the charge imbalance is called the **net charge**.

✓ *Check Point 2.6: Suppose there were* 2×10^{28} *protons in your body and* 2×10^{28} *electrons in your body (so you are neutral).*
(a) What (net) charge (in coulombs) would you carry if one-tenth of all the protons in your body suddenly disappeared?
(b) What (net) charge (in coulombs) would you carry if 90% of all the protons in your body suddenly disappeared?
(c) In which situation is there a greater net charge: the situation described in (a) or the situation described in (b)? Explain why that should be so.

[xiii] There are 6.2415×10^{18} protons per coulomb. Take the inverse of that to get the number of coulombs per proton. The process is the same for the electron, except the charge is negative. Note that the charge of a single electron or proton is only a tiny fraction of a coulomb, which should make sense since one coulomb represents the charge of a huge number of electrons and protons.

[xiv] Notice how the charge imbalance is still very small. Just think how tiny the electrons are that a million of them still don't add up to much charge!

2.6 Comparison with gravitational force

In general, for objects with charge, the electric force exerted on each (due to their charge) is much greater than the gravitational force exerted on each (due to their mass). One clue to this is the fact that the constant k is many orders of magnitude larger than the constant G.

To illustrate, let's suppose we have two protons that are 10^{-10} m apart (about the size of an atom). We know that the two protons will repel one another due to the electric force. At the same time, there is an attraction due to the gravitational force. Which is larger?

It turns out that the electric force of repulsion (for two protons) is over 36 orders of magnitude greater than the gravitational force of attraction (for the same two protons).[xv]

In fact, for the forces to be equal, the proton would have to have a mass of 2.2×10^{-9} kg. This seems small but it is still almost 10^{18} times bigger than the actual mass of a proton.[xvi]

For this reason, we will almost always neglect the gravitational force when dealing with individual protons and electrons.

✓ *Check Point 2.7: The nucleus of an atom contains no electrons, only protons and neutrons (recall that neutrons are neutral).*
(a) A fellow student argues that the electric force holds the nucleus together. Do you agree or disagree? Why?
(b) A fellow student argues that the gravitational force holds the nucleus together. Do you agree or disagree? Why?

[xv] I determined the value (36 orders of magnitude) by dividing the expression for the electric force by the expression for the gravitational force (see chapter 1). I then used the charge and mass of the proton (along with the conversion constants k and G).

[xvi] This can be shown by simply setting the expression for the electric force equal to the expression for the gravitational force, and then plugging in the charge of the proton and the conversion constants.

Summary

This chapter examined how objects containing charge interact via the electric force.

The main points of this chapter are as follows:

- Like particles repel and opposite particles attract.
- By utilizing the idea of positive and negative charge, we can explain both attraction and lack of an attraction.
- A force inversely related to distance explains how a charged object can attract a neutral object and how two dipoles attract.
- Protons are more than 1000 times more massive than an electron.
- Electrons carry negative charge and protons carry positive charge.
- The law of electric force describes how the magnitude of the electric force depends upon the charge on the interacting objects and their separation distance.
- Charge is measured in coulombs.

By now you should be able to describe the evidence used to support the charge model and use the model to predict the force on objects given their charge or imbalance of electrons vs. protons.

Frequently asked questions

If all objects are composed of positive and negative charges, why don't objects attract one another?

Most objects are neutral, which means they have equal numbers of positive and negative charges and so a second object, even if it has a net charge, would be attracted as much as it is repelled.

If an object is uncharged, that seems to imply that there are no charged particles present. Wouldn't it make more sense to say that the object has no *net* charge?

Yes, but since every object we are familiar with has *some* charged particles (protons and electrons), "uncharged" or "no charge" is typically assumed to mean "no net charge".

ARE THERE ANY OTHER WAYS FOR AN OBJECT TO OBTAIN A CHARGE OTHER THAN RUBBING?

Yes. Chapter 10 discusses other ways of giving an object an electric charge.

DOES THIS HAVE ANYTHING TO DO WITH STATIC ELECTRICITY?

Yes. An example of static electricity is when two pieces of clothing stick together after going through a dryer. The reason it is called "static" electricity is because electric charges are relatively stationary (on the clothes, balloon or wherever). The word "static" is a general term meaning stationary.

IF CHARGED OBJECTS ARE ATTRACTED TO NEUTRAL OBJECTS, WHY AREN'T PROTONS AND ELECTRONS ATTRACTED TO NEUTRONS?

A neutral object will be attracted to charged object only if the neutral object can be polarized (i.e., one side positive and one side negative). We consider the neutron to be a single particle and thus cannot be polarized.

IS TORQUE THE SAME THING AS FORCE?

No. While they are similar in some ways they are not the same thing and they have different SI units (N·m vs. N). See page 35.

Terminology introduced

Charge	Electric force constant	Neutron
Charged	Electrons	Polarized
Coulomb	Ion	Positive charge
Coulomb's law	Law of electric force	Protons
Dipoles	Negative charge	Torque
Electric force	Neutral	

Abbreviations introduced

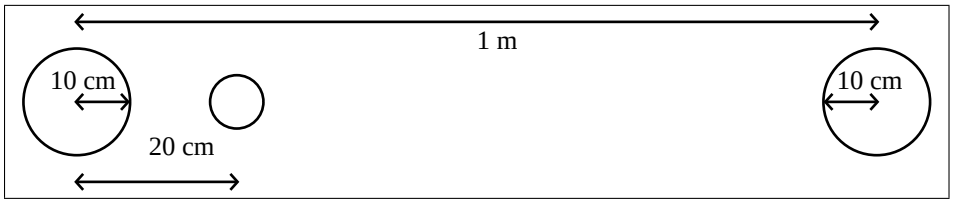
Quantity	**SI unit**
charge (q)	coulomb (C)
electric force constant k	newton meter squared per square coulomb ($\mathrm{N{\cdot}m^2/C^2}$)

Additional problems

Problem 2.1: Suppose a balloon and some paper both have a mass of 1 gram and the balloon is held 3 cm above the paper.
(a) What is the downward gravitational force on the paper *due to Earth*?
(b) What is the upward gravitational force on the paper *due to the balloon*?
(c) Given your answers to (a) and (b), does the paper rise upward? If not, how close must the balloon get to the paper for the paper to rise upward?
(d) A sheet of paper is about 0.05 to 0.10 mm thick. How does the answer in (c) compare to the thickness of a sheet of paper and what does that tell us about the validity of ignoring the gravitational force due to the balloon?

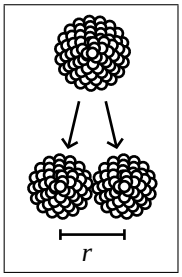
Problem 2.2: Two spheres of radius 10 cm and charge 1 C are arranged such that their centers are 1 meter apart.
(a) What is the magnitude of the electric force on one of the spheres due to the other? Is it attractive or repulsive?
(b) What would the magnitude be if the separation distance doubles?
(c) What would be the net electric force on a third object of mass 10 kg and charge 2 C if the third object was placed directly between the two spheres (50 cm from both spheres)?

Problem 2.3: Consider the two spheres from the previous problem (radius 10 cm, charge 1 C, 1 meter apart) but with a 10-kg 2-C object placed between the two spheres, 20 cm from one and 80 cm from the other, as in the figure below. What is the net electric force on the middle object?



Problem 2.4: The nucleus of uranium-238 contains 92 protons and 146 neutrons. Suppose the nucleus splits into two smaller spheres, each having 46 protons and 73 neutrons (see figure below right),[xvii] with the two smaller spheres separated by 10^{-14} m center-to-center (r in figure; about the

diameter of the nucleus).
(a) What is the magnitude of the electric force on one of the spheres due to the other? Is it attractive or repulsive?
(b) What is the magnitude of the gravitational force on one of the spheres due to the other? Is it attractive or repulsive?
(c) Which is stronger: the electric force or the gravitational force? What will happen to the two spheres (i.e., will they fly away from each other or will they recombine)?



[xvii] This is a highly simplified model of nuclear fission, which typically involves U-235, not U-238, because U-238 is less likely to break apart. In addition, the fragments are likely to be of unequal size, like Rb-90 and Cs-143 or Kr-89 and Ba-144. Several individual neutrons are also produced. Regardless of what actually happens, however, the same physics apply.

3. Nucleons and the Nuclear Force

Puzzle #3: If the electric force of repulsion between the two protons in a helium nucleus is so much greater than the gravitational force of attraction between the protons, how does the nucleus stay together?

Introduction

The puzzle raises the question of what keeps the protons together in the nucleus if they repel one another, being as they all have positive charge. It turns out that we cannot explain it with just the gravitational and electric forces because the gravitational force attracting the protons together is not strong enough to counter the electric force repelling them apart. This chapter, then, introduces a third fundamental force, the nuclear force, which we can use to explain the structure of an atom.

3.1 The nuclear vs. electric force

Experiments have shown that every atom is made up of a small positive-charged center, called the **nucleus**, surrounded by a negative-charged cloud. These experiments also show that the nucleus is much, much smaller than the cloud. Indeed, if the nucleus was the size of a penny, the electron cloud (and thus the size of the atom) would be the size of a football field. Most of the atom is just empty space.

We know from the previous chapter that an atom is made up of protons and electrons, so it makes sense that the electrons are in the negative-charged cloud and the protons are in the positive-charged nucleus. However, we also know that the gravitational force attracting the protons together is not strong enough to counter the electric force repelling them from each other.

✍ There *is* an attraction between the protons and the electrons but the electrons are outside the nucleus, meaning that the electrons add to the forces pulling the protons away from the nucleus.

To explain how the protons can stay together in the nucleus, we must introduce a third fundamental force: the **nuclear force**. The nuclear force acts like a glue in that it is an attractive force that is significant only when the protons are very, very close to each other. At very small distances, as in the nucleus, the nuclear force of attraction is greater than the electric force of repulsion and the protons are "stuck" to the other particles in the nucleus.

⁝ Like the gravitational and electric forces, the nuclear force is a non-contact force.

⁝ The nuclear force is so dependent on the separation distance that unless the particles are very, very close the nuclear force is insignificant.

If the nuclear force of attraction between two protons is so much greater than the electric force of repulsion, how is it that we ever get to see any electric repulsion at all?

As mentioned above, the nuclear force of attraction is stronger than the electric force of repulsion only at very, very small distances. Once we get separation distances much larger than the radius of the nucleus, the electric force overwhelms the nuclear force and we can ignore the nuclear force.

✍ In problem 2.4 the nuclear force was ignored. The nuclear force will add an attractive force to the problem. Still, with a separation of 10^{-14} m (the distance provided in the question), the nuclear force may be too small to keep the two fragments together.

This model of the atom predicts that the electrons are much more likely to transfer from atom to atom (by rubbing or chemical reaction) than the protons, as the protons are attracted to the nucleus by the strong nuclear force whereas the electrons are attracted to the nucleus by the relatively weaker electric force (though still much stronger than the gravitational force). In addition, the electrons are so much lighter than the protons that it doesn't take as much force to accelerate them.

Why does rubbing a balloon with hair or fabric cause electrons to transfer?

The balloon (i.e., rubber) has a significantly different affinity for electrons than the hair or fabric. When rubbed by the hair or fabric, the electrons will tend to leave for the object that has a stronger affinity for electrons. This does not necessarily mean they are extracted easily (i.e., they won't go on their own) but rather that they can be "rubbed" off if one material has a

higher affinity for electrons than the other.[i]

Do electrons transfer when any two materials are rubbed together?

No. You need to have two materials that have significantly different affinity for electrons. Otherwise, there would be no reason for the electrons to move from one object to the other. A **triboelectric table** (or triboelectric series) lists the relative affinities of various materials.

✓ *Check Point 3.1: Two protons, being of like charge, repel one another via the electric force. What force keeps the protons in the nucleus from flying away from each other?*

3.2 Neutrons

Additional experiments have shown that the mass of the nucleus is greater than what can be accounted for by protons alone. That finding suggests that, in addition to protons, the nucleus also contains neutral particles, which we call **neutrons**.

The general term for particles in the nucleus is **nucleon**, so a nucleon can be either a proton or a neutron. Be careful with the words "neutron" and "nucleon", as they look very similar but have different meanings. The latter (nucleon) is the general term used for any particle in the nucleus, and includes both protons and neutrons.

Why would there be neutrons in the nucleus?

It turns out that the attractive nuclear force, by itself, is not sufficient to keep two protons from repelling via the electric force. The neutrons are needed for additional nuclear attraction[ii] to counter the electric repulsion between the

[i] Not only does the rubbing involve pressing the objects together, ensuring close contact between the surface atoms that then allows the electrons to jump to the more favorable material, but the rubbing provides for contact with multiple atoms along the surface.

[ii] We can treat the nuclear force as purely attractive. However, it technically must become repulsive at very, very close distance, in order to prevent neutrons and protons from occupying the same location.

protons. This attractive force exists between *all* the nucleons, both protons and neutrons. This means that protons not only attract other protons but they also attract neutrons, and neutrons attract other neutrons.

Does there need to be equal numbers of protons and neutrons in a nucleus?

Not necessarily.

The need for neutrons is a little like the need for masks as a way to prevent the spread of respiratory diseases. If you are in a room by yourself, you don't need a mask. Similarly, a proton by itself doesn't need a neutron. If you are in a room with one or two other people, each person needs a mask. Similarly, in a small nucleus, each proton needs a neutron. If you are in a very crowded room, you may feel the need to wear a mask with multiple layers. Similarly, in a crowded nucleus with lots of protons, each proton needs lots of neutrons.

Indeed, if you consider the nuclei of naturally occurring elements, you'll find that hydrogen (with a single proton) doesn't need a neutron. With low-proton elements, like carbon and oxygen, there tends to be equal numbers of protons and neutrons in the nucleus. With high-proton elements, like uranium, there are many more neutrons than protons.

This pattern is seen in Figure 3.1, where each stable[iii] nucleus is indicated by a dot. As you can see, nuclei with 15 protons or less typically lie close to the diagonal line, which indicates that they have roughly the same number of protons as neutrons. For example, a helium nucleus typically has equal numbers of protons and neutrons (two of each).

As the nuclei get heavier, however, the dots lie within the range above the diagonal line, meaning a greater ratio of neutrons to protons tends to increase, with more and more neutrons compared to protons. For example, a platinum nucleus (78 protons) has 50% more neutrons than protons.[iv]

Why are there more neutrons than protons for the heavier nuclei?

Remember how it was mentioned that the nuclear attraction only acts at a short-range? That means that an individual proton is only attracted to

[iii] A stable nucleus is one that won't change or fall apart (see section 3.4 for more on this).

[iv] A platinum nucleus typically has 78 protons and 117 neutrons.

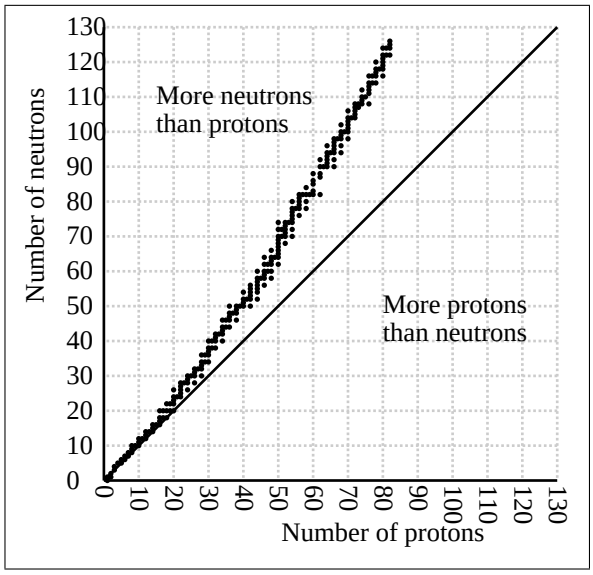


Figure 3.1: Each dot represents a naturally-occurring stable nucleus, based on the number of protons in the nucleus (horizontal axis) and number of neutrons in the nucleus (vertical axis). Dots above the diagonal line represent nuclei that have more neutrons than protons whereas dots below the diagonal line represent nuclei that have more protons than neutrons.

the immediate protons and neutrons around it. In comparison, the electric repulsion acts at farther distances. That means that each proton is repelled by all the other protons in the nucleus.

Consequently, every time you add a single proton to the nucleus, you increase the electric repulsion between *all* the protons, but every time you add a neutron, you only increase the nuclear attraction between nearby nucleons. That means we may need to add multiple neutrons throughout the nucleus every time we add a single proton. In other words, as we add protons, a greater and greater number of neutrons are needed as "glue" to keep them all together.

If neutrons act as additional glue for the nucleus, why not have a whole bunch more neutrons for all nuclei?

Neutrons are inherently unstable by themselves[v], so too many neutrons will result in an unstable nucleus (and decay; see section 3.4). For a particular

[v] An individual neutron by itself tends to last around 15 minutes before decaying.

element, there are particular ratios that are more "stable" than other ratios, and the stable ones are indicated in Figure 3.1. Indeed, once you get above lead (with 82 protons), the nucleus needs so many neutrons just to keep the nucleus together that there are no stable ratios.

✓ *Check Point 3.2: What is the problem with having too few neutrons in the nucleus? What is the problem with having too many neutrons in the nucleus?*

3.3 Isotopes

In the previous section, I mentioned that a helium nucleus *typically* has equal numbers of protons and neutrons (two of each). I used the word "typically" because a helium nucleus could also have unequal numbers (two protons and one neutron) and still be stable.

As another example, consider the nucleus of a carbon atom, which always has six protons. While there are typically six neutrons in the nucleus of a carbon atom (along with the six protons), there are atoms of carbon out there that have seven or eight neutrons in the nucleus instead of six. Only the carbon nuclei with six or seven neutrons are stable, but there still exists carbon nuclei with eight neutrons, which is unstable.[vi]

Nuclei of the same element (i.e., same number of protons) but different numbers of neutrons are called **isotopes** of each other. For example, one isotope of carbon has six neutrons while another isotope of carbon has seven neutrons. Both isotopes are carbon, though, because they both have six protons. Any nucleus with six protons is a carbon nucleus.

⁝ Each element is defined according to the number of protons in its nucleus.

Indeed, each element is defined by the number of protons in its nucleus, which is called the **atomic number**, not the number of neutrons or the total number of nucleons (both protons and neutrons).[vii]

[vi] Nuclei of carbon with eight neutrons are produced when a neutron interacts with a nitrogen nucleus in the atmosphere, kicking out a proton in the process. Since nitrogen has seven protons and typically seven neutrons, the result is a nucleus with six protons and eight neutrons.

[vii] The **atomic weight** (or **molecular weight**) is roughly equivalent to the average

WHY IS THE NAME OF THE ATOM BASED ON THE NUMBER OF PROTONS IN THE NUCLEUS RATHER THAN THE NUMBER OF NEUTRONS IN THE NUCLEUS OR THE TOTAL NUMBER OF NUCLEONS?

The name is based on the number of protons because it is the proton number that influences the electric properties of the nucleus (since the neutrons have no charge). For example, all oxygen atoms have eight protons in the nucleus, regardless of how many neutrons are present. This is why the periodic table (see last page of this book) is organized by number of protons (atomic number). Each name corresponds to a particular **element**. All isotopes of a particular element have the same number of protons.

✍ There are no electrons in the nucleus. Consequently, changing the number of electrons does not change the element type. Atoms that differ only in the number of electrons are called **ions** of the isotope.

✓ *Check Point 3.3: Carbon-12 and carbon-14 are both isotopes of carbon. Carbon-12 has twelve nucleons in its nucleus while carbon-14 has fourteen. If carbon-12 has six protons, how many neutrons does carbon-14 have?*

3.4 Decay

As mentioned in section 3.2, protons require neutrons to keep the nucleus together. It follows, then, that the nucleus needs a certain amount of neutrons. Too many or too few and the nucleus is **unstable**, which means that it likely won't remain that way.[viii] At some point, it will have to change into a configuration that is more stable. In this section, we'll examine the processes by which unstable nuclei become more stable.

⁉ A nucleus is unstable if it has too many or too few neutrons.

✓ *Check Point 3.4: Based on the discussion above, would you expect a nucleus of two neutrons and no protons to be stable?*

---

number of nucleons. It is not exactly equivalent because (1) protons and neutrons have slightly different masses and (2) the binding energy (discussed in chapter 9) has a non-negligible impact on the mass.

[viii] Isotopes with unstable nuclei are called **radioisotopes**.

Is an unstable nucleus dangerous?

It isn't the nucleus that is dangerous but rather the energy associated with the emitted particles that is dangerous. It is like the danger associated with a chemical reaction. In a chemical reaction, the individual atoms or molecules aren't dangerous. Rather, it is the energy associated with the reaction that can cause problems.

When a nucleus decays, whatever is emitted is energetic. Anything that the particle hits can be damaged. For this reason, it is important to identify what kind of particle is likely to be emitted.

During nuclear decay, the nucleus changes one or more nucleons in the nucleus are converted or emitted.

The change that an unstable nucleus undergoes is called nuclear **decay**. There are several types of decay, and three of the most common types are described in this section. However, regardless of what type of decay occurs, the nucleus doesn't disappear – it just becomes a different element because one or more protons are gained or lost. In addition, charge is *conserved*, meaning that we can't simply make a positive charge disappear or appear. Keep this in mind as we discuss the various ways an unstable nucleus becomes more stable.

3.4.1 Beta-minus decay – too many neutrons

If the nucleus has too many neutrons (i.e., too few protons), you might think that it would spit out a neutron. While this would make the nucleus more stable, a neutron itself is unstable. Consequently, unless there are way, way too many[ix] neutrons, one of the neutrons will likely decay into a proton and an electron (together, a proton and an electron is neutral, like a neutron). The proton will then remain in the nucleus and the electron is emitted from the nucleus. This type of decay is called **beta-minus** decay and is illustrated in Figure 3.2 (left side).

Why is it called beta-minus?

The "beta" is used because it is the second type of decay that was discovered (beta is the second letter of the Greek alphabet). The "minus" is used to indicate that the emitted particle has a negative charge.

[ix]For example, a nitrogen nucleus has seven protons. If it has 18 neutrons, it decays by emitting a neutron. However, that isotope of nitrogen is very unlikely.

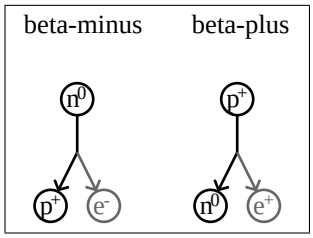


Figure 3.2: An illustration of two types of decay discussed in the text. Particles outside the nucleus are drawn in gray. Notice that in every case the total charge is the same before and after the decay.

There are two important things to notice about beta decay. The first is that the emitted electron does *not* come from the electron cloud. Rather, it comes from the decay of the neutron in the nucleus, as the neutron decays into a proton and an electron.

The second thing to notice is that this process not only lowers the number of neutrons in the nucleus but also raises the number of protons in the nucleus and, since the proton number is different, the result is a nucleus of a *different* element.

An example of beta-minus decay is what happens with carbon-14 (which means it has 14 nucleons in its nucleus). Since all forms of carbon have six protons, carbon-14 must have eight neutrons. Eight neutrons happens to be too many neutrons[x] for that nucleus so it decays via beta-minus decay, leading to a loss of a neutron and the gain of a proton. The result is the same number of nucleons (14) as before but with one less neutron (seven) and one additional proton (seven). Although the total number of nucleons is the same, the number of protons has changed so the nucleus is no longer carbon-14. Instead it is nitrogen-14.

✍ Practically all naturally-occurring isotopes of carbon are stable except for carbon-14. Consequently, a tiny amount (about one in every trillion) of carbon-14 ends up in the air we breath and food we eat. This is not considered dangerous, however, since the fraction is so small and the decay rate is so slow (see section 3.5).

Another nucleus that undergoes beta-minus decay is iodine-131. All isotopes of iodine have the same number of protons (53) and so they behave the

[x]The most common isotope of carbon is carbon-12, which has six neutrons.

same chemically (since chemical properties are due to electric properties). Consequently, iodine-131 is used in the body the same as any other isotope of iodine. By detecting the emission during the decay of iodine, we can figure out where the iodine is congregating in the body. In this way, iodine-131 can be used as a **radioactive tracer**.

✓ *Check Point 3.5: Carbon-14 undergoes beta minus decay. Does the atom remain as carbon? Why or why not?*

3.4.2 Beta-plus decay – too many protons

If the nucleus has too few neutrons (i.e., too many protons), you might think that it would spit out a proton. While this would make the nucleus more stable, it is unlikely unless there are way too many[xi] protons. It is more likely that one of the protons will decay into a neutron and a **positron**. This type of decay is called **beta-plus** decay and is illustrated in Figure 3.2 (right side).[xii] Notice that the neutron remains in the nucleus and the positron is emitted from the nucleus.

What is a positron?

A positron is a particle the size of an electron and the charge of a proton. This is why the positron is indicated by the e^+ in Figure 3.2.

You may not have learned about positrons because positrons don't last very long – they quickly combine with an electron. When the positron and electron combine, a type of radiation is produced that can be detected by a scanner. This is essentially how a PET (positron emission tomography) scan works.[xiii]

Why is this called beta-plus decay?

[xi] For example, a nitrogen nucleus has seven protons. If it has only three or four neutrons, it decays by emitting a proton. However, those isotopes of nitrogen are very unlikely.

[xii] The reverse of beta-minus is also possible, where an electron from the electron cloud surrounding the nucleus combines with a proton in the nucleus to form a neutron. This process, called **electron capture**, is less likely than beta-plus, however.

[xiii] Typically, with PET scans a short-lived isotope of carbon, nitrogen or oxygen is included into sugars that are then ingested. The PET scan can then identify where the sugars are congregating in the body.

Whereas beta-*minus* decay results in an *electron* being emitted, beta-*plus* decay results in a *positron* being emitted. This decay is called beta-plus because the emitted particles are similar to the electrons emitted during beta-minus except for their charge – negative for beta-minus decay and positive for beta-plus decay.

Note that in both cases the result is a nucleus of a different element since the number of protons in the nucleus has changed. However, with beta-plus decay, the number of protons is lowered and the number of neutrons is raised (as opposed to the reverse in beta-minus decay).

✓ *Check Point 3.6: Nitrogen has seven protons. A stable isotope of nitrogen has seven neutrons. An unstable isotope of nitrogen has five neutrons. Should that unstable isotope decay via beta-minus decay or via beta-plus decay? Why?*

3.4.3 Disintegration-type decay

Disintegration refers to when the nucleus spits out a nucleon or group of nucleons, without any protons changing to neutrons or neutrons changing to protons. Disintegration occurs when there are so many extra neutrons or extra protons that spitting out one or more nucleons is the fastest way to a more stable nucleus.

When discussing beta-minus and beta-plus decay, for example, it was mentioned that a nucleus with too many neutrons could become more stable by spitting out a neutron, and a nucleus with too many protons could become more stable by spitting out a proton. Such nuclei are rather rare but it can happen. For example, Be-13 (4 protons and 9 neutrons) and He-5 (2 protons and 3 neutrons) decay by spitting out a neutron while N-11 (7 protons and 4 neutrons) decays by spitting out a proton. It is even possible to spit out more than one nucleon. For example, H-5 (1 proton and 4 neutrons) decays by spitting out two neutrons while Fe-45 (26 protons and 19 neutrons) decays by spitting out two protons.

✍ It is common to use nuclear disintegration as a synonym for nuclear decay. I am using it here to refer to a particular type of nuclear decay, where individual nucleons don't change.

A more common instance of nuclear disintegration is with very heavy nuclei, with lots of protons and many more neutrons (like uranium). Such nuclei have so many protons that you'd think ejecting a proton would be the fastest way to make it more stable. However, as mentioned before, neutrons are inherently unstable, so the nucleus likely also has too many neutrons. Indeed, that is why very heavy elements like uranium have no stable isotope – they simply have so many neutrons (needed to keep the numerous protons in their nucleus from flying apart) that some will inevitably be very loosely connected to the protons and thus act almost like single neutrons (which are unstable). So spitting out a proton won't help address the overload of already unstable neutrons.

To help address both problems (too many protons and too many neutrons), heavy nuclei tend to undergo a process called **alpha decay**, where the nucleus spits out a group of four nucleons together: two neutrons and two protons.[xiv] Note that this group of four nucleons is the same as the nucleus of a helium atom.

Why is it called alpha decay?

The "alpha" is used because it is the first type of decay that was discovered (alpha is the first letter of the Greek alphabet).

With alpha decay, as with beta-minus and beta-plus decay, the element changes because the number of protons in the nucleus has changed (it has decreased by two).[xv] However, unlike beta-minus and beta-plus decay, none of the individual particles change. The protons and neutrons that are emitted from the nucleus remain as they were, and the protons and neutrons that remain in the nucleus remain as they were.

While alpha decay lowers the number of neutrons as well as the number of protons by two each, the *impact* on the number of protons is more significant, since there are less protons than neutrons in the nuclei of heavy elements (see section 3.2 and Figure 3.1).

For example, U-235 consists of 92 protons and 143 neutrons. A decrease of two is a greater *relative* portion of 92 than of 143. In a similar way, losing

[xiv] Another type of decay with heavy nuclei is **spontaneous fission**, where the nucleus breaks into two or more pieces.

[xv] Ejection of a proton also results in an element change. Ejection of a neutron does not.

five pounds may be cause of celebration for someone on a diet but taking five pounds of rock from the moon had a negligible impact on the moon's weight.

✍ Since heavy nuclei have more neutrons than protons, alpha decay won't help if there are too few protons. In those cases, the nucleus will either eject a neutron or under beta-minus decay.

✓ *Check Point 3.7: Suppose a nucleus of radium-226 undergoes alpha decay. Does the atom remain as radium? Why or why not?*

3.5 Half-life

⁝ The half-life is the length of time needed for half the nuclei to decay.

As mentioned above, if the number of protons within the nucleus changes, as with decay, then the element changes. How *quickly* a nucleus decays depends on how stable it is. Some nuclei decay very slowly, while others decay very quickly. The length of time needed for *half* the nuclei to decay is known as the **half-life**.

For example, carbon-14, which consists of 6 protons (characteristic of all carbon nuclei) and 8 neutrons, has a half-life of 5730 years. Uranium-235, which consists of 92 protons and 143 neutrons, has a half-life of about 700 million years. On the other hand, radium-202, which consists of 88 protons and 114 neutrons has a half life of only 0.7 milliseconds.

What is important to note is that half life indicates the time it takes for half of the nuclei to decay. It *doesn't matter* when you start the timer. Whenever you start the timer, half of the nuclei that were present when you started the timer won't be there after the half life.

This is a really important point that is deceptively easy to miss. For example, if it takes 1000 years for half of the nuclei to decay, that means half *of the remaining* (or an additional quarter of the original) will decay in another 1000 years.

In other words, the half life is really the "half of the remainder" life. It is the time it takes for half of *whatever you have at the moment* to decay. If I give you an unstable nucleus, it doesn't matter how long it has existed prior to

me giving it to you. One half-life after I give it to you, there is a 50% change it will have decayed (into something else).

✍ Remember that when a nucleus decays it stays as a nucleus, just of another element. For example, when a carbon-14 nucleus decays, it becomes a nitrogen-14 nucleus. The nucleus doesn't disappear.

✓ *Check Point 3.8: Suppose a particular sample has a half-life of 1 hour. After how much time has 75% of the sample decayed?*

3.6 Radiation

As mentioned on page 58, the "danger" with radiation is that when a nucleus decays the emitted particles can interact with other matter by destroying or modifying chemical bonds. If those bonds are within a cell, the cell can be destroyed or modified.

If the cell is a cancer cell then this is a good thing, which is why radiation is used for cancer treatment. This is also why food irradiation is used to get rid of pathogens in the food (as opposed to using heat).

On the other hand, if the cell is healthy, then destroying or modifying it is a bad thing.

WHICH KIND OF RADIATION IS MOST HARMFUL TO CELLS?

It depends. There are essentially four things to consider: the particle that is emitted, the energy of the particle, the rate at which particles are emitted and the location.

The heavier the particle, the more concentrated the damage. Consider, for example, the difference between an 11-pound bowling ball and a Ping-Pong ball. It turns out that their relative masses are the same as the difference between a proton and an electron. Since an alpha particle contains four nucleons (two protons and two neutrons), the difference between an alpha particle and an electron would be like the difference between four 11-pound bowling balls and a Ping-Pong ball.

A large particle is also unlikely to get very far into a material before interacting with it. For example, a sheet of paper (or clothing) is sufficient to

stop an alpha particle whereas a few millimeters of aluminum (or your skin) is needed to stop beta particles. There is a third type of radiation, called **gamma rays**, for which a thick slab of lead is needed.

The location of the source also makes a difference. Because alpha particles can be stopped relatively easily, there is little impact on us if the source of the alpha radiation is outside our body.

On the other hand, if ingested, inhaled or implanted, alpha decay can do more damage. For example, radon-222, which has a half-life of 3.8 days, is a gas. As a gas, it can be inhaled and, once inhaled, any decay will produce the alpha particles (helium nuclei). Those particles will be easily stopped by your insides but, in doing so, your insides can be damaged. That is the problem with radon.

What if we can get the radon out before it decays?

There is nothing dangerous about the radon itself, only the product of the decay. So, if it doesn't decay inside your body, there is no damage done. However, the half-life of Radon-222 is 3.8 days, which means half the radon decays in 3.8 days.

✍ In comparison, the half life of carbon-14 is 5730 years. Combined with its low concentration (1 part per trillion), carbon-14 isn't considered to be a hazard.

✓ *Check Point 3.9: Americium-241 is used in smoke alarms. As it undergoes alpha decay, the emitted helium nuclei ionize the oxygen and nitrogen in the air, which allow for current to flow. When smoke is present, the smoke particles neutralize the ions, preventing the flow of current and causing the alarm to sound. Why is the alpha radiation not dangerous to us?*

Summary

This chapter examined the nuclear force.

The main points of this chapter are as follows:

- Like the gravitational and electric forces, the nuclear force is a non-contact force.

- The nuclear force is so dependent on the separation distance that unless the particles are very, very close the nuclear force is insignificant.
- Each element is defined according to the number of protons in its nucleus.
- A nucleus is unstable if it has too many or too few neutrons.
- During nuclear decay, one or more nucleons in the nucleus are converted or emitted.
- The half-life is the length of time needed for half the nuclei to decay.

By now you should be able to explain why the nucleons in the nucleus stay together despite the electric repulsion between all of the protons, and be able to describe what happens if a nucleus has too many neutrons or too few.

Frequently asked questions

In an atom, what keeps the electrons from falling into the nucleus?

The reason the electrons don't fall into the nucleus, despite being attracted to the protons in the nucleus, is because of the kinetic energy of the electrons. It is similar to why the moon doesn't fall to Earth even though there is a gravitational force pulling the moon toward Earth.

Does the nucleus change during a chemical reaction?

No. Chemical reactions do not change the elements involved, only how they are bonded with other elements.

Why is a neutron unstable by itself, but stable when with a proton?

This question (along with why some isotopes are stable and others are not) remains unaddressed by our model but it has to do with the same reason why a ball rolls downhill rather than uphill. Certain configurations are associated with being "downhill".

If neutrons are neutral, why do they attract?

They attract due to the nuclear force. Neutrons are *electrically* neutral so they don't attract due to the electric force.

IF RADON DECAYS WHILE INSIDE YOU, WILL THAT MAKE YOU RADIOACTIVE?

It makes you radioactive only in the sense that something inside you is decaying. When something is **radioactive**, that means that it will decay and produce radiation (i.e., alpha, beta or gamma rays).

WHAT IF YOU ARE "HIT" BY THE EMITTED PARTICLES? DOES THAT MAKE YOU RADIOACTIVE?

Not typically. The damage the radiation makes to your body is not so great that the nuclei change and become unstable. Rather, it is the chemical structure that is changed, like in ionization, where an electron is removed or a chemical bond is broken.

IS ALPHA DECAY OR BETA DECAY AS DANGEROUS AS RADIATION?

The particles emitted by alpha decay and beta decay are radiation – they are two types of radiation (there are other types as well). Indeed, before scientists realized they were helium nuclei and electrons (or positrons), they were referred to as alpha and beta rays (there is also something called **gamma rays**, the third type that were discovered, that are sometimes emitted also during decay).

Terminology introduced

Alpha decay	Decay rate	Isotopes	Radioactive tracer
Atomic number	Electron capture	Molecular weight	Radioisotopes
Atomic weight	Element	Nuclear force	Stable
Beta-minus	Gamma radiation	Nucleons	Unstable
Beta-plus	Half-life	Positron	
Decay	Ions	Radioactive	

Additional problems

Problem 3.1: The most stable configuration of a nitrogen nucleus has seven protons and seven neutrons.
(a) If the number of protons in the nucleus changes, is it still considered a

nitrogen atom?
(b) If the number of neutrons in the nucleus changes, is it still considered a nitrogen atom?
(c) If the number of electrons surrounding the nucleus changes, is it still considered a nitrogen atom?

Problem 3.2: The nucleus of tin has 50 protons. According to the periodic table, the atomic weight of tin is 118.7. Based on this, what do you think is the number of neutrons in the most common isotope of tin?

4. Magnets and the Magnetic Force

Puzzle #4: Why do magnets attract?

Introduction

Almost everyone has experienced **magnets**. There are refrigerator magnets, for example, and some screwdrivers have a magnet at their tip so that they can attract screws. Since magnets *attract* some types of metal, it seems that the magnetic force is attractive like the gravitational force. However, magnets can both attract or repel other magnets. That property is like the electric force but it turns out that the magnetic force is neither gravitational nor electric. In this chapter, we'll examine a model that explains why magnets can attract or repel other magnets but only attract some types of metal.

4.1 Not the electric force

In order to explain how magnets attract and repel other magnets but only attract pieces of metal, we need to come up with a fourth force, one that is neither gravitational, electric nor nuclear.

Why can't we explain it with the electric force?

The magnetic force is not due to excess charge on the magnet.

The reason we can't explain it with the electric force is because magnets are electrically *neutral*, both overall and on each end.

This can be demonstrated by bringing a magnet near small pieces of paper or a rubbed balloon (see section 2.1). When we do so, we find that the magnets do not attract the rubbed balloon or small pieces of paper. In addition, it doesn't matter which end of the magnet we use.

This means that the magnet must be neutral and so the magnetic force is not due to excess charge on the magnets. This is why we say that the magnetic

force is another fundamental force, along with the gravitational force, the electric force and the nuclear force.

✍ Because magnets can both attract or repel one another, like electric charges, it is easy to mistakenly think that perhaps magnets have some excess charge on them. Be careful not to make this mistake!

✓ *Check Point 4.1: Suppose two magnets are found to attract. Does that mean they have opposite charge?*

4.2 Magnetic poles

Whether two magnets attract or repel depends on how the magnets are oriented. Every magnet has two sides, with opposite properties on each side.

For reasons we'll examine in section 4.6, the two sides are referred to as **north** and **south**. Furthermore, each side of a magnet is called a **magnetic pole** (or **pole**, for short).

⁉ The opposite sides of a magnet are called the north and south poles.

Here are the important properties of magnets:

1. When arranged with like poles facing, the magnets repel. When arranged with opposite poles facing, the magnets attract.
2. Unlike electric charge, you can't separate the north pole from the south pole. It is like a piece of paper – it has a top and a bottom – but you can't remove the top from the bottom as they are just two sides of the same thing.
3. All magnets have two sides, no matter how small the magnet. It is like a pad of sticky notes – there is always a top and bottom, regardless of how many sticky notes you take off the pad. If you split the pad in half, each set of sticky notes has a top and bottom.
4. The magnet is electrically neutral, Granted, magnets, like everything else, are made up of positive and negative particles, but they contain *equal* numbers of each. In addition, not only is a magnet, as a whole, electrically neutral, but so is each pole of the magnet.

Notice that the attraction and repulsion of magnets can't be due to electric charge, as magnets are electrically neutral. Since magnetic poles have no net

electric charge, it is important that you don't use the terms "positive" and "negative" to refer to the magnetic poles, lest you mistakenly interpret the force as electrical. The force isn't electrical.[i]

Also notice that every magnet has a north side and a south side, and you cannot separate the north pole from the south pole without each piece having its own north and south pole. Just as you can't have a front without a back, you can't have a north pole without a south pole.

This means there are not two types of magnet particles. That is unlike electric charge, which has two distinctly separate particles, a positive proton and a negative electron. As will be discussed in section 4.5, magnetic particles exist but each magnetic particle has *both* poles present.[ii]

? A magnet cannot have only a single pole.

✓ *Check Point 4.2: Suppose you bring a north pole of one magnet next to the south pole of another magnet. What will the two poles do: (a) attract or (b) repel?*

4.3 Magnetic poles always paired

As mentioned in the previous section, every magnet has both a south side and a north side. Unlike electric charges, where we can separate positive and negative charges, we cannot have a north pole without a south pole.[iii]

If we break a magnet in half, will one half only have a north pole and the other half only a south pole?

No. Each half will have *both* a north and a south pole.

Why?

Apparently, a magnet is made up of lots and lots of very tiny little magnets, each with their own north and south pole. Indeed, as we'll discuss in chapter

[i] Indeed, if it was electrical, we wouldn't need to have a separate chapter devoted to a fourth fundamental force.

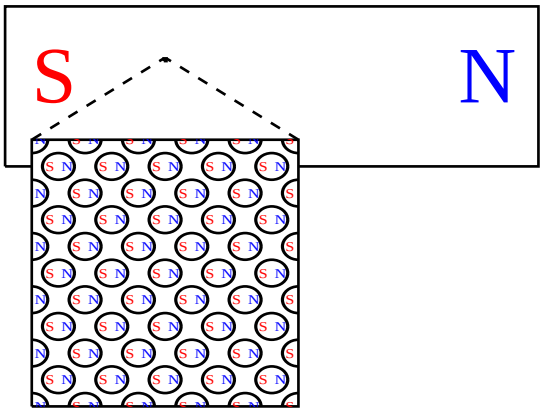
[ii] Some relationships in physics seem to imply that single magnetic poles can exist (called **monopoles**). However, attempts to find them have not been successful as far as I know.

[iii] I will tend to draw magnets as rectangles with the poles on opposite ends, but magnets come in different shapes and configurations.

12, each electron acts like a little magnet, with its own north and south pole (in addition to be negatively charged).[iv] Consequently, each part of a broken magnet is itself a magnet, consisting of lots and lots of little magnets.

⁝ A magnet can be thought of consisting of lots of tiny little magnets.

This is illustrated in the figure below, where a tiny piece of a magnet is magnified such that we can see the tiny little magnets that make it up.



The overall magnetic properties of a magnet, then, are due to the tiny little magnets inside being all aligned, with the north pole of the magnet being on the side that the tiny north poles face.

✓ *Check Point 4.3: Suppose you break a piece off of a magnet. If the piece came off the north pole end of the magnet, is it possible for the piece to only have a north pole (assuming the piece is small enough)?*

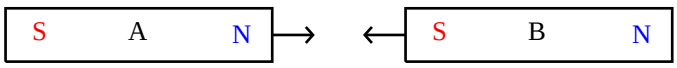
4.4 Magnetic force and torque

Just as objects can be electrically polarized, with one side negative and one side positive, magnets are magnetically polarized, with one side north and one side south. In fact, all magnets are magnetic *dipoles*. And, just like

[iv] Protons and neutrons also have magnetic properties, with their own north and south poles, but the magnetic properties are almost a thousand times weaker in protons and neutrons.

electric dipoles, two magnets will attract if they are aligned with opposite poles adjacent to each other.

For example, suppose we have two magnets, A and B, oriented as shown in the illustration below. The magnets face each other with opposite poles, so the magnets attract.



If two magnets are not aligned in a configuration where they attract then, just like with electric dipoles, there will be a torque acting to re-orient them into a position where they do attract (see section 2.2). As mentioned on page 38, they will either end up aligned or opposite, with the end result depending somewhat on the shape of the magnets.

Unlike electric dipoles, however, the torque on magnets isn't simply due to two forces acting on the magnet, one on each end. After all, the poles can't be separated. Rather, the torque acts on the magnet as a whole.

As an analogy, one can think of a person's "face" and "back" as analogous to the north and south poles of a magnet. If you have a bunch of magnets all aligned and then introduce a new magnet, that new magnet will likely orient itself to be aligned with the others. In a similar way, if all the people in a room (or elevator) are facing one way, then newcomers to the room will feel pressure to face the same way. It isn't that our "face" is forced one way and our "back" is forced the other way.

You can use the same analogy to see why you can't break a magnet into separate north and south pieces (see section 4.3). Again, consider a room full of people. If they all face one way then putting a divider down the middle doesn't split the faces from the backs. It just doesn't make sense.

✍ On the other hand, positive particles and negative particles are separate entities, like adults and children, and so we *can* separate them. For example, if the adults are on one side of a room and children are on the other side, we could put a divider down the middle of the room to create two separate rooms, one with adults and one with children.

✓ *Check Point 4.4: In the figure above, magnet B is oriented with south on the left and north on the right. Would the two magnets still attract if magnet B was oriented the opposite way, with north on the left and south on the right?*

4.5 Ferromagnetic materials

WHY IS THERE ONLY AN ATTRACTION BETWEEN MAGNETS AND PIECES OF METAL, NEVER A REPULSION?

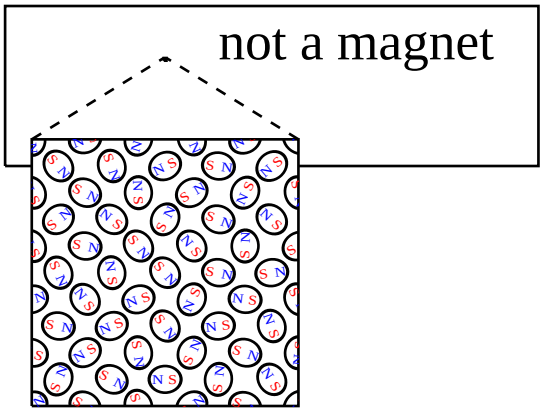
To answer this, we should first recognize that not all pieces of metal are attracted to magnets. Those that are attracted are referred to as being **ferromagnetic** and examples include iron, nickel, cobalt and neodymium.

We must also recognize that these ferromagnetic materials are not, themselves, attracted to each other. In other words, a piece of iron is not attracted to or repelled by another piece of iron.

To explain this, we extend our magnetic model from above.

Just as a magnet is made up of lots of tiny magnets, so too are other materials. The tiny magnets are electrons, so *all* objects are made up of lots of tiny magnets.

The difference is that in the non-magnet the tiny magnets are not aligned. An example of such non-alignment is shown below.



As illustrated in the figure, the tiny magnets are oriented in random directions. Consequently, some face one way and some face the opposite way. In this way, the tiny magnets cancel each other out and so the non-magnet does not exhibit any magnetic properties.

Only in magnets are the tiny magnets inside the magnet all aligned. In fact, magnets can eventually lose their magnetism if the tiny magnets inside somehow become non-aligned.[v]

What about ferromagnetic materials?

Ferromagnetic materials like iron are non-magnets, in that the tiny magnets inside the material are not aligned and thus they do not exhibit magnetic properties.

However, in ferromagnetic materials, the tiny magnets are free to align *if* a magnet is brought nearby. The magnet, then, acts to align the tiny magnets. Once aligned, the tiny magnets act together to make the object magnetic.

At that point, the ferromagnetic material can act like a magnet, attracting other magnets. In this way, it is attracted to the original magnet that was responsible for magnetizing it.

✍ It is attracted to that original magnet, not repelled, because the original magnet naturally caused the tiny magnets to align in a direction that led to an attraction.

What happens when the original magnet is moved away?

Once the original magnet is moved away, there is no longer anything acting to align the tiny magnets inside the ferromagnetic material. The tiny magnets inside the ferromagnetic material can become misaligned and, if so, the material loses its magnetic properties.

✍ We can distinguish between hard and soft ferromagnetic materials. Hard ferromagnetic materials retain their magnetic properties even when not around another magnet while soft ferromagnetic materials lose their magnetic properties when not around another magnet.

What about non-ferromagnetic materials?

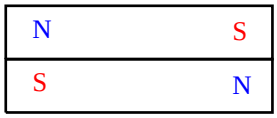
Not all metals are ferromagnetic. For example, aluminum and copper, which are used for the wires in a circuit, are not ferromagnetic.

[v] One way to do this is to heat the object.

In non-ferromagnetic materials, the tiny magnets inside the material don't align even if a magnet is brought nearby.

WHY NOT?

Just as two electric dipoles can end up in an opposite configuration (see page 38), so too can two magnets, ending up side-by-side and opposite in direction, as shown below. Since each end of the combination has both north and south poles, the paired magnets would not magnetically attract or repel a third magnet. In a sense, the paired magnets act like a non-magnetic object, even though each magnet that makes up the pair is still magnetic.



For this reason, we can consider a non-ferromagnetic material as having all of the tiny magnets paired up, and this is why bringing a permanent magnet nearby has no effect on a non-ferromagnetic material.

✍ Materials that are very weakly affected by magnets are **diamagnetic** while materials that are affected more strongly but not as strongly as ferromagnetic materials are called **paramagnetic**. To simplify things, for our purposes we'll treat materials as either being ferromagnetic or non-ferromagnetic.

✓ *Check Point 4.5: When the north pole of a magnet is brought near a ferromagnetic material like iron, it attracts the iron to the magnet. What happens if the south pole of that magnet is brought near the ferromagnetic material? Assume the ferromagnet is soft, so that its magnetic properties are temporary.*

4.6 Earth as a magnet

WHY DO WE REFER TO THE POLES AS "NORTH" AND "SOUTH"?

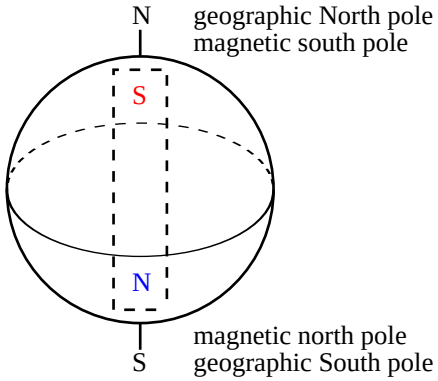
The reason for using the terms "north" and "south" has to do with the fact that Earth acts like a really big magnet, and if you place a small magnet on

a platform that is free to rotate, the magnet will orient itself with its north end pointed toward the geographic North pole.

In fact, a compass is just a very light magnet that is free to rotate upon a pivot. That way, the north pole of the compass (usually painted some color) ends up pointing toward Earth's geographic North pole.

SINCE THE NORTH POLE OF THE COMPASS MAGNET POINTS TOWARD EARTH'S GEOGRAPHIC NORTH POLE, DOES THAT MEAN THE SOUTH POLE OF THE "EARTH MAGNET" IS AT THE GEOGRAPHIC NORTH POLE?

Yes. This is illustrated below, where Earth acts as though there is a gigantic magnet inside it.



The Earth acts like a big magnet, with its magnetic south pole near the geographic North pole.

Since opposites attract, somewhere near Earth's geographic North pole there must be a magnetic south pole that is attracting the magnetic north pole of the compass.[vi] Admittedly, this difference between magnetic and geographic poles can be a little confusing.

DOES THIS MEAN THAT THERE IS A FORCE ON THE COMPASS PULLING IT NORTHWARD?

No. Earth is so big that there is little difference in how it influences one side of a small compass vs. the other side. Consequently, we can assume the force

[vi] The positions of Earth's magnetic poles wander slightly from year to year. While the magnetic pole is relatively close to Earth's geographic poles, the magnetic north pole is currently further from Earth's geographic South pole than the magnetic south pole is from Earth's geographic North pole.

on each side is the same, just opposite in direction, such that the net force is zero, but there is still a torque present to make it rotate if not aligned.[vii]

✓ *Check Point 4.6: Why is there no net force on a compass needle (due to Earth) but there is a net force on a magnet that is brought near another magnet?*

How come I don't notice a torque on a magnet if I hold it in the east-west direction rather than the north-south direction?

⁉ Compass needles align north-south because of the torque exerted on it, due to the opposite forces on each end.

The problem is that the torque is very weak. To see the torque, you need to place the bar magnet on something that is free to rotate without friction. Such a setup is called a **compass**.

✓ *Check Point 4.7: Why is it said that Earth's geographic North pole acts like the south pole of a giant magnet?*

Summary

This chapter examined the properties of magnets.

The main points of this chapter are as follows:

- The magnetic force is not due to excess charge on the magnet.
- The opposite sides of a magnet are called the north and south poles.
- The Earth acts like a big magnet, with its magnetic south pole near its geographic North pole.
- A magnet cannot have only a single pole.
- A magnet can be thought of consisting of lots of tiny little magnets.
- Compass needles align north-south because of the torque exerted on it, due to the opposite forces on each end.
- A ferromagnetic material can be thought of consisting of lots of tiny little magnets that will align when a magnet is brought nearby.

[vii] In a similar way, we are so far away from the center of Earth that we can take g to be 9.8 N/kg at both ends of a needle.

- A non-ferromagnetic material can be thought of having lots of tiny little magnets that are all paired up and thus a magnet will have no effect on it.

By now you should be able to describe how there is a torque exerted on magnets in terms of the attraction and repulsion of the magnetic poles.

Frequently asked questions

Is the attraction between magnets due to an excess of positive or negative particles?

No. A magnet has an equal amount of positive and negative particles. This can be shown by bringing a magnet near some pieces of paper. Like the unrubbed balloon, the magnet has no effect on the pieces of paper.

If the north pole of the compass needle is forced toward the North, the south pole should be attracted to the South. Is the net force on the compass needle zero?

Yes. While the compass needle's north pole is attracted to Earth's geographic North pole, the compass needle's south pole is attracted to Earth's geographic South pole. The net force on the compass needle is thus zero.

If the net force is zero, why does it point north?

Although the net force is zero, there can still be a net torque acting to rotate the compass needle. The torque will rotate the needle until the compass needle's magnetic north pole is pointing toward Earth's geographic North pole.

Why does a magnet pick up pieces of metal that aren't, themselves, magnets?

When a magnet is brought close to the piece of metal, though, a torque is exerted on the tiny magnets in the metal, aligning them. That makes the metal "magnetized", and the metal can be attracted to a magnet (or another piece of metal).

The effect is temporary, however, and the metal will lose its magnetization when the magnet is removed.

If every magnet is made of tiny magnetic electrons, what makes those electrons magnetic themselves?

This is explored further in chapter 12.

Terminology introduced

Compass	Magnetic poles	North	South
Diamagnetic	Magnets	Paramagnetic	Torque
Ferromagnetic	Monopoles	Poles	

Additional problems

Problem 4.1: Suppose a proton is placed near the north pole of a magnet. Will it be attracted to the magnet, repelled by the magnet, or neither?

Part B

Fields and Energy

5. Describing Fields

Puzzle #5: It was mentioned in chapter 1 that Earth's gravitational field strength is 9.8 N/kg near Earth's surface. What is a gravitational field?

Introduction

The puzzle raises the question of what we mean by the "field". While we have discussed the four fundamental forces (gravitational, electric, magnetic and nuclear), referring to "forces" is just one way to discuss the way objects interact. There are other ways, including fields and energy, which we will look at in this part of the book.

5.1 The gravitational field

Most people have some experience with the concept of "field," particularly in some science fiction films, where a spaceship might have some sort of "field" that protects it from harm, destroying any incoming weapons as they interact with the field prior to reaching the spaceship.

In a similar way to that of the spaceship's field, we can think of a planet as having a field around it that affects other objects that are nearby. We call that field the *gravitational* field because the planet interacts gravitationally with other objects that are nearby. It turns out that a charged object has an *electric* field around it that affects other charged objects that are nearby, and a magnet has a *magnetic* field around it that affects other magnets that are nearby. We'll start with the gravitational field simply because you are likely more familiar with the gravitational field, but the ideas are the same for all types of fields.

Every object has its own **gravitational field**. Indeed, even you and I have gravitational fields. However, our gravitational fields are very weak, so I'm going to focus on the gravitational fields of planets and stars.

We can think of an object like Earth as having a gravitational field around it that affects other objects (like us) that are nearby. For example, rather than saying that we gravitationally interact with *Earth* (which then pulls us downward), we can instead say that we interact with Earth's gravitational *field* (which then pulls us downward).

⁉ An object interacts with the *field* of another object not its own field.

If it means the same thing, why describe an interaction in terms of fields (instead of forces)?

Describing interactions with fields has several advantages.

First, we can describe an object's field without needing to know what that object is interacting with. For example, we can say that Earth's gravitational field strength is 9.8 N/kg near its surface. It doesn't matter what is actually interacting with Earth. It could be you or a rock. In comparison, we can't determine the force on an object unless we are given information about *both* objects that are interacting (like you and Earth, or the rock and Earth).

Basically, an object's field is something that is inherent to the object (unlike the force, which depends on both objects that are interacting). In fact, an object doesn't even need to be interacting with another object to have a gravitational field.

A second advantage of using field is that, because each object has its own field, we can use the fields of two different objects to compare how they'd influence a *third* object, without needing to know anything about that third object. For example, Earth's gravitational field strength is 9.8 N/kg near its surface and the moon's gravitational field strength is 1.6 N/kg near its surface. Since Earth's gravitational field strength is about six times greater than the moon's gravitational field strength, we can say that the gravitational force on an astronaut would be six times greater if they were on Earth than on the moon, regardless of the astronaut's mass.

We not only don't need to know the astronaut's mass but we don't even need to know any details about Earth or the moon (assuming we know the strength of the gravitational fields). Indeed, the third reason to use fields, and perhaps the strongest reason, is that we can use the language of fields to qualitatively predict the force on an object without needing to know the details about either of the objects that are interacting.

For example, if we are at a location where the gravitational field (not our own) is downward, we can predict that we'll be forced downward. It doesn't matter if we are on Earth or on the moon or anywhere else for that matter. Indeed, often we are able to directly measure the gravitational field at a particular location and, in those case, we don't need to know the details about what is *responsible* for the field we are measuring.

✍ Changes in the gravitational field travel at the speed of light. Consequently, if the Sun suddenly disappeared, the Earth would still be interacting with what was the Sun's gravitational field (and continue to be pulled toward where the Sun had been) for about eight minutes (the time it takes for the information to travel from the Sun to Earth).[i]

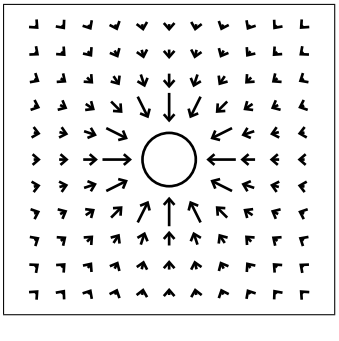
ARE THERE ANY DISADVANTAGES TO USING FIELDS?

The disadvantage has to do with language. It is important to distinguish between force and field, as they are not the same thing. Each object has its own field, and that field depends on the properties of the object, but that field doesn't tell us anything about what will happen to that object. It only tells us what will happen to *another* object. In other words, objects interact with the fields of *other* objects, not their own.

⁝ An object's field can be used to determine what will happen to another object, not itself.

✓ *Check Point 5.1: Does Earth's gravitational field push it toward the Sun? Why or why not?*

Like the gravitational force, the gravitational field has both a magnitude and a direction and thus can be represented by an arrow. For example, consider the illustration to the right, where the circle represents a planet and the arrows represent the planet's gravitational field and thus what would happen to *another* object if it were placed at that particular location.



⁝ Fields are represented by arrows called field vectors that indicate the direction an object will be forced when placed at a particular location.

Each arrow, called a gravitational **field vector**, points toward the planet because an object released at rest will start moving toward the planet (because

[i]So, if it weren't for the language of fields, physics students everywhere would spend the entire eight minutes struggling to describe what is going on instead of preparing for the end of daylight forever.

of the gravitational force of attraction).

Keep in mind that the arrows are not supposed to mean that there is *something else* pushing on the planet. Rather, the arrows represent the effect that the planet (represented by the circle) would have on *another* object if another object happens to be placed nearby.

✍ Because of this convention, stating that an object moves "toward the planet" is equivalent to stating that the object moves "with the direction of that planet's gravitational field." Conversely, moving "away from the planet" is equivalent to moving "*against* the direction of that planet's gravitational field".

What do the lengths of the gravitational field vectors represent?

⁝ The length of a field vectors indicates the strength of the field at that location.

The length indicates the *strength* of the field at that location. As you can see, the field vectors are bigger closer to the planet than farther away. This is consistent with the fact that the gravitational force on an object (due to the planet) is less the further away the object is from the planet.

✍ The strength of an object's gravitational field also depends on its mass. The more massive the object, the stronger its gravitational field.

Is an object influenced by its own gravitational field?

No.

Consider two planets, A and B. Each planet has its own gravitational field. Planet A "feels" the presence of planet B's gravitational field and, as a result, is forced toward planet B. Likewise, planet B "feels" the presence of planet A's gravitational field and, as a result, is forced toward planet A.

So the force on planet A has nothing to do with planet A's gravitational field?

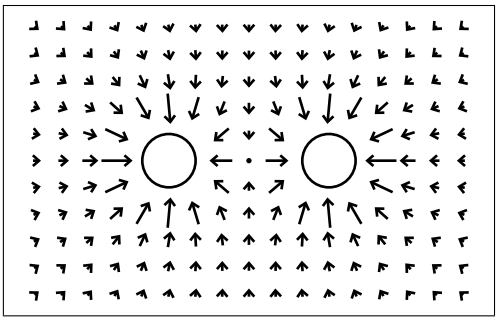
Correct. When determining the force exerted on an object due to another object, we need to identify which object is doing the "influencing" and which is feeling the effect of that "influence". If planet A is being influenced, then to find the force you need to know which planet is doing the "influencing" (planet B in this example). In other words, it is planet B's gravitational field that influences planet A.

Suppose we had two objects, A and B. Each object by itself

has its own gravitational field surrounding it. Are there two gravitational fields?

Yes. However, as mentioned above, each object is not influenced by its own gravitational field so the combination of the two gravitational fields would indicate what what would happen to a *third* object if it were placed at that location, not what would happen to the original two objects.

An example of a combined gravitational field is illustrated to the right, where the combined field at each point (represented by an arrow) is equal to the two individual gravitational fields added together.



⁉ The total field at a location corresponds to the sum of all the fields contributed by whatever objects are present.

What happens directly between the two planets?

I placed a dot, rather than an arrow, directly between the two planets because at that point the two individual gravitational fields are opposite in direction and cancel totally because I've assumed the two planets have the same mass.

✍ Planets were used in the examples above but any object with mass could be used. And the gravitational fields will cancel *somewhere* between them, although exactly where depends on the masses (i.e., it cancels right the middle only if the masses are the same).

✓ *Check Point 5.2: Two planets, A and B, are separated by* 10^{11} *m. Both planets have a mass of* 10^{24} *kg. Draw a picture, with circles for each planet.*
(a) Draw an arrow, with label "a," at the location of planet B, indicating the direction of planet A's gravitational field at that location.
(b) Draw an arrow, with label "b," at the location of planet A, indicating the direction of planet B's gravitational field at that location.
(c) Draw an arrow, with label "c," at a point right in between the planets, indicating the direction of planet A's gravitational field at that location.
(d) Draw an arrow, with label "d," at a point right in between the planets, indicating the direction of planet B's gravitational field at that location.
(e) Draw an arrow, with label "e," at a point right in between the planets, indicating the direction of the total *gravitational field at that location.*

5.2 The electric field

The electric field has many similarities with the gravitational field.

Just as objects with mass have a gravitational field, charged particles have an **electric field** that exerts an influence on other charged particles around it. An object's electric field is inherent to the object.

In addition, just as a rock is influenced by Earth's gravitational field, pulling the rock downward toward Earth, a proton is influenced by an electron's electric field, pulling the proton toward the electron.

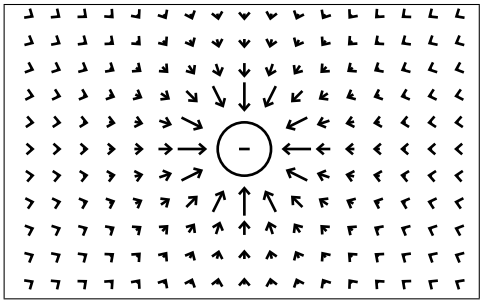
And, like the gravitational field, the electric field also has a direction. For example, as mentioned earlier, if you place an object in a location where the gravitational field (not its own) is downward, it will be forced downward. Similarly, if you place a proton in a location where the *electric* field (not its own) is downward, it will also be forced downward.

However, while the electric field has many similarities with the gravitational field, it has one major difference. Whereas objects are always forced in the direction of the gravitational field in which they are placed, objects aren't always forced in the direction of the *electric* field in which they are placed. Negative objects are forced *opposite* the direction of the *electric* field. The reason for this is because we know that charged objects, whether negative or positive, have the opposite effect on negative vs. positive objects.

Thus, if a particular electric field forces positive objects one way, that same electric field must force negative objects the opposite way. By convention, we say that the electric field forces positive objects in the direction of the electric field, and forces negative objects opposite that direction.

The direction of the electric field at a particular location is the direction that a positive charge will be forced if placed at that location.

For example, consider the drawing to the right, where the ⊖ represents a negative charge and the arrows represent the electric field vectors (as opposed to the gravitational field vectors). Notice that the electric field vectors point *toward* the negative charge.



Remember that the electric field vectors indicate what would happen to a *second* object, assuming the second object has a positive charge. The electric

field vectors therefore point toward the negative charge because it attracts positive objects.

For example, a positive object placed to the right of the ⊖ in the drawing will be pushed to the left (toward the ⊖), in the same direction as the electric field where the positive object was placed.

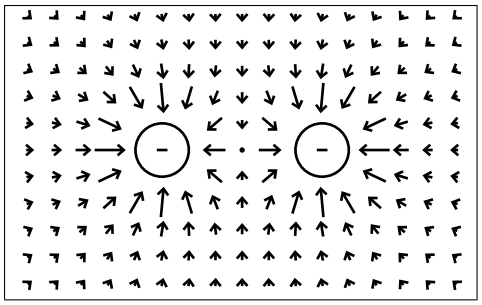
WHAT IF YOU PLACED A NEGATIVE OBJECT TO THE RIGHT OF THE ⊖?

We know that like charges repel, so the negative object will be pushed toward the right, away from the ⊖. Notice that this is *opposite* the direction of the electric field where the negative object was placed.

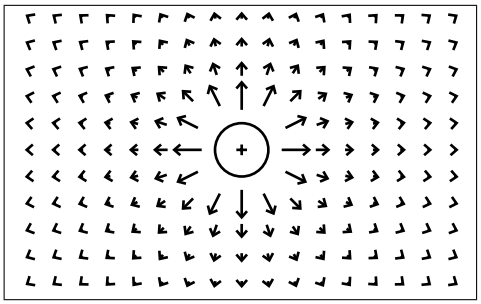
As stated before, the electric field at a particular location indicates which way a *positive* charge will be forced if placed at that location. A *negative* charge, if placed at that location, will be forced in a direction *opposite* the electric field at that location.

✓ *Check Point 5.3: Suppose we placed a positive object in a region where the electric field (due to a second object) is pointing toward the right. In which direction would that positive object be forced? What if we placed a negative object at that location instead?*

Now let's consider the case with *two* negative charges, illustrated to the right where the arrows represent the electric field of *both* charges together. Remember that the electric field indicates what would happen to a *third* (positive) object, not what happens to the two negative charges themselves.

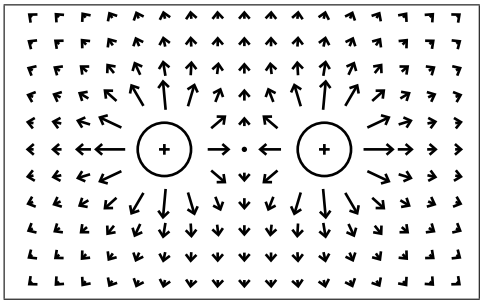


So, we know that the two ⊖ objects will repel one another and move apart (unless they are somehow fixed and unable to move). That isn't what the arrows in the drawing tell us. Notice, for example, that the electric field midway between the two charges is zero. That means that if we placed a *third* object midway between them there would be no force on that object. Basically, the positive object would be pulled equally in opposite directions, toward each ⊖, and the two forces would cancel.



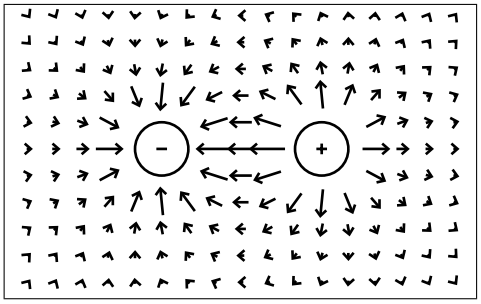
DOES A POSITIVE CHARGE ALSO HAVE AN ELECTRIC FIELD?
Yes. This is indicated in the drawing to the right, where the ⊕ indicates a positive charge. Notice that the electric field vectors point *away* from the positive charge (because it repels positive objects).

For example, a positive object placed to the right of the ⊕ in the drawing will be pushed to the right, away from the ⊕ and in the same direction as the electric field there. In comparison, a negative object placed to the right of the ⊕ in the drawing will be pushed to the left, toward the ⊕ and in a direction *opposite* the electric field there.



To further illustrate this, consider the case with two positive charges, as illustrated to the right. The arrows now represent the electric field of both charges together, and thus indicate what would happen to a *third* object that happens to be positive.

Again, it is important to keep in mind that the electric field vectors do <u>not</u> indicate what happens to the two positive charges. We know that those two will repel one another and move apart (unless they are somehow fixed and unable to move). The electric field vectors indicate what would happen to a *third* positively-charged object.



As a final example, consider the case illustrated to the right with a negative charge and a positive charge. A *third* object (positive) would be pushed away from the ⊕ and toward the ⊖, which is why the arrows point away from the the ⊕ and toward the ⊖.

A couple of things to keep in mind:

- As with the gravitational field, the electric field arrows are **not** meant to indicate the "flow" of anything (like the flow of rain or water). They only indicate the *direction* of the field at each location.
- The electric field vectors indicate the direction that a *positive* charge is forced. Negative charges are forced in the opposite direction.
- If we know the charges that are responsible for the field, we don't really need to draw the field since we already know how they will influence other charged objects. The real power of fields is when we know the field but don't know anything about the charged objects responsible for the field. In that case, we don't need to know anything about the charged objects responsible for the field. Knowing the field itself is sufficient for predicting what would happen to any objects that are placed in that field.

✓ *Check Point 5.4: Draw a picture of two particles, A (+2 μC) and B (−2 μC), 20 cm apart. Add five arrows indicating the following:*
(a) The direction of particle A's electric field at the location of particle B.
(b) The direction of particle B's electric field at the location of particle A.
(c) The direction of particle A's electric field midway between A and B.
(d) The direction of particle B's electric field midway between A and B.
(e) The direction of the total *electric field midway between A and B.*

What effect does an electric field have on neutral objects?

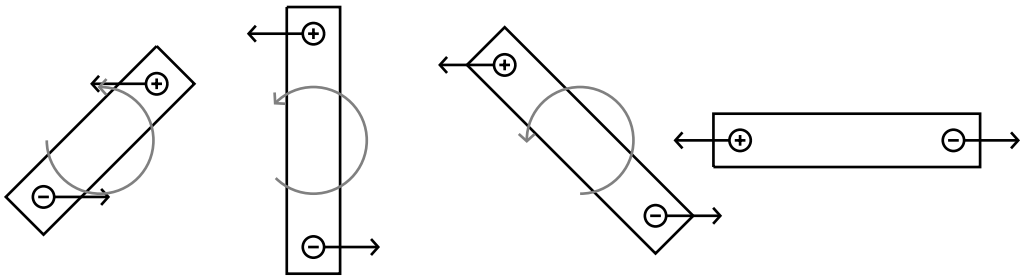
As mentioned in section 2.2, if an object is neutral but electrically polarized (an electric dipole), with one end positive and the other end negative, then a nearby charged object will make the dipole experience a torque, orienting it in such a way that it is then attracted to the charged object. Electric dipoles respond the same way to an electric field.

✍ If the neutral object is not already polarized, it can be polarized when placed in an electric field, which pushes the positive charges inside the neutral object in the direction of the field and pushes the negative charges the opposite way. If the neutral object cannot be polarized then it will be unaffected by the electric field.

To visualize this, let's first consider an electric field that is *uniform*, meaning that the electric field has the same direction and same strength everywhere.[ii]

[ii] While it may not be possible to create an electric field that is the same everywhere

The figure below illustrates the impact of a uniform electric field (same magnitude and leftward direction throughout) on an electric dipole of various orientations.



Notice how the positive side of the dipole is forced leftward (in the same direction as the electric field), regarding less of the dipole orientation. Conversely, the negative side of the dipole is forced rightward (opposite the electric field), regardless of the dipole orientation. As a result of the opposite-directed forces, the dipole experiences a torque, rotating it counter-clockwise until it is aligned with the field, with the positive side of the dipole on the left and the negative side of the dipole on the right.

Note that the net force on the dipole is zero the entire time, even when it is aligned with the electric field. In other words, the leftward force (on the positive side of the dipole) is exactly countered by the rightward force (on the negative side of the dipole). This is because the electric field is uniform.

In some cases, the electric field is close enough to uniform that we can consider it to be uniform. However, in many cases the electric field is not uniform. In a non-uniform electric field, the dipole will still experience a torque, rotating it to align with the electric field, but once aligned there will be net force pushing it toward the region where the electric field is stronger.

For example, let's consider the situation illustrated before, with the electric field directed toward the left, but this time let's suppose the electric field is stronger on the left than on the right. That means the force on the left

throughout the entire universe, it is certainly possible to create a field that is the same everywhere within a small region, and that is sufficient for our purposes. This can be accomplished by using a magnet that is large compared to the region we are considering, much like how Earth's magnetic field can be considered to be uniform for our neighborhood even though it may not be exactly the same as what it is a thousand miles away.

side of the dipole (pushing it leftward) is stronger than the force on the right side of the dipole (pushing it rightward). The result is a net force leftward, toward the region where the electric field is stronger.

Conversely, let's suppose the electric field is stronger on the right than on the left. That means the force on the right side of the dipole (pushing it rightward) is stronger than the force on the left side of the dipole (pushing it leftward). The result is a net force rightward, again toward the region where the electric field is stronger.

✓ *Check Point 5.5: In the illustration discussed above, the electric field was directed leftward. Describe what would happen to an electric dipole if the electric field was instead directed rightward.*

5.3 The magnetic field

Just as masses have an associated gravitational field and charges have an associated electric field, magnets have an associated **magnetic field**. And, just like the electric field on an electric dipole, the effect of the magnetic field on a magnet is to align the magnet with the magnetic field.

For example, if the magnetic field is directed leftward, a magnet placed in that magnetic field will experience a torque rotating it so that its north end is on the left and its south end is on the right. Furthermore, if the magnetic field is not uniform then the magnet will experience a force toward wherever the magnetic field is stronger.

Notice that the magnet is forced to align in a direction where its north end is pointed in the same direction as the magnetic field in which it has been placed. This is because our convention is to indicate the direction of the magnetic field as the direction that a magnet's north pole is forced.

Keep in mind, once again, that the magnetic field in this case is *not* the magnetic field of the magnet that is experiencing the torque. A magnet, like a charged object, is impacted by the field of *another* object, not its own field. However, just like every object has its own gravitational field and every charged object has its own electric field, every magnet has its own magnetic field. The magnetic field vectors for a magnet are drawn in Figure 5.1.

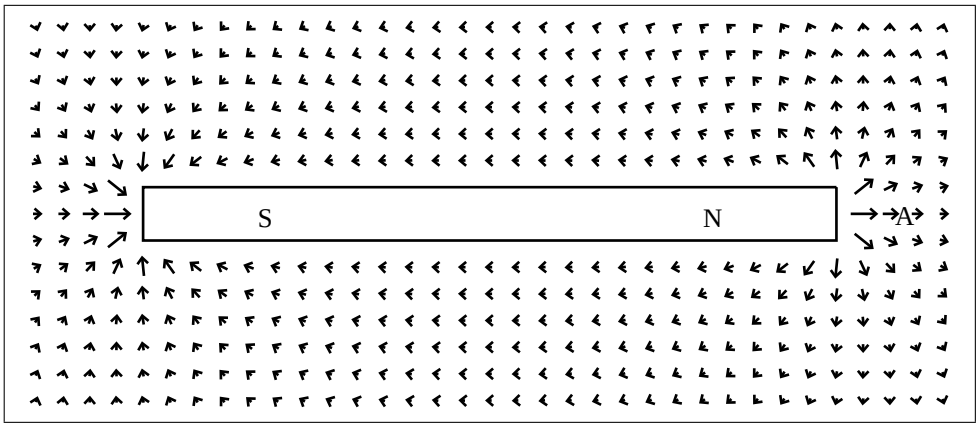


Figure 5.1: A magnet with its magnetic field indicated by arrows.

One thing you should notice is that the arrows point away from the north pole of the magnet and towards the south pole. This is because scientists define the direction of the magnetic field as the direction that a *north pole* is forced (south poles are forced in the opposite direction).

⁝ The direction of the magnetic field at a particular location is the direction that a north pole will be forced if placed at that location.

Again, keep in mind that the field vectors do not represent the *flow* of anything. Each arrow just indicates what would happen to *another* object *if* you placed it at that location. For example, if you placed a *second* magnet at the location marked as "A" in Figure 5.1, there will be a torque on that *second* object such that its north pole is forced rightward (as indicated by the arrows there) and its south pole leftward (opposite the arrows there). Also notice that the arrows are longer near the ends of the magnet than along the sides, and longer closer to the magnet than further away. This should be consistent with how the attraction and repulsion of magnets are greater near the ends than the sides, and greater closer to the magnet than further away.

✍ Remember that the magnetic field vectors represent how the magnet will influence *another* magnet. We have no way of knowing whether the force on that second magnet will be attractive or repulsive until we know something about the second magnet (the one being "influenced"). By convention, the arrows point in the direction a north pole will be forced. You simply need to remember that that is the convention.

✓ *Check Point 5.6: Suppose the south pole of a test magnet is placed at location "A" in Figure 5.1, just to the right of the magnet in the figure.*

(a) In which direction is the test magnet's south pole forced: [i] toward the magnet in the figure or [ii] away from from it?
(b) In which direction is the test magnet's south pole forced: [i] in the same direction as the magnetic field at location A due to the magnet in the figure or [ii] opposite that magnetic field?
(c) Do you need to know the direction of the test magnet's magnetic field?

Summary

This chapter examined how we use the concept of a field to indicate how one object might influence a second object when that second object is brought nearby.

The main points of this chapter are as follows:

- Fields are represented by arrows called field vectors. The direction and length of the arrow indicates the direction and strength, respectively, of the field at that location.
- The total field at a location corresponds to the sum of all the fields contributed by whatever objects are present.
- An object interacts with the *field* of another object not its own field.
- An object's field can be used to determine what will happen to another object, not itself.
  - The direction of the electric field at a particular location is the direction a positive charge will be forced if placed at that location.
  - The direction of the magnetic field at a particular location is the direction a north pole will be forced if placed at that location.

By now you should be able to predict and interpret the gravitational, electric and magentic fields associated with objects.

Frequently asked questions

Does any object have a gravitational field around it?

Yes. A planet was used in the examples but all objects with mass have an associated gravitational field.

Does an object's field still exist if there are no other objects around?

Yes. The field represents the potential, so to speak, of the object to exert a force, if another object were to be placed there.

Notice how the field "belongs" to the object that is doing the influencing, not the object being influenced.

Do the field vectors indicate something flowing through, toward or around the object?

No. The arrows do not indicate that something is flowing or even that the object itself is moving.

What do the little arrows represent?

Each little arrow indicates what the force would be on *another* object if placed at that location.

Do the electric field vectors indicate which way an electron will be forced?

The direction of the electric field vectors are the direction that a positive charge will be forced. An electron, having negative charge, will be forced in the opposite direction.

Which field do I use, the field created by the object or the field of some other object?

If you want to know the force exerted on the object, you can't use its own field. You need to know the field that is present at the location of the object due to the *other* objects that may be around.

What happens if you bring a magnet to a location where the magnetic field is directed rightward?

The north end of the magnet will be forced rightward. The south end of the magnet will be forced leftward.

If each magnet has both a north and a south pole, shouldn't the attraction of one pole cancel with the repulsion of the other pole?

If the magnetic field is the same on both sides then the force will be the same, just opposite in direction. The result is just a torque that rotates the magnet to be aligned with the magnetic field.

However, if the magnetic field is different then there will be a net force on the magnet. For example, the magnetic field of a magnet is greater closer to the magnet (see, for example, Figure 5.1; arrows are larger nearer the magnet). Consequently, when another magnet is placed nearby, the side of that other magnet closer to the first magnet experiences a greater magnetic field and thus is affected more (than the side farther away).

Terminology introduced

Dipole
Electric field
Field vector
Gravitational field
Magnetic field

6. Quantifying Fields

Puzzle #6: It was mentioned in chapter 1 that Earth's gravitational field strength is 9.8 N/kg near Earth's surface. Does the electric field have a value? What about the magnetic field?

Introduction

The puzzle raises the question of how we quantify the strength of fields and what the values mean.

As I did in the previous chapter, I'll start with the gravitational field and then move on to the electric field and then the magnetic field. Along the way, I'll show you where you might encounter the need to quantify the field.

6.1 Gravitational field

As we know from chapter 1, the strength of Earth's gravitational field is 9.8 N/kg near its surface. This is what we called the gravitational field strength g (see, for example, section 1.6).

⁉ The gravitational force has SI units of N/kg.

Notice how the units are *similar* to that for force (in N), but not the *same*. The difference has to do with the fact that a force is associated with the interaction between *two* objects whereas a field is associated with just *one* object.

To illustrate the difference, consider how we go about measuring the field. To measure Earth's gravitational field, for example, we first measure the gravitational force on an object due to Earth and then divide by the mass of that object (i.e., the mass of the object upon which the force is measured,

not Earth's mass). By dividing by the object's mass, we get a quantity that is *independent* of that object's mass and only depends on Earth's mass.

How can it be independent of the object's mass if we are using that object's mass to find the field?

I know it sounds backwards, so I'll explain by way of analogy. Basically, we have to recognize that the force itself depends on the object's mass, so dividing the force by the object's mass removes that dependence.

As an analogy, consider how we convey how expensive gasoline is, Basically, we do this by looking at the *price per gallon*, not the *total price* we paid for gas. This is because the total price depends on how much gas we put in our car, whereas the price *per gallon* does not. By dividing the total price by the number of gallons bought, we get a quantity (price per gallon) that only depends upon the gas station, not how many gallons we bought.

In a similar way, by dividing the gravitational force (which depends on both objects) by the mass of the *influenced* object, we get a quantity (the *gravitational* field) that only depends upon the *influencing* object, not the object that is experiencing the force. For that reason, the gravitational field is measured in units of newtons *per kilogram* (N/kg).

In this way, it doesn't matter what object we use to measure the gravitational field although for practical purposes we typically use something very small. The "test object" is called a **probe**. The gravitational force on the probe, divided by the mass of the probe, is equal to the gravitational field at that location due to all of the other objects present *other* than the probe.

✓ *Check Point 6.1: I hold a 2-kg ball in my hand. I find that the gravitational force on the ball, due to Earth, is 19.6 N downward. The mass of Earth is* 5.98×10^{24} *kg. To find the magnitude of Earth's gravitational field at the location of the ball, which mass would you divide the force by and why?*

6.2 The electric field

Whereas the gravitational field has units of newtons per *kilogram* (N/kg), the electric field has units of newtons per *coulomb* (N/C) because the electric

force depends on the charge (SI unit = coulomb) rather than the mass (SI unit = kilogram). As with the gravitational field, by dividing the electric force on a probe by the charge of the probe we get a quantity that is independent of the probe's charge and only dependent on the electric field of the other objects that may be present.

⁝ The electric field has SI units of N/C.

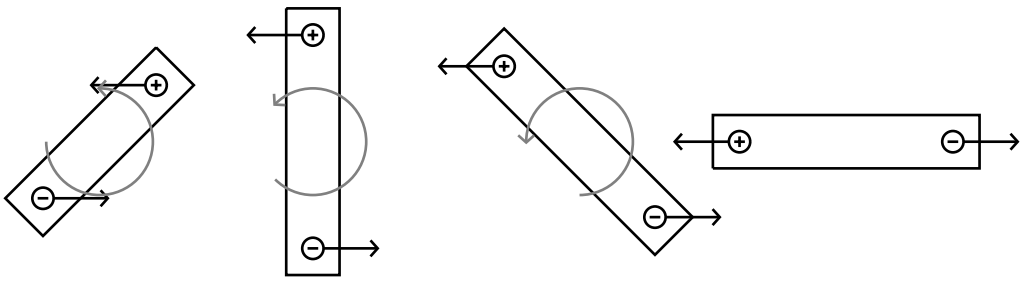
✍ By convention, we use E to represent the strength of the electric field.

For example, at about 30 cm away, appliances typically have an electric field strength between 5 N/C (light bulb) and 180 N/C (stereo receiver).[i] Even without electrical appliances, there is an electric field all around us due to the natural charge separation in the atmosphere. That natural field varies greatly from day to day but, on average, it is about 100 N/C downward, about the same strength as a refrigerator's electric field (about 30 cm away).

✓ *Check Point 6.2: An excess charge of +2 μC is placed on an insulated conducting sphere of radius 2 cm. According to section 10.5, the distribution of charge around the surface is such that no electric force is exerted on charged objects inside the sphere. Given this information, what is the electric field at the center of the sphere due to the distribution of charge around the surface?*

We can now explain why a spark occurs when we touch a doorknob and what produces lightning during a thunderstorm.

We know from section 5.2 that dipoles experience a torque that makes them align with the electric field within which they've been placed. Once aligned, the positive end of the dipole is forced one way and the negative end of the dipole is forced the other, as shown below.



[i]The electric field associated with various appliances was obtained from the Federal Office for Radiation Safety, Germany 1999.

As we know, every atom has both positive and negative particles, called protons and electrons, respectively. In an insulator, the electrons are fixed to the atoms. However, in a strong enough electric field, the opposite pull on the protons and electrons can not only polarize the atoms but pull an electron off each atom. Those now-free electrons can flow through the material, making it a conductor. At that point, we say that the insulator has "broken down."

⁝ The dielectric strength represents the maximum electric field an insulator can withstand without breaking down.

The maximum electric field that an insulator can withstand without breaking down (and conducting current) is called the **dielectric strength**. The dielectric strength varies from material to material. For example, the dielectric strength of air is about 3×10^6 N/C.[ii]

When a material breaks down, it is no longer an insulator (as there are now free electrons present). These electrons can accelerate to great speeds and heat up the material as it collides with the molecules that make up the material. This is essentially why[iii] a spark occurs. The material heats up much like the filament of an incandescent light bulb.[iv]

✍ The greater the charge on an object, the greater the electric field outside the object and, consequently, the greater the chance for the surrounding insulator to break down. Since charge naturally concentrates in corners of conductors (see section 10.3.1), that is where the electric field just outside the conductor will be greatest and, consequently, where the surrounding insulator is most likely to break down and create a spark.

✓ *Check Point 6.3: Suppose someone asserted that the electric field at your location was* 10^8 *N/C. What evidence could you point to that would suggest the actual electric field is likely much less than that?*

[ii] Technically, this value is less than what would be needed to actually pull an electron off an air molecule. This is because of the presence of individual free electrons that happen to be in the air. The electric field accelerates those electrons to a speed large enough to "knock off" another electron when it hits an air molecule. This starts a cascade production of free electrons, which is the same end result as a breakdown of the air.

[iii] A secondary mechanism of light generation is due to luminescence, where light is given off as electrons "rejoin" the atom. Only specific wavelengths are emitted during luminescence and the gas need not get hot.

[iv] The heating of the gas can be dangerous if there are combustible gases present.

6.3 Magnetic field

As discussed in the previous section, for many applications we care more about the *strength* of the electric field than its *direction*. This is also true for the magnetic field. For example, we might need to know the potential influence on an object due to the magnetic field associated with an MRI[v]. If the magnetic field is sufficiently strong, there will be a noticeable impact on any ferromagnetic material that may be present.[vi] By quantifying the magnetic field, we can identify what it means to be *sufficiently strong*.

⁉ The magnetic field has SI units of T (tesla).

There are a couple of differences between how we indicate the strength of the magnetic field vs. how we indicate the strength of the gravitational and electric fields. Whereas the gravitational and electric fields have units of force (N) per mass (kg) or charge (C), we use a totally new unit for the magnetic field: the **tesla**, which we abbreviate as T.[vii]

⁉ At our location, Earth's magnetic field is relatively weak, at about 50 microteslas.

At our location, the magnetic field of Earth is about 50 microteslas (i.e., 50 μT or 5×10^{-5} T), directed northward (for the most part). That is a pretty weak magnetic field, which is why it requires a very light compass needle, balanced on a pin, to sense it.[viii]

As you increase the magnetic field, however, it starts to have a more noticeable effect on things. At 3 milliteslas (i.e., 3 mT or 3×10^{-3} T), or roughly the magnetic field strength close to a refrigerator magnet, metal objects and instruments may be forced to move.

Still, such magnetic field strengths are unlikely to impact us[ix] until you get much higher.[x]

✍ In equations, B is used to represent the magnetic field strength.[xi]

[v] Magnetic resonance imaging (MRI) utilizes the natural resonance of particular atoms under a magnetic field.

[vi] Pacemakers are typically not ferromagnetic, so they shouldn't be a problem.

[vii] This unit honors Nikola Tesla (1856-1943), who was born in Croatia and emigrated to the United States in 1884. Tesla worked on motors that utilized alternating current.

[viii] Although weak, some animals apparently use it for navigation.

[ix] According to the World Health Organization (Fact sheet 299, March 2006), at about 500 microteslas (i.e., 500 μT or 5×10^{-4} T), it can start to influence people with pacemakers.

[x] A magnetic field strength of 16 T can be used to levitate a frog. As you might guess, that is quite a significant magnetic field.

✓ *Check Point 6.4: Earth's magnetic field (at our location) is about* 5×10^{-5} *T. In comparison, close to a small bar magnet, the magnet's magnetic field is about* 10^{-2} *T. What does the "T" stand for?*

The reason why we use T as the SI unit for magnetic field is because the *actual* SI unit is equal to a N/A·m. This mixture of units is a little complicated, which is why we use T (tesla) instead.

The mixture of newtons, amperes and meters, however, can be quite illuminating if you understand where the mixture comes from.

Since every magnet is a magnetic dipole, a uniform magnetic field can't exert a net force on a magnet – it can only rotate it (like Earth's magnetic field on a compass needle). Consequently, we can't determine the magnetic field by first measuring the force on a magnet probe.

However, we *can* measure the torque on a magnet probe. It turns out that the magnetic field is equivalent to the magnetic torque on a magnet divided by the **magnetic moment** of the magnet probe.

I'll explain in a little bit what the magnetic moment is, but first I want to point out that the magnetic moment has units of A·m^2. Since torque has units of N·m (since torque is the product of the force and the distance from the rotation axis), that means that the magnetic field, which is the torque divided by the magnetic moment, must have units of N·m divided by A·m^2, which is (N·m)/(A·m^2) or, simplified, N/(A·m), equivalent to what I mentioned earlier.

But what is the magnetic moment?

To answer that, let's first consider what mass and charge mean. We know that an object with mass is influenced by a planet's gravitational field. The

[xi] We can blame James Clerk Maxwell, a Scottish physicist who lived from 1831 to 1879, for why we use B instead of something more obvious like M. He published a paper in which he chose letter abbreviations for quantities alphabetically as he encountered them in his paper. The magnetic field just happened to be second. Some assignments were eventually changed, like using I for current instead of Maxwell's C but Maxwell's use of B for the magnetic field remains the convention today. Maxwell was also not that good at efficiency, so several of his variable abbreviations (like B and H) referred to somewhat similar concepts, leading to confusion among generations of students.

larger the object's mass, the greater it is influenced by the planet's gravitational field. Similarly, we know that an object with charge is influenced by a proton's electric field. The larger the object's charge, the greater it is influenced by the proton's electric field.

In a similar way, it is the magnet's magnetic moment that indicates how much it will be influenced by another magnet's magnetic field.

WHY DOES THE MAGNETIC MOMENT HAVE UNITS OF A·M^2?

To answer this, let's consider an electromagnet that consists of just a single current loop. The strength of this simple electromagnet depends on its cross-sectional area (measured in m^2) and the amount of current flowing through it (measured in A, for amperes).

The product of the two is called the magnetic moment (or **magnetic dipole moment**) and is measured in units of A·m^2. For example, if you have a loop of area 0.2 m^2 through which a current of 2 A flows, the magnetic moment would be 0.4 A·m^2.

⁝ The magnetic torque on a current loop depends on its magnetic moment (size times current) and the magnetic field in which it is placed.

WHAT IF YOU HAD AN ELECTROMAGNET WITH MANY LOOPS?

A coil with many loops is called a solenoid. To calculate the magnetic moment for a solenoid, simply calculate the magnetic moment for a single loop and then multiply by the number of loops. For example, if you have 100 loops of area 0.2 m^2 through which a current of 2 A flows, the magnetic moment would be 100 times 0.4 A·m^2, or 40 A·m^2.

DOES A PERMANENT MAGNET HAVE A MAGNETIC MOMENT?

Yes, permanent magnets have a magnetic moment, with the same SI units (A·m^2). While permanent magnets may not have a single current and area like electromagnets, we can think of them as consisting of lots of tiny little magnets, each with their own tiny little current loops and area. The magnet's magnetic moment, then, is the sum of all the magnetic moments of the tiny little current loops inside the magnet.

✓ *Check Point 6.5: At our location, Earth's magnetic field is about 50 μT, sufficient to rotate both a compass needle as well as a heavier, stronger magnet that is placed on a frictionless pivot. The torque exerted on the compass needle is less than the torque exerted on the stronger magnet. Given that, which of the two objects has a greater magnetic moment?*

Summary

This chapter quantified the strength of the gravitational, electric and magnetic fields.

The main points of this chapter are as follows:

- The gravitational force has SI units of N/kg.
- The electric field has SI units of N/C.
- The dielectric strength represents the maximum electric field an insulator can withstand without breaking down.
- The magnetic field has SI units of T (tesla).
- At our location, Earth's magnetic field is relatively weak, at about 50 microteslas.
- The magnetic torque on a current loop depends on its magnetic moment (size times current) and the magnetic field in which it is placed.

Frequently asked questions

What is the difference between a force and a field?

Force is measured in newtons while the units for the field depend on the kind of field (gravitational, electric or magnetic). Force is due to an interaction between two objects whereas field is a property of a single object.

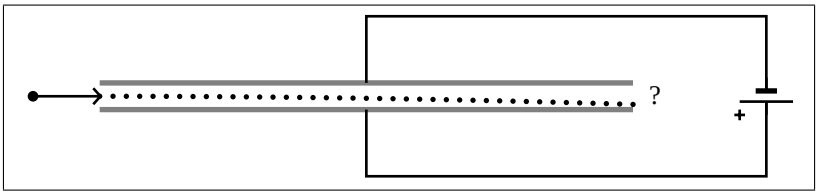
Terminology introduced

Dielectric strength	Probe
Magnetic dipole moment	Tesla
Magnetic moment	

Abbreviations introduced

Quantity	**SI unit**
gravitational field strength (g)	newton per kilogram (N/kg)
electric field strength (E)	newton per coulomb (N/C)
magnetic field strength (B)	tesla (T)[xii]

Additional problems



Problem 6.1: An electron is shot from the left midway through the gap between two parallel plates 1 cm apart and 20 cm long (see figure above). An electric field of 20 N/C upward is maintained between the two plates.
(a) Why does the electron path curve downward and not upward?
(b) What is the force on the electron when it is between the two plates?
(c) Suppose an electron is initially moving parallel to the plates with a speed of 400 m/s. If the electron is initially located directly between the two plates, will it make it though the gap? [note: this uses ideas from volume I]

Problem 6.2: At New York City, Earth's magnetic field is mostly horizontal (5.2×10^{-5} T toward the north) but it also has a small vertical component (1.8×10^{-5} T downward). Suppose we have an magnet whose axis is oriented north-south. Is there a torque exerted on the magnet?

Problem 6.3: The atmospheric electric field is about 100 N/C downward. What is the electric force on an electron that is present within the atmosphere?

[xii] A tesla is equal to a newton meter per ampere meter squared (N/A·m).

7. Conservation of Energy

Puzzle #7: When people tell us to turn off the lights to conserve energy, what energy is being conserved?

Introduction

You are probably very familiar with energy. There are energy drinks, energy companies and the energy crisis. There must be a reason why people like to speak in terms of energy rather than forces and fields, right? What is the reason?

The reason is that an accounting of the different types of energy (where is it, how do we get it in the form we want) is more straightforward than using forces and fields, particularly when dealing with temperature.

In this chapter, we examine what it means for energy to be conserved and provide the basic language or describing changes in energy. In chapters 8 and 9, we apply conservation of energy to chemical and nuclear reactions.

7.1 Conservation of energy

There are lots of ways that energy is manifested. For example, if something is moving, it has **kinetic energy**. The faster an object is moving, the greater its kinetic energy. Therefore, when something speeds up, its kinetic energy increases.

However, the overall energy must remain the same. For example, when an object falls, the gravitational energy decreases and the object's kinetic energy increases. As another example, when a gas-powered car speeds up, the chemical energy associated with the gasoline decreases and the car's kinetic energy increases.

Not only does the overall energy have to remain the same, with decreases in one form being balanced by increases in another form, but the 'exchange" of energy has to occur via adjacent objects, much like how forces are associated with interactions between objects.

This idea, that the *total* amount of energy in the universe is the same but just is transferred locally from one type to the other, is called **conservation of energy** (see Volume I).[i]

⁝ Energy is conserved (i.e., it is transferred locally such that the total amount is constant).

In chapters 8 and 9 (chemical and nuclear reactions, respectively), we'll focus on predicting how the change in thermal energy, which is the energy associated with temperature, is related to changes in the chemical and nuclear energy. In this chapter, we'll provide the background needed to properly interpret the energy changes that occur during chemical and nuclear reactions.

✍ Some types of energy are useful to us and others which are not. So, when people outside of physics say you need to conserve energy, they mean that you need to maintain *useful* forms of energy. In other words, non-scientists are using the word "conserve" to mean to save or maintain but they are only applying it to a subset of energy types, whereas in our usage energy is conserved whether we do anything about it or not. If anything, we can just move it around.

✓ *Check Point 7.1: Suppose an interaction occurs during which only three energy types are known to change: the kinetic energy, the gravitational energy and the thermal energy. If the kinetic energy decreases and the gravitational energy decreases, what does conservation of energy imply must happen to the thermal energy: increase, decrease or stay the same?*

7.2 Types of energy

In this section, we'll review the different types of energy that we'll be considering. Remember, if one energy type increases, at least one other energy type must decrease in order for energy to be conserved.

[i] Charge is also conserved. This is discussed in section 11.4.

I've already mentioned that **thermal energy** is the energy associated with temperature. The warmer an object is the greater the thermal energy. Since energy is conserved, any warming must be accompanied by a decrease in at least one other form of energy. Another energy type I've mentioned is **kinetic energy**. The faster an object moves, the more kinetic energy it has.

What we eventually want to do is predict the *change* in thermal energy or kinetic energy. Since energy is conserved, for an object to speed up and its kinetic energy to increase, some other form of energy must decrease. In fact, conservation of energy is essentially an accounting of the various *transfers* of energy, in much the same way that an accountant keeps track of money transfers to see where the money is going. To do that, we need to understand the *mechanisms* responsible for energy transfers.

It turns out forces act as "agents" for transferring energy to and from kinetic/thermal energy. As we know from part A, an object speeds up when the net force on it is in the direction of its motion. This corresponds to an increase in kinetic energy. That increase in energy is "delivered" by the force, transferring the energy from a "bank" of energy called **potential energy**[ii], so called because energy can potentially be transferred from that energy to kinetic energy.

For example, a ball released from rest speeds up as it falls due to the gravitational force. As it speeds up, the ball's *kinetic* energy increases while the gravitational *potential* energy decreases. Conversely, after being thrown upward, the rising ball slows down due to the gravitational force. As it slows down, the ball's *kinetic* energy decreases while the gravitational *potential* energy is increases.

Is there such a thing as "gravitational kinetic energy"?

No. Kinetic energy is associated with motion of individual objects, not the force responsible for that motion.

Is there such a thing as the "ball's potential energy"?

No. An individual object has kinetic energy (due to its motion) but the potential energy is associated with the potential for attraction or repulsion and thus requires a *system* of two or more interacting objects.

Does "potential energy" mean "gravitational potential energy"?

[ii] Also sometimes called the **interaction energy**.

Not necessarily. There are many reasons why an object changes speed. It need not be because of the gravitational force. In the example with the ball, the gravitational force is the "agent" for transferring energy and thus it is the *gravitational* potential energy that is increasing or decreasing. However, there are other types of potential energy as well, each corresponding to the "agent" responsible for the transfer to or from the object's kinetic energy. Every interaction, whether gravitational, elastic, electric, magnetic or whatever, is associated with a particular type of potential energy, like gravitational potential energy, elastic potential energy, electric potential energy, and so on.

In this section, we consider a couple of different forces and the particular type of potential energy "bank" that each force transfers energy to and from.

✓ *Check Point 7.2: Can an individual object have potential energy?*

7.2.1 Elastic potential energy

Consider two objects connected by a spring. Every spring has an equilibrium length or "unstretched" length where it neither pulls the objects together nor pushes them apart. If the spring happens to be stretched beyond its equilibrium length then the spring exerts an attractive force on the two objects, pulling them together. Conversely, if the spring is compressed shorter than its equilibrium length then the spring exerts a repulsive force on the two objects, pushing them apart.

In each case, whether stretched beyond the equilibrium length or compressed to a length shorter than the equilibrium length, we say that there is **elastic energy** "stored" in the spring. The greater the spring is stretched beyond the equilibrium length, or compressed beyond than the equilibrium length, the greater the elastic (potential) energy.[iii]

[iii] There is no need to say "elastic potential energy" rather than "elastic energy", since there is no other type of elastic energy (there is no "kinetic elastic energy," for example). However, sometimes it helps to add it in parentheses just to remind us of that fact. It is similar to how we often refer to the Canadian national flag as just the Canadian flag but we can include the word "national" if we want to emphasize that (as in "national flag of Canada").

The greater the elastic energy, the more energy that can be transferred to kinetic energy. For example, if we hold the two objects apart, with the spring stretched farther than the equilibrium length, then the objects will move together (due to the spring pulling them together) when we let go. When we let go of the objects, the stored elastic energy (associated with the spring) decreases and the kinetic energy of the objects increases.

Conversely, if we hold the two objects close together, with the spring compressed shorter than the equilibrium length, then the objects will move apart (due to the spring pushing them apart) when we let go. When we let go of the objects, the stored elastic energy (associated with the spring) decreases and the kinetic energy of the objects increases.

Notice that in both cases the force of the spring is responsible for speeding up the objects, and so in each case the elastic energy decreases as the force of the spring acts to speed up the objects, increasing their kinetic energy.

CAN THE SPRING ALSO ACT TO SLOW DOWN OBJECTS?

Yes, this happens when the force of the spring is opposite the direction the objects are moving. For example, if the objects are moving apart while the spring is stretched longer than the equilibrium length then the objects slow as the spring force acts to transfer energy from kinetic to elastic. The same is true when the objects are moving together while the spring is compressed shorter than the equilibrium length.

✍ Regardless of the interaction, you'll notice that the kinetic energy increases when the situation transitions from a high potential energy configuration to a low potential energy configuration.

✓ *Check Point 7.3: An example of elastic energy is the energy associated with a stretched rubber band. When you "shoot" a rubber band, you first stretch it between two fingers and then you let go of one end. The rubber band contracts to its equilibrium length and it flies across the room. As it flies, it has a non-zero kinetic energy.*
(a) While the rubber band contracts to its equilibrium length, what is happening to the elastic energy (increase, decrease or remain the same) and what is happening to the kinetic energy (increase, decrease or remain the same)?
(c) Explain how your answers to (a) and (b) are consistent with conservation of energy.

7.2.2 Gravitational energy

Gravity is like an invisible spring connecting all objects. And, just like springs that are stretched beyond their equilibrium length, gravity is an attractive force and there is a "stored" **gravitational energy** associated with that attraction.

Notice that the gravitational energy is associated with the attraction, which we can imagine as an invisible spring, rather than either of the two objects that are attracting gravitationally.[iv] For example, a rock falls to the ground because there is a gravitational attraction between the rock and Earth. And, just like a stretched spring, there is an energy associated with that attraction, which transfers to the kinetic energy of the rock as it falls, making it speed up.

⁉ Gravitational energy belongs to the interaction, not to either of the two interacting objects.

Does the gravitational energy always decrease when two objects come together?

Yes. The gravitational force is an attractive force. Consequently, the two interacting objects will speed up as they get pulled toward each other via the gravitational force. That leads to an increase in kinetic energy of one or both objects and a corresponding decrease in gravitational energy.

Does the gravitational force also decrease when two objects come together?

No. The gravitational force actually is stronger the closer the two objects are to each other. The gravitational energy decreases when two objects come together simply because the force is attractive. It is the same reason why the elastic energy decreases when two objects are pulled together by a spring (if stretched beyond the equilibrium length).[v]

[iv] It is quite common for people to incorrectly associate the gravitational energy with just the object that is being pulled toward Earth, rather than with the attraction between that object and Earth. While incorrect, they can get away with this as long as the focus is on predicting the motion of that single object near the surface of Earth. However, we are looking at energy with an eye toward predicting atomic and nuclear interactions, where both interacting objects (e.g., atoms) are moving, and so it is important that we follow the more correct approach (i.e., where the gravitational energy is associated with the interaction).

[v] Unlike the gravitational force, the spring force becomes stronger the further it is stretched.

✍ Gravitational energy is not the same thing as the gravitational force. The gravitational force is the agent that provides for the transfer of energy from gravitational to kinetic, much like a bank teller acts as an agent to transfer money from your bank account to you (and the reverse). The bank teller is unaffected by the transfer, just as the gravitational force is unaffected by the transfer of energy. If the gravitational force changes, it is due to something else.

What if there is air resistance?

If there is air resistance (drag), there is another interaction going on: between the rock and the air.

The drag force acts opposite the motion and, as such, it acts as an agent to *remove* kinetic energy. Consistent with conservation of energy, that energy must go somewhere, and for air resistance that energy gets transferred to thermal energy, meaning that the air warms up.

So, for the rock rising and slowing as it does, the rock's kinetic energy decreases more quickly because energy is not only being transferred to gravitational energy (by the gravitational force) but also to the air in the form of **thermal energy** (by the drag force).

What happens when two objects move apart?

Since the gravitational force is an attractive force, you might think that two objects won't naturally move apart if the gravitational force is the only force acting. However, objects in orbit around the sun move closer and farther from the sun during their orbit. When they move closer, the gravitational attraction acts to speed them up, increasing the kinetic energy as the gravitational energy decreases. Conversely, when they move apart, the gravitational attraction acts to slow them down, decreasing the kinetic energy as the gravitational energy increases.

✍ One can also consider a rock that has been thrown up in the air. As it continues to move upward (after being released), it *loses* kinetic energy as it slows. The gravitational force, by being opposite the rock's motion, acts as the agent for transferring energy *from* the kinetic energy to the gravitational energy.

✓ *Check Point 7.4: Halley's comet orbits the sun with a period of 76 years. (a) When the gravitational energy associated with the comet/Sun interaction*

increasing: when the comet is moving toward the sun or when the comet is moving away from the sun?
(b) When is the gravitational force on the comet, due to the Sun, increasing: when the comet is moving toward the sun or when the comet is moving away from the sun?

7.2.3 Electric energy

The electric force is like an invisible spring that connects all charged objects and there is stored **electric energy** associated with that invisible spring.

⁉ The electric energy associated with two opposite charges is larger when the charges are further apart.

Unlike the gravitational force, the electric force can be either attractive or repulsive, depending on whether we have opposite charges or like charges. If attractive, as with a spring stretched beyond its equilibrium length, the electric energy is greater the farther apart the two interacting objects. If repulsive, as with a spring that is compressed smaller than its equilibrium length, the electric energy is greater the closer together the two interacting objects.

⁉ The electric energy associated with two like charges is larger when the charges are closer together.

For example, a negatively charged object is attracted to a positively charged object. This is attractive, and just as a rock's kinetic energy increases as the rock and Earth are pulled together, so does the kinetic energy of the oppositely-charged particles as they are pulled together. The difference is that the gravitational force is acting as the agent in one case (transferring energy from gravitational to kinetic) whereas it is the electric force in the other case (transferring energy from electric to kinetic).

✓ *Check Point 7.5: (a) An electron is attracted to a positive ion, being of opposite charge. When they come toward each other, does the electric energy increase, decrease or stay the same?*
(b) Suppose we physically try to separate the electron from the positive ion. As they move farther apart, does the electric energy increase, decrease or stay the same?

7.3 Absorbing and releasing energy

Up to now, we've considered situations where one type of potential energy changes (elastic, gravitational, or electric) and the kinetic energy changes in the reverse way, consistent with energy conservation.

Let's now examine situations where the kinetic energy has the *same* value at the end as it had at the beginning. Note that the kinetic energy can change during the time period as long as it ends up having the same value it had at the beginning.

For example, consider a rock initially at rest that drops to the floor, coming to rest on the floor. At the beginning, the rock is at rest. At the end, the rock is again at rest. Consequently, the kinetic energy is the same at the end as at the beginning.

Note that the kinetic energy didn't *stay* at zero the entire time. As the rock fell, it sped up, which means its kinetic energy increased as it fell. However, once it hit the ground, the rock stopped, decreasing its kinetic energy back to zero. Although the kinetic energy was changing between the top and bottom, the kinetic energy at the end (zero) is the same as what it was at the beginning (zero).

Why should we consider situations like this?

It turns out that we'll consider something similar when examining chemical and nuclear reactions in chapters 8 and 9. In such situations, the kinetic energy is the same before and after the reaction but the **thermal energy** is not. Recall that thermal energy is the energy associated with an object's temperature. If the thermal energy increases, that means the object's temperature increases.

So, let's revisit the situation with the rock falling to the ground. Since the gravitational force is responsible for pulling the rock downward, we expect that the gravitational energy must be involved.

It is. The gravitational energy is *less* at the end than at the beginning. As mentioned in section 7.2.2, the gravitational energy decreases as two objects come toward each other because the objects are attracting gravitationally.

Has any other type of energy changed?

Since the kinetic energy hasn't changed (zero at beginning and end) and the gravitational energy has decreased, that means that there must be an increase in at least one other type of energy.

In this case, that energy is likely the thermal energy associated with the warming of the rock and ground as the rock hits the ground.[vi]

✓ *Check Point 7.6: Bungee jumping involves tying someone to a bungee cord and then having them jump off a bridge. As the person falls the bungee cord stretches and eventually stops the person from falling. After oscillating a couple of times, the person comes to rest somewhere between the bridge and the valley below. From the time the person was at rest on the bridge to the time the person is at rest again somewhere between the bridge and the valley below:*
(a) Was there an increase, decrease or no change in gravitational energy?
(b) Was there an increase, decrease or no change in kinetic energy?
(c) Based on your answers to (a) and (b), has any other type of energy changed? If so, what might that be and did it increase or decrease? If not, why not?

For situations where the kinetic energy is the same before and after, it helps to use a language that keeps track of the energy that is transferred between the system and the environment.

We'll treat the **system** to be the two objects that are attracting, which would be the rock and Earth in the example provided earlier. In comparison, we'll treat the **environment** as everything else. For the rock and Earth, the environment could be a person who happens to pick up the rock prior to dropping it. The environment could also be any of the other objects that happen to be present, like the dirt, air, trees, etc.

Since the rock and Earth are attracting gravitationally, we'll consider the gravitational energy to belong to the system. Conversely, since thermal energy can't be "contained" to just one object, as one warm object will warm nearby objects, we'll consider thermal energy to belong to the **environment**.[vii]

[vi] There is also sound energy, created when the rock hits the ground, but that sound energy is eventually absorbed by the air and surroundings warming them up.

[vii] This becomes more appropriate when we deal with point particles, as in chapters 8

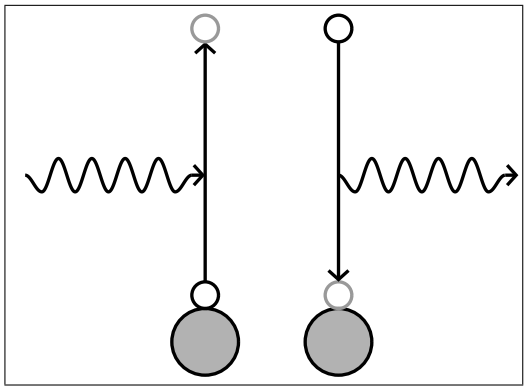


Figure 7.1: A rock that is raised off Earth [left] and a rock that has fallen to Earth [right]. The squiggly arrows indicate energy transfer from the environment (thermal) to the Earth and rock system (gravitational) [left] and from the Earth and rock system (gravitational) to the environment (thermal) [right].

So, consider the situation of the dropped rock. In that case, we can say that the energy of the system (gravitational) decreased and the energy of the environment (thermal) increased.

✍ If no energy transfers to the environment then the rock would bounce off the ground and return to the height from which it was released.

This is illustrated on the right side of Figure 7.1, where the solid arrow represented the rock falling down to Earth and the squiggly arrow represents the transfer of energy from gravitational (Earth and rock system) to thermal (environment).

As another example, let's consider the reverse situation, where a rock is at rest on the ground and then we come along, pick it up and move it to a position further from Earth (like on a shelf), where it is again at rest.

⁉ It takes energy (from the environment) to move two attractive objects apart and energy is released (to the environment) when they move toward each other. The reverse is true for repulsive objects.

As with the previous example, the rock's kinetic energy at the end is zero, just like what it was when the rock was on the ground. However, since the rock and Earth are farther apart, the gravitational energy of the Earth and rock system has increased. The transfer of energy is illustrated on the left side of Figure 7.1 where the squiggly arrow indicates the transfer of energy from thermal (the environment) to gravitational (Earth and rock system).

For our purposes, we can say that energy is *absorbed* by the Earth-rock system when the rock is raised up to the shelf and, conversely, energy is *released* to the environment when the rock falls to the floor (and comes to rest on the floor).

and 9. Point particles like electrons and protons can't warm up separately from the kinetic energy, so for the cases we will examine it will be sufficient to assign the thermal energy to the environment.

We can use the same language for charged particles. For example, consider an electron in a hydrogen atom. When the hydrogen atom receives a photon of light of a particular light energy, it can "bump" the electron to a "higher" electron shell. Because the electron and proton become further apart, the electric energy of the electron/proton system increases. This corresponds to the absorption of energy from the surrounding environment (light energy), a process illustrated in Figure 7.1 as the squiggly arrow on the left.[viii]

If the electron then "falls" to a lower electron shell, a photon is emitted. This corresponds to the release of energy from the electron/proton system to the surrounding environment, as illustrated by the right side of Figure 7.1.

✓ *Check Point 7.7: For the bungee jumper who is at rest before jumping off the bridge and also at rest after jumping off and coming to rest again somewhere between the bridge and the valley below:*
(a) Was energy transferred to the jumper/Earth system or was it transferred from the jumper/Earth system?
(b) Was energy absorbed by the jumper/Earth system or was it released to the surrounding environment?

7.4 Quantifying changes in energy

In chapters 8 and 9, we'll predict the energy released or absorbed in chemical and nuclear reactions. To do this, we need to quantify energy. Up until now, everything has been qualitative (without numbers).

[viii] The electron is moving "around" the proton in both states and thus has kinetic energy both before and after the transition. We can assume the kinetic energy is the same (no change in kinetic energy) or that the kinetic energy is part of the system. Either way, energy is transferred to the electron/proton system when the electron rises to a higher electron shell.

7.4.1 Kinetic energy

You may recall from the first semester of physics that **kinetic energy**, E_k, is defined as follows:[ix]

$$E_k = \tfrac{1}{2}mv^2 \tag{7.1}$$

As you can see, the kinetic energy is related to how fast an object moves (v). The faster it moves, the greater the value of v.

✍ The one-half factor is needed to make the value consistent with the units of mass, velocity and energy. How it comes about is not particularly relevant to our discussion here, which is focusing on the concept of energy conservation and the different energy types.

According to the definition of kinetic energy (equation 7.1), kinetic energy has SI units of kg·m^2/s^2. I obtained that messy group of units by multiplying the SI unit of mass (kg) by the square of the SI units for velocity (m/s).

⁉ The SI unit of energy is the joule (J).

Since the group of units is messy, we usually replace it with a single unit called the **joule** (J). The following examples should suffice for illustrating how the units work:

- If I walk down the street at a speed of 1 m/s then, since my mass is 75 kg, my kinetic energy is 37.5 J (square 1 m/s, multiply it by 75 kg then divide by 2).
- Let's suppose I go bowling. I take a 6-kg bowling ball, initially at rest, and then roll it down the bowling lane. If the ball ends up rolling at 3 m/s then it has gained 27 J of translational kinetic energy (square 3 m/s, multiply it by 6 kg then divide by 2).[x]

✓ *Check Point 7.8: A newton is equivalent to a kg·m/s^2. Show that a N·m is equivalent to a joule.*

7.4.2 Power ratings

As you might already know, light bulbs are rated by their **power**, which has SI units of watts (e.g., a 60 watt light bulb).

[ix] Try not to confuse the abbreviation for energy with the abbreviation for electric field. Both are represented by the letter E.

[x] There is also about 11 J of rotational kinetic energy.

Why are light bulbs measured in watts?

A light bulb transfers electric energy into thermal and light energy. The power (or **wattage**, when measured in watts) is the rate at which energy is transferred from one energy type to another. So a bulb's wattage is intimately tied to how bright the bulb is when current flows through it (see next section).

Basically the bulb "extracts" energy from the circuit in much the same way that a wind turbine wheel extracts energy from the wind. Just as the wind turbine doesn't remove air from the atmosphere, the bulb doesn't remove electrons from the circuit.

✍ You may at times hear people mention a power "loss" or a "drain" of power. This just refers to situations where energy is being transferred to energy types like light and thermal, which can't be recovered back. There is no loss of *charge* (nor any loss of electrons, for that matter) during the energy transfer process. The electrons pass through the bulb, leaving at the same rate they enter.

⁉ Power, measured in SI units of watts, is defined as the rate at which energy is transferred to a different form.

Mathematically, the definition of power can be expressed as follows:

$$P = \frac{\Delta E}{\Delta t} \tag{7.2}$$

where P is used to indicate the power, ΔE is the energy that is transferred, and Δt is the time.

⁉ A watt is equivalent to a joule per second.

Since the power is the energy transferred per time, it is measured in joules per second (J/s). A **watt** (W) is equivalent to a joule per second. For example, a 60-W light bulb transfers 60 J of electric energy to heat and light every second.[xi]

✓ *Check Point 7.9: Calculate the energy used (in joules) for a 1.2 kW iron that is used for half of an hour.*

Is a higher wattage bulb necessarily brighter?

For a given construction, we can assume that the higher the wattage, the greater the amount of light that is generated per second.

[xi] If you were to shine a 1-mW laser pointer on a liter of water, it take almost 10 days to warm up the water by 1°C if the water was able to absorb all of the light.

However, we need to keep in mind that the electric energy need not be transferred entirely into light energy. A portion could be transferred to thermal energy. How bright a bulb is depends on the amount of light energy, so if most of the energy is being transferred to thermal energy, not light energy, the bulb won't be as bright.

An incandescent bulb, for example, gets a lot warmer, for the same wattage, than an LED bulb. Consequently, for the same wattage, an LED bulb would be brighter than an incandescent bulb.[xii]

Aren't LED bulbs just as bright as incandescent bulbs?

Yes, but that is because LED bulbs have a lower wattage.

For example, an 8-W LED bulb can be just as bright as a 60-W incandescent bulb if 45% of the energy emitted by the LED bulb is light energy (as opposed to thermal energy) compared to only 6% for the incandescent bulb.[xiii] This just means that they emit light energy at the same rate but the incandescent bulb generates a great deal more thermal energy.

Most of the energy dissipated by an incandescent bulb is in the form of heat, not light.

✍ Because not all of the energy is transferred to light energy, most bulbs are also rated in terms of their **luminosity**, which indicates how bright they get. For example, an 8-W LED bulb and a 60-W incandescent bulb may have the same luminosity.

✓ *Check Point 7.10: An incandescent bulb is hot to the touch whereas an equally bright LED bulb is not. Why?*

7.4.3 Paying for electricity

A 60-Watt light bulb converts 60 J of electric energy to heat and light every second? That sounds like a lot of energy! How much does that cost?

[xii] This is because an LED bulb uses a different process to create light. An LED bulb uses luminescence, where the light is emitted when an electron falls to a lower orbital level within an atom. An incandescent bulb uses incandescence, where the light is emitted because the material is very hot.

[xiii] Six percent of sixty is equal to forty-five percent of eight.

It turns out that 60 J is not expensive, probably just a tiny fraction of a penny. To determine the actual cost is a bit tricky, however, because the utility companies do not use units of joules when calculating the cost. Instead, they use a unit called the **kilowatt·hour**.

What's a kilowatt·hour?

A kilowatt·hour is equal to the *product* of a kilowatt and an hour. To see what this means, let's first rewrite equation 7.2 as follows:

$$\Delta E = P\Delta t.$$

As you can see, energy usage can be written as the product of the power (the rate at which it is transferred to another form) and the elapsed time. For example, if an object transfers energy at a rate of P and does so for a length of time equal to Δt, the total amount of energy transferred is $P\Delta t$.

The unit of kilowatt·hour is obtained when the power has units of kilowatts and the time is in units of hours. So, for example, if you transfer energy at a rate of 100 kW for 3 hours, the total amount of energy transferred is 300 kWh (i.e., multiply 100 kW by 3 h).

⁉ A kilowatt·hour is the energy transferred by a one kilowatt process in one hour.

Since the power companies measure power in kilowatts and measure time in hours, it is simpler for them to simply use the kilowatt·hour (or kWh·hr or kWh) as their unit of energy rather than converting it to joules.

The cost per kilowatt·hour varies from place to place but it is around 18.4 cents/kWh for residential customers in Pennsylvania[xiv].

✓ *Check Point 7.11: Calculate the energy transferred (in kilowatt-hours) for a 1.2 kW iron that is used for half of an hour. At 20 cents/kWh, how much would that usage cost?*

[xiv] This was the rate in June 2023. The rate varies depending on the costs of generating the electricity. Coal is typically cheaper than nuclear, for example. In 2023, the cost per kilowatt·hour was around 11.2 cents/kWh in Louisiana whereas it was 41.7 cents/kWh in Hawaii. Commercial, industrial and transportation rates are typically lower.

Summary

This chapter examined how the energy of a system depends on how far apart the two objects are.

The main points of this chapter are as follows:

- Energy is conserved (i.e., it is transferred locally such that the total amount is constant).
- Gravitational energy belongs to the interaction, not to either of the two interacting objects.
- The electric energy associated with two opposite charges is larger when the charges are further apart.
- The electric energy associated with two like charges is larger when the charges are closer together.
- It takes energy (from the environment) to move two attractive objects apart and energy is released (to the environment) when they move toward each other. The reverse is true for repulsive objects.
- The SI unit of energy is the joule (J).
- Power, measured in SI units of watts, is defined as the rate at which energy is transferred to a different form.
- A watt is equivalent to a joule per second.
- Most of the energy dissipated by an incandescent bulb is in the form of heat, not light.

You should now be able to describe the energy changes that occur during gravitational and electrical interactions.

Frequently asked questions

Since energy is conserved, why do some people keep urging us to conserve it?

The reason is that some types of energy are easily converted to the energy we want (thermal energy to heat our homes in winter, kinetic energy for our cars) while other types are not so easy to convert (for example, all of the thermal energy during the summer doesn't help us at all).

When people refer to the need to conserve energy, they are just referring to

the need to slow the rate at which energy is transferred to the non-useful types.

Can the gravitational or electric energy be negative?

For our purposes, the actual value of the gravitational energy is irrelevant – we just need to know the *change* in gravitational or electric energy. In fact, there is no meaning in the value of the gravitational or electric energy. Only the *change* is meaningful.

Still, some physics books talk about the *value* of the gravitational energy and electric energy and, based on their definition, they may find that the gravitational energy and/or electric energy is negative in certain cases. A negative energy value just means that the energy is less than what it would be at some reference situation. Usually that reference situation is when the two objects in the system are infinitely far apart. On the other hand, for situations restricted to objects on or near Earth's surface, that reference is typically taken to be when the object is on the ground.

Because of the ambiguity of the reference and because it isn't necessary for using conservation of energy, I think it is better to avoid trying to find "the" value.

If an electron is attracted to a proton, why doesn't the electron simply "fall into" the proton?

As the electron and proton move toward each other (due to their attraction), the kinetic energy of the electron eventually becomes so high that it moves *around* the proton rather than toward it. Because of this, there is a limit to how close the proton and electron can be. The separation (as in a hydrogen atom) is essentially the radius of the electron's "equilibrium orbit."

Where does potential energy come into this?

There is a class of energy types that are called potential energies. Examples include gravitational energy, electric energy and magnetic energy. If you want, you can call them gravitational potential energy, electric potential energy and magnetic potential energy. Simply saying "potential energy" doesn't tell you which type of energy you are referring to.

The advantage of adding the word "potential" when referring to these energy types is that it emphasizes that the energy can "potentially" be transferred to some other energy (although that is not its only property).

Terminology introduced

Conservation of energy	Interaction energy	System
Elastic energy	Joule	Thermal energy
Electric energy	Kinetic energy	Watt
Environment	Luminosity	Wattage
Gravitational energy	Power	Work

Abbreviations introduced

Quantity	**SI unit**
Energy (E)	joule (J)[xv]
Power (P)	watt (W)[xvi]

Quantity	**non-SI unit**
Energy (E)	kilowatt hour (kWh)

Additional problems

Problem 7.1: A 1-kg object is dropped (from rest). Suppose there is a loss of 50 J of gravitational energy (potential). Assuming no energy is lost via other means, how much kinetic energy does the object gain?

Problem 7.2: Which of the following situations has the greatest electric energy: (a) two protons that are close together, or (b) two protons that are far apart? Explain.

[xv] A joule is equal to a newton meter.
[xvi] A watt is equal to a joule per second.

8. Chemical Reactions

Puzzle #8: ATP is a molecule used in cellular respiration. It consists of another molecule, ADP, bonded to a phosphate. During cellular respiration, the phosphate is removed from ATP, leaving ADP, and the phosphate bonds with another molecule, like water. In the process, energy is "released" to be used for cellular processes. Where does the energy come from?

Introduction

Seeing the puzzle, you might be wondering why I'm bringing up biology when this is a physics book. The answer is because a chemical bond is simply an electric attraction between two polarized atoms or molecules. Consequently, we can apply the ideas of energy conservation, introduced in the previous chapter, to situations where chemical bonds are formed or broken. In other words, we can apply the ideas of energy conservation to chemical reactions.

8.1 The idea of separation

As discussed before, energy is transferred to kinetic energy when a force acts to speed up the object, as when two objects move together due to an attraction or two objects move apart due to a repulsion. It turns out that the amount of energy transferred by a force is equal to the product of the force exerted and the distance the objects have moved together or apart.[i]

$$\Delta E = F_{\text{avg}} \Delta s \tag{8.1}$$

[i] This quantity is also known as the **work** done by the force.

With this relationship, we can determine how much energy is transferred.

Unfortunately, it turns out that this expression is of limited use for us since we'll be examining situations where the objects are moved such a great distance that there is no single value of the force that we can use in the equation.

For example, suppose we totally remove an electron from an atom and, in so doing, ionize the atom. To totally remove the electron, we have to take the electron to a location so far away from the atom that there is no longer an electric attraction between the now-positive atom and the negative electron.

How can two objects be so far apart that they no longer exert a force on each other?

Technically, this is not possible. However, as we know, objects can be far enough apart that their interaction is insignificant. For example, there is a gravitational force between, say, two chairs, but we consider the force to be insignificant compared to all the other forces that are present.

The key thing is to recognize that the force changes between when the electron is "attached" to the atom and when the electron is "separate". To use equation 8.1, we'd need to know the *average* force value, which is not something we're going to consider.[ii]

Instead of using an equation to figure out how much energy is needed to separate two attracting objects (or bring together two repelling objects), we instead will assume that we are provided with the value. After all, that value will depend on what the two objects are and the type of force involved. For example, the energy needed to separate a phosphate from an ATP molecule, as described in the puzzle, will be different than the energy needed to separate two magnets. So, we won't go through the hassle of trying to figure out how much energy is needed – we'll just use whatever value other people have already determined is needed.

✓ *Check Point 8.1: What does it mean to separate an electron from an atom and why won't we be using an equation to figure out the amount of energy needed to do so?*

[ii] It turns out that the average gravitational force is $Gm_1m_2/(r_i r_f)$ (obtained from the universal law of gravitation; see equation 1.1), where r_i and r_f are the initial and final separation distances. A similar relationship applies for the average electric force: $kq_1q_2/(r_i r_f)$. See chapter 2 for the meaning of k and q.

8.2 Ionization energy

In the previous section, an example was described where an electron is removed from an atom, leaving the atom positively-charged. An atom that loses an electron, thereby having a net charge, is called an **ion** and the process of creating an ion (by removing electron) is called **ionization**.[iii]

? The ionization energy refers to the amount of energy required to ionize an atom.

Much like how it takes energy to separate two magnets, it also takes energy to remove an electron from an atom.[iv] The **ionization energy** is the amount of energy needed to remove the electron from the neutral atom. Note that it is not a *type* of energy but rather an *amount* of energy.[v]

Opposite charges attract, so an electron is attracted to the rest of the atom (which is a positive ion without that electron). As we know from the previous chapter, we (the environment) need to provide energy to separate the two. For an electron and the ion, the ionization energy is what we call the *amount* of energy that needs to be transferred to separate the two.

The ionization energy value depends on the atom, as the electron affinity depends on the atom. Based on measurements made by scientists over the years, we have found that the ionization energy ranges from about 8×10^{-19} J to 4×10^{-18} J. It is larger for noble gases (where the electrons are more tightly bound) than alkali metals (which have only one electron in the outer shell). This should make sense, since more energy would be required to remove an electron that is strongly attached to the atom.

The ionization energy is also less for larger atoms (i.e., more nucleons), where the outer electron is farther from the nucleus. This should also make sense, as it starts out being further from the rest of the atom.

[iii] One can also create a negatively-charged ion by adding an electron to a neutral atom. We are focusing on the more common process, where an electron is removed from the neutral atom.

[iv] It is interesting to note that the kinetic energy of an electron in "orbit" around an atom (see discussion on page 126) is roughly equal to the amount of energy needed to separate the electron from the atom, in much the same way that the kinetic energy of a planet in orbit around the sun is roughly equal to the amount of energy needed to separate that planet from the sun.

[v] In much the same way, a dozen apples is an *amount* of apples, not a *type* of apple (like Macintosh, Golden delicious, etc.).

✓ *Check Point 8.2: As mentioned in chapter 7, the system consists of the two objects that are attracting and the environment is everything else. With ionization, the system is the electron and ion. Would it be more appropriate to say that the ionization energy is the amount of energy that is (a) "released" to the environment when the electron is removed or (b) "absorbed" by the system when the electron is removed?*

8.3 Bond dissociation energy

Just as it takes energy to remove an electron from the rest of the atom, it also takes energy to separate two atoms that are bonded as part of a molecule. Recall that many atoms and molecules are electric dipoles, and electric dipoles attract. Atomic and molecular bonds are simply the result of an electric attraction between atoms and molecules.

Atoms that are more strongly bonded (as in double and triple bonds) require more energy to separate than atoms that are weakly bonded. The **bond dissociation energy** is the amount of energy needed to break the bond. As with ionization energy, the bond dissociation energy is not a type of energy. Rather, it is an *amount* of energy.

⁉ The bond dissociation energy refers to the amount of energy required to break a bond

As with ionization energy values, we don't use an equation to determine the bond dissociation energy. Instead, we just use what scientists have measured.[vi] Based on those measurements, we have found that typical bond dissociation energies range from 2×10^{-19} J to 1.6×10^{-18} J. These are about one-quarter to one-third the values for the ionization energy, meaning that it takes less energy to break a bond than to pull an electron off an atom. This is because a chemical bond is not a bond between two ions but rather a bond between two neutral but polarized atoms. Such bonds will require less energy to "break."

What happens when a bond forms?

Forming a bond is like having a rock fall to the ground. In both situations, energy is "released" (to the environment). To see why, we can treat the two

[vi] The bond dissociation energy associated with various bonds is shown in Table 8.1 on page 134.

atoms as two elastic balls that would normally just bounce off each other (rather than bond) unless we could somehow remove their energy and slow them down enough so that they can "stick". That removal of energy is performed by the surroundings.

You might need to reread the last paragraph again. Notice that bond dissociation energy is the amount of energy released (to the environment) when a bond forms (usually in the form of thermal and light energy) and also the amount of energy required (from the environment) to break that same bond.

✓ *Check Point 8.3: In the language of chemical bonding, which situation is associated with the larger "bond dissociation energy": (a) the rock/Earth system when the rock is on the ground outside of a building or (b) the rock/Earth system when the rock is on the roof of the building? Explain.*

8.4 Units

Table 8.1 lists the bond dissociation energy for a couple of bonds. There are a couple of things one should notice about the values in the table.

First, the bond dissociation energies are less for single bonds and more for double and triple bonds, as single bonds tend to be weaker, as discussed in section 8.3.[vii] Second, notice the bond dissociation energy is given in terms of kJ/mol, which stands for kilojoules per mole of bonds.

The reason the units are kJ/mol is because of the minuscule amounts of energy associated with the bond between just two individual atoms. A mole is a huge number. It is equal to about 6.022×10^{23}, a number known as **Avogadro's number**. Whereas it takes about 2×10^{-19} J to 1.6×10^{-18} J to break a bond, it would take 6.022×10^{23} times more to break a mole of bonds. Multiplying 2×10^{-19} J/atom and 1.6×10^{-18} J/atom by 6.022×10^{23} atoms/mole, we get a range of 1.2×10^{5} J/mol to 9.64×10^{5} J/mol.

[vii] With a double or triple bond, the bond dissociation energy refers to the two or three bonds together, which is stronger than a single bond. The two or three single bonds that make up double or triple bond may be weaker than an individual single bond, however.

Table 8.1: Approximate average bond dissociation energies. Source: The Wired Chemist.

Bond	**Bond dissociation energy (kJ/mol)**
O–O	142
N–N	167
P–P	201
N–C	305
P–O	335
N–H	386
C–H	411
N=N	418
H–H	432
O–H	459
O=O	494
C=C	602
N=C	615
O=C	799
C≡C	835
N≡C	887
N≡N	942

Since these numbers are rather large, the bond dissociation energy is usually written in terms of *kilojoules* per mole, in which case the range would be 120 kJ/mol to 964 kJ/mol.[viii]

To find the energy required for an individual bond, simply divide the value in kJ/mol by Avogadro's number. For example, according to the table, the bond dissociation energy associated with the O=O bond[ix] is 494 kJ/mol. Divide this by 6.022×10^{23} molecules per mole to get 8.203×10^{-22} kJ per molecule. Since 1000 joules equals one kilojoule, this is equivalent to 8.203×10^{-19} J per molecule.

✓ *Check Point 8.4: (a) According to Table 8.1, how much energy, on average, does it take to break apart a mole of N–N bonds?*

[viii] Note that mol is the unit abbreviation for mole, which is 6.022×10^{23} atoms or molecules. It is not the unit abbreviation for a single molecule.

[ix] This is an average value. For comparison, the energy required to break the O=O bond in diatomic oxygen, O_2, is 495 kJ/mol.

(b) According to Table 8.1, how much energy, on average, does it take to break apart a mole of N≡N bonds?
(c) Why is the value in (b) so much more than the value in (a)?

8.5 Application

In a chemical reaction, the atoms undergo a rearrangement. The atoms are arranged one way within the **reactants**, which are the molecules we have before the chemical reaction, and another way within the **products**, which are the molecules we have after the chemical reaction. For example, when a hydrocarbon fuel is burned (a process called **combustion**), the reactants are oxygen and a hydrocarbon whereas the products are carbon dioxide and water.

⁉ In a chemical reaction, atoms are rearranged and some bonds are broken (requiring energy) and some bonds are formed (releasing energy)

The same atoms are present in the products as in the reactants. They are simply arranged differently. We can determine the energy released or absorbed by comparing how much energy it takes to break the bonds in the reactants and then compare that to how much energy is released when the bonds in the products form.[x] The amount of energy in both cases is the bond dissociation energy.

In terms of our book falling from a higher shelf to a lower shelf, the process is analogous to comparing the energy needed to raise the book from the higher shelf all the way out to space and then comparing that energy with the energy released when the ball falls from space onto the lower shelf.

The process is illustrated in Figure 8.1, where an object is initially on a higher shelf and then raised to some point (during which energy must be absorbed from the environment) and then dropped down to a lower shelf (during which energy is released to the environment). The net result is a release in energy.

For example, let's calculate the amount of energy released when methane burns. Methane consists of a carbon atom and four hydrogen atoms and is represented as CH_4. To burn methane, the methane (CH_4) is combined with diatomic oxygen (O_2) to form carbon dioxide (CO_2) and water (H_2O)

[x]The bond dissociation value is an estimate and depends on other factors like temperature.

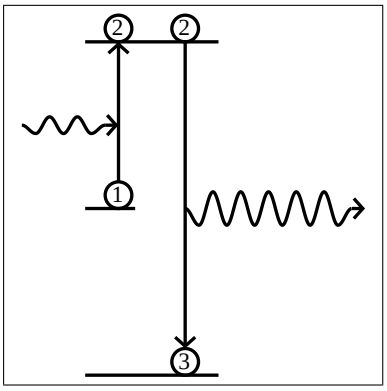


Figure 8.1: A rock that is raised off one shelf and placed on a higher shelf [left] and a rock that has fallen from the shelf onto the floor [right]. Squiggly arrows indicate energy absorbed by the system [left] and released to the environment [right].

as follows:

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$

Each hydrogen atom in a methane molecule is bonded to the carbon atom via its own single bond, so that there are four single bonds in the molecule, one to each hydrogen atom (from the carbon atom). According to Table 8.1, it takes 411 kJ to break a mole of C–H bonds. Since there are four such bonds in each CH_4 molecule, that means it takes 1644 kJ of energy to break the C–H bonds per mole of methane (i.e., four times the 411 kJ/mol).

Diatomic oxygen consists of two oxygen atoms connected by a double bond. From Table 8.1 we find that it takes 494 kJ to break a mole of O=O bonds. There is one such bond in each O_2 molecule but the reaction requires two moles of O_2 for each mole of methane (see equation). Consequently, it takes 988 kJ of energy to break the O=O bonds per mole of methane (i.e., double the 494 kJ/mol).

At the same time that these bonds are breaking, we also have bonds forming.

Carbon dioxide consists of one carbon atom and two oxygen atoms, with each oxygen atom connected to the carbon atom via a double bond (O=C=O). During the reaction, 799 kJ is released when a mole of C=O bonds form. Since there are two such bonds in each CO_2 molecule, 1598 kJ of energy is released per mole of methane (i.e., double the 799 kJ/mol).

Water consists of a single oxygen atom and two hydrogen atoms, with each hydrogen atom connected to the oxygen atom via a single bond (H–O–H).[xi]

[xi] The atoms in a water molecule are not aligned in a straight line. It is more like $^{H}{}_{-O-}{}^{H}$).

From Table 8.1 we find that 459 kJ is released when a mole of H–O bonds form. Since there are two such bonds in each H_2O molecule and the reaction produces two moles of H_2O for each mole of methane (see equation), 1836 kJ of energy is released per mole of methane (i.e., four times the 459 kJ/mol).

So, to find the total energy released, we subtract the total amount needed to break apart the reactants (1644 kJ/mol + 988 kJ/mol = 2632 kJ/mol) from the total amount released when the products form (1598 kJ/mol + 1836 kJ/mol = 3434 kJ/mol). That produces 802 kJ/mol (3434 kJ/mol − 2632 kJ/mol).[xii]

✍ Alternatively, you could be given the standard enthalpy (ΔH) or heat of reaction of each reactant and product (rather than the energy of each bond). The enthalpy is not the same as the bond dissociation energy[xiii] but the end result would be the same.

If every chemical reaction involves the breaking of bonds in the reactants, and breaking bonds requires energy, where does the energy come from to do the initial breaking?

The amount of energy needed to break the first bond is known as the **activation** energy. This is why combustion requires an initial spark or match.

Net energy is released to the surroundings (i.e, more released than absorbed) only if that activation energy is smaller than the energy released during the bond formation that follows. In other words, whether more energy is released or absorbed during the reaction depends on which is greater: the energy released when the new bonds formed or the energy absorbed when the old bonds are broken.

This is illustrated in Figure 8.2, where a rock is initially on the ground and then raised to the top shelf of a bookcase (requiring the absorption of energy) and then dropped to a middle shelf (requiring the release of energy). Since the distance of the fall was less than how far the rock was raised, there is a

[xii] The bond dissociation energy associated with each bond will vary somewhat depending on the molecule involved. For reference, the actual bond dissociation energies for the bonds in methane, oxygen, carbon dioxide and water are as follows: 416 kJ/mol (C–H), 495 kJ/mol (O=O), 803.5 kJ/mol (C=O) and 484.5 kJ/mol (H–O). Using the actual bond dissociation energies rather than the average values in the table, one gets 890 kJ/mole (as opposed to 802 kJ/mole).

[xiii] The enthalpy value (ΔH) is given relative to some reference. Consequently, the enthalpy can be zero, as in the case of O_2, or even negative.

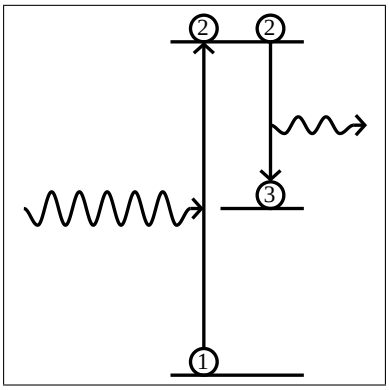


Figure 8.2: A rock that is raised off the floor and placed on a shelf [left] and a rock that has fallen from the shelf onto a lower shelf [right]. Squiggly arrows indicate energy absorbed by the system [left] and released to the environment [right].

smaller amount of energy released to the environment than absorbed. The net result is an absorption in energy.

✓ *Check Point 8.5: For the combustion of methane, the bond dissociation energy for the O=O bonds in the reactants is greater than the H–O bonds in the products. Why, then, is energy released in such a reaction (instead of absorbed)?*

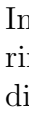
8.6 Language

As we know from previous sections, some bonds are "stronger" than others, in that they require a greater energy to break (i.e., a greater bond dissociation energy). The stronger bonds are usually associated with a smaller **bond length**.

Just as it takes energy to move two attracting objects from a small separation distance to a larger separation distance, it also takes energy (from the environment) to move two atoms from a smaller bond length arrangement (requiring a higher bond dissociation energy to separate) to a greater bond length arrangement (requiring a lower bond dissociation energy to separate). Conversely, energy is released (to the environment) when two atoms move from a greater bond length arrangement (lower bond dissociation energy) to a smaller bond length arrangement (higher bond dissociation energy).

Note that the language is opposite what we'd say when referring to the po-

tential energy associated with the two situations. As discussed in chapter 7, the potential energy is greater when two attracting objects are farther apart. That means the potential energy is smaller when the bond length is smaller, even though that situation is associated with a *higher* bond dissociation energy (more energy needed to separate). Conversely, the potential energy is larger when the bond length is larger, even though that situation is associated with a *smaller* bond dissociation energy (less energy needed to separate).



In biology, when people refer to a "high energy" bond, they are referring to one with high potential energy (large bond length, small bond dissociation energy).

Just as energy is released (to the environment) when a book falls from a higher shelf of a bookcase to a lower shelf, energy is likewise released when atoms rearrange from a larger bond length (higher potential energy bond and lower bond dissociation value) to a smaller bond length.

This is why energy is released during the metabolic process (see puzzle). The ATP molecule (adenosine triphosphate) consists of a bunch of atoms with three phosphates attached to it (hence the TRI-phosphate in the name). The phosphate groups are only weakly bonded, with a high bond length, so it doesn't take much energy to break one off (leaving ADP, or adenosine diphosphate). The broken-off phosphate then combines with water and the resulting bonds are stronger, with a smaller bond length (smaller potential energy; higher bond dissociation energy). The net result is a release in energy, much like how energy is released when a book falls from a higher shelf to a lower shelf.

Notice how the initial bond (phosphate with ADP) is weaker than the resulting bond (phosphate with water). Basically, with a relatively weak bond, like that in ATP, less energy is required to break it (i.e., small bond dissociation energy). Once broken, though, the atoms can be recombined in some other way that produces a stronger bond. During the formation of that bond, energy is released. The greater the difference in bond length, the more energy that is released.

Such a reaction is called an **exothermic** reaction because of the warming that results (where the exo- prefix means outwards).[xiv] An example of a common exothermic reaction is combustion. During combustion, a hydrocarbon

[xiv] The energy transfer need not result in the environment warming up, although that

molecule (hydrogen and carbon) combines with oxygen (O_2) to form water (hydrogen and oxygen: H_2O) and carbon dioxide (carbon and oxygen: CO_2).

The hydrocarbon contains a lot of C–H bonds. As indicated by the bond dissociation values in Table 8.1, the C–H bonds in hydrocarbons are relatively weak (411 kJ/mol bond dissociation value) compared to the C=O bonds in carbon dioxide (799 kJ/mol bond dissociation value). Like the book falling from a higher to lower shelf, the transition from the weaker C–H bonds (higher bond length, higher potential energy, lower bond dissociation energy) to stronger C=O bonds is accompanied by a release of energy to the environment.

What about the bonds in oxygen and water?

Whereas there is a significant difference in the bond dissociation energy of the C=O bonds in carbon dioxide compared to the C–H bonds in the hydrocarbon, the O=O bonds in oxygen and the O–H bonds in water have roughly the same bond dissociation energy (494 kJ/mol vs. 459 kJ/mol, respectively). In other words, it takes about the same energy to break the O=O bonds in oxygen as the O–H bonds in water. Since it takes about the same amount of energy, the energy released by forming the O–H bonds is roughly balanced by the energy absorbed by breaking the O=O bonds, resulting in an insignificant net transfer to or from the environment.[xv]

Will a reaction always result in warming?

No, not all reactions are exothermic. With some reactions there is cooling, due to energy being absorbed from the surroundings. Such reactions are called **endothermic** (where the endo- prefix means inwards).[xvi]

Photosynthesis is an example of a chemical reaction where energy is absorbed. During photosynthesis, which is roughly the opposite of combustion, water

---

is what we'll assume for our purposes. An **exergonic** reaction is one where energy is released to the surroundings, and so an exothermic reaction is a type of exergonic reaction where the temperature increases during the reaction.

[xv] Technically, it takes slightly more energy to break a O=O bond than is released when a H–O bond is formed.

[xvi] As with exothermic reactions, the energy transfer need not result in the environment cooling down, although that is what we'll assume for our purposes. An **endergonic** reaction is one where energy is absorbed from the surroundings, and so an endothermic reaction is a type of endergonic reaction where the temperature decreases during the reaction.

and carbon dioxide are combined to produce oxygen and a hydrocarbon-containing compound (i.e., glucose). Because the reactants have stronger bonds than the products, energy is required during the process. That is why solar energy is required for photosynthesis.

How do we know if a particular chemical reaction will release energy or absorb energy?

If the new bonds (in the products) have smaller bond lengths (lower potential energy, higher bond dissociation energy) than the old ones (in the reactants), energy is "released" (i.e., transferred to the environment).

✓ *Check Point 8.6:During combustion, a hydrocarbon reacts with oxygen to produce water and carbon dioxide. Energy is transferred to the environment in such a reaction. In which molecules do the bonds require more energy to break: the reactants (hydrocarbon and oxygen) or the products (water and carbon dioxide)? Explain your reasoning.*

Summary

This chapter examined how the electric energy of a system depends on the charges involved and how far apart they are.

The main points of this chapter are as follows:

- The ionization energy refers to the amount of energy required to ionize an atom.
- The bond dissociation energy refers to the amount of energy required to break a bond.
- In a chemical reaction, atoms are rearranged and some bonds are broken (requiring energy) and some bonds are formed (releasing energy).

You should now be able to do the following:

- Describe how energy is released or absorbed during chemical reactions.
- Predict how much energy needs to be provided to break a bond or released when a bond forms.
- Given the bond dissociation energies associated with the bonds involved in a chemical reaction, predict how much energy is released or absorbed during the reaction.

Frequently asked questions

ISN'T ENERGY RELEASED WHEN BONDS ARE BROKEN?

No.

Just like it takes energy to break apart two magnets, it takes energy to break apart a molecule.

For example, you may have noticed that you feel cold stepping out of a pool on a dry day. This is due to evaporation of the water on your skin. To evaporate, the water molecules must break free of the other water molecules.[xvii] It takes energy to break those bonds. Evaporation is a cooling process because it takes the energy from the surroundings.

Energy is released only when bonds are formed. For example, to condense, water molecules must form bonds with other water molecules. During this process, energy is released. Condensation is a warming process because it releases energy into the surroundings.

IN CHEMISTRY, IT IS COMMON TO REFER TO THE ENERGY "STORED" IN THE BONDS. DOESN'T THIS MEAN THAT ENERGY IS RELEASED WHEN THE BONDS ARE BROKEN?

While it is common to say that the energy is stored in the bonds, that can be very misleading. The energy is associated with the electric interaction of the individual atoms. The closer or stronger the objects, the *lower* the electric energy associated with the configuration.

IF ENERGY ISN'T RELEASED WHEN BONDS ARE BROKEN, WHY DO SOME CHEMICAL REACTIONS RELEASE ENERGY?

In a chemical reaction, some bonds are broken and some bonds are formed. Whether energy is released or absorbed depends on whether the old bonds are weaker or stronger than the new bonds. If the old bonds are weaker, less energy is needed to break them than is released when the new bonds are formed and the reaction is exothermic.

[xvii] Technically, this is due to *intermolecular* forces (i.e., forces between molecules) rather than *intramolecular* forces (i.e., forces between atoms that make up the molecules). However, the forces are electrical in both cases.

Terminology introduced

Avogadro's number	Endothermic	Ionization energy
Bond dissociation energy	Exergonic	Mole
Chemical energy	Exothermic	Products
Combustion	Ion	Reactants
Endergonic	Ionization	

Additional problems

Problem 8.1: Each molecule of ethane (C_2H_6) has one C–C bond and six C–H bonds. The total energy associated with the seven bonds is 2829 kJ/mol. Calculate the energy released or absorbed when ethane is burned via:
$2C_2H_6 + 7O_2 \rightarrow 4CO_2 + 6H_2O$

Problem 8.2: The bond dissociation energy of a C–H bond in methane (CH_4) is 416 kJ/mol (i.e., 416 kJ for a mole of bonds). What is the bond dissociation energy associated with a single C–H bond?

9. Nuclear Reactions

Puzzle #9: What is the difference between fission and fusion, and why do nuclear reactors only use fission and not fusion?

Introduction

Just as energy is "released" when a rock falls toward Earth or an electron "bonds" with a positive ion, energy is also released when two or more **nucleons** (protons and neutrons) combine to form a nucleus (recall that nucleons are the protons and neutrons that make up the nucleus). And to calculate the amount of energy released when nuclei form, it turns out that we can use the same process as we used for calculating the amount of energy released (or absorbed) when molecules form.

The only difference is that the *nuclear* energy decreases as the nucleons come together, whereas it is the *gravitational* and *electric* energies that decrease in the other examples. That means there are a couple of differences in the language and units we use, which will be discussed in this chapter.

9.1 Chemical vs. nuclear reactions

Before getting to the terms and units used with nuclear reactions, we first need to discuss the difference between chemical and nuclear reactions.

Are fission and fusion chemical reactions?

No. In a *chemical* reaction, the nucleus of each atom remains exactly the same – it is just the bonds between them that change.[i] In a nuclear reaction, the nuclei themselves change (i.e., break apart or combine).

[i] In other words, if carbon atoms make up part of the reactants, carbon atoms must make up part of the products as well.

How does a nucleus change?

It is typically much harder to change a nucleus than to change a molecule. This is because it is much harder to break the nuclear bonds between the nucleons than the chemical bonds between atoms.

Recall from chapter 3 that each nucleus consists of a certain number of protons and neutrons (e.g., a carbon atom has 6 protons and 6 neutrons in its nucleus). When protons and neutrons are very close together, as in the nucleus, there is a nuclear force of attraction between them. This force is stronger than the electric force of repulsion between the protons, which is why the nucleus stays together.

✍ Remember that once we get separation distances much larger than the radius of the nucleus, the electric force overwhelms the nuclear force and we can ignore the nuclear force.

If it is so hard to break a nuclear bond, how does a nuclear reaction release energy?

Just like in a chemical reaction, nuclear reactions tend to involve a *rearrangement*, where bonds are broken and new bonds form, and energy is released when the new bonds are stronger than the old bonds.

Unlike a chemical reaction, however, the bonds are not between atoms but rather the bonds between the nucleons in the nucleus. In a chemical reaction, the nuclei don't change, meaning that the elements involved are the same before and after a chemical reaction. In a nuclear reaction, on the other hand, the nuclei (and thus the elements) change.

⁝ Energy is released in a nuclear reaction if the nuclear bonds in the new nuclei are stronger than the ones in the old nuclei.

There are two basic ways this can occur.

- A single nucleus, made up of a bunch of nucleons, can split into two, each with a portion of the original bunch of nucleons. This is called nuclear fission.
- Two separate nuclei, each representing a different element, can combine to form one nucleus. This is called nuclear fusion.

In each case, energy can be absorbed or released (in the form of thermal and light energy) depending on whether the bonds in the new nucleus or nuclei are weaker or stronger than the bonds in the old nucleus or nuclei. Recall that it takes energy to break a bond, whereas energy is released when a bond forms. So if the new bonds are stronger than the old bonds, more energy is released when those new bonds form than is required to break the old bonds.

✍ The energy released during a nuclear reaction is mostly in the form of thermal and light energy. We shouldn't refer to the thermal and light energy as nuclear energy, since technically the nuclear energy is the potential energy associated with the nuclear attraction force.

✓ *Check Point 9.1:During fusion, a deuterium nucleus (one proton and one neutron) combines with tritium (one proton and two neutrons) to form helium (two protons and two neutrons) and an extra neutron. Energy is released in such a reaction. In which are the nuclear bonds stronger: the reactants (deuterium and tritium) or the products (helium and neutron)? Explain.*

9.2 The electron-volt

⁝ The electron-volt represents a tiny amount of energy compared to the joule.

Before we get into the details of fission and fusion, we should first get familiar with the units involved.

As pointed out in the last chapter, the amount of energy released when two atoms bond is pretty small. As such, for chemical reactions, we tend to look at the energy released per *mole* of interactions. In comparison, for nuclear reactions, the convention is to identify the amount of energy released for each nucleus that forms, rather than a mole of them. Since the amount of energy per nucleus is very small, we tend to use a different unit of energy, called the **electron-volt** (abbreviated as eV), rather than the joule. One eV is equal to 1.6×10^{-19} joules, which is a tiny amount of joules.

Why is it called an electron-volt?

It is called an electron-volt because it represents the kinetic energy obtained by a single electron when it experiences a voltage difference of 1 volt (as in a one-volt battery). Voltage will be discussed in part D. Rather than go into it now, all you need to know at this point is that the electron-volt is a unit of energy and a very small amount of energy.

Is an electron-volt equal to the charge on an electron?

The *number* is the same but the *units* are different. The charge on an electron is 1.6×10^{-19} C. The number is the same because of how the electron-volt is

defined.[ii] If you can remember the charge on an electron, you can remember the energy associated with an electron-volt.

To convert joules to electron-volts, then, simply divide by 1.6×10^{-19} J/eV. For example, in section 8.2, it was mentioned that the ionization energy tends to range from about 8×10^{-19} J to 4×10^{-18} J. Dividing each by 1.6×10^{-19} J/eV, we can express the range as 5 eV to 25 eV. A typical bond dissociation energy, on the other hand, ranges from 1.25 eV to 10.0 eV (see section 8.3 for corresponding values in joules).

✓ *Check Point 9.2: Is 1 joule equal to a large number of electron-volts or a tiny fraction of an electron-volt?*

9.3 Binding energy

⁝ The binding energy refers to the amount of energy released when nucleons come together in a nucleus (or required to break them apart).

Just as energy is released when bonds form between atoms during chemical reactions, energy is also released when bonds form between nucleons during nuclear reactions. The terminology is different, however. Whereas the amount of energy released when *chemical* bonds form is called the bond dissociation energy, the amount of energy released when *nuclear* bonds form is called the **binding energy** value.

For example, suppose it takes X amount of energy to break a chemical bond. We'd say that the amount X is the *bond dissociation energy* of that particular chemical bond. We have to provide an amount of energy equal to X to break that bond, and the chemical potential energy would go up by the amount X when the bond is broken. Conversely, when the bond forms, an amount of energy equal to X would be released (to the environment), and the chemical potential energy would go down by the amount X when the bond forms.

Now suppose it takes X amount of energy to break a *nuclear* bond. We'd say that the amount X is the *binding energy* value of that particular nuclear bond. We have to provide an amount of energy equal to X to break that bond, and the nuclear potential energy would go up by the amount X when

[ii] In a similar way, there are 12 months in a year, 12 hours on a clock, 12 inches in a foot, and 12 eggs in a dozen. Just because the number is the same doesn't mean that a month is equal to an hour or an inch or an egg.

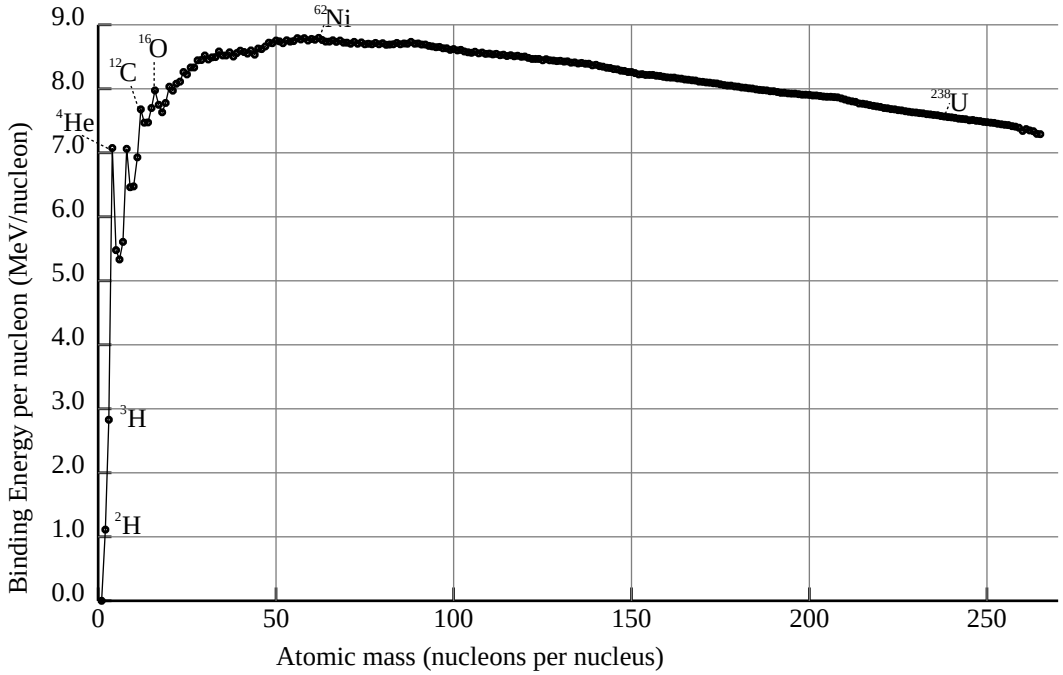


Figure 9.1: Binding energy, in MeV, per nucleon for each element (from G. Audi and A.H. Wapstra Experimental Mass Data, Nucl. Phys. A565, 1, 1993).

the bond is broken. Conversely, when the bond forms, an amount of energy equal to X would be released (to the environment), and the nuclear potential energy would go down by the amount X when the bond forms.

DOES IT TAKE THE SAME AMOUNT OF ENERGY TO BREAK A NUCLEAR BOND AS A CHEMICAL BOND?

No. Actually, it is quite different. Figure 9.1 shows the binding energy value per nucleon for each element.

✓ *Check Point 9.3: (a) According to Figure 9.1, which element has the higher binding energy per nucleon: helium (*^{4}He*; four nucleons) or deuterium (*^{2}H*; two nucleons)?*
(b) Hydrogen (^{1}H*; one nucleon) is not listed in figure 9.1. What do you think its binding energy is and why?*

There are two things I want you to notice about the values provided in Figure 9.1.

The first is that the binding energies are given in units of Mega-electron-volts (MeV). Mega is the metric prefix for a million (10^6), so one MeV is equal to a million electron-volts.

✍ Since one electron-volt is equal to 1.6×10^{-19} J, one MeV is equal to 1.6×10^{-13} J.

As seen in the graph, a typical binding energy is about 7 MeV per nucleon. That means 7 MeV of energy is typically released when a nucleon, like a proton, bonds with the nucleus. Conversely, it typically takes about 7 MeV to remove a nucleon from the nucleus.

In comparison, it was shown in the previous section that a typical ionization energy is between 5 and 25 eV, which is much smaller than 7 MeV. This means it is much, much harder to extract a proton from the nucleus than it is to extract an electron from the atom (almost a million times harder). A typical bond dissociation energy is even smaller, so it is also much, much harder to break a nuclear bond than break a molecular bond.

The second thing you should notice about the graph is that it provides the average binding energy *per nucleon*.

There can be many nucleons in a nucleus. To find the binding energy associated with the entire nucleus, you need to multiply the value in Figure 9.1 by the number of nucleons in the nucleus.

For example, according to the figure, the binding energy of helium is 7.1 MeV/nucleon. Since there are four nucleons in a helium nucleus (2 protons and 2 neutrons), that means the total binding energy associated with helium is 28.4 MeV (i.e., four times 7.1 MeV).

In other words, 28.4 MeV of energy is released when the helium nucleus forms from the combination of the four nucleons. Conversely, it would take 28.4 MeV of energy to break apart the helium nucleus (i.e., separate all four nucleons).

✓ *Check Point 9.4: Use figure 9.1 to estimate how much energy it would take to break apart the entire nucleus of uranium-238 (i.e., 238 nucleons).*

WHY ARE THE BINDING ENERGIES LARGER AT MID NUCLEON NUMBERS AND SMALLER AT LOW AND HIGH NUCLEON NUMBERS?

Elements in the middle (like iron and nickel) have the highest binding energy per nucleon because the nuclear bonds within mid nucleon nuclei are, on average, stronger (i.e., it takes more energy to take apart those nuclei).

WHY WOULD THE MIDDLE ELEMENTS HAVE A HIGHER BINDING ENERGY PER NUCLEON THAN LIGHTER ELEMENTS?

Because the nuclear force is so strong, adding additional nucleons increases the "binding force".[iii]

WHY THEN WOULD THE BINDING ENERGY PER NUCLEON DECREASE FOR HEAVIER ELEMENTS?

Once a certain nucleus size is reached, additional nucleons tend to *decrease* the average binding energy per nucleon because the additional nucleons are only weakly attracted to the nucleus. Recall that the nuclear force acts over a shorter range than the electric force. If we add a neutron/proton pair to a large nucleus, only those nucleons nearby feel the effect of the pair's nuclear force. On the other hand, all of the protons feel the repulsion due to the electric force. As the nucleus gets bigger, then, the nuclear force (being short ranged) is less able to keep the nucleus together against the electric force. Since there is less force keeping the nucleons together, it is easier to separate the particles, leading to a lower binding energy per nucleon.[iv]

✍ What I have been referring to as the nuclear energy is actually the total of the nuclear and electric energies, as the both nuclear and electric forces are involved in the interactions among the nucleons in the nucleus.

✓ *Check Point 9.5: Is it easier to extract a proton from a uranium nucleus or a nickel nucleus? What is it about Figure 9.1 that suggests this is the case?*

[iii] Such a pattern works for the most part. Occasional "peaks" in the binding energy curve represent particularly strong configurations like helium, carbon and oxygen.

[iv] Keep in mind that Figure 9.1 shows the average binding energy *per nucleon*. To calculate the total binding energy associated with a nucleus, one must multiply the binding energy per nucleon with the number of nucleons present. Consequently, the total binding energy of heavy nuclei still tends to be greater than the total binding energy of light nuclei.

9.4 Applications

What use is knowing the binding energy per nucleon?

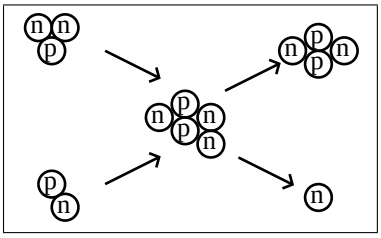
In order for a nuclear reaction to release energy, the bonds in the nuclei beforehand must be weaker (i.e., require less total energy to break apart) than the bonds in the nuclei afterwards. To determine whether the reaction releases energy, then, we simply need to compare the total binding energy before and after. If the total binding energy after is greater, energy is released. Otherwise, energy is needed.

When the binding energy associated with the fragments is greater than the binding energy associated with the original nucleus, breaking the nucleus into the fragments *releases* energy. As mentioned earlier, such a process is called **fission**.

Conversely, when the binding energy associated with the fragments is *less* than the binding energy associated with the original nucleus, then *combining* the fragments into a single nucleus releases energy. Such a process is called **fusion**.

These two processes are described in the next sections.

9.4.1 Fusion



In a *fusion* reaction, two nuclei are fused together to form one nucleus. An example is shown in the illustration to the right. In this example, two nuclei (one of three nucleons and one with two nucleons) fuse together to form one nucleus (of five nucleons).

For fusion to release energy, two smaller nuclei are combined to form a larger nucleus that has stronger bonds than the original two nuclei.

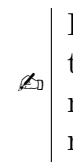
As shown in the illustration, this particular result is somewhat unstable so it quickly rearranges into a 4-nucleon nucleus and a lone neutron.

Still, this example serves to illustrate how energy can be released due to the fusion of two nuclei. Indeed, we can use Figure 9.1 to determine how *much* energy is released.

In this case, the two initial nuclei are isotopes of hydrogen. The three-nucleon nucleus is a tritium nucleus ($^3_1$H) and the two-nucleon nucleus is a deuterium

nucleus ($^2_1$H). The combine and form a helium nucleus ($^4_2$He) with one extra neutron left over.

Just as we did for chemical reactions, to find the energy released, we have to compare the energy associated with the nuclei before and after the reaction. We get the values from Figure 9.1.



I know it is difficult to get precise values from the figure. For the examples that follow, I will use the values that the figure was based upon, which means they will be more specific than what you can get simply from reading the graph.

Before the reaction, we have a 2-nucleon nucleus at 1.15 MeV/nucleon, for a total of 2.3 MeV, and a 3-nucleon nucleus at 2.86 MeV/nucleon, for a total of 8.58 MeV. The sum of these together, 10.9 MeV, represents how much energy it takes to take apart the two nuclei.

After the reaction, we have a 4-nucleon nucleus at 7.08 MeV/nucleon, for a total of 28.32 MeV. This represents the energy that is released upon forming that nucleus.

The difference, about 17.4 MeV, represents the energy that is released in the reaction.

What about the binding energy of the lone neutron at the end?

There is no binding energy associated with that neutron, as it is not "binded" with another nucleon. Remember that the binding energy indicates how much energy is needed to break the nucleon from the others.

Why not just combine two deuterium nuclei ($^2_1$H)? That way there are less nuclear bonds to break and we still form the same number of bonds as before when helium forms.

The extra neutron seems like it is superfluous and unnecessary but it actually serves a purpose – it carries away the energy to the environment. In other words, combining deuterium and tritium results in two objects – a helium nucleus and a neutron – and those two objects can fly away from each other.[v]

Is this how we get energy in nuclear power plants?

[v]The extra neutron is lighter than the helium nucleus and so carries more of the kinetic energy, just like a falling rock vs. Earth (rock gains the kinetic energy).

No, fusion is not *currently* used as an energy source because it requires a lot of energy to overcome the repulsive electric force and get the two nuclei to fuse together. Remember that each nucleus is positively-charged (being made up of positive protons and neutral neutrons). Only once we can overcome the electric repulsion, will energy be released (as the attractive nuclear force takes over).[vi]

At the present time, scientists have not yet figured out a reliable and energy-efficient way of getting the two nuclei close enough together to fuse.

How do we know nuclear fusion is possible?

Nuclear fusion has been accomplished under very limited instances but a lot of energy was needed to do so. We also know that nuclear fusion occurs in the sun, where the gravitational force is large enough and the temperature is high enough to bring the nuclei sufficiently close together.

✓ *Check Point 9.6: As indicated above, the energy released during the fusion of a single helium nucleus is 17.4 MeV. For a mole of helium nuclei, that would be* 1.05×10^{25} *MeV (multiply by Avogadro's number), which is equal to* 1.68×10^{9} *kJ. How does this compare to the energy released when a mole of methane undergoes combustion in a chemical reaction (around 800 kJ; see page 137)?*

9.4.2 Fission

The example with fusion used isotopes of hydrogen, which is a very light element. Will fusion of heavier elements also release energy?

Not really.

The greater the number of protons in the nuclei involved, the greater the repulsion and the harder it is to get them close enough together for the nuclear force to take over. That decreases the productivity of the fusion process. Indeed, above some nucleon number, it actually takes more energy to bring them together than we get out when the nuclear force takes over.

[vi] This is analogous to the activation energy in chemical reactions.

To illustrate this, consider the following reaction:

$$^{92}_{36}\text{Kr} + ^{141}_{56}\text{Ba} + 2\text{n} \rightarrow ^{235}_{92}\text{U}$$

In this case, we take a 92-nucleon nucleus (krypton) and fuse it with a 141-nucleon nucleus (barium) to create a 235-nucleon nucleus (uranium). We also have to include two extra neutrons to make the total nucleon number work out.

Just as we did before, we use Figure 9.1 to determine the energy associated with the nuclei before and after the reaction.

Before the reaction, we have a 92-nucleon nucleus at 8.51 MeV/nucleon, for a total of 782.92 MeV, and a 141-nucleon nucleus at 8.33 MeV/nucleon, for a total of 1174.53 MeV. The sum of these together, 1957.45 MeV, represents how much energy it takes to take apart the two nuclei.

After the reaction, we have a 235-nucleon nucleus at 7.57 MeV/nucleon, for a total of 1778.95 MeV. This represents the energy that is released upon forming that nucleus.

Whereas before we had more energy released than was supplied, in this case we have the reverse. The difference in this case, about 178.5 MeV, represents the energy that is *absorbed* in the reaction.

So, fusion of heavy elements won't work to release energy.

On the other hand, if this process *absorbs* energy then the reverse process should *release* energy. In other words, reversing the process,

$$^{235}_{92}\text{U} + 1\text{n} \rightarrow ^{92}_{36}\text{Kr} + ^{141}_{56}\text{Ba} + 3\text{n}$$

should be accompanied by a release of energy. Indeed, the process described here should release 178.5 MeV by the same calculation we just did.

Energy is released because the products (Kr and Ba) have stronger nuclear bonds than the reactants (U). This process is called **fission**.

For fission to release energy, a large nucleus is split into two smaller fragments that each have stronger nuclear bonds than the original nucleus.

✍ The actual energy released in a fission reaction is more than this, because the daughter fragments aren't stable and thus subsequent reactions produce additional energy.

Why is there an extra neutron with the reactants?

The neutron is there to "help" the uranium break apart. Firing a neutron at a uranium atom is sort of like striking a match against a match box to start it.[vii] Regardless, individual neutrons before and after the reaction do not contribute to the binding energy.

WHY IS ENERGY RELEASED BY FUSION IN ONE CASE (FUSION OF HELIUM) BUT BY FISSION IN THE OTHER CASE (FISSION OF URANIUM)?

This is because we are dealing with both repulsive and attractive forces. With repulsive only, energy is released when the objects separate. With attractive only, energy is released when the objects come together. With both acting, as with the nucleus, it depends on the situation. With lots of protons, as with uranium, the repulsive nature leads to energy being released when it breaks apart into two pieces. With few protons, as with hydrogen and helium, the attractiove nature leads to energy being released when they combine.

The key point in both cases is that energy is released whenever the prior nuclei have weaker nuclear bonds (i.e., have a lower binding energy) than the resulting nuclei.

According to figure 9.1, nuclei around the middle (like nickel) have the highest binding energies (per nucleon). Consequently, energy can be released either by taking a very heavy nucleus and breaking it into two (as with fission) or it can happen by taking two very light nuclei and combining them together (fusion). Either way, it is the *difference* in binding energies that tells us how much energy we get out (per nucleus).

✓ *Check Point 9.7: In the fission reaction described above, the U-235 nucleus is split into Ba-141 (56 protons and 141 total nucleons) and Kr-92 (36 protons and 92 total nucleons). There are two less neutrons in the products than in the original uranium nucleus, which means there are fewer nuclear bonds in the products. Why does this fission reaction still release energy when there are less nuclear bonds in the product?*

[vii] The uranium-235 nucleus is particularly likely to "capture" a neutron (what physicists call a large *cross-section*. The nucleus then briefly becomes a uranium-236 nucleus. Normally, such a nucleus wouldn't be so unstable that it spontaneously decays but since it has the added energy that the neutron had (before being captured) it is unstable enough that it undergoes spontaneous fission.

Summary

This chapter examined how the nuclear energy of a system depends on the number of nucleons in the nucleus.

The main points of this chapter are as follows:

- The electron-volt represents a tiny amount of energy compared to the joule.
- Energy is released in a nuclear reaction if the nuclear bonds in the new nuclei are stronger than the ones in the old nuclei.
- The binding energy refers to the amount of energy required to break apart the nucleons in a nucleus (or released when the come together).
- For fission to release energy, a large nucleus is split into two smaller fragments that each have stronger nuclear bonds than the original nucleus.
- For fusion to release energy, two smaller nuclei are combined to form a larger nucleus that has stronger bonds than the original two nuclei.

You should now be able to do the following:

- Describe how energy is released or observed during nuclear reactions.
- Predict how much energy needs to be provided to break apart or form a nucleus.
- Given the binding energies associated with the various nuclei involved in a nuclear reaction, predict how much energy is released or absorbed during the reaction.

Frequently asked questions

Are there any electrons in a nucleus?

As far as we can tell, the nucleus only consists of tightly-packed positive and neutral charges. The electric charge is limited to a "cloud" of electrons surrounding the nucleus.[viii]

[viii] The nucleus is thought to be about 10^{-14} m in diameter. The atom, on the other hand, is thought to be about 10^{-10} m in diameter. If we scaled the nucleus to be the size of a marble (about 10^{-2} m in diameter), the atom would be the size of a football field (about 10^{2} m in diameter). In other words, the size of the nucleus is very tiny compared to the size of the electron "cloud" that surrounds it.

Does the nucleus change during a chemical reaction?

No. In a chemical reaction, the number of protons and neutrons in the nucleus does not change. The number of electrons can change during a chemical reaction but the electrons are outside the nucleus.

In a *nuclear* reaction, however, the nucleus changes. In other words, a particular nucleus may lose or gain some protons and/or neutrons.

Why does it take energy to break the nuclear bonds – don't the protons repel one another?

Yes, the protons repel one another. This is due to the electric force. However, there is also an attractive force between the protons. The attractive force between the protons is called the *nuclear* force.

As mentioned in section 3.1, the reason why we normally ignore the nuclear force between protons is because the nuclear force (attraction) is stronger than the electric force (repulsion) only at very, very small distances. Once we get separation distances much larger than the radius of the nucleus, the electric force overwhelms the nuclear force and we can ignore the nuclear force.

In addition, there is no electric force of repulsion between the neutrons. There is only a nuclear force of attraction.

For these reasons, it takes energy to break the bonds in the nucleus.

Why do the binding energy values seem smaller than the bond dissociation energy values associated with chemical bonds?

That is because the binding energy in Figure 9.1 is per *bond* whereas the bond dissociation energy in Table 8.1 is per *mole of bonds*.

In addition, we use electron-volts for binding energy vs. kilojoules for bond dissociation energy.

To properly compare, you need to use the same units.

For example, when we convert a typical chemical bond dissociation energy to electron-volts we get values between 1 and 8 eV per bond (see page 148 for the conversion). The bond dissociation energy for the O=O bond in O_2 is equal to 495 kJ per mole of bonds and 5.14 eV per individual.

In comparison, the nuclear binding energy is typically between 5 and 8 MeV per nucleon. This means that the energy associated with a typical nuclear

bond is about a million times greater than the energy associated with a typical chemical bond. Even though a nuclear bond is the result of competing forces (an attractive nuclear force and a repulsive electric force), the nuclear bond, on average, is stronger.

DOES THE EQUATION $E = mc^2$ HAVE ANYTHING TO DO WITH NUCLEAR REACTIONS?

The expression has to do with the fact that energy and mass are related. When energy is released in a reaction, the mass of the various constituent parts is actually a little less than what it was before. In chemical reactions, the amount of energy involved is small compared to the masses involved. However, in nuclear reactions we are dealing with energies that are more significant. The difference in mass, before and after a nuclear reaction, is actually measurable. The expression $E = mc^2$ relates the difference in mass, called the **mass deficit** or **mass defect**, to the energy that is released.

Terminology introduced

Binding energy	Mass defect
Electron-volt	Mass deficit
Fission	Nucleons
Fusion	

Abbreviations introduced

Quantity	**SI unit**
Speed of light (c)	meter per second (m/s)

Quantity	**non-SI unit**
Energy (E)	electron volt (eV)

Additional problems

Problem 9.1: Which nucleus has the higher total binding energy: helium or uranium? Explain your choice.

Problem 9.2: (a) What is the binding energy of helium in MeV/nucleon?
(b) Convert this into units of kJ/mol. A mole is equal to 6.022×10^{23} nuclei.
(c) How does the binding energy of helium (in kJ/mol) compare to a typical bond dissociation energy (in kJ/mol; see chapter 8)? Is your result consistent with the discussion in section 9.1 (about how it is typically much harder to change a nucleus than to change a molecule)?

Problem 9.3: The nucleus of uranium-238 contains 92 protons. Let's assume that, somehow, the nucleus is divided into two separate nuclei: Ba-141 (56 protons and 141 total nucleons; binding energy of 8.33 MeV/nucleon) and Kr-92 (36 protons and 92 total nucleons; binding energy of 8.51 MeV/nucleon).
(a) How much energy is gained by this fission process?
(b) Convert your answer to Joules.

Problem 9.4: Show how the energy released during the fusion of a single helium nucleus is equal to 1.68×10^9 kJ/mol.

Problem 9.5: Nuclear fission releases a great deal of energy to the environment. In nuclear fission, a large uranium nucleus is broken into two fragments. The fragments will both be positively-charged, separated by a small amount (e.g., 10^{-14} m apart, about the size of a nucleus). As they move apart, indicate for each of the following whether the indicated type of energy increases, decreases or stays the same:
(a) gravitational energy
(b) electric energy
(c) nuclear energy

Problem 9.6: Based on your answers to the previous problem, which type of energy (of the three listed) changes the most? Explain.

Part C

Current

10. The Flow of Charge

Puzzle #10: In the puzzle for chapter 2, it was mentioned that a balloon, after rubbing it with hair or fabric, will attract neutral pieces of paper and the paper will stick to the balloon. The rubbed balloon will also attract tiny pieces of aluminum foil but, after touching the balloon, the foil then jumps away.[i] Why does the foil act differently than the paper?

Introduction

To explain the phenomenon described in the puzzle, we need to explore how charged particles seek an "equilibrium" in a material. When we place charged particles, like electrons, on a material (perhaps by rubbing with another material), sometimes the electrons just sit there and sometimes they seem free to move throughout the material.

In this chapter we'll distinguish between the two types of materials and expand our electric model to explain how charge moves within a material.

10.1 Insulators vs. conductors

When an object has a net charge, meaning that it has more positive than negative or more negative than positive, we say that it is "charged". In other words, a charged object is an object that has a charge imbalance. When we are changing an object from having a charge balance (neutral) to a charge imbalance (positive or negative), we say we are **charging** the object.

[i] It is easier to see if you hang a thin strip of aluminum foil from something and then bring the balloon near the strip of foil.

One way to charge an object is to rub it with another material, as discussed in chapter 2. In this chapter, we'll discuss two other ways of charging an object.

✍ When we give an object a charge imbalance, we may say we are giving it a charge but we aren't "charging" the object the same way we might "charge" a battery or cell phone. It is unfortunate that we use the same term for both situations, but charging a battery doesn't give the battery a charge imbalance. The battery remains neutral.

When we charge an object, as with rubbing with another material, we are essentially transferring electrons to or from the object. In this chapter we want to know what happens to those electrons and why.

Before embarking on this investigation, though, I want to point out that the reason I'm focusing on the *electrons*, rather than *protons*, is because the electrons are much lighter and less tightly bound to the nucleus than the protons. This makes them much easier to move around. For that reason, I'll assume that it is the electrons that will move (if they are able), not the protons.

Next, we need to recognize that the movement of the electrons in a material depends on the material itself. In some materials, the electrons are somewhat fixed to the atoms, just as the protons are. In other materials, however, some electrons are free to wander around the material, as though they were unattached to the atoms.

⁝ Insulators are materials in which electrons are not free to move.

The former (electrons fixed to atoms) are called insulators. In an **insulator**, electrons are not free to move through the object, and the added electrons will just stay where they are placed.

⁝ Conductors are materials in which electrons are free to move.

Contrast this with something like a **metal** rod, which is a conductor. A **conductor** is a material in which there are electrons that are free to wander and spread out through the material. Metals are conductors because they have outer electrons far from the nucleus, which allows the electrons to shift easily from atom to atom.

One way to illustrate the difference is to hang two thin strips from a ruler, one strip of paper and one strip of aluminum foil, and then bring a negatively charged balloon (after rubbing it with hair or a sweater) near the two strips. You'll find that the foil is attracted more than the paper, even though both are neutral.

As explained in chapter 2, both neutral objects will be attracted to the negatively charged balloon because the electrons in the neutral objects are pushed away from the balloon and, being further, there is less of a repulsion between the electrons (in the neutral object) and the balloon than the attraction between the protons (in the neutral object) and the balloon.

However, the foil is attracted more because the electrons in the paper are "tied" to the individual atoms and can only move to the far side of each atom, whereas the electrons in the foil are free to move through the foil to the "far" side of the foil. The foil, with its electrons being further away and thus experiencing less of a repulsive force as a result, experiences a greater attraction to the balloon.

Is air a conductor or is it an insulator?

Air is an insulator. While a "free" electron, separate from any air molecules, may be free to move through the air, the electrons in the air molecules themselves are *not* free to move independently of the air molecules. More is said on this in section 10.3.2.

✍ In a solid insulator, like plastic, even "extra" electrons are not free to move through the material. Instead, they just sit where they've been deposited. Apparently this has something to do with what is called the band structure of the material.

In a conductor, are all of the electrons free to move or just some?

Only some electrons are free to move. In general, in a solid material, the atoms are fixed. However, in a conductor there are some "free" electrons (far from the nucleus) that are free to move.[ii]

✓ *Check Point 10.1: Is air considered to be a conductor or an insulator? What about metals?*

[ii] Metals are conductive because they have loosely attached electrons in their outermost shells. Being further from the nucleus, there is less electric force attracting the negative electrons to the positive nucleus.

10.2 The process of charging a conductor

In chapter 2, it was explained that one can give an object a charge imbalance by physically "scraping" electrons on or off the object, as we would when rubbing a balloon with fur or fabric. Let's now examine two other ways of charging an object.

The two methods we'll examine will only work with conductors. Fortunately, there are plenty of materials that are conductors (even ourselves), so it is not much of a restriction.

The two methods are called charging by **contact** and charging by **induction** (i.e., non-contact). As you will soon see, in each method we "coax" the electrons to move onto (or off of) the conductor rather than "grab" the charge (as we would do when rubbing).

10.2.1 Charging by contact

The simplest way to charge an initially neutral conductor is to *touch* the conductor with a second conductor that is charged. We call this "charging by contact."

For example, suppose we have two conductors, A and B, as indicated in the left side of Figure 10.1. Being conductors, the electrons are free to move through each object.

Conductor A has a net negative charge (excess of electrons), as indicated by the ten "minus signs" ($-$). Conductor B is neutral (equals amounts of positive and negative), as indicated by the lack of either "plus signs" ($+$) or "minus signs" ($-$).

If we then touch conductor B and conductor A (see right side of figure), some of the electrons in conductor A will flow into conductor B. This is because the electrons in conductor A repel one another and the neutral conductor B provides a "safe" area to escape into.

How many electrons will flow to conductor B?

As electrons flow into conductor B, conductor B becomes negative, just like conductor A. The flow stops when the electrons are equally repelled by each conductor. At that point, both conductors have negative charge.

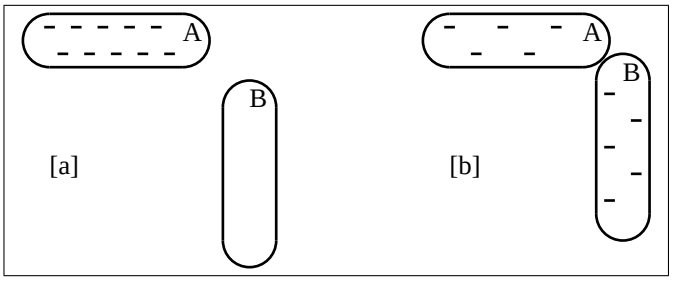


Figure 10.1: [a] Two conductors, indicated as A and B, where A has negative charge and B is neutral. [b] The same two conductors after they touch.

Notice how conductor B was initially neutral and, by touching it with conductor A, conductor B became negative. This is what we mean by charging a conductor "by contact."

Indeed, this is what happens with the aluminum foil in the puzzle. The aluminum foil initially is attracted to the charged balloon, just like the paper, but once the foil touches the balloon some electrons flow into the aluminum foil, making the foil negative as well. At that point, the aluminum is repelled by the balloon rather than attracted.

✍ Since the balloon is an insulator, we can assume that the only electrons that transfer are the extra electrons at the location where the foil touches the balloon.

What is a real-life example of charging by contact?

During the winter, when it is particularly dry, you might find that you are more likely to get shocked when you touch a doorknob or something metal.

The phenomenon is a result of two charging mechanisms. The first involves the process of charging by rubbing, which was examined in chapter 2. When you walk along a carpet, you can pick up electrons just as a balloon does by rubbing it with fur.

The second charging mechanism involves the process of charging by contact. When you touch a neutral conductor like a piece of metal, the electrons will flow from you into the previously-neutral conductor. As the electrons flow from your finger to the metal, you feel as a spark.

✍ As we will learn later, when charge flows through an object, that object can get warm (as the electrons bump into atoms). That is why you feel a sharp pain when you get a shock. The electrons are flowing through a small area on your skin and it gets hot for a brief moment.

Of course, both mechanisms occur in summer as well as winter. However, in winter there is a lot less water vapor in the air.[iii] Drier air is a better insulator, allowing more charge to build up on materials before discharging through the spark.

✓ *Check Point 10.2: Suppose conductor B was just as negatively charged as conductor A is positively charged. What would happen when the two touch?*

10.2.2 Ground

To properly do the transfer described above, we need to make sure that the electrons that moved into conductor B are not free to move *out of* conductor B into *another* object. For example, if conductor B is attached to a third conductor, electrons would be free to move from conductor B into that third conductor as they are repelled by the other extra electrons in conductor B.

To prevent such a flow between conductor B and a third object, we can surround conductor B with an insulator. One example of a good insulator is air (others include the rubber of a balloon).

What about sticking the conductor in the ground?

The ground is actually a relatively better conductor than insulator. Not only that, the ground is part of the larger Earth, which means that we can consider the ground to be one big reservoir of potentially free electrons, so many in fact that the removal or addition of a few here or there doesn't really cause it to lose its neutrality. In other words, it is so big that the density of the excess charge in the reservoir remains essentially zero (i.e., it remains

[iii] This is because the colder it is the less water vapor that can exist without condensing. Since the air inside buildings comes from the air outside the buildings, the water vapor content of the air inside the buildings matches the water vapor content outside the buildings unless there is a dehumidifier (in summer, lowering the water vapor content inside) or a humidifier (in winter, raising the water vapor content inside).

neutral) even with the addition or removal of charge. It is similar to water in the ocean in the sense that the height of the water (sea level) remains the same even if we take a bucket of water out of the ocean or we add a bucket of water to the ocean.

Consequently, if conductor B in Figure 10.1 was somehow connected to the ground, electrons would be free to flow between conductor B and the ground. What this means is that, if conductor was still connected to conductor A, then all of the extra electrons trying to flee conductor A (because of the other extra electrons on conductor A), would likewise flee conductor B and head into ground. The end result would be that both conductors would end up being neutral, just like ground.

In this sense, it is similar to a small pond at sea level that is attached to the ocean, also at sea level. If we add some water to the pond, there will be a flow of water from the pond to the ocean. Conversely, if we remove water from the pond, there will be a flow of water to the pond from the ocean.

For this reason, we say that a conductor that is connected to the ground is considered to be *grounded*[iv]. In many cases, any excess or deficiency of charge on the conductor will naturally flow into or out of the **ground** to ensure that the conductor remains (or becomes) electrically neutral. A notable exception to this is discussed in the next section.

⁝ An object is grounded when it is connected to the ground, which acts like a neutral conductor, regardless of how much positive or negative charge it provided to it.

✍ As mentioned in a previous note, when charge flows through an object, that object can get warm (as the electrons bump into atoms). For that reason, we typically need to be careful about touching charged objects while we are grounded, since the charge will flow through us to ground. To prevent this, we can either insulate ourselves from ground or we can ground the object we want to touch, which will keep the object from accumulating excess charge.

✓ *Check Point 10.3: The third plug in an outlet is connected to ground. It is there so that the container of some electronic device (like a computer) can be connected to ground. Why might that be a useful thing?*

[iv] Sometimes the word "ground" is used for any object that has characteristics similar to Earth ground. For example, the chassis of a car is sometimes referred to ground, as when we connect the negative end of a jumper cable to the chassis of the dead battery's car. In that case, though, the chassis is connected to the negative end of the (dead) battery (as well as the starter), not the Earth ground, as the rubber in the tires are insulators.

10.2.3 Charging by induction

As mentioned at the end of the previous section, we can consider ground to be always neutral. Consequently, for most cases, any object connected to ground will also be neutral.

Notice that I write "in most cases." A notable exception is when we are actually are using ground to charge an object by *induction*[v]. We'll consider two ways of charging an object by induction, and it is via the second way that an object can be charged even when connected to ground.

According to *The Random House Dictionary* (1980 by Random House, Inc.), one definition of the word *induce* is to "lead by persuasion." In a similar way, one can produce charge on an object without actually touching the object but rather by "persuading" the charged particles to enter or leave the object. We call this "charging by induction."

For example, consider the neutral conductor illustrated in part [a] of Figure 10.2. When we bring a negatively-charged object close to, *but not touching*, the neutral conductor (see part [b]), the electrons in the neutral conductor are repelled, producing a charge separation (see part [c]). We say that the neutral conductor has been **polarized** by induction. This is similar to what happens to the aluminum foil when the charged balloon comes near, but doesn't touch, the foil.

Even though charge can't flow between the negative conductor and the neutral conductor (since there is air between the two conductors, and air is an insulator), the electrons in the neutral conductor still feel the force from the excess charge on the negative conductor. In other words, the insulator doesn't "block" the force, just the flow of charge.[vi]

We can then create two charged objects by breaking the polarized rod in two, as in part [d], without the original two rods touching at any point in the process.

---

[v] As is mentioned shortly after this, we are using the word "induction" in much the same way one would use it for placing someone into elected office. In both cases, the word "induct" has to do with leading into, and comes from the Latin *ducere*, which means "to lead."

[vi] The electric force, like the gravitational force, is a non-contact force. Two charged objects do not need to *touch* each other for there to be an electric force exerted on each.

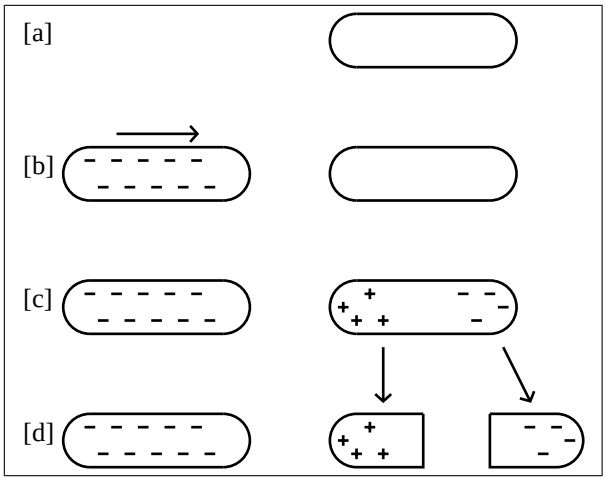


Figure 10.2: [a] A neutral conductor. [b] A negatively-charged conductor is brought toward the neutral conductor. [c] A charge separation is induced in the neutral conductor. [d] The neutral conductor is split in two, producing a positive object and a negative object.

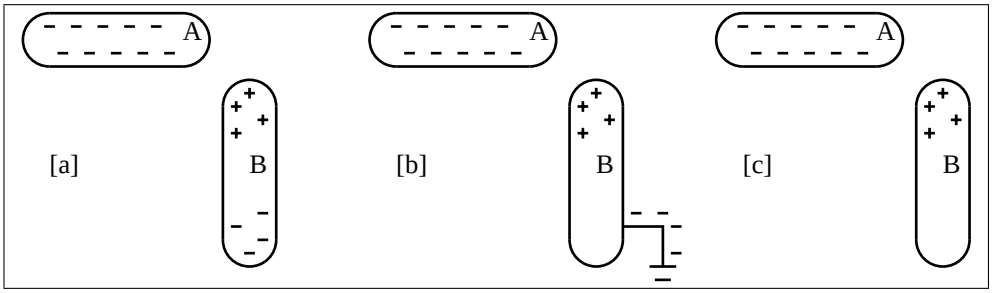


Figure 10.3: [a] A negatively-charged conductor (A) near, but not touching, a neutral conductor (B). [b] Conductor B attached to ground (the negative signs along the attached wire indicate the flow of electrons into the ground). [c] Conductor B removed from ground, with conductor A still nearby.

A second way of charging by induction is to use the ground. This process is illustrated in Figure 10.3. It starts with the same polarization by induction process illustrated in Figure 10.2 in that a charged object is brought near a neutral conductor, inducing a charge separation within the neutral conductor (see part [a]).

However, rather than breaking the polarized conductor B in half, we instead connect one side of conductor B to ground (see part [b]). The electrons

on conductor B that are repelled by conductor A can then leave conductor B into ground. The loss of electrons leaves conductor B with net positive charge, which it retains when the ground is disconnected (see part [c]).

To better understand what is going on, consider the following analogy. Prior to connecting to ground, object A is like a teacher and object B is like a classroom full of students who hate the teacher. Since the students can't leave the classroom, they move to the back of the room, leaving a bunch of empty seats in front.[vii] The students are only able to leave the room when the door is opened. In a similar way, when conductor B is grounded (i.e., it is connected to ground) the electrons (those that are free to move) are able to leave the conductor, leaving conductor B with a positive charge.

✍ Notice how grounding an object doesn't necessary force the entire object to become neutral, just the side connected to ground. As shown in Figure 10.3, it depends on what else is around the object.

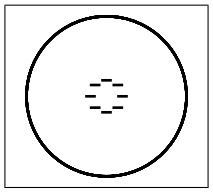
✓ *Check Point 10.4: (a) Describe how one can "charge by induction" to change a neutral object into one that has negative charge.*[viii]
(b) During which part of the process in Figure 10.3 is the conductor grounded?

10.3 Distribution of charge on a conductor

Now that we've answered the puzzle, let's examine in more detail where excess electrons go within the conductor. Where do they move *to*? In particular, given the ability to wander, how will excess charge distribute itself on a conductor?

The answer is that the excess charge will be spread around the surface.

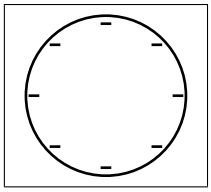
To understand what that means, suppose we place a bunch of electrons on a conducting sphere at a particular location, as indicated in the figure to the right, where the "minus signs" (−) represent an excess of negative charge.



[vii] Hmmm... some of my classes have students who only sit in the back...
[viii] As opposed to positive charge, as in Figure 10.3.

✍ A neutral region is indicated by the lack of "minus signs" since a neutral region has just as many electrons as protons.

In a conductor the excess electrons will not remain concentrated at that spot. Instead, they will separate and spread themselves uniformly around the surface, as shown in the figure to the right.



WHY DO THE EXCESS ELECTRONS UNIFORMLY DISTRIBUTE THEMSELVES AROUND THE SURFACE OF THE CONDUCTOR?

Charges of the same sign repel, just like a room full of people who hate each other. People who hate each other, if placed in the same room, tend to migrate (move) to opposite walls of the room rather than come together in the center. In the same way, extra electrons migrate toward the outer surfaces.

WHEN YOU SAY "OUTER SURFACES" DOES THAT MEAN THE CHARGE IS "OUTSIDE" THE OBJECT?

⁘ Excess charge on a conductor will distribute itself around the surface of the conductor.

No. The (excess) electrons are still inside the object. They are just very near the "surface". Consider the room full of people who hate each other. Everyone will remain "inside the room" – they will just be along the walls.

HOW CAN WE TELL THAT THE EXCESS ELECTRONS ARE ON THE SURFACE RATHER THAN CONCENTRATED AT THE CENTER?

From the outside, we can't tell.

In fact, from the outside, the electric force on another object is the same whether the electrons are spread uniformly throughout the surface or concentrated in the center.[ix]

However, from the inside, we can tell the difference. This is explored in section 10.5.

✓ *Check Point 10.5: What does it mean to say that charge resides on the surface of a conductor (rather than in a conductor)?*

[ix] For this reason, we can use the law of electric force (Coulomb's law) for non-point objects, just by assuming all the charge is concentrated at the center of the object.

10.3.1 Corners

❢ Charge will congregate more at corners and sharp tips.

Another important property of excess electrons on a conductor is that they tend to prefer corners or pointed parts of the conductor.

In terms of the analogy used before, if you put a bunch of people in a room and they all hate each other, they will migrate first to the corners of the room. The same holds true for the excess electrons. If you have an excess of electrons, their density will be greatest at any pointed edges of the conductor.

I CAN SEE WHY THE CORNERS ARE PREFERRED, BUT IF THE CORNERS ARE MORE CROWDED, WHY WOULD THE ELECTRONS CONTINUE TO MIGRATE THERE?

This is where the analogy between electrons and people break down. People tend to focus only on those who are closest to them. Electrons, on the other hand, tend to "feel" the effect of electrons both near and far. While there may be more electrons at a corner, there is still a benefit to being far from the rest of the electrons.

✍ For objects with no corners (like a flat plate or a sphere), the surface charge density should be the same everywhere on the surface.

✓ *Check Point 10.6: For a sphere, the excess charge is distributed evenly around the surface of the sphere. If the object was a cube instead, how would the distribution be different?*

10.3.2 The insulating nature of a vacuum

I CAN SEE THAT THE PEOPLE IN THE ROOM ARE CONSTRAINED TO STAY IN THE ROOM BECAUSE OF THE WALLS. WHAT ABOUT THE EXCESS ELECTRONS IN A CONDUCTOR? WHAT PREVENTS THEM FROM SIMPLY JUMPING OFF THE CONDUCTOR INTO THE SURROUNDING AIR?

To answer this, let's assume that the conductor has an excess number of electrons and there is *nothing* surrounding the conductor, not even air. Such a region of space, devoid of all material, is called a **vacuum**.

Even in this case, with *nothing* surrounding the conductor, electrons still won't leave the conductor, despite being repelled by the other excess electrons.

The reason why the electrons stay on the conductor has to do with the fact that materials are made up of protons as well as electrons.

Those protons exert a great attractive force on the electrons, so great that even the excess electrons don't leave the material. So, even though an individual electron may be free to move throughout the material (as in a conductor), it requires a lot of force to actually extract the electron *from* the material.

⁉ Air, like a vacuum, acts like an insulator, preventing excess charge from leaving the conductor.

✍ This is similar to why you and I don't fall apart even though we are neutral. If any molecule attempts to "leave," there would be a large electric force exerted on it keeping it "attached" to our body.

A similar thing happens even with the pieces of aluminum foil and the charged balloon in the puzzle. As discussed before, when the foil touches the balloon it picks up some of the excess electrons that were residing on the balloon and the foil jumps away as it is then negatively charged like the balloon. However, even in this case, if you again bring the balloon close enough to the now-charged foil, at some point the pieces of foil will *still* attract to the balloon, despite some initial repulsion.

Even though there is a net negative charge on the foil, just as there is a net negative charge on the balloon, the excess electrons can be pushed so far to the other side of the foil that there is still a net attraction between the foil and the balloon, and the foil will "stick" to the balloon. Just like excess electrons on a conductor, the negative-charged foil is still "stuck" on the negatively-charged balloon.

✍ The attraction only happens at close distances. At further distances the two negatively charged objects repel, just as you would expect. In addition, this is easily observable in this case because the demonstration uses a conductor (the aluminum foil) and an insulator (the balloon). It would be harder to show this with two conductors.

For this reason, we consider a vacuum to be the perfect[x] insulator. Even though a free electron, if it happens to be within the vacuum, is able to

[x] Actually, it is perfect only in the sense that it inhibits the transfer of electrons into the vacuum. This also inhibits the transfer of energy via conduction and convection, so a vacuum is also considered an ideal *thermal* insulator (which is why thermos bottles tend to have an evacuated space between an outer and inner casing). However, a vacuum

freely move, there are no atoms in the vacuum for an electron to "hop onto" and thus is inhibited from entering the vacuum in the first place (for reasons discussed above). For the same reason, air is considered to be a very good insulator.[xi]

✓ *Check Point 10.7: Is a vacuum (i.e., empty space) an insulator or a conductor? Explain.*

10.4 The movement of positive charge

As mentioned before, I've assumed that it is the electrons that are moving, since protons are heavier than the electrons and attached to an even heavier nucleus. However, instead of referring to how the *electrons* were moving, I could have instead described how the *negative charge* was flowing.

For example, when a negatively-charged conductor touches a neutral conductor, I mentioned how excess electrons will move from the negatively-charged conductor to the neutral conductor, making them both negatively charged. However, I could have instead described it follows: where some of the negative charge flows from the negatively-charged conductor to the neutral conductor, making them both negatively charged.

Notice how I refer to the negative charge as though it is a fluid, even though it really is just a property of the electrons, which are the actual particles that are moving.

Furthermore, rather than referring to how the *negative* charge flows *from* the negatively-charged conductor to the neutral conductor, I could have instead stated that the *positive* charge flows from the neutral conductor *to* the negatively-charged conductor.

⁝ We can consider positive charge as moving in a direction opposite the motion of the electrons.

The equivalence is easier to see when you recognize that a positive charge is equivalent to a "lack" of negative charge. For example, consider a conductor

doesn't prevent transfer of energy via radiation, and some materials have been found that are overall better thermal insulators than a vacuum due to their ability to also block radiation.

[xi] It is also a good thermal insulator, although only if the air isn't flowing or mixing, as that allows the transfer of energy through convection.

where there is a deficiency of electrons. For that material, there will be "holes" in the material where there is excess positive charge. As the electrons move in one direction, then, the "holes" (positive charge areas) move in the opposite direction.

As an analogy, consider a classroom with fixed chairs. The chairs are like protons whereas the students, who are free to wander throughout the room, are electrons. Locations where seats are taken by students will be "neutral." Empty seats will look like "extra" protons.

Now consider what happens if we let the students move to another seat. When a student switches chairs, an "empty" chair now appears where the student had been. In other words, the chair hasn't moved but the "emptiness" has been transferred, and that transfer is opposite the motion of the students. In the same way, the protons haven't moved but the positive charge has been transferred, and that transfer is opposite the motion of the electrons.[xii]

For example, consider the case where a positively-charged conductor A touches a neutral conductor B. When they touch, some of the electrons in the neutral conductor B will flow into conductor A. This is because the electrons in conductor B are attracted to the positively-charged conductor A.

As electrons flow out of conductor B, conductor B becomes positive, just like conductor A. At some point, the electrons are equally attracted to each conductor and the flow stops. At that point, both conductors will have positive charge.

Notice how the end result is equivalent to what would happen if we simply described the flow of positive charge. In other words, the positive charge flowed out of the positively-charged conductor A and into conductor B. Also notice how much easier it is to say it that way.

Keep in mind that I'm not saying that positive *particles* moved. They didn't. But the positive *charge* did, just as the negative charge flowed the other way.

In terms of our analogy with students and chairs, conductors A and B are like two classrooms, where conductor A has 10 empty chairs and conductor

[xii] Another potentially useful analogy is to think of positive charge as student loan debt. In the same way that conductor A obtains positive charge from conductor B when electrons are transferred from conductor A to conductor B, student A can take on the debt of student B by transferring money from student A to student B.

B has every chair filled with a student. When the two conductors touch, it is like opening a door between the classrooms. The students, who hate each other, would much rather be spread out and thus a portion of them move to the empty classroom. The result is that there are now five empty seats in each classroom. The empty chairs in conductor A decreased and the empty chairs in conductor B increased, but no chairs physically moved from conductor A to conductor B.

As another illustration of the language, let's revisit the distribution of charge on a conductor, as in section 10.3, but this time consider what happens when we have a *deficiency* of electrons rather than a *surplus*.

We know from before that a surplus of electrons will distribute themselves around the surface of the conductor. In terms of charge, this is equivalent to excess negative *charge* distributing itself around the surface of the conductor.

If, instead, the conductor has a *deficiency* of electrons then the conductor has a net positive charge. And, just like with excess negative charge, the excess positive charge distributes itself around the surface of the conductor. In other words, the second picture in section 10.3 would be the same except with "plus signs" spread around the surface instead of "minus signs".

✓ *Check Point 10.8: Suppose students are negative charges and empty chairs are positive charges. As students enter a previously-empty classroom, can we say that the empty chairs are leaving the classroom?*

10.5 Faraday cages

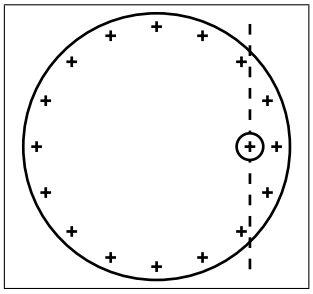
To further illustrate how we can describe the distribution of excess *charge* instead of particles, let's consider a very interesting property of conductors that we haven't yet considered.

In the previous section, it was mentioned that excess charge tends to distribute itself on the outside surface of a conductor (and in a uniform way if there are no corners). This was actually first noted by Benjamin Franklin in 1755, when he observed that a neutral object did not pick up any charge when placed inside a charged, conducting container. Based on this observa-

tion, he concluded that the inside of the charged container must be neutral, with all of the excess charge on the outside surface.

Not only does the inside object not pick up any charge but it turns out that it experiences no electric force, even if the inside object is itself charged. It doesn't matter where the inside object happens to be in inside the container, it doesn't matter if the container is charged, and it doesn't matter what is outside of the container.

To see why, let's first consider the situation illustrated in the figure to the right, where a positive particle, indicated by the $\oplus$, is placed inside a larger conducting sphere that is itself positive.

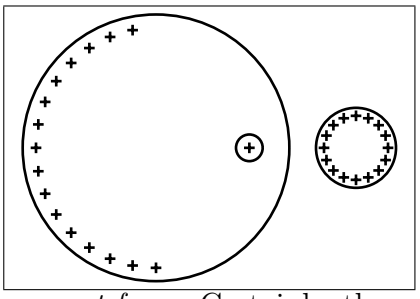


The charges to the right of the dashed line force the particle to the left. Conversely, the charges to the left of the dashed line force the particle to the right.

Although the charges to the right are closer and thus exert a greater effect (because they are closer) there are many more charges to the left. In other words, although individually the force from each charge on the further side is weaker, the total force exerted by the charges to the left is the same as the total force exerted by the charges to the right, just opposite in direction.

The end result is that the electric force on the small charged particle inside is zero, no matter where it is placed.

Now let's consider what happens if we bring a second charged object nearby but outside the container. This is illustrated in the figure to the right. Due to that outside charge on the right, the charges on the sphere are repelled to the left (effect exaggerated for illustrative purposes).



It turns out the inside particle still experiences *no net force*. Certainly, there is a leftward electric force on the inside particle due to the outside particle (since like charges repel). However, the positive charge on the surface of the container now exerts a rightward electric force on the small inside particle, since the positive charges on the surface have migrated over to the left. The end result is that the net force on the inside particle is zero.

What if the outside particle is negatively charged?

It doesn't even matter what objects are *outside* the container. There will still be no force on the inside object, because the charges on the container would naturally distribute themselves in such a matter that would negate the effect on the inside particle.

What if the container is neutral?

Again, it doesn't matter, because the charges on the container would naturally distribute themselves in such a matter that would negate the effect on the inside particle.

✍ It is like the conducting shell "blocks" the outside charges from influencing the inside when, in reality, it is the charges in the shell itself that produce an opposite force, negating the effect of the outside charges. In comparison, the force on the container is not necessarily zero (and usually not zero), nor is the force on the outside object necessarily zero.

Because this is true for any conducting container and for any object inside the container, it is as though the surrounding container acts like a "cage," protecting the inside object from experiencing an electric force. For this reason, a conducting container is called a **Faraday cage**.[xiii]

⁉ Objects placed inside a conducting container will not experience an electric force on it from charges on or outside the container.

✍ This is one reason why being inside a car during a thunderstorm is somewhat safe. The charge will flow through the outer surface of the car to the ground, and not pass through the occupants on the inside.[xiv]

✓ *Check Point 10.9: An excess charge of +2 μC is placed on a hollow conducting sphere of radius 2 cm (that is insulated from other objects). For each question, provide a rationale in support of your answer.*
(a) What is the electric force on a proton placed at the center of the sphere?
(b) What is the electric force on a proton placed 0.2 cm away from the center of the sphere (but still within the hollowed-out center)?
(c) Suppose an additional object, with charge -2 μC is near (but still outside) the hollow conducting sphere. Do your answers to (a) or (b) depend on whether this additional object is present or not? Why or why not?

[xiii] Named after Michael Faraday who, like Franklin, observed this effect (albeit about 80 years after Franklin).

[xiv] Given the huge amount of charge in a lightning bolt, the rubber in the tires will break down and allow current to flow to the ground, even though rubber is normally an insulator.

Summary

This chapter examined how charge distributes itself in a conductor.

The main points of this chapter are as follows:

- Insulators are materials in which charged particles are not free to move.
- Conductors are materials in which charged particles are free to move.
- We can consider positive charge as moving in a direction opposite the motion of the electrons.
- Excess charge on a conductor will distribute itself around the surface of the conductor.
- Objects placed inside a conducting container will not experience an electric force on it from charges on or outside the container.
- Charge will congregate more at corners and sharp tips.
- Air, like a vacuum, acts like an insulator, preventing excess charge from leaving the conductor.
- An object is grounded when it is connected to the ground, which acts like a neutral conductor, regardless of how much positive or negative charge it provided to it.

By now you should be able to predict how excess charge will move though a conductor, and use that to explain how to charge a conductor by contact and induction.

Frequently asked questions

What is a conductor?

A conductor is an object in which some portion of the protons and/or electrons are free to flow within the object (without being constrained by being attached to an atom).

What is different about a conductor that allows the electrons free to wander?

Conductors tend to be made out of materials that have a relatively high atomic weight. Those atoms have a lot of electrons. Most of the electrons are attached to the atom (by being attracted to the very positive nucleus). However, the outer electrons are less tightly bound and can be free to roam.

Does an electrical insulator have anything to do with the thermal insulation we put in the attic?

Yes. A good electrical insulator tends to be a good thermal insulator as well. If we heat up an insulator on one end, the energy is transferred through a slow process whereby each atom knocks into another atom, like a classroom full of students passing notes, from one student to the next. A conductor is used for pots because the energy in one place can be "carried" to another place by the more mobile electrons in the material. It is like a classroom full of students *throwing* the notes around the room.

If a conductor is already charged, what happens to new charge we add to a point inside the conductor?

The new charge will migrate to the outside surface of the conductor, despite the presence of the charge that is already present on the surface. To understand why, let's consider the case where the conductor is negatively charged. If the new charge is made up of a single electron then, from that electron's point of view, it is surrounded by negative charge and repelled by them all. That repulsion, being on all sides, cancels out and the electron stays where it is. However, if the charge we place in the center is made up of lots and lots of electrons then, from the point of view of those electrons, all they see are the other *new* electrons and therefore spread apart. The negative charge that was already there on the outside surface have no influence on what happens to the inside charge since their impact cancels out.

Does a conductor (i.e., Faraday cage) shield inside objects from any outside electric forces?

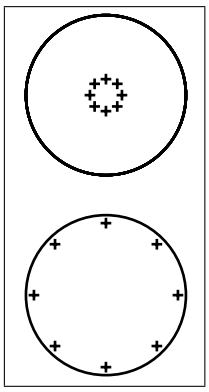
There are still forces and those forces still adhere to the law of electric force (Coulomb's law). However, the forces cancel such that the *net* force on the inside object is zero.

Terminology introduced

Conductor	Ground	Metal
Contact	Induction	Polarized
Faraday cage	Insulator	Vacuum

Additional problems

Problem 10.1: Two metal spheres, identical in shape, size and material, but with different amounts of charge (same sign), repel one another with a force of 2×10^5 N when they are 4.5 cm apart. They are then brought together, touch briefly and then are returned to their original separation distance. Do they now attract, repel or neither? If neither, why not? If they attract or repel, is the force less than 2×10^5 N, equal to 2×10^5 N or greater than 2×10^5 N?



Problem 10.2: On the right are two spheres with excess positive charge.
(a) Of the two spheres, which sphere more accurately represents the distribution of positive charge if the spheres were conductors?
(b) For the sphere chosen in (a), what caused the region of excess charge: free protons moving to that region or free electrons moving away from that region?

Problem 10.3: An insulated conducting sphere of radius 2 cm has a charge of $+2$ μC. Given that a charged object inside a charged conducting sphere experiences no electric force (see section 10.5), what is the electric field at the center of the sphere due to the distribution of charge around the surface?

11. Electric Current

Puzzle #11: What do we need to get a light bulb to light and why do we need those things?

Introduction

In chapter 10, we examined how charge distributes itself in response to forces from other charges. The focus was on the distribution of the charges. In this chapter, we're going to focus on the *movement* of the charges. It is moving charges that everyone associates with **electricity**. The path taken by the charges is called a **circuit**.

The context for exploring the movement of charges will be the process behind turning on a light. One of the simplest circuits is one in which a small bulb is lit by connecting the bulb to a **battery** via some wire. So the answer to the first part of the puzzle is just some wire and a battery.

To understand *how* the wire and battery must be connected and *why* it works to light the bulb, though, we need to extend our conceptual model of positive and negative charge so that it can be used to explain batteries, wires and bulbs.

11.1 Batteries

Although we cannot see what is happening inside a battery, a couple of observations of the battery can tell us that certain things are not happening.

First, in order to light a bulb, we find that we need to connect *both* sides of the battery to the bulb. In other words, the bulb won't light if only a single side of the battery is connected to the bulb.

This tells us that the battery isn't supplying light to the wire, which then flows through the wire until it "escapes" through the bulb. In other words,

the battery isn't like a spigot that provides water to a hose, which then escapes through something like a lawn sprinkler. If that was the case, connecting only a single side of the battery to the circuit would be sufficient. It isn't.

Second, we find that the mass of the battery doesn't change while the battery is being used. This tells us that the battery *can't* be providing any material to the bulb. If this was the case, the battery would get lighter and lighter as it is used. That doesn't happen.[i]

Third, we find that the ends of the battery are essentially neutral, in that they do not attract neutral pieces of paper, unlike a charged balloon (see[ii] chapter 2). In fact, the entire battery, as far as we can tell, is neutral. This tells us that a battery, even a so-called "charged" battery[iii], does not contain an imbalance of positive vs. negative particles and so can't be providing one or the other (or both) to the circuit.

Is the battery an electric dipole, with one end positively charged and the other end negatively charged?

While there is indeed two ends of a battery, which we call the positive and negative **terminals**, any dipole nature of the battery is not noticeable in the sense that neither end acts like a charged object, and neither end can attract pieces of paper. Consequently, as far as we can tell, each battery terminal is neutral, not charged.[iv]

If the battery doesn't provide any material to the circuit and the battery isn't charged, what does it do?

---

[i] A common misconception is that one end of the battery is providing electrons while the other end of the battery provides protons, with the electrons and protons combining in the bulb to produce light. This can't be happening, since the battery doesn't get lighter as it is used.

[ii] As discussed in chapter 2, neutral pieces of paper are attracted to a charged object because the paper becomes polarized (with one end positive and one end negative) when a charged object is nearby. Since the pieces of paper are not attracted to either end of the battery, this tells us that each end of the battery must be neutral.

[iii] When we "charge" a battery, we are providing it with energy, not charge. The use of the term "charge" in this way is unfortunate, as it can be misleading. It would be more accurate to say that we "energize" a battery.

[iv] In reality, there are probably a couple of extra electrons on the negative end, with an equal deficiency of electrons on the positive end. However, as mentioned in the text, it isn't noticeable so we can consider both ends to be neutral.

The battery is essentially a pump, much like a water pump or a fan. A fan makes air flow around a room but it doesn't *provide* the air – the air is *already* there.[v] In a similar way, the battery makes electrons flow through the circuit but it doesn't *provide* the electrons – the electrons are *already* present in the circuit.[vi]

In addition, just as the fan pushes the air out the fan at the same time it pulls the air into the fan, the battery simultaneously sucks electrons into one end of the battery and pushes electrons out the other end.[vii]

Indeed, this is why the two ends (terminals) of a battery are marked "+" and "−". The "−" side is where the battery pushes out electrons and the"+" is where the battery pulls in electrons. In terms of charge (instead of electrons; see section 10.4), we can say that the "+" side is where the battery pushes out positive charge and the "−" side is where the battery pulls in positive charge. The battery remains neutral during its operation because for each electron that leaves the battery another one simultaneously enters.

How does the battery do this?

The "power" for this pumping comes from the chemical reaction that occurs in the battery. However, in order for the chemical reaction to take place, an electron needs to be provided to the battery (via the positive terminal) while, *at the same time*, an electron is removed from the battery (via the negative terminal).

This is why the battery has to be connected on both sides. The circuit provides both a source of electrons (to the battery) and a destination of electrons (from the battery).[viii]

---

[v] In a similar way, the heart doesn't produce blood and the blood isn't used up in the capillaries. Instead, the blood is already there, and the heart just pushes on it.

[vi] As mentioned in chapter 10, protons are much heavier than electrons and are not as free to flow. Consequently, we can consider the protons as remaining fixed in place within the materials that make up the circuit. The materials are conductors, which means that there are some electrons that are free to move and those are the particles that flow through the circuit.

[vii] The heart, in comparison, consists of two sets of chambers, and the dual chamber nature of each set means the input and output alternate, unlike the battery where the input and output of electrons occur simultaneously.

[viii] While there are electrons present on the negative terminal, even without the circuit being connected, it would take too much energy to extract them without the electrons simultaneously being replaced from elsewhere.

Without that connection, the reaction would stop, much like how blood flow stops when the arteries or veins are blocked. This is why each end of the battery has no noticeable charge, as evidenced by the fact that neither end attracts neutral pieces of paper.[ix]

✍ When you "charge" a battery, you aren't creating a charge imbalance (see earlier footnote). Instead, you are running the chemical reaction backwards. As with any chemical reaction, the chemicals remain neutral the entire time you are "charging" the battery. You are just rearranging the atoms inside the battery to a higher potential energy state (that can then be used to power a device, like your phone). More information on chemical reactions and energy was given in chapter 8.

Do all of the free electrons in the circuit start moving simultaneously?

⁝ All of the free electrons throughout the circuit start moving simultaneously.

Yes, as long as they are in a part of the circuit that is connected to both sides of the battery. Since the battery pulls an electron into the battery at the same time it pushes an electron out the battery (on the other side), all free electrons[x] in the circuit must move when the battery is connected to simultaneously make space for the electron that is provided to the circuit (by the battery) and fill in the space left by the electron provided to the battery (by the circuit).

⁝ The electrons flowing through the wire come from the material in the wire itself.

Notice how the electrons flowing through the wire are generally those already present in the wire. It is similar to what happens when someone's heart starts after momentarily stopping. The blood is already present in the circulatory system. The heart simply makes that blood start moving.

What happens when a battery dies?

The power for the process comes from the chemical reaction that occurs inside the battery. Once the reactants have been converted into the products of that reaction, the reaction stops and the battery is no longer able to power the flow of electrons.

[ix] Inside the battery, things are not as simple as this. Certainly, as an electron leaves an atom, it leaves behind a positive ion, which then gains an electron as a new electron comes along. Whether this occurs practically instantaneously or not (during which time there would be a "flow" of positive ions moving opposite a "flow" of electrons), the key concept remains – the battery provides energy to the circuit, not charge or mass.

[x] As mentioned in chapter 10, in most metals there are so many electrons per atom that some electrons are relatively "free" to roam through the metal.

✓ *Check Point 11.1: According to our model, is the battery essentially neutral and, if so, does it remain neutral while it is connected to the circuit? Why or why not?*

11.2 Wires and bulbs

As we know from chapter 10, charge won't flow through an insulator. Thus, we must make sure that we have a complete conducting path through the circuit from one terminal of the battery to the other. In other words, there can't be any "breaks" in the circuit.

This doesn't mean that it has to be *easy* for the electrons to flow. Certainly, the circuit must be made up of conductors, not insulators, but some materials are really good conductors and some are poor conductors. The charge will flow, even with poor conductors, just not as well as with good conductors.

An example of a really good conductor is copper, which is why many wires are made of copper. Another popular choice for wires is aluminum, which is lighter and less expensive than copper. Aluminum isn't as good a conductor as copper but it is still a very good conductor.

Because aluminum and copper wires are such good conductors, we can assume that the electrons can flow through those wires without being impeded by the atoms in the wire. Poorer conductors, on the other hand, will impede the flow of electrons somewhat because the electrons will "bump" into the atoms of the material, making the material warmer in the process.

An example of a poor conductor is the **filament** inside an **incandescent** bulb. The filament is just a thin wire within the bulb casing. It is common to have the filament made out of tungsten, which is not as good a conductor as aluminum. In addition, by being thin, it is harder for the electrons to pass through it, thus warming it up more than if the wire was thicker (for the same current flow). In fact, the filament in an incandescent bulb gets so hot that it glows, which is how the light is generated in an incandescent bulb.[xi]

[xi] For the same current flow, better conductors don't get as hot. This is why the wires

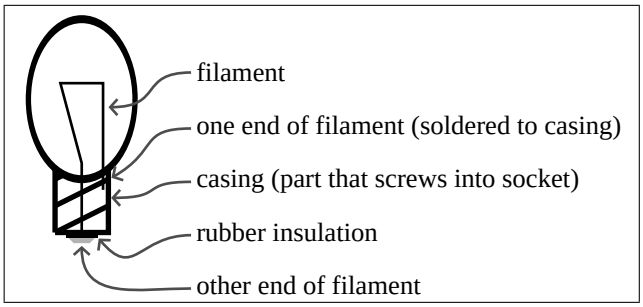


Figure 11.1: A drawing of a light bulb, showing the path of the filament inside the light bulb.

When you turn off the light, you stop the flow of electrons and the filament cools and stops glowing.

What happens to the electrons that pass through the filament?

There is nothing special about the filament except that it is there that the energy of the circuit is dissipated in the form of heat and light. Remember that electrons are not energy. Electrons pass *through* the filament and remain as electrons when they come out.[xii]

You can think of the electrons as animals running blindly through a forest. As the animals bang into the trees, the branches vibrate but the trees remain trees and the animals remain animals (they just bounce off and then resume running through the forest). In this way, the energy of the animals gets transmitted to the trees. In a bulb, the atoms of the filament are like the trees and get energized by the electrons banging into them. The atoms then emit the energy in the form of light and thermal energy.

✓ *Check Point 11.2: According to our model, when the electrons enter the filament in the bulb, do they disappear (and turn into energy) or do they pass through the filament and come out the other end?*

---

leading up to the bulb don't normally glow. However, better conductors allow more current to flow (for the same push) so if you had a circuit with just wire and no bulb then so much current would flow that the wire itself would get hot and glow. This is called a short circuit, which is a fire hazard. This is why we use fuses – to "stop" the current before it gets too high.

[xii] The bulb is constructed in such a way to prevent the electrons from flowing through the outer casing of the bulb rather than through the filament. This is accomplished by including a small ring of insulation (usually black) around the base of the bulb.

11.3 Electric current

You may know that we quantify blood flow by the volume of blood that flows in a given amount of time, not the number of red blood cells that pass by. Similarly, we quantify water flow (as in a river) by the volume of water that flows in a given amount of time, not the number of molecules that pass by. Consequently, it probably wouldn't surprise you to learn that we quantify the flow of charge by how much charge flows in a given amount of time, not the number of electrons that flow by.

In this section, I introduce how we quantify the electric current, which is the rate charge flows. Keep in mind that we'll assume the positive charge, being the heavier nuclei of the atoms, will stay fixed and only the free electrons will flow. Despite the movement of the electrons, the material stays neutral at every point, as just as many electrons flow out of a region as flow into it, and the negative charge of the electrons is balanced by the positive charge of the protons (in the nuclei of the atoms).

11.3.1 Drift velocity

As mentioned before, as the electrons move through the circuit, they repeatedly bump into other particles, stopping and starting again as they move along the wire. The rate at which electrons "drift" along the wire is called the **drift velocity**.

What is a typical drift velocity?

A typical drift velocity is on the order of millimeters per second.

That seems slow. Why does a light come on as soon as you turn on the switch?

Because all of the free electrons in the circuit start moving together.

✓ *Check Point 11.3: When a battery is connected to a particular 1-m long wire, the electrons in the wire are made to travel along the wire at a drift velocity of 0.2 mm/s. How long does it take for an individual electron to pass through the entire wire?*

11.3.2 Definition of current

In circuits, we describe the flow of electrons in terms of the **electric current** rather than the drift velocity. This is because the drift velocity, by itself, does not represent how much charge is flowing.

It is similar to describing the flow of water in a river. To know how much water is flowing by you, you not only need to know how fast the water is flowing but you also need to know how wide and deep the river is. The current takes into account both of those factors.

✍ I will usually write "**current**" in place of "electric current" when it is apparent we are talking about the electric current as opposed to some other type of current.

⁉ Electric current is defined as the rate at which charge flows past a point.

We define the electric current, I, as the amount of charge flowing past a point divided by the time required.[xiii]

$$I_{\text{average}} = \frac{\Delta q}{\Delta t} \tag{11.1}$$

where Δq represents the amount of charge that flows through the wire in the time interval Δt.

Why is I used to represent current?

Apparently it is from the word "intensity" (as in "intensity of the flow").[xiv]

Since a typical drift velocity is so small, does that mean a typical electric current is small also?

It depends on your units. Although the drift velocity is very small, there are lots and lots of free electrons in a typical wire. In copper, for example, the density of free electrons[xv] is about 8.5×10^{28} free electrons per m^3, which means that even for a small drift velocity there can be lots and lots of electrons passing a point in the wire each second.

[xiii] The instantaneous current (i.e., the current at any given instant), is defined as the same ratio but in the limit as Δt goes to zero.

[xiv] The phrase was originally in French, written by French physicist André Marie Ampère.

[xv] The density of free electrons can be estimated from the density of copper (8.92 g/cm^3) and the molar mass of copper (63.54 grams of copper per mole; see page 507). Divide the density by the molar mass to get 0.140 moles of copper per cm^3, and then multiply this by Avogadro's number (6.022×10^{23} atoms/mole) to get the number of atoms per cm^3. I then assumed that there was one free electron per atom.

In fact, this is the reason why we don't define the electric current in terms of the rate at which electrons pass through the wire. That would be like measuring the current of water by counting the number of molecules that pass a point in the river each second. Like counting molecules, counting electrons yields very large numbers. For example, if we plug a 100-Watt light bulb into the wall, and count the number of electrons that pass a point on the wire each second we get about 5×10^{18} electrons/second.

To avoid these huge numbers, we instead express the electric current in terms of coulombs per second (or C/s), which is a *charge* per time. For example, a current of 5×10^{18} electrons/s corresponds to only 0.8 C/s, since each electron only has -1.6×10^{-19} C of charge (see chapter 2).

⁝ Electric current is measured in units of amperes.

This ratio (coulomb per second) is known as an **ampere** (or amp, for short) in honor of the French physicist André Marie Ampère (1775-1836) who examined the properties of electricity. The ampere is abbreviated as A.

Example 11.1: Suppose 2 C of charge flow through a wire in 10 seconds. What is the current?

Answer 11.1: The current is the ratio of charge per time: (2 C)/(10 s) = 0.2 C/s = 0.2 A.

✓ *Check Point 11.4: A copper wire with a cross-sectional area of 0.50 mm^2 (like the wire connecting a study lamp to an outlet) carries a current of 1.50 A (a value similar to that for a study lamp).*
(a) How much charge flows through the wire each second?
(b) How many electrons flow through the wire each second?

11.3.3 Direction of current

In the example, why do you say that positive 2 C of charge flows through the wire? Wouldn't it be more accurate to say that negative 2 C of charge flows through the wire, since electrons carry negative charge?

Perhaps, but as discussed in chapter 10, −2 C flowing one way is the same as +2 C flowing the other way.

Indeed, when circuits were first investigated, they didn't even know which charge was flowing. It turns out that it doesn't make any difference since a positive current moving one way is equivalent to a negative current moving the other way.

⁉ The direction of electric current is defined as being in the direction that positive charge is flowing.

To simplify matters, we say that positive current in the wire moves from the "+" terminal to the "−" terminal. In other words, we'll treat the flow of charge as though only the positive charge was moving (away from the positive terminal of the battery and toward the negative terminal of the battery) and define that as the direction of the electric current.

ISN'T IT ARBITRARY WHICH PARTICLES WE CALL POSITIVE AND WHICH WE CALL NEGATIVE? COULDN'T WE HAVE CALLED THE ELECTRONS POSITIVE INSTEAD OF NEGATIVE?

Yes, we could have. But once the choice was made[xvi] it was simpler to stick with it.

✓ *Check Point 11.5: For each of the following, identify the direction as either (i) from the positive terminal of the battery through the light bulb to the negative terminal of the battery, or (ii) from the negative terminal of the battery through the light bulb to the positive terminal of the battery.*
(a) Which way do electrons flow through the circuit shown in Figure 11.2?
(b) According to convention, which way does the electric current flow through the circuit shown in Figure 11.2?

11.4 Neutrality of circuits

IS THE AMOUNT OF CURRENT FLOWING INTO A BULB THE SAME AS THE AMOUNT OF CURRENT FLOWING OUT OF THE BULB?

Yes. As mentioned in section 11.2, electrons pass *through* the filament and remain as electrons when they come out. Indeed, for every electron that enters the filament, another electron must *simultaneously* exit the filament.

[xvi] Apparently, Benjamin Franklin was the one who decided the direction of current (and thus positive charge).

In addition, remember that in our model the material itself provides the electrons. The battery is not the "source" of the electrons any more than it is the "sink" of electrons. Consequently, all of the electrons throughout the circuit move at the same time.

Since the circuit is neutral initially (equal numbers of protons and electrons), this means it must remain neutral while current flows. As mentioned in section 11.1, the chemical reaction within the battery only occurs if an electron is taken into the battery for every electron it spits out.

We can go further and assume that *every part* of the circuit remains neutral.[xvii] That means that the current flowing into any part of the circuit must equal the current flowing out of that part, not just for the battery and the bulb but also for the wires.

If this were not the case and more electrons went in than came out, that part of the circuit would gain electrons, leading to increased mass and a net negative charge. It does not.

✍ We can show that each part of the circuit is neutral via our balloon test (i.e., by bringing each part of the circuit near some pieces of paper).

WHY CAN'T THE ELECTRONS JUST DISAPPEAR?

While there is nothing in our model that suggests they can't disappear, we are going to adhere to a principle called **conservation of charge**, which means that the total amount of it in the universe remains the same, no matter what happens, and that there is a local path through which charge is transferred. This means the negative charge associated with the electron cannot disappear without it being transferred to some adjacent location.

⁉ In a circuit, we assume every part of the circuit remains neutral, even as charge flows through it. Consequently, the current entering a location must be equal to the current leaving that location.

IS THE CURRENT THROUGH THE BULB FILAMENT THE SAME AS THE CURRENT IN THE WIRE LEADING TO THE BULB?

Yes. Just because the bulb heats up and the wire doesn't, does not mean the current is different. It only means the materials react differently to the amount of current that flows. In this case, the filament impedes the current and grows hot as a result. The wire, on the other hand, lets the charge flow easily through it and thus does not warm up.

IF THE FILAMENT IMPEDES THE CURRENT, SHOULDN'T THE CURRENT BE SLOWER THERE?

[xvii] We will modify this assumption in chapter 17.

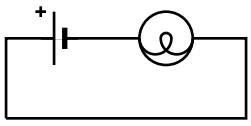


Figure 11.2: A schematic of a bulb connected by wire to a battery.

No. To keep everything neutral, the electrons cannot "bunch up". Otherwise, one part of the circuit would be negative or positive and that would produce a force to quickly even things out again.

It is important to realize that electrons in a wire are not like cars on a highway. Cars can bunch up when the cars ahead are slowing down and, conversely, cars can spread out when the cars ahead are speeding up. Electrons, on the other hand, experience an electric force of repulsion that prevents them from bunching up. If one part of the circuit slows down, the *entire* circuit is affected and the *entire* circuit slows down at the same time. So, the current in the filament cannot be different than the current in the wire leading up to the filament because, if it was, the charge would "bunch" up like cars approaching a construction zone.

✓ *Check Point 11.6: Is the current in the wire leading up to the light bulb any different than the current flowing* through *the light bulb?*

11.4.1 Circuit schematics

We can consider more complicated circuits to help illustrate the idea of circuit neutrality. However, before doing so, I want to introduce some short-hand figures for representing circuit elements. Doing so makes it a lot easier to draw a circuit.

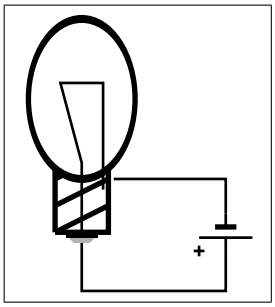
For example, a simple circuit arrangement is shown in Figure 11.2. The bulb is indicated by a circle with a loop inside (to represent the filament). The battery is indicated by two parallel lines with a space between them; the long thin line indicates the "+" side and the short fat line indicates the "−" side. Wires are indicated by thin lines connecting the battery with the bulb.

One important thing to recognize about the circuit schematic is that it doesn't matter how I draw the wires. The only important thing is that

they illustrate which elements are being connected by the wires. So, all we know from the schematic is that one wire connects the bulb to one end of the battery, and the other wire connects the bulb to the other end of the battery.

Another important thing to recognize about the circuit schematic is that it doesn't show *how* the wires are connected to the bulb. We assume that the wires are connected in such a way that the current must pass through the bulb filament.

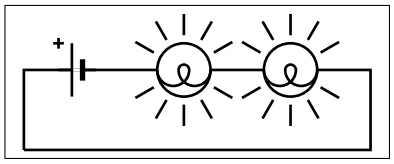
The illustration to the right shows how the wires should actually be connected to the bulb, with one wire connected to the bulb casing and the other connected to the base of the bulb. That way the wires are connected to the two ends of the filament. The base and the side casing are separated by a small rubber insulator, which prevents the current from bypassing the filament.



✓ *Check Point 11.7: In the circuit illustrated in Figure 11.2, the bulb is connected to two wires – one connecting it to the battery's negative terminal and another connecting it to the battery's positive terminal. Is one wire necessarily longer than the other?*

11.4.2 Single paths and split paths

Now that we have a way of illustrating how the circuit elements are connected, consider the schematic to the right, which shows two bulbs connected to a battery.



If we do this with identical bulbs, we find the two bulbs are equally bright, which means the current through each is the same, as the only way identical bulbs could be equally bright is if the current through them is identical.

This finding is consistent with the idea that the current must be the same through each part of the circuit. Notice that it does not matter which bulb is closer to the battery (the left one in this case) or which way the current is flowing (from the right bulb to the left bulb in this case). Both are equally

bright because they are identical bulbs and the current is the same through each.

What if the bulbs were not identical?

The current through each would still be the same. However, if the bulbs were not identical, their brightness could be different. Some bulbs have less resistance (see chapter 14) and thus don't get as hot for the same current. Those bulbs won't be as bright.

If the flow of electrons heat up the filament, wouldn't they also heat up the wires leading to and from the light bulb?

As mentioned earlier, the wires are very good conductors, so fewer collisions occur between the moving electrons and the atoms that make up the wire. In fact, we'll assume that they don't heat up at all.[xviii]

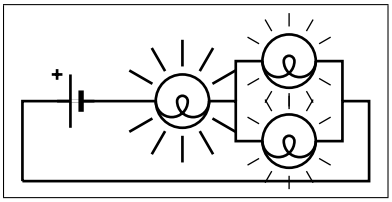
Is the current the same in every part of a circuit then?

It depends on the circuit.

In the circuit we've been considering, the current is the same through each bulb because of the way they are placed in the circuit. In particular, they are placed in such a way that any current that flows through one bulb *has* to flow through the other. We say that elements in a circuit are arranged in a **single path** when they are arranged in this way.[xix]

⁉ The current is the same for two elements in arranged in a single path.

However, let's consider the circuit illustrated in the diagram to the right. This circuit is the same as the previous one but with a third identical bulb placed in a **split path** with one of the bulbs.



We say the two bulbs are create a "split path" because the current flowing through the circuit has to *split*, with some going through the top bulb and

[xviii] Even very good conductors like copper and aluminum can heat up if the current is high enough. Wires are rated by how much current can safely flow through them without heating too much. If too much current flows through them they can get hot enough to cause a fire. To prevent this, houses have circuit breakers that automatically "break" the circuit if too much current is sensed.

[xix] Most physicists refer to this arrangement as having the elements **in series**. I am describing the arrangement as being a "single path" because I think that is clearer.

some going through the bottom bulb.[xx] Since the bulbs are identical, the split is equal, with half flowing through the top bulb and half through the bottom, and those two bulbs are equally bright.

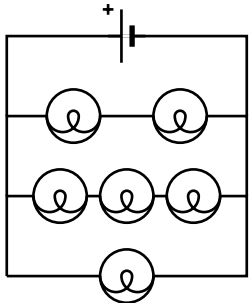
WOULD THE CURRENT STILL BE EQUAL THROUGH NON-IDENTICAL BULBS?

If the top and bottom bulbs are not identical then the current may not split equally. More will go through the bulb that has a lower resistance. Whereas bulbs along a single path *must* have the same current, bulbs in a split path do not.

⁉ The current need not be the same for two bulbs in a split path.

Even with identical bulbs the paths may be non-identical if there are a different *number* of bulbs in each path.

For example, consider the circuit to the right, where a battery is connected to six identical bulbs. There are three paths for the current: the top path contains two identical bulbs along that path, the middle path contains three identical bulbs along that path, and the bottom path contains a single bulb. Which bulb or bulbs are brightest? Which bulb or bulbs are dimmest?



We know that the current must split between each path. In this case, the split isn't even, however. Fewer bulbs is easier to flow through, and thus more current flows through the bottom path, with less current flowing through the middle path.

At the same time, the current through *each* bulb in the middle path is the same as the current through the others in that *same* path, as they are all exist along the same route toward where the splits paths come back together. Thus, each bulb in the middle path is dimmer than the bulbs in the top path or the bulb in the bottom path. And, the bulb in the bottom path is brightest.

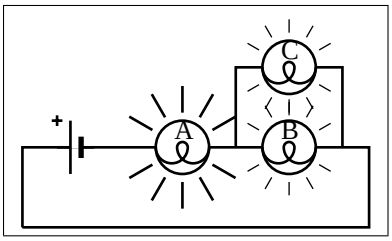
WHAT IF ONE OF THE BULBS BURNS OUT?

If the filament inside a bulb breaks, there can no longer be current through that bulb, it no longer lights and we say that the bulb "burned out." When

[xx]Most physicists refer to this arrangement as having the elements **in parallel**. I am describing the arrangement as being a "split path" because I think that is clearer, especially since the difference between a single path and a split path is based on what the current is doing, not the physical layout of the circuit.

that happens, there is no longer any current through *any* bulb that is along that same route as the burned out bulb. For example, if the middle bulb in the middle path burns out, then all three bulbs in that path go off. The other bulbs (top and bottom paths) are unaffected and remain lit.

As you can see, particularly in the more complicated circuit with six bulbs, some bulbs may lie along the same path with some bulbs but not others. To investigate this in more detail, let's revisit the simpler circuit from before, this time with the bulbs labeled as A, B, and C, as illustrated to the right.



Bulbs B and C are in a split path, since the current has to split between them. If all three bulbs are identical then the current splits evenly between bulbs B and C, with half flowing through bulb B and half through bulb C, which is why they are shown as equally bright. On the other hand, bulbs A and B are not in a single path (with those two bulbs alone) and so the current through A is not the same as the current through B, and is actually brighter than the other two.

Why is bulb A brighter than the other two?

Bulb A is brighter because the currents from bulbs B and C come together, producing a greater current through bulb A. Not only is the current greater through bulb A but it must be equal to the *sum* of the currents through bulbs B and C. Otherwise, each part of the circuit wouldn't remain neutral.

I'll refer to this idea, that the current flowing into a particular location must equal the current flowing out of that location, as the **current rule**.[xxi] It follows from the idea that every part of the circuit remains neutral, even while current flows through it.

Any location where two wires come together or split apart is called a **junction**. Consistent with the current rule, the total current flowing into a junction equals the total current flowing out of the junction.

[xxi] Some people also refer to this as the **Kirchhoff's current rule** or **Kirchhoff's junction rule**. Gustav Robert Kirchhoff (1824-1887) was a German physicist.

✓ *Check Point 11.8: Two* different *light bulbs are placed in a circuit. We find that one light bulb is dimly lit and the other is brightly lit.*
(a) If the bulbs are arranged in along a single path, is it possible that the current through each light bulb is different? Why or why not?
(b) If the bulbs are arranged in a split path, is it possible that the current through each light bulb is different? Why or why not?

11.4.3 Ammeters (measuring current)

Now that you have a sense of how current flows in a circuit, you can understand how we go about measuring current.

We measure current with an **ammeter**, so called because electric current is measured in units of amperes or milliamperes ("A" or "mA" for short).

An ammeter has two sockets. Current is sent in one socket and out the other. The ammeter itself measures how much current goes through the meter. For information about how to set up an ammeter, see the supplemental readings. Here we'll focus on how to place the ammeter in a circuit.

For example, suppose we want to use an ammeter to determine the current going through a light bulb.

Since the ammeter measures how much current flows through the ammeter, we have to arrange the ammeter such that the current through the meter is the *same* as the current through the light bulb.

How can that be arranged?

The answer is to place the ammeter along the *same* path as the bulb.

For example, consider the two circuits drawn in Figure 11.3. In each schematic, I have indicated the ammeter as Ⓐ.[xxii] In which of the two arrangements will the current through the ammeter be the same as the current through the light bulb?

⁉ In circuit diagrams, an ammeter is indicated as an "A" inside a circle.

From before, we know it is the one where the ammeter is along the same path as the bulb (because charge is conserved). Thus, it is that circuit (the one

[xxii] Some people replace the "A" by a "mA" (for milliamperes) or a "μA" (for microamperes), depending on the scale indicated on the ammeter.

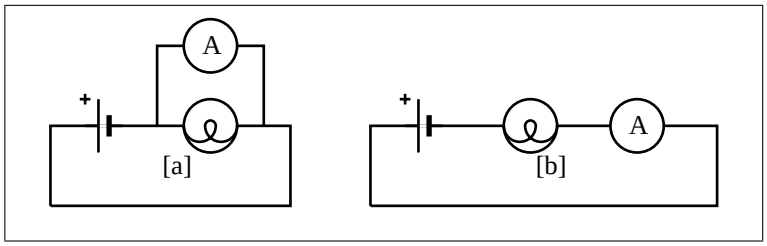


Figure 11.3: Two circuits. (a) Ammeter is connected in a split path with the light bulb. (b) Ammeter is connected along the same path as the light bulb.

on the right in the figure) where the reading on the ammeter will accurately reflect the current through the bulb.

Won't the placement of the ammeter in the circuit change the current, much like the current is less when there are two bulbs along a single path compared to a single bulb?

Unlike a bulb, it is very easy for electrons to flow through an ammeter. Otherwise, using an ammeter in a circuit would change the current that you are trying to measure.

What about the left circuit?

In that circuit, the current through the ammeter can be different than through the light bulb.

Indeed, since it is very easy for electrons to flow through an ammeter, much more current will flow through the ammeter than the light bulb when arranged in that split configuration.

This is why you don't want to place an ammeter and bulb in a split path configuration since a lot of current can flow through ammeter (and circuit) and bypass the bulb. Too much current can damage the meter (or other parts of the circuit) so most meters contain a fuse that melts if too much current flows through the meter, stopping the flow of current (see supplemental readings).

✓ *Check Point 11.9: In which of the circuits in figure 11.3 would the ammeter be measuring the current through the light bulb? Explain your choice.*

Summary

This chapter examined how we quantify and measure the current going through a simple circuit.

The main points of this chapter are as follows:

- The charged particles flowing through the wire come from the material in the wire itself.
- All of the free electrons throughout the circuit start moving simultaneously.
- Electric current is defined as the rate at which charge flows past a point, and is measured in units of amperes.
- The direction of electric current is defined as being in the direction that positive charge is flowing.
- The wires and each element in the circuit remain neutral, even as the charge flows through them. Consequently, the current entering a location must be equal to the current leaving that location.
- The current is the same for two elements in arranged in a single path.
- The current need not be the same for two bulbs in a split path.
- In circuit diagrams, an ammeter is indicated as an "A" inside a circle.

By now you should be able to do the following:

- Construct a simple circuit to light a light bulb and describe why it works.
- Explain why the current is produced practically instantaneously throughout the wire.
- Quantify the electric current in appropriate units.
- Use electric circuit symbols for circuit elements in schematics of circuits.
- Use the current rule (neutrality of circuit) to predict whether the currents at two points are the same or not.
- Use an ammeter to measure current through a wire.

Frequently asked questions

Why do electrons move through the wire and not protons?

Not only are the protons much more massive but they are attached to the

even more massive nucleus and thus are not free to move separately from the atoms.

Does the battery supply electrons?

The electrons that make up the current are *already* in the wire. As the wire's free electrons move through the wire, the battery provides additional ones at one end (the "−" terminal) and receives them at the other (the "+" end). This is necessary to not only keep the circuit as a whole neutral but also to keep the chemical reaction within the battery going.

Is the light the result of electrons combining with the protons within the filament?

No. The electrons and protons do not combine. See page 190.

Does the presence of the charged particles (flowing through the filament) make the filament charged?

No. See section 11.4.

Do electrons flow until the circuit is "full" of electrons?

No. As mentioned in section 11.4, every part the circuit (like the battery and the wires) remains neutral. The electric repulsion between the electrons prevents them from bunching up. You can't put an electron in without an electron coming out the other end at the same time.

When you say "current" do you mean "electric current"?

Yes. Since our focus is on electric current only, I will frequently write "current" rather than "electric current." They mean the same thing.

Terminology introduced

Ammeter	Current	Filament
Ampere	Current rule	Fuse
Battery	Drift velocity	Incandescent
Circuits	Electric current	Single path
Conservation of charge	Electricity	Split path

Abbreviations introduced

Quantity	**SI unit**
Electric current (I)	ampere (A)[xxiii]

Additional problems

Problem 11.1: A copper wire with a cross-sectional area of 0.50 mm^2 (like the wire connecting a study lamp to an outlet) carries a current of 1.50 A (a value similar to that for a study lamp). Calculate the average drift velocity of the electrons traveling through the wire. Hint: the number of free electrons in copper is about 8.5×10^{28} per m^3.

Problem 11.2: A student argues that the circuit should get negatively charged when a battery is connected to it because the battery provides charge to the circuit. Do you agree? If so, why? If not, what is wrong with the student's reasoning?

[xxiii] An ampere (or amp) is equal to a coulomb per second (C/s).

12. Electromagnets

Puzzle #12: When people mention how high power lines or cell phones might cause cancer, they always speak about the electromagnetic field. What is the electromagnetic field?

Introduction

You'll probably run into the term "electromagnetic field" outside of class. For example, when scientists investigate the effect of cell phones or high-voltage power lines on cancer rates, they speak of the electromagnetic field associated with those devices.

The use of this term suggests there is some relationship between electric charge and magnets. Even though we have treated them as distinctly different, it turns out there is something relating the two. Remember how in chapter 4 it was mentioned that a magnet is actually made up of lots and lots of tiny magnets? At the time, it was mentioned that each electron acts like a little magnet. To see how a magnet can be electrically neutral yet have its magnetism be due to electrons, we need to examine the **electromagnet**, which we describe in this chapter.[i]

12.1 Neutrality of the electromagnet

An electromagnet, as far as we are concerned, acts like a magnet in the sense that it has all of the properties associated with regular magnets, including being electrically neutral. The reason it is called an "electro"-magnet is because it is created by sending current through a coil of wire.

⁉ An electromagnet is a coil of wire with current.

[i] We'll just describe their properties, not explain *why* they have the properties they have.

When discussing electromagnets, it is crucial that you recognize that there is nothing special about the wire except for its coiled shape.[ii] Another important feature is that the electromagnet only acts like a magnet when current passes through the wires of the coil. An example of a coil is illustrated on page 210.

It is also crucial to recognize that the coil by itself, as with any wire in a circuit, is electrically *neutral*, whether current flows through it or not.

Apparently, when charge moves in loops or circles, a magnet is created. An electromagnet is just a bunch of those current loops.

✍ Just because there is a link between moving charge and magnets does not mean that magnets are electrically charged. As far as we can tell, electromagnets are electrically neutral, like permanent magnets.

✓ *Check Point 12.1: An electromagnet is created by sending current through a coil of wire.*
(a) Is it necessary to have insulation on the wire? Why or why not?
(b) Is the electromagnet, or any part of the electromagnet, electrically neutral?

12.2 The value of the core

Typically, an electromagnet is created by wrapping the coil of wire around a piece of ferromagnetic metal, like a nail. While that makes it easier to get the wire in the form of a coil, the metal **core** is not necessary to create an electromagnet. However, the ferromagnetic core *does* make the electromagnet stronger. In other words, the coil, by itself, is an electromagnet, and the coil *with* the ferromagnetic core is *also* considered to be an electromagnet.

⁝ A metal core is not necessary but makes the electromagnet stronger.

The reason it makes the electromagnet stronger has to do with the fact that the electromagnet acts to align the little magnets in the ferromagnetic core, making it into a magnet (see section 4.5). This increases the overall magnetic

[ii] To visualize what I mean by a coil, imagine taking a wire and wrapping it around your finger several times, making several loops. That would be a coil of wire. I suppose a ring would be a coil of only one loop.

strength of the electromagnet (even though the current through the wire is the same as before).[iii]

Keep in mind that the current does not pass through the metal core. The current stays in the wires surrounding the metal core. The metal core becomes a magnet because it is in the presence of another magnet – in this case the electromagnet associated with the coil of current.

✍ The "tiny little magnets" inside objects that were discussed in chapter 4 are associated with what we call the "spin" of the electrons. Most electrons are paired with an electron of opposite spin, negating the magnetic properties. Only unpaired electrons can align to produce a magnet.

Similarly, the bar doesn't strengthen the current through the wire either. The bar simply becomes magnetic when it is near another magnet, and since the coil is magnetic (because current flows through it), the bar becomes magnetic also. The two together, the bar (which is now magnetic) and the coil (which is magnetic because of the current), produce a stronger magnet overall, like having two magnets instead of one.

✓ *Check Point 12.2: According to our model, why does the strength of an electromagnet depend on whether a ferromagnetic bar is used as a "core"?*

12.3 Describing an electromagnet

Since electromagnets act just like magnets, we can use the same language for electromagnets that we used for magnets (e.g., north and south poles).

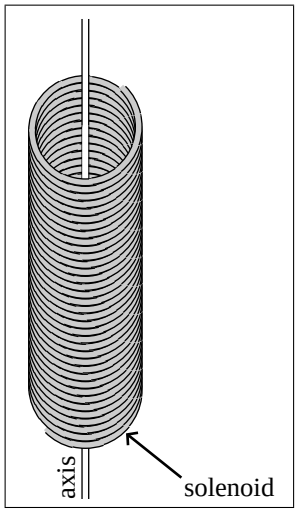
However, since electromagnets are constructed out of wire, it is useful to agree on some terms to describe the structure of an electromagnet so that when I refer to various parts of the electromagnet there is no confusion regarding which part of the electromagnet I am referring to.

[iii] I suppose one can think of the coil with a ferromagnetic core as *two* magnets – the electromagnet associated with the coil and the temporary magnet associated with the ferromagnetic core.

As mentioned in the previous section, an electromagnet is just a coil of wire. The general term for a coil of wire is a **solenoid**.[iv]

The location of the north and south poles depends on how the current circulates within the solenoid. To adequately describe the relationship, we first need to have a language for describing the orientation of the solenoid.

To describe how the solenoid/electromagnet is oriented, I'll refer to its **axis**. The axis is the imaginary line that runs down the "inside" of the electromagnet (see figure to right).



When current is sent through the solenoid, it acts as a magnet with an orientation along the axis of the electromagnet, with the north pole at one end and the south pole at the other. For example, for the electromagnet in the figure the south pole could be at the bottom with north at the top, or the south pole could be at the top with south at the bottom.

WHAT DETERMINES WHICH END IS THE NORTH AND WHICH IS SOUTH?

As mentioned earlier, it depends on which way current circulates around the solenoid axis.

⁉ The north pole is at one end of the solenoid, depending on how the current circulates around the solenoid axis.

Imagine yourself looking down the axis of the solenoid, as though you were looking through the solenoid "tube" like a telescope. If, from your perspective, the current is flowing around the axis in a *counter-clockwise* manner then you are looking into the *north* end of the electromagnet. If, on the other hand, you see the current flowing *clockwise* around the axis, then you are looking into the *south* end.

This relationship is illustrated by the drawing in Figure 12.1.

In the drawing, the fingers are curled in such a manner as to "point" in a *counter-clockwise* manner. In this pose, the thumb points back to us.

So, if we take the fingers as "representing" the direction of the current around the axis of the solenoid, the thumb indicates which part of the solenoid is the north side.

[iv] From the Greek *solen* (pipe, channel) and *eidos* (form, shape). Similar to trapeziod and spheroid.

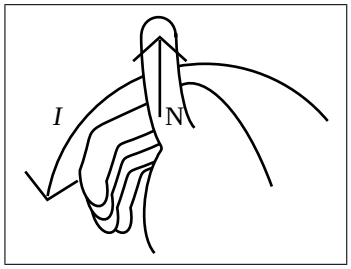


Figure 12.1: A drawing of one's right hand with the thumb extended and the fingers curled in a direction that represents the way the current through a solenoid curls around the axis of the solenoid (see curved line labeled "I"). When the current flows through the solenoid in this way, the north pole of the solenoid/magnet is at the end indicated by the extended thumb (indicated by the arrow labeled "N").

Since this uses the right hand, this method is called the **right-hand rule** for remembering which side of an electromagnet will be the north end.

The right-hand rule can be used to remember which side of the electromagnet is the north end.

If you use your left hand instead of your right hand, your thumb will point in the opposite direction (i.e., the side that is the south end).[v]

For example, consider the solenoid shown in the figure on page 210. Suppose the current comes in from the bottom. Due to the way the windings are in the solenoid, that would make the current circulate in such a way that the north end of the electromagnet would be at the top (of this particular solenoid). If we switch the direction of the current, the north end would be at the bottom (of this particular solenoid).

It is important to realize that the location of the north end has to do with how the current *circulates* around the axis. We control which way the current circulates via the winding.

To more easily see how the wires wind around the axis, consider the six solenoids drawn in Figure 12.2. The ones on top only have five loops, compared to fifteen loops for the middle ones and 25 loops for the bottom ones.

The only difference between the left and right solenoids is the direction in which the wire is wound around the axis. The solenoids look the same but, actually, the solenoids on the right have a winding that is opposite the winding of the solenoids on the left.

So, suppose the current enters each solenoid in Figure 12.2 from the left and exits to the right. The solenoids on the left will have their north ends on the *right* whereas the solenoids on the right will have their north ends on the *left*.

[v] I guess that would be the *left-hand* rule for remembering which side is the *south* end. As suggested by Joshua Chambers (spring of 2013), this can also be called the "southpaw" rule, as southpaws in baseball are left-handed pitchers.

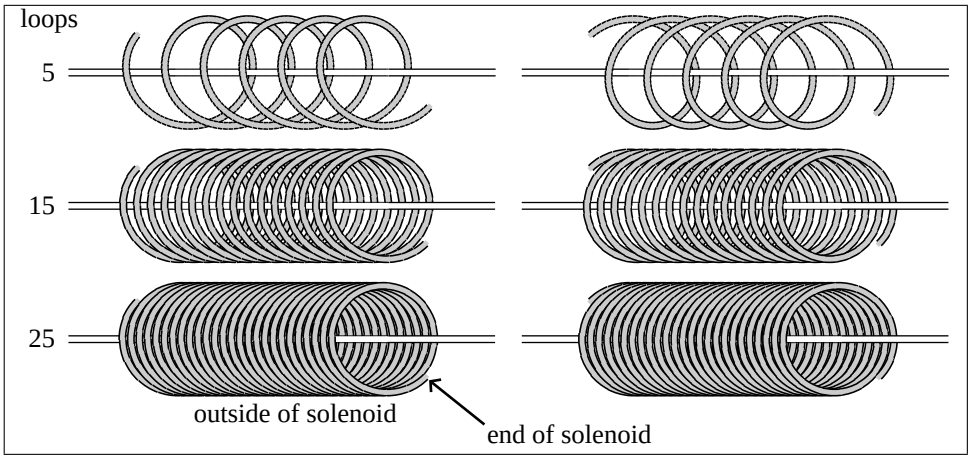


Figure 12.2: A drawing of six solenoids. The three on the right have an opposite winding as the ones on the left. The top solenoids have five loops, the middle ones have 15 loops and the bottom ones have 25 loops.

Notice that in both situations the current enters on the left and exits on the right but the north-sound orientation is different because the circulation is different. From the point of view looking at the solenoids from the right, the current is going counter-clockwise through the solenoids on the left but clockwise through the solenoids on the right.

✓ *Check Point 12.3: Suppose current is sent through the lower right solenoid of Figure 12.2 from the right to the left.*
(a) From the perspective of someone looking down the axis of the solenoid from the right side of the page, which way is the current flowing: clockwise or counter-clockwise?
(b) Based on your answer to (a) and the right-hand rule, which side of the solenoid is magnetic north: the left end or the right end?

12.4 Electric motors

One application of electromagnets is the **electric motor**, which is like the internal combustion engine inside a gasoline-powered car except that an electric motor runs on electricity, not a fuel like gasoline. Both make something

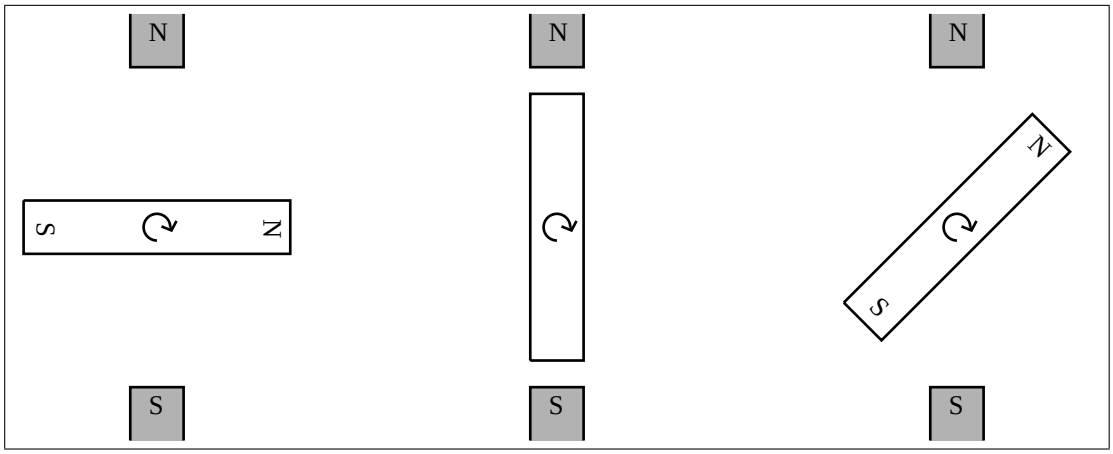


Figure 12.3: A schematic of an electric motor.

turn, like the wheels on a car. If you have a CD player or hard drive inside your laptop, a little electric motor is used to turn it. A similar motor is used to run an electric fan.

To make something spin, there needs to be a torque exerted on the object. As we know from chapter 4, magnets exert torques on one another (to make them align with each other) so it makes sense that perhaps we can use magnets to make something spin. The problem with magnets, though, is that, once aligned, they no longer rotate. In an electric motor, we want it to continue spinning. This problem is fixed by using an electromagnet.

This is illustrated in Figure 12.3, where the electromagnet is indicated as the white rectangle. In the left panel, the electromagnet is oriented left-right with south on the left and north on the right. Two permanent magnets are set up, one on top and one on the bottom, such that they exert a torque on the electromagnet, pulling the left end of the electromagnet up and pushing the right end of the electromagnet down.

Now comes the ingenious part. Since the electromagnet can be turned off simply by stopping the current through it, we set up the motor so that the electromagnet turns off just as it becomes aligned. This is illustrated in the middle panel of Figure 12.3. Since it was already spinning when it reaches that orientation, its inertia keeps it spinning. It does not get "stuck" in that orientation.

Then, once the electromagnet has moved passed that orientation, the current

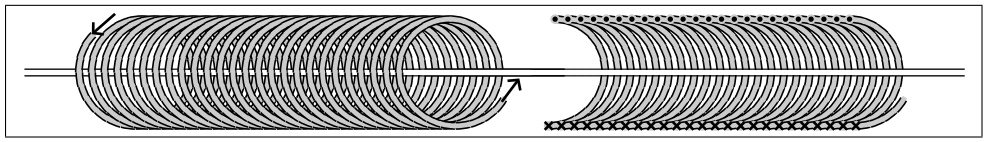


Figure 12.4: Two illustrations of an electromagnet. Left: as in Figure 12.2. Right: Cross section as though sliced along the axis.

turns on again, again with south on the left end and north on the right, making the electromagnet continue to spin in the same direction.

Thus, a motor is just an electromagnet that is being rotated by a permanent magnet, with the direction of the current through the electromagnet changing so that a torque is always applied in the same direction.[vi]

It is an amazing device when you think about it. And, the fact that it can be explained just knowing what we know about electromagnets makes it just that much more amazing.

12.5 Magnetic field

Since an electromagnet acts like a magnet, it should come as no surprise that an electromagnet also has a magnetic field and that the magnetic field of an electromagnet looks the same as the magnetic field of a magnet. It is somewhat hard to illustrate this using the electromagnet illustrations shown earlier (with the loops) so instead I will draw the electromagnet as though we are looking at a cross-section of it. This is illustrated in Figure 12.4. On the left is the electromagnet as illustrated in Figure 12.2. On the right is the same electromagnet but with only the back half shown.

It is important to keep in mind that the right illustration, like the left illustration, represents a complete solenoid, not just the back half. To illustrate that current is still flowing through the solenoid, I'll use dots and crosses to show the direction of the current through the solenoid. For example, for the left solenoid in Figure 12.4, I added two arrows to show how the current is

[vi] Alternately, we could switch the permanent magnet and electromagnet, with the permanent magnet spinning and the electromagnet alternating its orientation so that the permanent magnet keeps spinning. However, it turns out it is easier to switch the direction of the electromagnet if it is the one that is spinning.

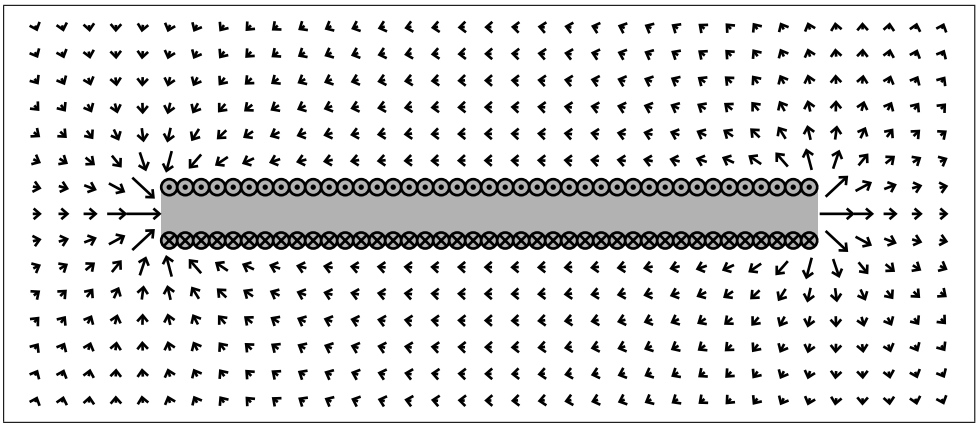


Figure 12.5: An illustration of the magnetic field around a solenoid that is set up when the current flows as indicated (into the page at the bottom and out of the page at the top). Each arrow (called a magnetic field vector) represents the direction of the magnetic field at that location. The length of the arrow represents the strength of the magnetic field at that location.

flowing into the left end and out the right end. For the right solenoid, though, it is easier to just show how the current is flowing "into" the bottom part of each loop out "out" the top part of each loop. Each $\odot$ (dot) represents a cross-section of a wire with current coming "out of the page" (as though we are viewing the head of an arrow pointed out of the page) and each $\otimes$ (cross) represents a cross-section of a wire with current going "into the page" (as though we are viewing the tail of an arrow pointed into the page).

⁉ Dots and crosses indicate the directions "out" of the page and "into" the page.

For the rest of this chapter, then, I will indicate the electromagnet via the cross section, using the dots and crosses to show how the current is flowing through the electromagnet. This is illustrated in Figure 12.5 along with the electromagnet's magnetic field, showing how the electromagnet's magnetic field looks the same as a permanent magnet's magnetic field.

⁉ The magnetic field of an electromagnet looks the same as for a permanent magnet.

✍ If you use the right-hand rule described on page 211 with the current going in the direction indicated in Figure 12.5, you should find that the electromagnet's north end should be on the right.

Does the electromagnet also have an electric field?

No. Like everything in an electric circuit, an electromagnet is neutral, so it doesn't have an electric field. Neither does a magnet, since a magnet is also

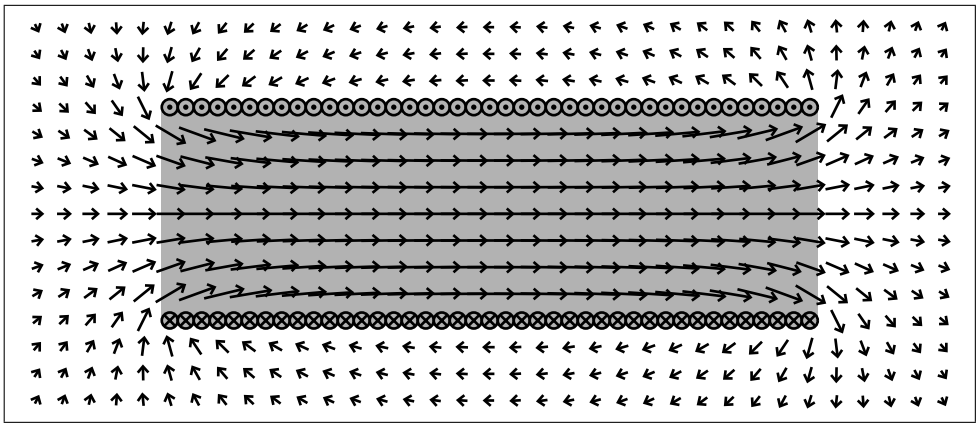


Figure 12.6: An expanded illustration of the magnetic field both inside and around a solenoid, with current flowing as in Figure 12.5.

neutral.[vii]

What about the magnetic field within the cavity of the electromagnet?

Within the cavity of the electromagnet, the magnetic field vectors point toward the right. I didn't draw them in Figure 12.5 just because I thought it made the figure too crowded.

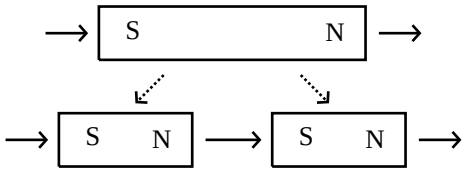
However, in Figure 12.6 I have drawn an expanded view of the electromagnet in order to illustrate the field vectors both within the solenoid and outside the solenoid.

Notice how the magnetic field inside the solenoid points toward the right (in this case). If you compare the magnetic field vectors inside a solenoid (Figure 12.6) with the electric field vectors between an electric **dipole** (a positive charge and a negative charge; see figure on page 90), you'll see that the directions are *opposite*.

This is because an electromagnet is a series of current loops, each acting like a magnet (with north and south oriented the same way), rather than a single north pole at one end and a single south pole at the other.

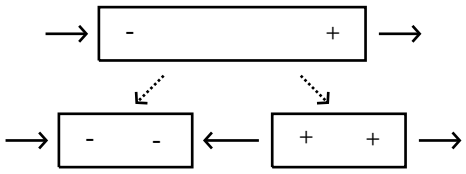
[vii]Technically, there is no electric field *outside* the wires. There is indeed an electric field *inside* the wires, which is what pushes the electrons along the wire and produces the current, but the arrows are *outside* the wires, so they represent the magnetic field only.

Consider, for example, what happens if you break a magnet in half. This is illustrated in the figure below. On top, a magnet is indicated by its "N" and "S" ends. Its magnetic field is indicated by the two arrows, both pointing toward the right. One is on the left side pointing toward the south end and one is on the right side pointing away from the north end.



If we break the magnet in half, we get two smaller magnets. This is indicated on the bottom of the figure. Notice that between the two magnets the magnetic field is still rightward, pointing from the north end of one magnet toward the south end of the other magnet.

Let's compare this what happens if you break an electric dipole in half. This is illustrated in the figure below. On top, an electric dipole is indicated by its "−" and "+" ends. Its electric field is indicated by the two arrows, both pointing rightward. One is on the left side pointing toward the negative end and one is on the right side pointing away from the positive end.



If we break the electric dipole in half, we don't get two smaller dipoles. Instead, we are essentially separating the negative end from the positive end. This is indicated on the bottom of the figure. Notice that between the two ends the electric field is *leftward*, pointing from the positive piece toward the negative piece.

With electromagnets, then, the right-hand rule not only tells which side of the electromagnet is the north end but it also tells us the direction of the magnetic field vectors *inside* the solenoid. In other words, when the fingers of the right hand are curled such that they mirror the direction of the current,

the thumb points in the direction of the magnetic field vectors along the axis of the electromagnet.

What if the current is sent through the other way?

Then the magnetic field would point the opposite way (not shown).

✓ *Check Point 12.4: Suppose current is sent through the lower right solenoid of Figure 12.2 (on page 212) from the right to the left. Inside the solenoid, which way does the solenoid's magnetic field point?*

As with a permanent magnet, the magnetic field is stronger closer to the electromagnet (see, for example, how the arrows in Figure 12.5 are larger nearer the electromagnet) and so the impact of the electromagnet on another magnet is stronger on the side of the magnet closer to the electromagnet. This is why, once aligned with the electromagnet's magnetic field, a magnet is attracted to the electromagnet and pulled into the solenoid.

Once the magnet is pulled inside the solenoid, it stays there. Why?

Notice in Figure 12.6 how the magnetic field vectors are all the same size within the solenoid. This means that once the magnet is sucked into the solenoid, each end of the magnet experiences the same magnetic field and the net force on the magnet becomes zero (since the force on each end will be equal and opposite). Once the magnet slows down due to friction, it will stay there.

✓ *Check Point 12.5: When a permanent magnet is brought near a solenoid (with current flowing through it), the magnet is forced into the solenoid. However, when placed in the solenoid with both ends of the magnet sticking out, there is no force on the magnet. Why is this?*

12.5.1 Single loop

Does a single loop of current have a magnetic field?

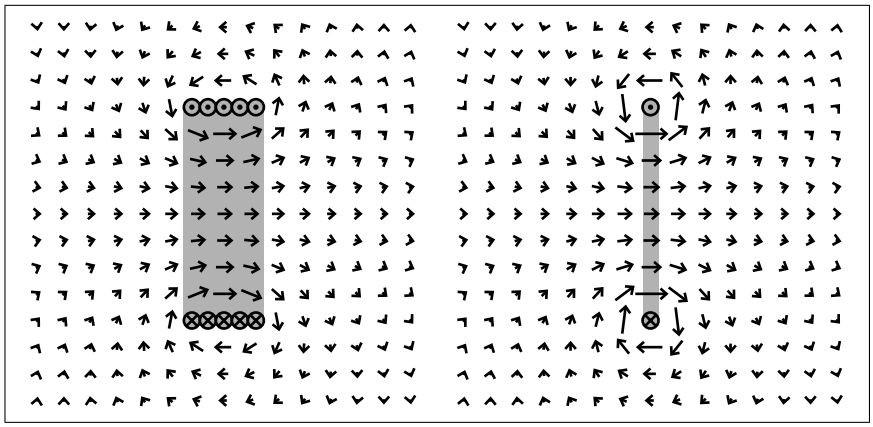


Figure 12.7: A cross-section through a five-loop solenoid (left) and single loop (right) showing the magnetic field that is set up when the current flows as indicated (into the page at the bottom and out of the page at the top).

Yes. Magnets and electromagnets are just lots and lots of current loops. So, a single loop, like a magnet, has a magnetic field, which we can indicate with the magnetic field vectors.

However, the magnetic field of a single current loop is very weak, so it can be kind of hard to measure. If we could measure it, though, the magnetic field vectors associated with a single current loop would look like what I've drawn on the right side of Figure 12.7.

For comparison, I've also drawn the magnetic field associated with a 5-loop solenoid (with one-fifth the current, so that the strength of the magnetic fields are comparable).

Notice how the current direction is the same as before (out of the page at the top and into the page at the bottom) and so, according to the right-hand rule, the magnetic field vectors are directed toward the right *along the axis of the loop*. As you can see, away from the axis the magnetic field can be in a different direction.

✓ *Check Point 12.6: Suppose one wanted to model Earth's magnetic field by using current through a single wire loop wound around the equator. In which direction should the current be flowing, toward the east or toward the west? Remember that Earth's North pole is a magnetic south pole.*

12.5.2 Straight wire

Does a wire have to be in a loop to have a magnetic field?

No. As long as there is current flowing through the wire, the wire will have a magnetic field. However, a straight wire doesn't have a defined north end or south end, so the wire's magnetic field won't look like the magnetic field of a solenoid or magnet. In fact, the wire's magnetic field is directed *around* the wire, as can be seen in Figure 12.7 (see, for example, the magnetic field near the bottom of the single loop).

⁉ A straight wire's magnetic field is directed "around" the wire.

How can the field be directed around the wire?

While it may seem strange, it is a consequence of how we defined the direction of the magnetic field to be the direction that a north pole is forced. North poles (and south poles) are the result of current loops. If you straighten out the wire, you don't get the poles. To see this, go back to the solenoid and notice how the solenoid's magnetic field is parallel to the solenoid axis, which is *perpendicular* to the direction of the current in the wires (since the wires go in circles around the axis). With a straight wire, the wire's magnetic field is still perpendicular to the direction of the current in the wire, but since the wire is straight that means its magnetic field must be circular.

Regardless of the direction, the key point here is that even straight wires, like those in transmission and distribution lines, have a magnetic field. Indeed, every electrical device has a magnetic field because there is current flowing through the device.

Is the magnetic field from electric devices and power lines dangerous?

The short answer is no, since electrical devices tend to have magnetic fields much less than the magnetic field of a typical permanent magnet. For example, in chapter 6 it was mentioned that the magnetic field of a refrigerator magnet is a few milliteslas close to the magnet. In comparison, the magnetic field near electric devices is around a couple of microteslas (about a thousand times smaller).[viii] The magnetic field of power lines is larger but we

[viii] The magnetic field is only 4 to 8 microteslas near a microwave (from one foot away), about 0.3 to 5 microteslas near an electric blanket (from one inch away), and only about 0.1 microtesla near a television (from three feet away). The magnetic field is stronger closer to an appliance, so a hair dryer's magnetic field could be larger since you are typically much closer to a hair dryer (6 to 20 microtesla from one inch away).

much further away from them. High-voltage transmission lines, for example, typically have a magnetic field less than 1 microtesla at a location 100 feet from the edge of the right of way.[ix]

Notice how the magnetic field of electrical devices is even less than Earth's magnetic field (around 50 microteslas), so there should be no danger, but even stronger magnetic fields should not be dangerous. Whole body scanning (MRI) can use magnetic fields as high as 60,000 times stronger than Earth's (3 teslas). We just need to be careful that we don't have any ferromagnetic objects around, since such will experience a strong magnetic force and it can be dangerous to be in a room where a ferromagnetic material is flying toward the magnet.

WHY DO SOME PEOPLE SAY THE MAGNETIC FIELD FROM POWER LINES ARE DANGEROUS?

The concern has to do with the fact that the transmission line's magnetic field oscillates (like your blood but more quickly) because the current through the wires oscillates (see chapter 16). Earth's magnetic field doesn't oscillate. Since an oscillating field will "jiggle" charged objects like electrons and such, some people may fear that such oscillating fields may cause health issues like cancer. However, according to the National Cancer Institute[x], no consistent evidence has been found relating cancer with the magnetic field from transmission lines (or cell phones).[xi]

✍ An oscillating magnetic field, it turns out, produces an oscillating electric field (more on this in chapter 18), and the combined effect of both fields is called the **electromagnetic field**. So, the electromagnetic field is just a fancy way of referring to the electric and magnetic fields.

✓ *Check Point 12.7: Does a straight wire with current have a magnetic field? If so, does it point toward it, away from it, or neither?*

[ix] Source: lipower.org

[x] Go to https://www.cancer.gov and search for "electromagnetic fields fact sheet".

[xi] While there is no evidence that these magnetic fields are a danger, there may be some studies that show an unexplained correlation between some diseases and one's long term proximity to power lines. In addition, an oscillating magnetic field can be dangerous if the oscillating frequency is very high, much higher than what it is for transmission lines. For example, X-rays and gamma rays are associated with very high frequency magnetic fields.

Summary

This chapter examined the relationship between moving charge and magnets.

The main points of this chapter are as follows:

- An electromagnet is a coil of wire with current.
- A metal core is not necessary but makes the electromagnet stronger.
- The north pole is at one end of the solenoid, depending on how the current circulates around the solenoid axis.
- The right-hand rule can be used to remember which side of the electromagnet is the north end.
- The magnetic field of an electromagnet looks the same as for a permanent magnet.
- A straight wire's magnetic field is directed "around" the wire.
- Dots and crosses indicate the directions "out" of the page and "into" the page.

By now you should be able to create an electromagnet and predict the locations of the north and south ends.

Frequently asked questions

Does the metal core of an electromagnet have to be a permanent magnet?

No. A metal core isn't necessary but, if there is one, the metal only needs to be a ferromagnetic material so that the little magnets inside it can be aligned, thus strengthening the electromagnet.

Why does current produce a magnetic force?

For our purposes we can leave this as an unknown, just as we don't know why protons and electrons interact electrically. All we know is that the wires are apparently still neutral, so it isn't due to excess charge.

While a model for explaining it does exist, the model involves a concept called special relativity, which we have not yet examined (and won't examine in this book). Thus, for us, it will remain an unexplained phenomenon. We can still *describe* the phenomenon, however.

Terminology introduced

Axis	Electromagnetic field
Core	Ferromagnetic
Electric Motor	Right-hand rule
Electromagnet	Solenoid

Additional problems

Problem 12.1: An electromagnet is created by wrapping wire around an iron bar and sending current through the wire.
(a) Is the iron bar necessary? Why or why not?
(b) Is it necessary to have insulation on the wire? Why or why not?
(c) Is the electromagnet, or any part of the electromagnet, electrically neutral?

Problem 12.2: Suppose the magnetic field inside an electromagnet points rightward. Does that mean the current through the electromagnet flows rightward as well? Why or why not?

Problem 12.3: Suppose we have a loop with its axis directed into the page and current flowing clockwise.
(a) If this loop is placed in a region where the magnetic field is directed into the page, would there be a magnetic torque exerted on the loop? Why or why not?
(b) What if the magnetic field was directed rightward? Why or why not?

13. Fluids

Puzzle #13: Why are the walls of the aorta (the blood vessel through which blood *leaves* the heart) thicker than the walls of the venae cavae (the blood vessel through which blood *enters* the heart)?

Introduction

When looking at how charged particles move through wires, you might be reminded of fluid flow, and you wouldn't be far off. Charged particles do indeed move through wires in much the same way liquids flow through pipes (like blood flow through veins and arteries) and air flows through the atmosphere.

13.1 Describing fluids

What is a fluid?

The word **fluid** is used for both **liquids** and **gases** because both flow. In fact, the word *fluid* comes from the Latin word *fluere*, which means *to flow*. This ability to flow is the main feature that distinguishes fluids (liquids and gases) from solids.

Both gases and liquids are considered to be fluids.

13.1.1 Liquids vs. gases

There are some differences between liquids and gases. Most notably, gases can be compressed whereas liquids are relatively **incompressible**, which means you can't change their volume by compressing it. Since electrons in a circuit, like the atoms and molecules in liquids, keep their same spacing throughout, we'll be focusing on the properties of liquids.

We can consider liquids to be incompressible.

WHY ARE LIQUIDS RELATIVELY INCOMPRESSIBLE?

All objects are made up of tiny particles called atoms and molecules. In a liquid or solid, those tiny particles are "attached" to one another, as though they were connected by strong, tiny springs. In comparison, the tiny particles in a gas are disconnected, and don't interact until they "bump" into each other, at which point they bounce off each other like little rubber balls.

When I say that the tiny particles in solids and liquids are connected by "strong" springs, I mean that they resist compression and expansion. For example, when you step on the floor, the tiny springs (connecting the molecules in the floor) compress a tiny bit, as they push back on your foot, but the compression is so tiny it isn't noticeable. We consider the floor to be rigid but in reality it must compress a tiny bit when we walk on it.

Similarly, suppose you take a syringe, fill it with water and then cap the end of the syringe so no water can get out. You can press on the syringe plunger really hard and the plunger doesn't go anywhere – the volume of water doesn't seem to change. In comparison, if the syringe was filled with air, instead of water, then the air would be compressed as you press on the syringe plunger because air, being a gas, can be compressed.

✍ Liquids (like water) *are* compressible, as evidenced by the fact that cold water sinks, and colder water must have its water molecules more closely packed for this to happen. However, the compression is very slight, which is why we can consider liquids to be *relatively* incompressible.

✓ *Check Point 13.1: The brake pedal in a car is connected to the brakes via a tube filled with brake fluid. Pressing on the brake pedal pushes the brake fluid against the brake pads, pushing them against the wheel to slow them down. If there is air bubbles along with the brake fluid in the tube, the brake pedal feels "spongy", not firm. Why might that be?*

13.1.2 Density

To quantify what we mean by incompressible, it helps to define a quantity called **density**.

For incompressible liquids and solids, the mass is proportional to the volume (i.e., the larger the volume, the larger its mass). Because the mass is proportional to the volume, the ratio is independent of the object's size. We thus define a property called *density* that is the ratio of the object's mass m to its volume V:

$$\rho = \frac{m}{V} \tag{13.1}$$

where ρ (the lower-case Greek letter "rho") is used to represent density.

⁇ An object's density is defined as its mass divided by its volume.

WHY IS ρ USED FOR DENSITY?

I don't know why ρ is used but it is so common that I will also use it here.[i]

✍ Be careful: the abbreviation for density looks like the lower-case Roman letter "p". It isn't. It is a Greek letter. This is important because we will also be talking about *pressure* and we use a "P" (capital) to represent that.

WHAT ARE THE UNITS OF DENSITY?

Density is measured in units of kg/m^3. There is no special unit specifically for density.

Example 13.1: (a) The density of water is 1 g/cm^3. What is this in kg/m^3? (b) The density of air at sea level is about 1.3 kg/m^3. What is this in g/cm^3?

Answer 13.1: (a) Multiply the density by (1 kg)/(1000 g) and by (100 cm)^3/(1 m)^3 to get

$$\begin{aligned} 1\ \text{g/cm}^3 &= 1\ \text{g/cm}^3 \\ &= 1\frac{\cancel{\text{g}}}{\cancel{\text{cm}}^3}\frac{1\ \text{kg}}{1000\ \cancel{\text{g}}}\frac{(100\ \cancel{\text{cm}})^3}{(1\ \text{m})^3} \\ &= 1000\ \text{kg/m}^3 \end{aligned}$$

(b) There are two general ways to convert from kg/m^3 to g/cm^3. One way is to multiply the density by ratios equal to one but which are written in units

[i] Some people use d for density but I'm following the convention in physics, which is to use Greek letters for density-type quantities. For example, it is common to use λ (lambda) for mass per length, σ (sigma) for mass per area, and ρ (rho) for mass per volume. Since lambda and sigma can be thought as the *linear* and *surface* densities, respectively, maybe rho was chosen because it represents the *regional* density? Also d conflicts with the use of d in calculus (representing a differential).

Substance	**Density (kg/m^3)**
Air (STP[ii])	1.225
Gasoline	700
Ice (0°C)	920
Water (20°C)	998
Water (4°C)	1,000
Aluminum	2,700
Iron	7,860
Lead	11,300
Mercury	13,600
Gold	19,300

Table 13.1: Densities of some common materials.

such that the previous units are replaced by the desired units. In this case, we would multiply by (1000 g)/(1 kg) and by (1 c)^3/(100 cm)^3 to get

$$\begin{aligned} 1.3 \text{ kg/m}^3 &= 1.3 \text{ kg/m}^3 \\ &= 1.3 \frac{\cancel{\text{kg}}}{\cancel{\text{m}}^3} \frac{1000 \text{ g}}{1 \cancel{\text{kg}}} \frac{(1 \cancel{\text{m}})^3}{(100 \text{ cm})^3} \end{aligned}$$

which gives an equivalent value of 1.3×10^{-3} g/cm^3. The other method would be to replace "kg" by "1000 g" and replace "m" by "100 cm" (which would then be cubed). Both ways should give you the same answer.

What are typical values of densities?

Each substance has its own density. A list of sample substances and their densities is shown in Table 13.1.

Example 13.2: What is the mass of an iron sphere of radius 50 cm? Note: the volume of a sphere is $\frac{4}{3}\pi r^3$. See table 13.1 for the density of iron.

Answer 13.2: We can rewrite $\rho = m/V$ as $m = \rho V$. Plug in 7,860 kg/m^3 for ρ and $\frac{4}{3}\pi$ (0.50 m)^3 for V and solve. I get 4120 kg (pretty massive!).

[ii] Standard temperature (15°C) and pressure (1 atm).

✓ *Check Point 13.2: Consider a syringe capped at the end and filled with water (which we can consider to be incompressible). As you press on the syringe plunger, does the density of the water change?*

13.1.3 Pressure

We are assuming that liquids, like water, are incompressible. That is why, when we have a syringe filled with water and capped at the end (to keep the water in the syringe), we say that the water volume is unchanged regardless of how hard we press on the syringe plunger. This means the water density doesn't change either.

However, the force exerted by the water (on the plunger) *does* change, as it must equal the force being applied by the plunger (on the water). This is, after all, consistent with the law of interactions.

Technically, the increased force is due to a very tiny compression in the invisible springs. The compression is so tiny that it has no impact on the density, but the compression must be there nonetheless if the force has increased. In addition, the springs throughout the liquid must all experience that same very tiny compression, as the forces must balance out everywhere. Otherwise, the liquid would be moving (we'll get into this later).

This uniform tiny compression throughout the liquid results in an increased force throughout the liquid. What, then, can we say about the force exerted by the liquid?

The answer is that the force exerted by the liquid depends on the number of springs that acting, and that number will depend on the area of liquid that is pressing against whatever object we are considering.

The force exerted by the liquid on an object, then, is proportional to the surface area of the liquid that is pressing against the object. Because the force is proportional to the area, the ratio (F/A) is independent of the area. We thus define a property called **pressure** that is the ratio of the force (exerted by the liquid on the object) to the area (of contact between the liquid and the object):

$$P = \frac{F}{A} \tag{13.2}$$

⁘ The pressure of a fluid is equal to the force exerted by the fluid divided by the area of contact.

where P is used to represent pressure.

✍ The SI unit of pressure is N/m^2. This is usually replaced by the unit pascal[iii] , abbreviated as Pa.

Whereas the force (exerted by the liquid on an object) depends on the area of contact between the liquid and the object, the pressure does not. Since the pressure doesn't depend on the size of the contact area, we can say that the pressure is the same even if the contact area is extremely tiny, like a point. Indeed, the pressure value is the same for every point in the liquid, as long as the liquid isn't moving and we ignore gravity.

What happens if there is gravity?

If there is gravity then the pressure will be greater at the bottom than at the top, since the particles on the bottom are being compressed by the weight of the particles above them. As mentioned earlier, the compression is extremely tiny, so we don't notice it, but it is enough to change the pressure, making the pressure increase with depth.

⁉ Because of gravity, the pressure of a fluid decreases with height.

This is also why the air pressure is lower the higher one goes in the atmosphere. This may cause your ears to "pop" when ascending or descending in a plane (or going up or down a mountain), as the air inside your ear tries to escape or the air outside your ear tries to enter in order to equalize the pressure inside and outside your ear.

The decrease of pressure with height is what leads to **buoyancy**, the upward force on objects that are submerged in the fluid, since the pressure on the bottom of the object (pushing it up) is greater than the pressure on the top of the object (pushing it down). In air, the buoyancy is slight because the density of air is small and so buoyancy in air can usually be ignored for solid objects unless the object has a very low density itself, like a balloon. Water, though, has a higher density than air and thus the buoyancy in water is more significant than the buoyancy in air, and can no longer be ignored. The buoyancy in water is why things like wood float.

What happens if the liquid is moving?

If the liquid is moving, there can be different pressure values at different locations. In the next section, we'll examine why this is.

[iii] The pascal is named for Blaise Pascal, who was born in France in 1623 and did experiments in atmospheric pressure. Pascal died at age 39.

✓ *Check Point 13.3: Consider a syringe capped at the end and filled with water (which we can consider to be incompressible). Is the pressure the same throughout the water? As you press on the syringe plunger, does the pressure of the water change?*

13.2 Moving liquids

Sometimes, when opening the door to a house or building, you might notice that the air rushes in or out. This is due to the air pressure inside the building not being equal to the air pressure outside. Barring any other forces (like gravity[iv]), fluids flow from regions of high pressure to regions of low pressure (much like how objects are forced from high potential energy configurations to low potential energy configurations).

⁉ Fluids flow from regions of high pressure to regions of lower pressure.

It is important to recognize that it is the *difference* in pressure that forces the fluid to flow. If the pressure is the same everyone, the fluid doesn't flow, regardless of what the pressure value is. The greater the pressure *difference*, the faster the fluid flows.

⁉ Fluid flow is related to the pressure difference, not the pressure value itself.

Is this consistent with the law of force and motion?

Yes, it is consistent, although at first glance it may not appear to be. For fluids, the law of force and motion states that a difference in pressure forces the fluid to *accelerate*, so a fluid *can* flow when the pressure is the same everywhere – it just can't *accelerate*. However, for the situations we'll be examining (namely water flow in pipes, blood flow through veins, and the like) there is often enough mixing and drag to cause the fluid to quickly[v] slow down and stop when there isn't a pressure difference present to keep the fluid flowing. In these cases it is the *speed* of the fluid flow, rather than its acceleration, that depends on the pressure difference.

[iv] We know that gravity causes the air pressure to be less at higher altitudes. In that case, the air doesn't necessarily flow upward from higher to lower pressure. Thus, the discussion in this section focuses on *horizontal* flow and *horizontal* differences in pressure.

[v] The fluid won't stop instantaneously, as evidenced by how your coffee continues to swirl in your cup for a short time after you stop stirring it with a spoon. Within pipes this results in a phenomenon known as *hydraulic shock* or water hammer (a knocking noise that occurs when the water is turned off quickly).

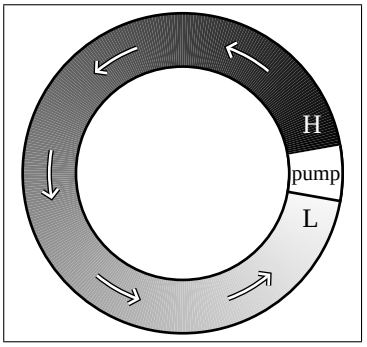
✍ As we'll see in part D, electrons flow through an electric circuit in much the same way as liquids through a pipe.

What is an example of a pressure difference?

To get an intuitive understanding of what we mean by a pressure *difference*, consider what happens when you stick your hand out the window of a car that is moving down the street or highway. The air ahead of your hand is being "squashed", leading to a higher air pressure ahead of your hand and a corresponding lower air pressure behind your hand. The difference in pressure values is what pushes your hand backwards.[vi]

For water pipes, we can create a pressure difference by using a water **pump**. A water pump makes the pressure lower at the inlet (where the water is drawn in) and higher at the discharge tube (where the water is pumped out). The resulting difference in difference (between the inlet side and the discharge side) is what sucks water into the pump (drawing it into the low pressure area) while simultaneously pushing water out of the pump (out of the high pressure area) and through the pipes connected to the pipe.[vii]

This is illustrated to the right as a circular pipe, with a pump. The amount of pressure is indicated by the shading, with darker shading representing higher pressure values. The highest pressure (indicated by the "H" and darkest shading) is on the discharge side of the pump, and the water flows out of the discharge side of the pump, through the pipe, to the inlet side of the pump, where the pressure is lowest (indicated by the "L" and lightest shading).



Notice how the water flows through the pipe from higher pressure (at pump discharge) to lower pressure (at pump inlet). Also notice how the pressure decreases from higher (darker shades) to lower (lighter shades) as one goes "downstream" along the pipe.

[vi] In the previous section, we discussed buoyancy and how, due to gravity, the pressure of a fluid is greater at the bottom than at the top. That is why there is a buoyancy force on a piece of wood when it is submerged in water.

[vii] For the wind, the difference in air pressure is usually related to differences in temperature and/or the amount of air above, since more air above will press more on the air below it, leading to a greater pressure.

This is similar to the circulatory system. Blood moves through the circulatory system because the heart creates a pressure difference – the highest pressure in the aorta (where blood leaves the heart; the **systolic** pressure) and lowest pressure in the venae cavae (where blood enters the heart; the **diastolic** pressure). This is why the aorta have thicker walls than the venae cavae – to handle the greater pressure being exerted on the walls.

Just as with the water pipe, the blood pressure is less the further the blood is from the aorta (heart exit), decreasing as it travels through the body (from the arteries through the capillaries to the veins).

✓ *Check Point 13.4: When you turn on faucet in a sink, water flows out of the faucet into the sink. Assuming the pipe is horizontal, where is the water pressure greater – at the start of the pipe or at the end (where the water leaves the faucet)?*

13.3 Current

⁝ Fluid current indicates the rate at which a quantity of fluid flows past a point.

As with the flow of electric charge, with fluids we tend to focus on the *current* rather than the fluid *speed*. This is because the current represents *how much* fluid (or charge) flows past a point divided by time.

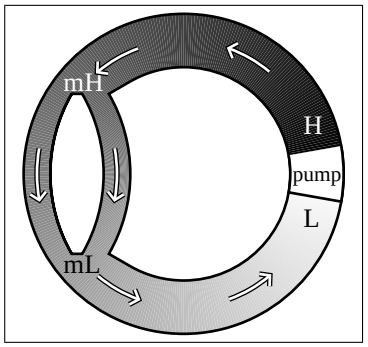
To appreciate the difference between current and speed, consider water flowing down a river. Certainly, the faster the flow, the greater the current. However, the current also depends on how wide and deep the river is so, for the same speed, a tiny stream has less current than a large river.

Why pay attention to current instead of velocity?

⁝ For a pipe full of liquid, the amount entering the pipe must equal the amount leaving it, meaning the current is the same throughout.

The advantage of current is that, for single paths, the current in one part must be the same as another part since the mass can't disappear and we are assuming the fluid is incompressible (which is a pretty good assumption; see section 13.1.1). So, if we had a single pipe with water flowing through it, the amount entering the pipe must equal the amount leaving it (assuming water fills the entire pipe).

A single path is a simple example but the same is true for more complicated situations, like the one illustrated to the right, where the left part of the pipe splits into two parts and then rejoins.



As with electric current, the current splits evenly at the split if each of the two pipes that make up the split path is identical, with half going through each.

Even though the value of the current is half within each split path, the speed is not. Rather, the speed is likely the same within each split as along the rest of the piping. This is because in this pipe the pressure uniformly decreases (as shown by the shading) along the split in the same way it decreases along the rest of the pipe. Even though the pressure at the start of the split (indicated in the illustration by mH, a medium high pressure) is less than the pressure at the pump outlet (indicated by high pressure H), it is still larger than the pressure at the end of the split (indicated in the illustration by mL, a medium low pressure), which in turn is still greater than the pressure at the pump inlet (indicated by low pressure L).

How do we calculate the current?

Since the current depends on both the fluid speed and the width and depth of the flow, it likely comes as no surprise that the current is equal to the product of the fluid speed and cross-sectional area. Since the SI units of velocity and area are m/s and m^2, respectively, that means the SI unit of current is m^3/s (cubic meters per second).[viii]

✍ In chapter 11, the amount of electric current was indicated by the letter I. We can use the same notation for fluid flow and also call it **current**, since they represent the same idea. However, because electric current refers to the flow of charge, not volume, the units are different (coulombs per second rather than cubic meters per second).

Since the current must be the same throughout a single path, we can use the definition of current to determine how the speed changes when the diameter

[viii] This definition provides the rate at which a volume of fluid flows past. Current can also be expressed in terms of the mass, in which case you'd also multiply by the density (in kilograms per cubic meter) to get the current with SI unit kg/s.

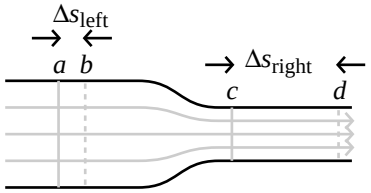


Figure 13.1: A pipe with liquid flowing inside it (either from left to right or from right to left). The cross-sectional area of the pipe is smaller on the right than on the left. Correspondingly, the liquid inside the pipe moves faster on the right part.

of the pipe changes. For example, consider the pipe illustrated in Figure 13.1, where the liquid flows from left to right through the pipe.

Since there is a single path, the current through each part must be the same. Furthermore, since the current is the product of the speed (v) and the cross-sectional area (A), we have:

$$A_{\text{left}} v_{\text{left}} = A_{\text{right}} v_{\text{right}} \tag{13.3}$$

This means that the flowing is flowing faster through the narrower part of the pipe on the right compared to the wider part of the pipe on the left. If we tagged some fluid at a and c then some time later we'd find that the water at a had moved a shorter distance (to b) compared to how far the water at c had moved (to d).

Equation 13.3 is called the **equation of continuity** and can be used to identify how fast a fluid flows in different regions, as in the following example.

⁝ The equation of continuity relates the speed of a liquid with the cross-sectional area of the pipe.

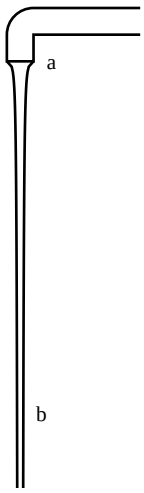
Example 13.3: Suppose the height of the water in a container is decreasing at 0.1 cm/s. That means the water at the top of the container has a speed of 0.1 cm/s. Suppose further that the diameter of the circular container is 10 cm and the diameter of the opening at the bottom is 0.2 cm. How fast is the water coming out the opening? Note: the area of a circle is πr^2.

Answer 13.3: The water at the top has a speed of 0.1 cm/s and a cross-sectional area equal to $\pi(5\text{ cm})^2$. The water coming out the opening has a speed v (unknown) and a cross-sectional area equal to $\pi(0.1\text{ cm})^2$. According to the equation of continuity,

$$\pi(5\text{ cm})^2(0.1\text{ cm/s}) = \pi(0.1\text{ cm})^2 v$$

Solve for v to get 2.5 m/s. Notice that I divided the diameters by two to get the radii. Also note that you can decrease the number of calculations you have to do by canceling the π's first.

✓ *Check Point 13.5: Water flows straight down from an open faucet (see illustration). At the top (point a), the stream has a radius equal to the faucet opening (0.75 cm) and has a speed equal to 0.85 m/s. At a distance of 0.10 m below the faucet (point b), the speed of the water is 1.6 m/s. What is the radius of the water stream at point b?*



Summary

This chapter described how liquids flow through pipes in terms of pressure, current and resistance.

The main points of this chapter are as follows:

- Both gases and liquids are considered to be fluids.
- We can consider liquids to be incompressible.
- An object's density is defined as its mass divided by its volume.
- The pressure of a fluid is equal to the force exerted by the fluid divided by the area of contact.
- Because of gravity, the pressure of a fluid decreases with height.
- Fluids flow from regions of high pressure to regions of lower pressure.
- Fluid flow is related to the pressure difference, not the pressure value itself.
- Fluid current indicates the rate at which a quantity of fluid flows past a point.

- For a pipe full of liquid, the amount entering the pipe must equal the amount leaving it, meaning the current is the same throughout.
- The equation of continuity relates the speed of a liquid with the cross-sectional area of the pipe.

Terminology introduced

Buoyancy	Equation of continuity	Liquids
Current	Fluid	Pressure
Density	Gases	Pump
Diastolic	Incompressible	Systolic

Abbreviations introduced

Quantity	**SI unit**
area (A)	meter cubed (m^3)
density (ρ)	kilogram per meter cubed (kg/m^3)
distance (Δs)	meter (m)
fluid current (I)	cubic meter per second (m^3/s)
pressure (P)	pascal (Pa)[ix]
speed (v)	meter per second (m/s)

[ix] A pascal is equal to a newton per meter squared (N/m^2).

Part D

Circuits

14. Voltage

Puzzle #14: Why are batteries rated in terms of voltage, like a 1.5-V battery, instead of energy or current?

Introduction

In this part of the book we examine electric circuits, much like what we did in chapter 11, but with an eye toward how a circuit works. A big piece of that is the battery, which acts in a way that is analogous to the role of pumps within water pipes. The puzzle mentions that the battery strength is given in terms of its voltage. In this chapter, we'll explore what voltage is and why we use it.

14.1 Properties of electric current

Before explaining what voltage is and how we use it, let's first review a little bit about current, which you learned about in chapters 11 and 13.

In section 11.3.2, *electric* current was defined as the rate at which charge flows past a point (see equation 11.1 on page 192). Notice how similar that is to the concept of water current, which is the rate at which water flows past a point, or the blood flow rate, which is the rate at which blood flows through the body.

Because each element in a **circuit** remains neutral as current flows through it, the current flowing into a particular element must equal the current flowing out of that particular element, a relationship we call the **current rule** (see section 11.4).[i] Let's review this rule by applying it to single paths and split paths.

[i] For fluids, the reasoning is based off the incompressibility of liquids.

14.1.1 Current along a single path

When there is only one path, the current rule means the amount of current that flows through one element along that path *has* to equal the current flowing through the next element along that same path.

⁝ For bulbs along a single path, the current must be the same through each bulb.

As mentioned on page 11.4.2 in section 11.4, some physicists refer to a single path arrangement of elements as having the elements **in series**.

This is why ammeters (which measure current) must be placed in such a way that the ammeter is along the same path as the element through which we want to measure the current. Only then will the current through the ammeter (which is really what the ammeter is measuring) be the same as the current through the other element.[ii]

Consistent with the current rule, if you have a single path for electrons to flow through, an ammeter will read the same amount of current regardless of where it is placed along that path.

If we don't have an ammeter, can we tell that the current is the same throughout that single path?

⁝ For an ideal bulb, the greater the current flowing through it the brighter it is.

If you have a set of identical bulbs, you can use the bulbs to show that the current is the same throughout a single path. This is because the brightness of the bulb is related to how much current is flowing through it.[iii] By placing a bunch of identical bulbs along a single path, one can see that the current is the same everywhere along that path because all the bulbs have the same brightness.

What about the wires leading up to the bulb? Is the current the same through the wire as through the bulbs?

Yes, the current is the same through the wire since the wire and bulbs are

[ii] And, since ammeters have very low resistance (meaning it is easy for electrons to flow through them; see section 11.4.3), the current with the ammeter along the path is the same as what the current would be without the ammeter.

[iii] As mentioned in section 11.2, this is because electric current consists of electrons flowing through the wire and, as the electrons flow through the wire, the electrons bump into the atoms in the wire. This is much like how children running through a crowded room will pump into the people in the room. This results in the wires getting hot, and light can be emitted if the material gets hot enough. As we know, an incandescent bulb won't light at all if the current is low enough, but as long as it is "hot enough" then we can use the brightness to get a qualitative sense of the current.

along the same single path. It is just that the wires have so little resistance that they don't get hot when the electrons flow through them.[iv]

✓ *Check Point 14.1: In which case would a particular bulb be brighter, assuming there is enough current flowing for it to be lit: when 3 A of current flows through it, when 5 A of current flows through it, or doesn't it matter?*

What if the bulbs aren't identical?

If the bulbs aren't identical then they'll have different brightness but the current through each must still be identical since there is only one path. The difference is simply that some bulbs will get hotter than other bulbs for the same current. After all, the current in the wires connected to a bulb must be the same as the current in the bulb itself, but the wires don't get hot at all since they have no resistance. Just because the resistance is different doesn't mean the current is different.

✓ *Check Point 14.2: Suppose we have several identical bulbs all arranged along a single path with a battery. Should the bulbs have identical brightness or should the first in line be brightest, followed by the next one, and so on down the line? What if they were not identical?*

14.1.2 Current when the path splits

As mentioned on page 200 (section 11.4), when a wire splits into two, or two wires combine into one, we call that a **junction**. Because each element in the circuit remains neutral, the total current flowing into a junction equals the total current flowing out of the junction, consistent with the current rule (see section 14.1.1).

⁝ The total current flowing into a junction equals the total current flowing out of the junction.

When the circuit splits at a junction, some of the current takes one path and some takes the other path. So, for example, if a wire carrying 10 A of

[iv] Although **superconductors** do exist, which have no resistance at all, our wires are not superconductors. However, their resistance is so much smaller than the resistance of the filaments in our bulbs that we can ignore the resistance of the wires. In chapter 15 we'll examine situations where we can no longer ignore the resistance of the wires.

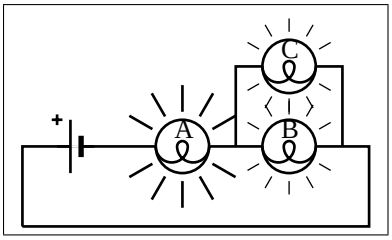
current splits into two parts then, according to the current rule, one of the parts could carry 3 A while the other carries 7 A (for a total of 10 A).

The actual split will depend on the elements that are present along each path. However, the current leading up to the split (e.g., 10 A) must be greater than the current through any path after the split (e.g., 3 A and 7 A). This also means that, for identical bulbs, bulbs on the path prior to the split must be brighter than bulbs on the separate paths after the split.

The reverse occurs for situations where the current in two wires combine into one. Again, the total current flowing prior to the combination must equal the total current flowing after the combination. So, for example, if a wire carrying 3 A of current combines with another wire carrying 7 A, the combined wire would carry a current of 10 A. This also means that, for identical bulbs, bulbs on the path after the junction must be brighter than bulbs on the separate paths prior to the junction.

✍ As mentioned on page 11.4.2 in section 11.4, some physicists refer to a split path arrangement of elements as having the elements **in parallel**.

✓ *Check Point 14.3: Suppose we have the circuit illustrated to the right, with three identical bulbs. Why is bulb A brightest? Would it still be brightest if the bulbs were not identical?*



14.2 Why voltage?

From the previous section, we know that along a single path, when the electrons only have a single path to flow through with no splits anywhere along the path, the current in one part of that path must be the same as the current in any other part of that path. However, the *value* of that current will depend on how strong the battery is, what **elements** (like bulbs) are along the path and how the elements are arranged. As you might already be aware, the battery strength is given in terms of its voltage.

While the terms we've used previously to predict interactions (forces, fields and energy) can still be used with circuits[v], the reason we use voltage is

[v] For example, we can say that the battery exerts a *force* on the electrons, pushing them

because voltage, unlike the others, is solely a property of the battery and does not depend on what is attached to the battery.[vi]

The battery **voltage** essentially indicates how hard the battery pushes and pulls on the electrons. Just like a pump pushes water out one end of the pump and pulls water in the other end of the pump, a battery pushes electrons out one end of the battery and pulls electrons in the other end of the battery. Certainly, whether the electrons actually move, and how much they move, depends on what is connected to the battery, but how hard the battery pushes/pulls, and thus the battery voltage, depends solely on the battery, not what is connected to the battery.

For example, a water pump can be pushing really hard but if the pipe is plugged, no water will flow. Similarly, a battery can have a high voltage but if there is a break in the circuit no electric current will flow. Also, a narrower wire (like a narrower pipe with water) is harder for current to flow through than a wider wire, so for the same battery (pushing just as hard through each wire) more current will flow when a wider wire is connected to it than when a narrower wire is connected to it.[vii]

So there are *two* things that influence how much current flows: the voltage of the battery (how hard the battery pushes on the electrons) and the resistance of the circuit (how hard it is for the electrons to flow through the circuit). In chapter 15, we'll examine how we can predict the current when given the voltage and the resistance. For now, let's explore qualitatively how the voltage and current differ in various situations.

✓ *Check Point 14.4: Does the battery voltage depend on what is connected to the battery? What about the current flowing to/from the battery?*

---

through the circuit. We could likewise say that the battery creates an electric *field* within the circuit. And we could say that the battery provides *energy* to the circuit, which is then dissipated in the various elements within the circuit (like bulbs).

[vi] While we will take this as true for the purpose of this examination, and it is true enough to stand as the reason for using voltage in the first place, in chapter 15 we'll find that the voltage of *real* batteries, and particularly *cheap* batteries, does depend slightly on what is connected to it.

[vii] Note that if a single wire is wider in one place and narrower in another, the current will be the same through both parts of the wire (in keeping with the current rule). However, the value of that current depends on the wire structure, as well as the battery voltage.

14.3 Electric potential

The water pump analogy can help give us a better sense of what voltage is.[viii] As mentioned in chapter 13, a water pump creates a pressure difference between the inlet (where the water is drawn in) and the discharge tube (where the water is pumped out), with the pressure being higher in the water leaving the pump than in the water entering the pump.

In a circuit, the voltage is analogous to the pressure difference across the pump, in that the "electric pressure" is greater on the side of the battery that is pushing the charge through the circuit, with lower "electric pressure" on the other side of the battery, which is simultaneously sucking in the charge back to the battery. I've put the phrase "electric pressure" in quotes because, for a battery, the proper term is the **electric potential**, not electric pressure. Still, the idea is the same.[ix]

WHY IS IT CALLED THE ELECTRIC POTENTIAL INSTEAD OF THE ELECTRIC PRESSURE?

To understand why we call it electric potential, let's compare the current in a circuit to the flow of water in a river. In a river, there is no pump. Instead, the water flows due to a difference in elevation between the beginning and the end of the river. That difference in elevation corresponds to a difference in gravitational potential energy per mass (higher at the higher elevation). Using the same language, then, we can say that electrons flow in a circuit due to a difference in electric potential energy per charge between the beginning and end of the circuit.

Regardless of whether you prefer the phrase "electric potential" or the phrase "electric pressure", we don't actually need to know their values. All we really need to know is the *difference* in electric potential/pressure between the two sides of the battery because it is that difference that drives the charge through the circuit. Indeed, the voltage of the battery is exactly that – it is the *difference* in electric potential/pressure between the two sides of the battery.[x]

⁉ The SI unit of voltage is the volt.

[viii] Like any analogy, it isn't perfect but it does capture the most essential characteristics.

[ix] And, unit-wise, the electric potential is equivalent to a pressure per charge density.

[x] There is a slight difference between voltage and the electric potential difference but there is no difference as long as we restrict our examination to batteries, as we are doing in this part of book.

In the SI system, voltage has units of volts, which we abbreviate as V.[xi] Electric potential is also measured in volts, as is the *difference* in electric potential, which is the voltage.

✍ Recall that an electronvolt (eV) is a unit of energy, not voltage, and is equivalent to $+1.6\times10^{-19}$ J (note: the electron charge is -1.6×10^{-19} C). We'll explain this curious happenstance in section 14.6.

✓ *Check Point 14.5: Suppose a 1.5-V battery is connected to a bulb with some wires and makes the bulb light. What is the difference in electric potential between the two sides of the battery?*

14.4 Voltage across elements

The water analogy not only helps us understand what is meant by the voltage but also helps us predict in a qualitative sense both the voltage across the various elements in a circuit and the current through them.

As mentioned before, current refers to something traveling *through* an object, whether it is water passing through a pipe or electric charge traveling through a wire. Voltage, on the other hand, refers to the electric potential (pressure) difference between the two sides of the object. That is why we refer to current flowing *through* an element but we talk about the voltage *across* an element.

The greater the battery voltage, the greater the current that would flow across a particular circuit when the battery is connected to it. However, the amount of current depends not only on the battery voltage but also on what is connected to the battery, much like how the water flow through a pipe depends not only on the strength of the pump but also on what constrictions are present in the pipe.

Suppose we have a battery connected to a single bulb that is burned out. When a bulb burns out, the filament inside it breaks. Since the two sides of the bulb are no longer connected, no current can flow (since space between

[xi] The volt is named in honor of the Italian physicist Count Alessandro Guiseppe Antonio Anastasio Volta (1745-1827) who studied electricity.

the wires, like air, is an insulator). This is similar to a pipe that is blocked so that water cannot pass by, even with the pump present.

Just as a pressure difference is still produced across the blockage (as the pump tries to push the water onto one side of the blockage and simultaneously tries to suck the water in from the other), there is a non-zero voltage across the burned out bulb (as the battery tries to "push" electrons onto one side and simultaneously "pull" them out the other). No current flows through the bulb, despite the voltage across it, just as no current flows through the blockage, despite the pressure difference across it.

WHAT IF THE BULB WASN'T BURNED OUT?

As mentioned in chapter 13, the same thing occurs if we had a narrowing of the pipe or some similar constriction rather than a complete blockage, in that there would still be a pressure difference across the constriction equal to the pressure difference across the pump. Similarly, the voltage across a single working bulb would be the same as the voltage across the battery.

So, if you had a 1.5-V battery connected to a single bulb, the voltage across the bulb would be 1.5 V. We don't know the current through the bulb, though, as that depends upon the bulb's resistance.

By convention, the electric potential is higher at the positive **terminal** of the battery, so for our 1.5-V battery we could say that the electric potential is 1.5 V at the positive terminal and 0 V at the negative terminal. Similarly, we could say the electric potential is 1.5 V on the "upstream" side of the bulb (the side closer to the positive terminal of the battery, from which the current is flowing) and 0 V on the "downstream" side of the bulb (the side closer to the negative terminal of the battery, toward which the current is flowing). However, the actual values of the electric potential are somewhat arbitrary – all we really know is that the *difference* is 1.5 V.

✓ *Check Point 14.6: (a) Suppose a battery is connected to a bulb and makes it light. Is the electric potential before the bulb different than the electric potential after it?*
(b) Is the current greater before the bulb than after it?
(c) Is the current leaving the battery greater than the current entering the battery?

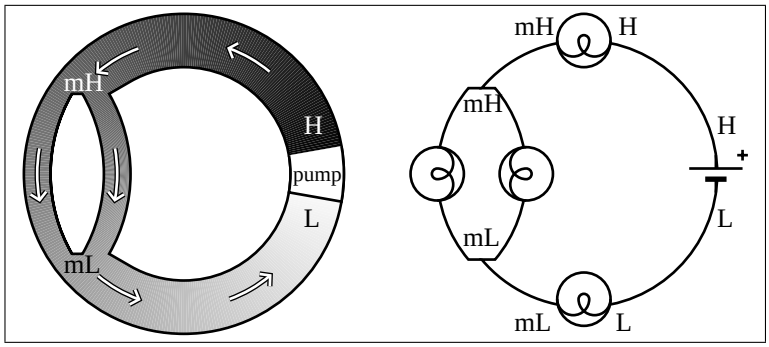


Figure 14.1: An illustration of a pump and pipe alongside an equivalent circuit of a battery and bulbs.

What if we had more than one bulb?

Just as the amount of water flowing into a constriction equals the amount of water flowing out of that constriction, the electric current flowing into a bulb must equal the electric current flowing out of that bulb. And that same amount of current must flow into the next bulb along the line (assuming no split paths). Thus, for a single path, the current must be the same through each bulb along that path.

As for the voltage, it helps to first interpret the situation in terms of the electric potential and water pressure. To illustrate, examine Figure 14.1, which compares the pipe of Figure 13.3 with the equivalent circuit.

In the water pipe, the water pressure is highest at the pump outlet (H) and lowest at the pump inlet (L), and the water is forced from higher pressure to lower pressure through the pipe. In the same way, the electric potential is highest at the positive terminal of the battery (H) and lowest at the negative terminal (L), and positive charges are forced from higher electric potential to lower electric potential through the circuit.

In addition, the pressure is lower the further along the pipe we are, where mH indicates a high pressure but not as high as the pressure at the pump outlet and mL indicates a low pressure but not as low as the pressure at the pump inlet. The reason for the steady decrease in pressure is because of the resistance along the pipe.

Similarly, the electric potential is lower the further along the circuit we are. However, the electric potential only changes where there is a resistance, and

since we are treating wires as having very little resistance that means that there is only a electric potential *difference* across the bulbs. That is why each single piece of wire has the same electric potential at both ends of the wire (indicated by the same letter).

For example, suppose the battery has a voltage of 9 V across its terminals. This means the electric potential could be 9 V at the H locations, 6 V at the mH locations, 4 V at the mL locations and 0 at the L locations. The actual values will depend on the resistances of the bulbs but if the battery voltage is 9 V then there must be a *total* electric potential decrease equal to 9 V from H to L.

That means that across each bulb, the potential difference between the "upstream" side of the bulb and the "downstream" side of the bulb will *not* be 9 V (unless there was only a single bulb in the circuit). For the example numbers provided above, the voltage across the top bulb would be 3 V (subtract 6 V from 9 V), 2 V for the each of the two bulbs on the left side of the circuit (subtract 4 V from 6 V) and 4 V for the bottom bulb (subtract 0 V from 4 V).

Note that the total drop in electric potential, as the current flows from the positive battery terminal to the negative battery terminal, is equal to 9 V. However, the voltage across each element along that path is not 9 V. In general, if you imagine yourself on a boat drifting downstream through the circuit from the positive battery terminal to the negative battery terminal, you'd encounter a total drop in electric potential equal to that across the battery. Indeed, if there was only a single path (no split paths) then adding up the voltages across each element along that single path must equal the voltage across the battery.

❢ For a single-path circuit, the voltage across the battery equals the sum of voltages across the elements along that path.

✓ *Check Point 14.7: Suppose the battery in Figure 14.1 has a voltage of 1.5 V. Would the voltage across the first wire (in upper right part of the circuit; between the battery and the top bulb) be equal to 1.5 V? What about the voltage across the top bulb?*

Are the voltages across each bulb the same?

It depends. The left two bulbs in Figure 14.1 must have identical voltages across them because, for both, the difference in electric potential across the bulb is mH minus mL. We say that those two bulbs "share the same end

points" in the sense that each bulb has a top end and a bottom end, and in the configuration in Figure 14.1 the top ends are connected and the bottom ends are connected.

The voltage across each path (that shares the same end points) is the same.

✍ The idea that the voltage across each path (that shares the same end points) is the same is known as the **voltage rule**.

However, just because the voltages across the two left bulbs must be identical does not mean that voltage value is the same as the voltage values across the other bulbs or even that the voltages across those other bulbs are equal to each other. In other words, H minus mH need not be the same as mH minus mL or mL minus zero.

Compare this to how the currents are related. The current through the battery must equal the current through the top bulb, since there is nowhere else for the current to flow. However, the current splits when it gets to the left part of the circuit, and the current won't split evenly if the bulbs are different. So, the current through each of those two left bulbs need not be identical and will certainly be less than the current through the top bulb. However, the *sum* of the two currents through those two bulbs will necessarily equal the current through the top bulb. And, that total current must also equal the current through the bottom bulb.

Basically, if there is a *single path* of current then the current though each part of that path must be the same but the voltages across each of the elements along that path need not be the same (and won't be if the bulbs have different resistances). Conversely, if the current *splits* and then recombines, as it does on the left side of the circuit in Figure 14.1 then the voltage across each path (from the point where the paths split to the point where the paths recombine) must be the same but the current through each of those paths need not be the same (and won't be if the bulbs have different resistances).

✍ Try not to memorize relationship which is which. Instead, you should be able to come to the proper conclusions by applying the model of what we know about voltage and current.

To illustrate the difference between current and voltage, consider the circuit shown in Figure 14.2. In this case, current flows leftward out of the positive terminal of the battery and then splits at point A, with some current flowing into the bottom bulb and some current flowing into the top bulb. The current then recombines at point B and flows back into the negative terminal of the battery.

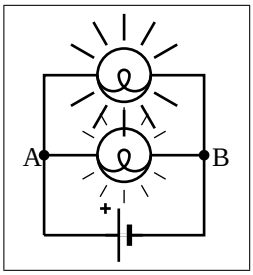
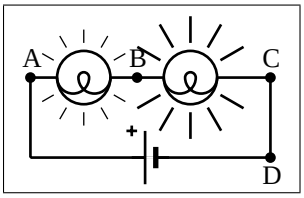


Figure 14.2: Schematics of two non-identical bulbs connected in split paths with a battery.

In this case, the two bulbs are not identical. The top bulb is brighter, suggesting it has less resistance and more current flowing through it. Yet, the voltage must be the same across each bulb since they share the same end points A and B. For example, if the positive and negative terminals of the battery are at electric potentials H and L, respectively, then points A and B would likewise be at electric potentials H and L, respectively, meaning that the voltage across each bulb (H minus L) would equal the voltage across the battery (also H minus L) even though the currents are *not* through each bulb (and not the same as what flows through the battery).

For comparison, with the situation illustrated in the figure to the right we again have two non-identical bulbs but they are not along the same path with the battery. We know they are not identical because they have different brightness even though, by the current rule, the currents must be the same through each.



Basically, the bright bulb has a greater resistance (to be discussed in chapter 15) and this makes it hotter (for the same current) and thus brighter. The voltage across each bulb will be some fraction of the voltage across the battery, and the voltages across the bulbs will not be equal, with a larger voltage across the higher resistance bulb (since the drop in electric potential depends on the resistance).

✓ *Check Point 14.8: Suppose the voltage of the battery is 9 V. If there is a single path with three identical bulbs, what is the voltage across one of the three bulbs?*

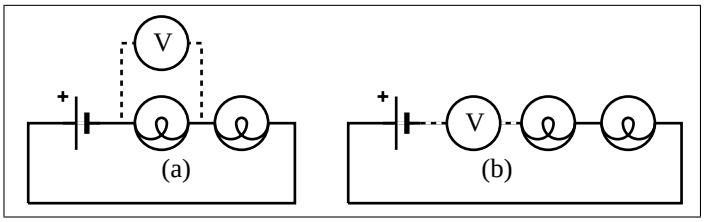


Figure 14.3: Two schematics of two circuits, each with two bulbs connected to a battery. (a) A voltmeter connected in a split path with one of the bulbs. (b) A voltmeter connected along the same path as the bulbs.

14.5 Using a voltmeter

Voltmeters are used differently than ammeters, and you can figure out how to use each by applying what we know about voltage and current. The supplemental readings provide additional information on how to set up a voltmeter but we can figure out how to use a voltmeter by using what we know so far.

Recall that an ammeter measures the current through itself, so if you want to use the ammeter to determine the current through something *else* then you need to place the ammeter in a way such that the current through the ammeter is necessarily the same as the current through that other thing. Similarly, a **voltmeter** (an instrument that measures voltage) measures the voltage across itself so if you want to use a voltmeter to determine the voltage across something *else* then you need to place the voltmeter in a way such that the voltage across the voltmeter is necessarily the same as the voltage across that other thing.

⁉ Voltage is measured by a voltmeter, which is placed such that the voltage across the voltmeter must be the same as the voltage across the element.

To ensure that the voltage across the voltmeter is necessarily the same as the voltage across the element, voltmeters are connected such that the two ends of the voltmeter are the same as the two ends as the element, as illustrated in part (a) of Figure 14.3. In contrast, the voltage would not be the same if the voltmeter was placed along the same path, as in part (b) of the figure. Actually, placing the voltmeter along the same path would would stop the current from flowing, since a voltmeter has a very high resistance (which is why I've used dashed lines in the figure for the voltmeter connections).

Why does a voltmeter have a very high resistance?

This is necessary so that additional current isn't drawn when a voltmeter

is used. For example, our homes have the lights (and appliances) wired in split paths. That way, the voltage across each appliance is the same (120 V), regardless of how many elements we add, each in its own separate path, and the lights don't dim as you turn on more and more lights. As additional paths are added, each with a single bulb, the battery provides the same push to all the paths and so the same current flows through each path. That means the total current (to/from the battery) becomes greater as we add paths.

⁉ The greater the number of paths, the greater the current flowing through the battery (for the same battery).

✓ *Check Point 14.9: It was stated that as more bulbs are added along a single path, the lower the current that flows, despite the voltage across the battery remaining the same. Does the current also lower when additional paths are added, each with a single bulb? What about the voltage across the battery?*

14.6 Voltage as energy per charge

It was mentioned earlier that V (volt) is a unit of voltage but eV (electronvolt) is a unit of energy. Let's examine why that is by first explaining why a battery provides a voltage that is independent of what is connected to it.

The battery drives the electrons through the circuit.[xii] As mentioned on page 185, the energy for this process comes from a chemical reaction inside the battery. For the chemical reaction to take place, electrons need to be taken in through the positive terminal (from the circuit) and *at the same time* electrons must be spit out through the negative terminal (to the circuit). During this process, both the battery and the circuit remain *neutral*.[xiii]

For each electron that is "passed through" the battery as a result of the chemical reaction, a certain amount of energy is provided to the circuit. Thus, the energy *per charge* is independent of how much charge flows (i.e., it is independent of the amount of current that flows or how long it flows).

[xii] Remember that there are *already* free electrons in the circuit. The chemical reaction in the battery doesn't "provide" electrons to the circuit any more than it "takes" electrons from the circuit.

[xiii] Technically, there is a slight charge on each terminal of the battery, which produces an electric field within the wire that pushes the electrons throughout the circuit toward the positive terminal. However, the amount of charge is so slight that the terminals are essentially neutral.

The energy per charge is actually the voltage, so this is saying that the voltage is solely dependent on the chemical reaction inside the battery, which is why the voltage of a D-cell battery is the same as the voltage of a much smaller AAAA-cell battery. Both batteries use the same chemical reaction (e.g., alkaline) and thus provide the same energy to the circuit per electron that passes through it. The only difference is that the D-cell battery has more of the chemical and thus lasts longer (if the current is the same).

⁉ Voltage is equal to the energy per charge.

Using V, E and q for voltage, energy and charge, respectively, the relationship can be written as follows:[xiv]

$$V = \frac{E}{q} \tag{14.1}$$

Since voltage is an energy per charge, multiplying the voltage by the charge of the electron gives the energy provided to each electron. An electronvolt (eV), then, is simply the energy provided to an electron by a 1-V battery. One eV happens to equal $+1.6\times10^{-19}$ J because the electron charge is -1.6×10^{-19} C, and the product of 1.6×10^{-19} C and 1 V equals 1.6×10^{-19} J.

✓ *Check Point 14.10: What is the voltage across a bulb if 1 eV of energy is dissipated for each electron that passes through it?*

The relationship between voltage and energy is useful for determining how much energy is being provided by the battery. Normally, however, the expression is rewritten in terms of power, the rate at which energy is converted. If we take equation 14.1 and divide both E and q by the time t, we get:

$$V = \frac{E}{q} = \frac{E/t}{q/t}$$

We then recognize that the power is equal to the energy provided/dissipated divided by the time, and that the current (i.e., the rate at which charge flows through the element) is equal to the charge divided by the time:

⁉ Voltage is equivalent to the power per current.

$$V = \frac{E}{q} = \frac{E/t}{q/t} = \frac{P}{I}$$

[xiv] Notice that the abbreviation for the voltage value is the italicized V, whereas the abbreviation for the volt unit is the non-italicized V.

This states that the voltage across an object is equal to the rate that energy is transferred by the object divided by the current flowing through the object. Rearranging for power, we get:

$$P = IV \tag{14.2}$$

Appliances (like a TV or a bulb) are rated in terms of the wattage. Given that all of these appliances are used with 120 V outlets[xv] (in the United States), we can use this relationship to determine the current that flows.

Example 14.1: Calculate the current drawn for a 60 W light bulb when connected to a 120-V outlet.

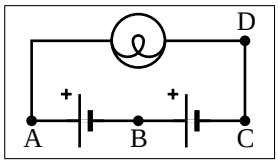
Answer 14.1: Since the power is the product of the current and the voltage, divide both sides to get that the current is the power divided by the voltage. That means that the current through the bulb is the power (60 W) divided by the voltage (120 V), which is 0.5 A.

✓ *Check Point 14.11: Calculate the current drawn for each of the following items and compared to the current drawn by a 60-W bulb. Assume that each item runs on 120 V.*
(a) 150 W TV set
(b) A 1.2 kW iron

14.7 Batteries in different configurations

Now that we've examined a circuit with bulbs in different configurations, let's consider batteries in different configurations.

To the right I've drawn a schematic of two batteries along the same path with a light bulb. In this case, the two batteries each provide a certain amount of energy to the circuit and the voltage from A to C is the sum of the voltages across each battery.



[xv]The 120 V value isn't really provided by a battery; we'll learn about the difference between batteries and standard electrical outlets in chapter 16.

For example, if the left battery had a voltage of 3 V across it (from A to B) and the right battery had a voltage of 5 V across it (from B to C) then the total voltage from point A to point C would be 8 V.

If such a "dual" battery is connected to a bulb, the bulb would be brighter than if used with either battery individually.

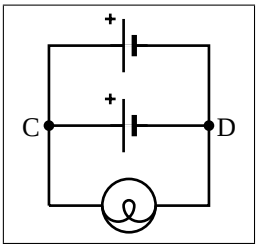
What if the orientation of one of the batteries was switched?

In that case, the voltages subtract.

✍ Many batteries aren't designed to have charges flow through it in a direction opposite the intended direction, so I don't suggest you orient the batteries in opposing directions.

✓ *Check Point 14.12: Two light bulbs are arranged along the same path with two batteries, such that the "+" terminal of one battery is connected to the "−" terminal of the other, as in the schematic above.*
(a) If the first battery has a voltage of 3.0 V across it and the second battery has a voltage of 2.0 V across it, what is the voltage across both together?
(b) What would be the voltage across both batteries if one of the batteries was flipped such that its "+" terminal was connected to the "+" terminal of the other?

Now let's consider two identical batteries arranged in two separate paths, as illustrated in the schematic to the right. There are three paths between points C and D. Not only do we have the two batteries but we also have the bulb. Since all three share the same endpoints the voltage across the bulb is the same as that across each battery.



Would the bulb be brighter with the two batteries arranged in two separate paths than with a single battery?

For ideal[xvi] batteries, the bulb brightness would be the same since the voltage across the bulb is the same either way. The only difference is that each battery would only need to provide half the current and half the energy, thus extending the life of each battery.

[xvi] As will be discussed in section 15.4.2, the bulb may be a little brighter due to the internal resistance of the batteries.

✍ To simplify things, we won't consider situations with batteries of unequal voltage arranged as shown in the figure. However, even in that situation, the voltage across each must be the same since the two batteries share the same end points. It is just that the current may flow "backwards" through one the batteries.[xvii]

✓ *Check Point 14.13: Suppose each battery in the split-path circuit on page 257 had a voltage of 1.5 V. What would be the voltage across the bulb?*

Summary

This chapter defined what voltage and how it impacts a circuit. The main points of this chapter are as follows:

- For bulbs along a single path, the current must be the same through each bulb.
- For an ideal bulb, the greater the current flowing through it the brighter it is.
- The total current flowing into a junction equals the total current flowing out of the junction.
- The SI unit of voltage is the volt.
- For a single-path circuit, the voltage across the battery equals the sum of voltages across the elements along that path.
- The voltage across each path (that shares the same end points) is the same.
- Voltage is measured by a voltmeter, which is placed such that the voltage across the voltmeter must be the same as the voltage across the element.
- The greater the number of paths, the greater the current flowing through the battery (for the same battery).
- Voltage is equal to the energy per charge.
- Voltage is equivalent to the power per current.

[xvii] This may destroy the battery or "recharge" it, depending on its structure.

Frequently asked questions

Does the battery have the same voltage across it whether it is connected to a circuit or not?

Unless otherwise stated, we will assume that the battery is able to maintain the same voltage across its two terminals regardless of the rate at which current flows.[xviii]

Why do the electrons move in the wire and not the protons?

As mentioned in chapter 10, the proton is almost two thousand times heavier than the electron. In addition, according to our model, protons are found in the nucleus and not free to roam around by themselves. These two aspects of our model suggest that the electrons will move rather than the protons.

Do larger batteries have a higher voltage?

Not necessarily. The voltage depends on the chemical reaction occurring inside the battery, not the size of the battery. A larger battery has more chemicals, though, so it could provide a larger amount of energy to the circuit. So, for the same circuit, a larger battery would last longer.

What's the difference between a volt and an electron-volt?

An electron-volt (see chapters 8 and 9) is a unit of energy, not voltage.[xix]

Does any current go through the voltmeter when it is used?

We've been assuming that no current flows through the voltmeter, so that the current through the circuit is unaffected by its use and the circuit continues to run as it did before. This is why we do not consider the voltmeter to be "part" of the circuit and why adding the voltmeter in a separate path does not "short out" the battery (unlike an ammeter which provides an easier path for current to reach the other side of the battery).[xx]

Is the negative terminal of a battery negatively charged?

[xviii] Typical batteries cannot maintain the same voltage when current is drawn and, as such, the voltage across the terminals tends to be a little lower while current is being drawn. This is explored in chapter 15.

[xix] You can think of one electron-volt as energy gained by a single electron when experiencing a voltage of 1 Volt. Mathematically, rearranging equation 14.1 shows that the energy of an electron is equal to the product of its charge and the voltage experienced ($E = qV$).

[xx] In reality, a little bit of current goes through the voltmeter (which allows it to make

We've been assuming the battery terminals are neutral. Indeed, if we use the balloon test from chapter 2, we'll find that pieces of paper are not attracted the sides of the battery.[xxi]

What happens if two batteries share the same end points and the orientation of one of batteries is switched?

This would essentially be a **short-circuit**, as the current would just be going in a circle through the batteries. This would be a safety hazard, as the wires can get hot and start a fire. Either that or the batteries would die out. More information on short circuits is provided in chapter 15.

If all of the energy supplied by the battery is lost to heat in the resistor, does that mean the electrons stop as soon as they pass through the resistor?

No. The battery provides the energy to the circuit. That energy is dissipated through the resistors and bulbs and such. The same chemical reaction that provides the energy also provides electrons through one terminal, but it also *collects* electrons through the other terminal. So, the battery doesn't really provide electrons to the circuit.[xxii]

When a path splits in a circuit, does current get split in half?

Not necessarily. It depends on which path has less resistance.

Does 100% of the current follow the path of less resistance?

No. The fraction of the current that follows one path or the other depends on the relative resistances of the two paths.

---

the measurement). Indeed, if too much current flows through the voltmeter (i.e., > 300 mA or so), it can destroy the meter. For this reason, the meter usually has a little fuse along the same path as the voltmeter. If too much current flows through the fuse, it melts and breaks the path through the voltmeter.

[xxi] Much like how the density of water is probably slightly different at the pump outlet than at the pump inlet, there is probably a tiny amount of charge on each terminal since the chemical reaction inside the battery produces an electron on one terminal and removes an electron from the other. However, the difference is insignificant and we can consider each terminal to be neutral.

[xxii] Some confusion might result from the money analogy I used in chapter 7 to account for the energy. As we follow a charge in its path through the circuit, I refer to it collecting and dissipating the energy, like money. However, that energy doesn't *belong* to the individual charge we are riding. In other words, the charge doesn't go faster when we collect charge and slower when we dissipate it.

Terminology introduced

Circuit	In parallel	Terminal
Current	In series	Voltage
Current rule	Junction	Voltage rule
Electric potential	Superconductors	Voltmeter
Elements	Short-circuit	

Abbreviations introduced

Quantity	**SI unit**
voltage (V)	volt (V)[xxiii]
power (P)	watt (W)[xxiv]

Additional problems

Problem 14.1: A 5-V battery is applied to a bulb and makes it light. The bulb is then replaced with a second bulb, which when placed in the circuit is dimmer than what the first bulb was. Which of the following is the same for both cases? Choose all that apply.
(a) The voltage across the bulb.
(b) The current flowing through the bulb.
(c) The rate at which energy is dissipated by the bulb.
(d) The ratio of (c) divided by (b).

Problem 14.2: Two non-identical light bulbs are arranged along a single path with a 5-V battery. If bulb A is brighter than bulb B, then:
(a) Through which bulb is the current greater or are they equal?
(b) Which bulb has a higher power value (i.e., is converting electric energy to heat and light energy at a greater rate) or are they equal?
(c) Across which bulb is the voltage greater or are they equal?

Problem 14.3: Suppose a 1.5-V battery is connected to a circuit such that 50 mA of current flows. During one minute, how much energy does the battery

[xxiii] A volt is equal to a joule per coulomb (J/C).
[xxiv] A watt is equal to a joule per second (J/s).

provide to the circuit? (hint: first find out the amount of charge that flows through the circuit)

Problem 14.4: A wire is used to connect the positive terminal of a battery with its negative terminal. Assume that a voltage of 1.5 V exists across the positive and negative terminals of the battery.
(a) How much kinetic energy would an electron obtain by traveling along the wire from the negative to the positive terminals? Assume no other forces are doing work on the electron (note: This assumption is most likely not valid as electrons in the wire lose their kinetic energy by bumping into the atoms in the wire; the wire heats up as a result).
(b) Suppose the wire connecting the two ends of the battery is 1 m long. How fast would an electron be traveling by the time it reached the end of the wire? Again, assume no other forces are doing work on the electron.
(c) A typical drift velocity (of the electrons) is less than a millimeter per second. How does the speed obtained in (b) compare to the typical drift velocity? Why is it so different?

Problem 14.5: Suppose that a proton was sent through a circuit instead of an electron. Furthermore, suppose that, as with the electron in checkpoint 14.10 but from the opposite direction, 1 eV of energy is dissipated by the light bulb for each proton that passes through it. What is the voltage across the light bulb?

15. Resistance

Puzzle #15: A typical car battery has a voltage of 12 V. There are also 12-V lantern batteries. Why can't you use a (much cheaper) 12-V lantern battery in a car? Or eight 1.5-V flashlight batteries?

Introduction

If all batteries were like the ideal batteries described in chapter 14, there would be no difference between a 12-V car battery and a 12-V lantern battery.

Real batteries have something called **internal resistance**, which impacts how well the battery can provide the voltage under different conditions. We'll examine internal resistance in section 15.4.2 but first we have to define resistance. This will not only allow us to explain internal resistance but will also help us make quantitative predictions about voltage and current (as opposed to the purely qualitative predictions we made in chapter 14).

So, once resistance is defined, I'll apply it to various circuit configurations so that you can get comfortable with how it is related to current and voltage. I'll then examine real wires, which have a small resistance, and how the wire's resistance impacts the circuit. Once I am done with all of that, I can examine the internal resistance of real batteries.

15.1 Definition of resistance

We have used resistance before but only in a qualitative sense. Basically, we recognized that the higher the resistance, the less current that flows for a given applied voltage.

✍ Since current also depends on the applied voltage (more current when voltage is greater), we need to be careful to specify that we are not changing *both* the voltage *and* the resistance. That is why I add the phrase "for a given applied voltage," to ensure that we are only changing the resistance and seeing the impact of only that change on the current, not *also* changing the voltage.

For example, good conductors have a low resistance and allow a lot of current to flow for a given applied voltage. Good insulators, on the other hand, have a high resistance and allow little current to flow for a given applied voltage.[i]

✍ The resistance, in turn, depends on the material, the length and thickness of the material and the temperature of the material.

⁝ An element's resistance is defined as the ratio of the voltage across it to the current through it.

We define the **resistance** of an element as the ratio of the *voltage* across the element to the *current* through the element.

$$R_{\text{of element}} = \frac{V_{\text{across element}}}{I_{\text{through element}}} \tag{15.1}$$

Notice how the current, I, is in the denominator. That way, a large current means a low resistance.

What are the units of resistance?

The ratio of voltage to current has units of volts per ampere (V/A). However, we hardly ever indicate the resistance with those units. Instead, we replace this ratio with the unit **ohm**.[ii]

⁝ Resistance is measured in ohms, abbreviated as Ω.

The ohm unit is abbreviated as Ω, the Greek letter omega. It may seem strange to use a Greek letter for a unit, but abbreviating ohm with the letter "O" would be confusing, since it looks like a zero.

✓ *Check Point 15.1: It has been found that currents of 100 mA through the body can be fatal. What resistance will allow 100 mA to flow when a voltage of 120 V (as in a standard electrical outlet) is applied?*

---

[i] Some materials fall somewhere in between and still others are either conductors or insulators depending upon the temperature or direction of current (e.g., semiconductors).

[ii] This unit is in honor of Georg Simon Ohm (1787-1854), a Bavarian (Germany) physicist who studied electricity.

WHY WRITE "OF ELEMENT", "ACROSS ELEMENT" AND "THROUGH ELEMENT" AS SUBSCRIPTS IN EQUATION 15.1? IS IT REALLY NECESSARY?

Normally, people do not write the subscripts. There are two reasons I do so.

First, I want to remind you that these are three different quantities. Resistance is a property of the element (thus why I write *of* element), current is a flow of charge through the element (thus why I write *through* element) and voltage is like force on the element (thus why I write *across* element, as we need to know the force on each side to know the net force on it).

Second, it is deceptively easy to apply this equation inappropriately if you don't make a conscious effort to apply each quantity in the equation to the *same* element. A given circuit can have lots of different things in it and the resistance, current and voltage can be different depending on which part of the circuit we are examining. You need to be careful that you don't plug into the equation the value of the voltage across one part of the circuit and the value of the current through a *different* part of the circuit.

For example, consider the following scenario:

> Two bulbs (A and B) are arranged in a single path with a battery. Suppose the current through the circuit is 30 mA, the voltage across bulb A is 0.5 V and the voltage across bulb B is 1.0 V. If we want to find the resistance of bulb A, which V do we use in the equation $R = V/I$: the voltage across bulb A, the voltage across bulb B, or the voltage across both light bulbs together?

Here we are given *two* voltages, one across bulb A and one across bulb B. Which we use depends on which element we are applying the resistance relationship (equation 15.1). The subscripts remind us that *bulb A*'s resistance is related to the voltage *across bulb A* and the current *through bulb A*.

In this case, if we want the resistance of bulb A, we need to use the current through bulb A (30 mA; since the current is the same throughout the circuit) and the voltage across bulb A (0.5 V).

For clarity, after this section I will write the equation without subscripts. However, make sure you still "say" the subscripts to yourself when you write out the equations.

✓ *Check Point 15.2: If I want to use the definition of resistance (equation 15.1) to determine the resistance of a particular bulb in a circuit, should I use the voltage of the battery, the voltage across the entire circuit or the voltage across the bulb? Why?*

How do we use the definition of resistance equation?

When using the definition of resistance equation, you need to recognize that it is a relationship between *three* quantities. Usually you are given two of the three (or can obtain two of the three) so that you can then use the equation to get the third. However, there can also be the situation where one of the three is held constant, a second is changed, and the task is to figure out what happens to the third.

For example, if R is held constant, then a larger V would mean that I would have to be larger also, since R is equal to the ratio V/I and, if R is held constant, then so must the ratio V/I, and the only way that can happen (if V is larger) is if I is larger also.

In a similar way, if V is held constant, then a larger R would mean that I would have to be smaller, since R is equal to the ratio V/I and if the ratio is larger (because R is larger) then I must be smaller since I is in the denominator.

Finally, consider the case where I is held constant, which means a larger R must correspond to a larger V since R is equal to the ratio V/I and if R is larger then the ratio V/I must be larger and so V must be larger as well since V is in the numerator.

How do we know which of the three is held constant in a given situation?

It depends, of course, but there are three general scenarios.

Each object has a specific resistance and, for a certain class of objects called resistors, we can assume their resistance is constant and doesn't depend on the situation. In that case, if the current is different that means the voltage must be different also and, conversely, if the current is different that means the voltage must be different (with larger voltage meaning larger current, consistent with $R = V/I$ and what was discussed in the first case above). Resistors are discussed in section 15.2.

The second scenario we can consider is the split path, where the voltage is the same across each path. If the resistance is different for each path, the current must be different as well (with larger current through the path with smaller resistance, consistent with $R = V/I$). This situation is discussed in section 15.2.1.

The third scenario we consider is the single path, where the current must be the same through every object along the path. For different objects along the path, with different resistances, the voltage must be different as well (with larger voltage across higher-resistance objects, consistent with $R = V/I$). This situation is discussed in section 15.2.1.

Be careful! The relationship between resistance, voltage and current is like the relationship between density, mass and volume. Just because density is proportional to mass (both in numerator) and inversely proportional to volume does *not* mean that higher-density objects have more mass and a lower volume than lower-density objects. After all, a penny has a higher density than wood, but it has LESS mass than a tree (since its volume is so much less). All we know is that higher-density objects have a greater *ratio* of mass to volume than lower-density objects. In a similar way, higher-resistance objects do not necessarily correspond to higher voltage and lower current values. All we know is that the *ratio* is higher.

✓ *Check Point 15.3: Suppose we had three objects in a single path with resistances of 1 Ω, 2 Ω and 3 Ω.*
(a) Would the current be the same through each object? If not, through which object would the current be greatest?
(b) Would the voltage be the same across each object? If not, across which object would the voltage be greatest?

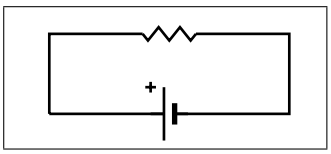
15.2 Resistors

In circuit schematics, I've been using lines to represent wires. Since we've been assuming the wires have negligible resistance (compared to the resistance of the elements like bulbs and such), it really doesn't matter how long the wires are drawn. Their drawn length is arbitrary. In addition, for clarity

they are typically drawn as straight lines whether they are crooked, curved or straight in the circuit.

Since wires have very little resistance, a circuit needs to have elements that have some significant resistance to prevent the current from being very high.

Such elements are called **resistors** and in schematics are indicated by a jagged line (see schematic to the right of a battery connected to a resistor).



Resistors are designed such that their resistance (defined as V/I) remains the same regardless of how much voltage is applied. If the voltage goes up, the current goes up in the same proportion (assuming we haven't changed something else, like the temperature[iii]).

For the purpose of our analysis of circuits, then, we will assume that every resistor has a resistance that never changes, regardless of the current flowing through it or the voltage across it.[iv]

⁉ We will assume that the resistance of a resistor is always the same, regardless of the current.

✍ While it is reasonable to make this assumption for resistors, it would not be a reasonable assumption in all cases. For example, a material's resistance depends on its temperature, and the temperature of an incandescent bulb is significantly different when it is on (and hot) vs. when it is off (and cold). Resistors tend to be thicker than wires and significantly thicker than the filament of an incandescent bulb, so that the heat can be "spread" through a larger volume of material and thus dissipated more quickly, minimizing the changes of significant temperature changes.

If an object's resistance has a particular value that never changes (which we will assume is true for resistors in general) then we can use the definition of resistance, $R = V/I$ to determine a resistor's resistance. Alternatively, if we already know the resistor's resistance then we can use variations of the definition, $V = IR$ and $I = V/R$, to determine the voltage across the resistor and the current through the resistor.

For example, if we are given the resistor's resistance and the voltage across the resistor, we can use $I = V/R$ to determine the current through the

[iii] We happen to know that the temperature of a light bulb changes a great deal between when it is on as opposed to off and this will change its resistance.

[iv] The idea that the resistance is independent of the voltage is known as **Ohm's law** in honor of the person who first discovered the relationship. If the resistance of a material follows Ohm's law then we say that the material is **ohmic**.

resistor. Likewise, if we are given the resistor's resistance and the current through the resistor, we can use $V = IR$ to determine the voltage across the resistor.

✓ *Check Point 15.4: Suppose that when placed in a circuit the current through a 2-Ω resistor is 0.6 A. What is the voltage across the resistor?*

Not only can we use the resistance and its definition to predict the current (given the voltage) but we can also predict the rate of energy dissipation and, correspondingly, how bright a particular bulb may be.

For example, in chapter 14 we saw how two bulbs in separate paths may shine with different brightness even though they have the same voltage across them (see page 252). At the time, we explained it by saying that more current passed through the brighter one. Now we can explain it in terms of the resistance.

As shown before, the **power** or rate at which energy is provided to the circuit by the battery (or dissipated by some other element) can be related to the current through the element and the voltage across the element (see equation 14.2):

$$P = IV$$

Using the definition of resistance, we can replace I by V/R to get[v]

$$P = \frac{V^2}{R}$$

For two bulbs in a split path (sharing the same end points), the voltage across each is the same and so this expression tells us that the only thing affecting the power (brightness) is the resistance. Furthermore, since the power is inversely proportional to the resistance, it shows that for bulbs in this configuration, the lower resistance bulb will dissipate energy at a greater rate and thus be brighter.

This is consistent with what we had before. Basically, the current splits unevenly, with more current flowing through the one with lower resistance.

[v] Although I've removed the subscripts, the equation still relates the power lost through a *particular element* with the resistance *of that particular element* and the voltage across *that element*. I do the same for the equation after this (equation 15.2).

Since the voltage across each bulb is the same, the one with more current will burn brighter.

Example 15.1: A 1-V battery is applied to a 2-Ω resistor. How much energy is lost to heat in 1 minute?

Answer 15.1: The power loss P is $V^2/R = (1\text{ V})^2/(2\ \Omega) = 0.5$ W. Multiply by time to get the total energy lost = $(0.5\text{ W})\times(60\text{ s}) = 30$ J.

In chapter 14 we saw the opposite occur with two bulbs in a *single* path, where the higher resistance bulb is brighter. This is because the current has to be the same through each bulb when placed along a single path. The higher resistance bulb gets hotter and thus burns brighter. It is like having two brake pads on the same tire. The one that is held tighter against the wheel will get hotter.

This can be shown mathematically by taking the $P = IV$ expression and using the definition of resistance to replace V by IR:

$$P = I^2R \tag{15.2}$$

For two bulbs along the same path, the current through each is the same and so this expression tells us that the only thing affecting the power (brightness) is the resistance. Furthermore, since the power is directly proportional to the resistance in this case, it shows that a high-resistance bulb will dissipate energy at a greater rate and thus be brighter.

Example 15.2: A battery is applied to a circuit containing two bulbs along the same path. Which bulb is brighter, the one with the higher resistance or the one with the lower resistance?

Answer 15.2: The energy loss (per time) via each bulb depends upon the current through the bulb and the resistance of the bulb. Since the current is the same through each bulb (being along the same path), the difference is due to the resistance. The power loss will be greater through the bulb with the higher resistance. This means that the bulb with the higher resistance will most likely be brighter (i.e., the energy is lost via light).

✓ *Check Point 15.5: Two bulbs, A and B, are along the same path, which means the current through them is the same. Bulb A is brighter than bulb B.*
(a) Which is larger: the voltage across bulb A or the voltage across bulb B? Why?
(b) Which is larger: the resistance of bulb A or the resistance of bulb B? Why?

15.2.1 Elements in split paths

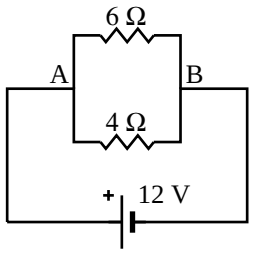
To illustrate the use of resistance when dealing with circuits, I'm going to apply the idea to circuits in various configurations.

We basically have three ideas that govern the voltage and current in a circuit:

1. **Current**: the current flowing into a particular element or location must equal the current flowing out.
2. **Resistance**: The resistance of an element is constant and equal to the voltage across it divided by the current through it ($R = V/I$).
3. **Voltage**: Each path between two points must have the same voltage across it.

With these ideas, we can solve for the current. We've done this so far with simple circuits. Now let's try it with more complicated circuits.

To illustrate, let's consider a circuit containing a 4-Ω resistor and a 6-Ω resistor in a split path configuration, as illustrated to the right. What is the current through the 4-Ω resistor?



⁝ All resistors that share the same end points have the same voltage across them.

Regarding voltage, we know that the voltage across the 4-Ω resistor must be the same as that across the battery (12 V).

Knowing the voltage across the resistor (12 V) and the resistance of the resistor (4 Ω), we plug into $V = IR$ (from the definition of resistance) to get the current through the resistor (which turns out to be 3 A).

We can solve for the current through the other resistor the same way. The voltage across that resistor is also 12 V. Knowing the voltage across the

resistor (12 V) and the resistance of the resistor (6 Ω), we plug into $V = IR$ to get the current through the resistor (which turns out to be 2 A).

Although it wasn't asked, we can apply what we know about current to conclude that the total current must be 5 A, the sum of 3 A and 2 A, because the two paths combine into one.

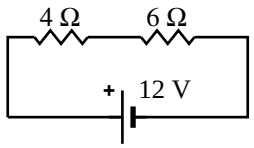
✍ Adding resistors in separate paths makes the current *greater* than it would be with any *single* one of the resistors (the opposite of what happens when adding resistors in a single path). In other words, the overall resistance *decreases* with additional resistors in separate paths.

✓ *Check Point 15.6: Suppose the battery in the circuit above (with two resistors in separate paths) has a voltage of 1.5 V and the two resistors have resistances of 2 Ω and 3 Ω. What is the total current in the circuit?*

15.2.2 Elements along a single path

Determining the current when the resistors are in separate paths is pretty straightforward since the voltage across each resistor is the same, and that voltage equals the voltage across the battery. We don't have that luxury when the resistors are along a single path, which makes the math a little more complicated. However, the physics is the same.

For example, consider the 4-Ω resistor and 6-Ω along a single path, as illustrated to the right. What is the current through the 4-Ω resistor?



To solve the problem, let's think about how the three ideas of current, resistance and voltage apply. With current, we know that the current must be the same through both resistors (since there is only one path for the charge to follow).

⁉ You cannot arbitrarily assign values of V, I and R when using $V = IR$. All three must correspond to the *same* element.

From the definition of resistance we know that $V = IR$ for each individual resistor. We want to find the current I through the 4-Ω resistor. We have the resistance R of the 4-Ω resistor but we do not have the voltage across the 4-Ω resistor.

The voltage across each resistor is not 12 V. That is the voltage across the *battery*, not across the 4-Ω resistor.

Indeed, based on the idea of voltage, we know that the voltage across the 4-Ω resistor must be *less* than 12 V since the dissipation is spread between both resistors. More specifically, the applied voltage by the battery (12 V) must equal the *sum* of the voltages across the individual resistors.

To solve this, let's set I to be the unknown current value.

From the definition of resistance, we know that the voltage across each resistor is $I(4\ \Omega)$ and $I(6\ \Omega)$, respectively. Adding these together, that means the total voltage must be $I(10\ \Omega)$. Since the total voltage is given as 12 V, we can set $I(10\ \Omega)$ equal to 12 V and solve for I to get a current of 1.2 A.

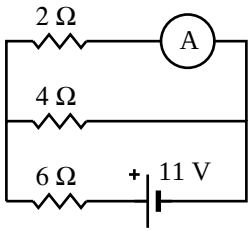
A common short-cut is to add resistances together when resistors are along the same path, as in this case. When you add them together you get a total resistance of 10 Ω. This brings you to the last step, and you can get the current (1.2 A) by dividing the voltage (12 V) by the total resistance (10 Ω).

✍ Adding resistors along the same path making the current *less* than it would be with any *single* one of the resistors. In other words, the overall resistance *increases* with additional resistors along the same path.

✓ *Check Point 15.7: A 1.5-V battery is connected to two resistors, of resistance 1-Ω and 2-Ω, along a single path. What is the voltage across each resistor?*

15.2.3 Mixed circuits

A "mixed" circuit is one that has multiple paths with some paths having more than one resistor or bulb. An example is illustrated to the right. In this case, the current passes through the 6-Ω resistor and then splits between the 2-Ω and 4-Ω resistors.



What is the current through the 4-Ω resistor now?

What makes this difficult is that we don't immediately know the voltage across the 4-Ω resistor. Since the 2-Ω and 4-Ω resistors share the same end points, we know that the voltage across each of them must be the same. However, that voltage is not 11 V. In fact, it must be *less* than 11 V, since

the 11 V must be split between the 6-Ω resistor and the group of two resistors (since voltages add for elements along the same path).

Note that we can't treat the 4-Ω as being in the same path as any other resistor, since the current through the 4-Ω isn't necessarily the same as that through any other resistor.

This can be solved with a little bit of mathematics, which I describe in the footnote.[vi] For the purpose of this discussion, let's suppose that we already know that the current through the ammeter is 1 A (see footnote for details). How do we then get the current through the 4-Ω resistor?

The simplest way to do this is to first determine the voltage across the 2-Ω resistor. We can use $V = IR$ for that one, since we know its resistance (2 Ω) and we know the current flowing through it (1 A, since the ammeter is along the same path as the resistor). That gives us a voltage of 2 V across the 2-Ω resistor.

Now that we know the voltage across the 2-Ω resistor, we also know the voltage across the 4-Ω resistor, since the two resistors share the same end points. The voltage across the 4-Ω resistor must likewise be 2 V.

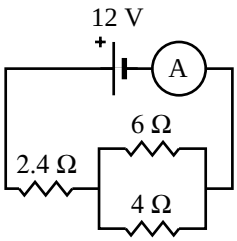
We can use $I = V/R$ for 4-Ω resistor, since we know its resistance (4 Ω) and we know the voltage across it (2 V). That gives us a current of 0.5 A through the 4-Ω resistor.

Note that the current through the 6-Ω resistor must be 1.5 A, since the two currents come together. We can then use $V = IR$ for the 6-Ω resistor, since we know its resistance (6 Ω) and we know the current through it (1.5 A). That gives us a voltage of 9 V across the 6-Ω resistor.

Consistent with how voltage works, the voltage across the bottom path (including the battery and the 6-Ω resistor) is 2 V (subtract 9 V from 11 V), the same as that across the top path and the middle path.

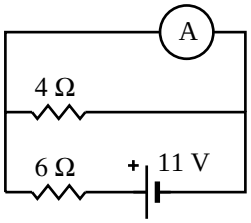
[vi] Let I be the current through the 2-Ω resistor. The current through the 4-Ω resistor will then be half that since the voltage is the same across each and the ratio of their resistances is 2. That means the current through the 6-Ω resistor is $1.5I$, since the two currents add (i.e., $I + 0.5I = 1.5I$). The voltage across the whole circuit is 11 V, which must be split between the voltage across the 6-Ω resistor, $(1.5I) \times (6\ \Omega)$, and the voltage across the 4-Ω resistor, $(0.5I) \times (4\ \Omega)$, or the voltage across the 2-Ω resistor, $(I) \times (2\ \Omega)$, as they'd both be the same. Solve for I to get a current of 1 A through the 2-Ω resistor, which means there must be 0.5 A through the 4-Ω resistor and 1.5 A through the 6-Ω resistor.

✓ *Check Point 15.8: In circuit to the right, the current passes through the 2.4-Ω resistor and then splits between the 4-Ω and 6-Ω resistors, which are in separate paths. Suppose the ammeter measures a current of 1 A. (a) Through which resistors, if any, is the current 1 A? (b) Across which resistors, if any, is the voltage 12 V?*



15.2.4 Short circuits

As a final example, let's suppose we replace the 2-Ω resistor in the previous example with a wire, as illustrated to the right. What is the current through the 4-Ω resistor now?



We can guess that all of the current will take the top path since the ammeter has no resistance, with no current flowing through the 4-Ω resistor. We can show that this is what happens by repeating our analysis from before using a zero resistance for the top path instead of 2 Ω. Since the top path has no resistance, the voltage across it must likewise be zero ($V = IR$). Since the top and middle paths share the same end points, the voltage across the middle path must likewise be zero. Using $I = V/R$, with a voltage of zero and a resistance of 4 Ω, we get a current of zero through the 4-Ω resistor.

This is what is called a **short circuit**.[vii] Placing a wire across an element will cause the current through the element to fall to zero.

Is the current through the ammeter 1 A, as before?

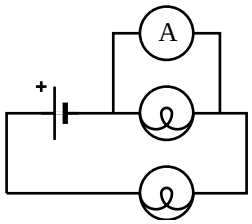
No.

With no current flowing through the 4-Ω resistor, the circuit is essentially a circuit with a single 6-Ω resistor along the same path as an ammeter and an 11-V battery. Using $I = V/R$, we get a current of 1.83 A through the resistor and ammeter. This is larger than what was obtained with the 2-Ω resistor rather than the wire.

[vii] We don't mean *physically* short, as in a smaller circuit. Rather, we mean that we've provided an alternate path around the resistor.

Both findings (increased current through the ammeter and zero current through the element that was bypassed) is why we don't connect ammeters across elements, as we would with voltmeters. Ammeters are designed to have very low resistance.

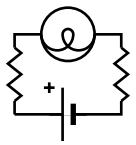
✓ *Check Point 15.9: In circuit to the right, an ammeter is placed across one of the bulbs. (a) One light is off. Which one and why? (b) Do either of the bulbs have the same current as that measured by the ammeter? If so, which?*



15.3 Real vs. ideal wires

We've been assuming that all wires have zero resistance.

That is not technically true. The resistance of wires may be small and insignificant, but it isn't zero. Materials with zero resistance are called **superconductors** and usually require very low temperatures. The wires that we use are not superconductors. We just treat them as such because relatively little energy is lost as the electrons travel through the wires.

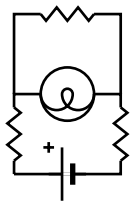


To illustrate the impact that this small resistance has on our measurements, consider the circuit drawn to the right. The resistors to the left and right are meant to represent the small resistance that is in each wire that connects to the bulb. Let's suppose the bulb has a resistance of 6 Ω and the battery has a voltage of 1.5 V. We'll first calculate the current flowing through the bulb with zero-resistance wires and then we'll re-calculate the current assuming a small resistance of 0.1 Ω in each wire.

With no resistance in the wires, the voltage across the bulb is equal to 1.5 V, the voltage across the battery. We can then use $I = V/R$ with $V = 1.5$ V and $R = 6\ \Omega$ to get a current of 0.25 A.

Now let's calculate what the current would be with 0.1 Ω of resistance in each wire. We'll use the short-cut introduced in section 15.2.2 and first calculate the total resistance of the wires and bulb by adding up their resistances (since they are along the same path): $6\ \Omega + 0.1\ \Omega + 0.1\ \Omega = 6.2\ \Omega$. We can then use $I = V/R$ with $V = 1.5$ V and $R = 6.2\ \Omega$ to get a current of 0.242 A.

As you can see, it doesn't make much of a difference.



However, consider what happens if we add a wire across our bulb (to allow the current to bypass the bulb), as illustrated in the circuit drawn to the right. As before, I've added resistor symbols to each wire to represent the small resistance that is in each wire.

Again we'll treat each wire as having a resistance of 0.1 Ω. We now have a split path, with some of the current flowing through the bulb and some flowing through the extra wire at the top of the circuit. However, since the resistance of the bulb is so much more than the resistance of the wire, much more of the current takes the path through the top wire segment rather than through the bulb.

Because of this, the current is practically the same through each of the three wire segments. Furthermore, since the resistance of each wire is the same, the voltage gets split between the three equally. For a 1.5 V battery, that means there is 0.5 V across each wire segment.

Since the bulb shares the same end points as the top wire segment, that means there is also 0.5 V across the bulb, not 1.5 V (as it would be if there was no resistance in the wires). With less voltage across the bulb, it won't be as bright.

✍ As we will see in the next section, with real batteries the bulb likely won't light at all.

One risk of adding a wire to bypass a bulb is that doing so allows a great deal of current to flow through the wires. Ignoring the bulb for the time being, the current through the circuit can be determined by using the total resistance of the wires (0.3 Ω, assuming 0.1 Ω for each wire) and the voltage of the battery (1.5 V). Using $I = V/R$, I get a current of 15 A.

A current of that magnitude through a 0.1 Ω wire dissipates energy at a rate of 22.5 W (obtained by using $P = I^2R$). That is almost as much as a 25-W light bulb when connected to a 120-V outlet.

If you think about the heat and light generated by a 25-W bulb, you can imagine that the wires can get quite warm when something is shorted out, even with a 1.5-V battery. Imagine the fire that might result if the battery was 120 V, as in a typical household outlet.

It is for this reason that houses are equipped with circuit breakers that "trip" and stop the current when the current gets too high, preventing the wires from getting too hot.[viii]

✓ *Check Point 15.10: Suppose we use wires that have a resistance of 0.01 Ω instead of 0.1 Ω as in the example that was discussed. How would that impact the current flowing through the wires in the two situations discussed above (with and without the extra wire across the bulb)?*

15.4 Real vs. ideal batteries

In the previous section, it was mentioned that placing a wire to bypass a bulb will cause the current through the bulb to be slightly less. However, with real 1.5-V batteries, doing this will likely cause the current to be so low that the bulb won't light at all.

To explain why, we need to consider what distinguishes a real battery from the ideal battery we've been assuming up to now.

15.4.1 Open circuit voltage

Up to now, we've been assuming that a battery maintains a constant voltage. While mostly true, careful measurements will reveal a strange thing: when the battery is connected to the circuit and current is allowed to flow, the battery voltage is lower than it is when it is disconnected from the circuit and no current flows.

So a 1.5-V battery doesn't provide 1.5 Volts?

If you measure the voltage across a new 1.5-V battery then, with the battery not connected to anything, you'll likely find that the voltage is actually *greater* than 1.5 V. Then, when you connect the battery to the circuit and generate current, you'll find the battery voltage is about 1.5 V. Then, when the battery is again disconnected from the circuit, the battery voltage *returns* to what it was before.

⁝ The battery voltage tends to be lower when current is being drawn.

[viii] A fuse is a little piece of metal that melts if too much current flows through it, breaking the circuit and stopping the current from flowing. You then need to replace the fuse.

✍ I don't mean that the battery voltage goes down *as* current flows. Rather, it has a lower value, and maintains that lower voltage, *when* current flows. If no current flows, it has a higher value, and maintains that higher voltage.

This may appear to be very strange, considering that the same chemical reaction that produces the voltage across the battery when little or no current is flowing is the same chemical reaction that produces the voltage when a larger current is flowing. For hand-wavy explanation, see the footnote.[ix]

In any event, to distinguish between the voltage when the battery is disconnected from the circuit and the voltage when it is connected, I'll refer to the "disconnected" voltage (i.e., when no current flows) as the **open circuit voltage**.[x]

WHY DO YOU CALL IT THE "OPEN CIRCUIT VOLTAGE"?

Much like an open draw bridge prevents traffic from moving across the bridge, no current is being drawn from the battery when the circuit is "open."

DOESN'T "OPEN" MEAN THAT CURRENT CAN FLOW?

Not in this context. Think of an open door. If you were an ant crawling along the wall, you'd have to stop when you reach an open door. You wouldn't be able to continue along the wall until someone closes the door. In a similar manner, electrons can't flow until someone "closes the switch."

✓ *Check Point 15.11: Suppose a fresh AA battery has an open circuit voltage of 1.60 V. What is the voltage of the battery before it is placed in a circuit: greater than 1.60 V, less than 1.60 V or equal to 1.60 V?*

[ix] To explain why the voltage goes down, consider an analogy between electric potential and water pressure discussed in section 14.2. In particular, consider what happens to the water pressure in the pipes when you turn on the faucet. The water pressure may be larger when the faucet is closed (no water running) than when you open the faucet. In a sense, the plumbing cannot maintain the same pressure as the water is removed. The greater the current, the lower the pressure becomes. In a similar way, the battery voltage goes down when current is drawn because the chemical reactions cannot proceed quickly enough to maintain the voltage.

[x] Among physicists, it is common to call the open circuit voltage the **emf** (pronounced ee-em-ef, as in the letters). This is because the voltage was originally called the **electromotive force**. It is not a force, though. It is a voltage. That is why I prefer to call it the open circuit voltage.

15.4.2 Internal resistance

We can usually ignore the difference between the battery voltage when connected to the circuit and the battery voltage when it is disconnected. However, for those times when there is a significant difference, we need a way of quantifying *how much* the voltage will go down.

It turns out that the voltage decreases by an amount proportional to the current. In other words, the larger the current, the more significant the difference and the harder it is to ignore it.

In fact, this is why shorting out a bulb is likely to make it turn off. Shorting it out leads to a large current, and this leads to a significant decrease in the battery voltage.

In any event, the voltage across a resistor is *also* proportional to the current (i.e., $V = IR$), so it is like there is a tiny resistor inside the battery that is dissipating energy, countering the voltage being generated by the battery.

⁉ The internal resistance refers to the way the battery's voltage is lower when current is being drawn

Thus, we say that a real battery has an **internal resistance** that causes the battery voltage to be less when it is connected to a circuit and current flows.

✍ We'll assume that each battery has a particular internal resistance that is constant for that battery. However, the internal resistance of a real battery can increase as the battery is used because the product of the chemical reaction can "impede" the flow of the current. Regardless, the the open circuit voltage should still remain the same.[xi]

Since the voltage drop across a resistor is just the product of the current through the resistor and the resistance of the resistor, we can calculate how much the battery voltage drops when current is flowing by simply multiplying the current by the internal resistance of the battery.

For example, one difference between a 12-V lantern battery and a 12-V car battery is that the car battery has a much smaller internal resistance (less than 0.01 Ω compared to about 2 Ω for an Eveready 732 lantern battery[xii]),

[xi] Although we've been assuming that voltmeters have an infinite resistance, in reality then have a finite resistance and draw a tiny amount of current in order to make their measurement. That tiny amount of current will result in a voltage reading that is lower than *actual* voltage.

[xii] The actual internal resistance varies with temperature. See http://data.energizer.com/pdfs/732.pdf.

which means that only the car battery can provide the current of 200 A that a car needs to start. At 200 A, the car battery experiences a temporary lowering of 2 V (multiply the current by the internal resistance of 0.01 Ω) whereas the lantern battery would experience a lowering of 400 V (multiply by 2 Ω instead of 0.01 Ω), which is not possible, given that its open circuit voltage is 12 V. Fortunately, lanterns only require 0.5 A or less.

✍ Another advantage of the car battery is that it is easily rechargeable, so that the car itself can recharge the battery while driving.

In a sense, the higher resistance of the 12-V lantern battery is safer, as the internal resistance prevents a large amount of current from flowing if you short out the terminals.[xiii]

✍ On page 257 it was mentioned that internal resistance can be responsible for why a light bulb may be brighter with two batteries side-by-side (in a split path configuration) than with either individually. This is because when used side-by-side, less current is drawn from each battery. With less current, the internal resistance of each battery has less of an effect.

✓ *Check Point 15.12: Suppose a typical fresh AA dry cell has an open circuit voltage of 1.60 V. If the battery voltage drops by 0.2 V when the battery is connected to a circuit and 0.1 A of current flows through the circuit, what is the battery voltage when it is connected to a different circuit, where 0.2 A of current flows?*

Summary

This chapter examined how the resistance of wires and the internal resistance of batteries influence the current that flows through the circuit.

The main points of this chapter are as follows:

- The amount of current that flows in a circuit depends on its resistance.

[xiii] In comparison, although a car battery is only 12 Volts, it can provide a lot of current if the two terminals are connected (via, say, the car frame) and this current can create a dangerous spark. Simply touching the two terminals of a car battery, though, is relatively harmless assuming your body has a high resistance (e.g., dry skin).

- An element's resistance is defined as the ratio of the voltage across it to the current through it.
- Resistance is measured in ohms, abbreviated as Ω.
- We will assume that the resistance of a resistor is always the same, regardless of the current.
- All resistors that share the same end points have the same voltage across them.
- The current through each resistor along the same path is the same.
- You cannot arbitrarily assign values of V, I and R when using $V = IR$.
- The battery voltage tends to be lower when current is being drawn.
- The internal resistance refers to the way the battery's voltage is lower when current is being drawn (as the current is flowing through an imaginary resistor inside the battery).

You should now be able to relate the voltage, current and resistance for a particular element or battery in a circuit.

Frequently asked questions

If the battery voltage is less than the open circuit voltage, does that mean the battery is going dead?

No. We blame the lower battery voltage on the internal resistance. This happens even for a new battery. How much it goes down depends on battery's internal resistance.

Does the open circuit voltage go down when the battery is going dead?

It shouldn't, as long as there are enough chemicals inside the battery to carry out the reaction. However, even a voltmeter draws some current to make its measurement and, as such, it may read progressively lower voltages as the battery dies.

Terminology introduced

Electromotive force	Ohm's law	Short circuit
EMF	Open circuit voltage	Superconductors
Internal resistance	Resistors	Terminal voltage
Ohm	Resistance	

Abbreviations introduced

Quantity	**SI unit**
resistance (R)	ohm (Ω)[xiv]

Additional problems

Problem 15.1: Suppose a typical fresh AA dry cell has an open circuit voltage of 1.60 V. It is hooked up to a 100-Ω resistor and 14 mA of current is found to flow through the circuit. While current is flowing,
(a) What is the voltage across the resistor?
(b) What is the voltage of the battery?

Problem 15.2: Suppose someone measures the resistance of a light bulb when it is disconnected from the circuit and then uses that resistance value to predict how bright the light bulb will be when part of the circuit (using $P = V^2/R$ to predict the power dissipated by the bulb). It turns out that the resistance is greater when the temperature is greater. Based on that, how will the actual brightness of the bulb compare to the predicted value (i.e., will it be less than predicted, greater than predicted or equal to the predicted)? Explain your choice.

Problem 15.3: Suppose a battery is a typical fresh AA dry cell with an open circuit voltage of 1.6 V and an internal resistance 0.31 Ω.
(a) What is the voltage of the battery when the battery is connected to a resistor of unknown resistance R such that a current of 58 mA flows through the resistor?
(b) What is the voltage across the unknown resistor?

[xiv] An ohm is equal to a volt per ampere (V/A).

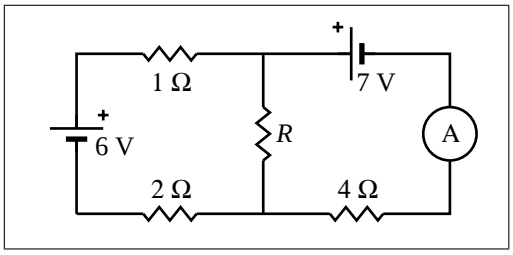
(c) What is the resistance of the unknown resistor?
(d) Suppose the resistor was replaced by a wire of negligible resistance. What current will flow through the wire?

Problem 15.4: A 2-V battery is applied to a 3-Ω resistor. How much energy is lost to heat in 4 seconds?

Problem 15.5: Suppose we only had two resistors of resistances 1 Ω and 2 Ω. Is it possible to combine these in a way to produce the same current as what would be produced if we had a 1.5 Ω resistor? Explain.

Problem 15.6: Two light bulbs (A and B) are arranged along a single path with a battery. It is found that there is a voltage of 0.5 V across light bulb A and a voltage of 1.0 V across light bulb B. If the current through each light bulb is 30 mA, what is the resistance of each light bulb?

Problem 15.7: A simple circuit is set up as shown below.
(a) Is the current through the 2-Ω resistor necessarily the same as the current through the 1-Ω resistor? How about through the 4-Ω resistor? Why or why not?
(b) If the ammeter reads 1 A with the current going up the page through the ammeter, what is the voltage across the 4-Ω resistor?
(c) Given the situation in (c), what is the voltage across the resistor R?
(d) What is the voltage across the 1-Ω and 2-Ω resistors?
(e) What is the resistance R? Hint: First solve for the currents in all three paths.



16. Describing AC Circuits

Puzzle #16: What is the difference between AC and DC, and why should we care?

Introduction

Up to now, we've been considering DC circuits, which can be used with light bulbs, motors and computers. However, our homes are supplied with AC, not DC. Since AC is so common, the remainder of this part focuses on what AC voltage is and how it impacts the current (or, equivalently, how it impacts an organism that comes into contact with an AC voltage source).

16.1 Period and frequency

Up to now, when we spoke of a voltage being applied to a circuit, the voltage was assumed to be steady and unchanging. Consequently, a particular, steady current was set up. Such a current is called "DC" (for direct current, where "direct" means steady and unchanging).

However, the transmission and distribution power lines that carry electricity to our homes provide **AC voltage**, not DC voltage.[i] AC voltage **oscillates**[ii], which means it switches directions in a periodic way, undergoing a periodic **cycle** that is repeated over and over again.

⁉ An AC voltage is a voltage that oscillates in value.

An example of an oscillating AC voltage is illustrated in Figure 16.1. In this particular example, the voltage starts with a value of zero and then increases up to +3 V and then returns past zero to −3 V and then back again.

[i] The reason AC is used for these purposes is because it was easier to transport long distances through transmission lines.

[ii] Oscillations were also examined in volume I but with things like pendulums and springs.

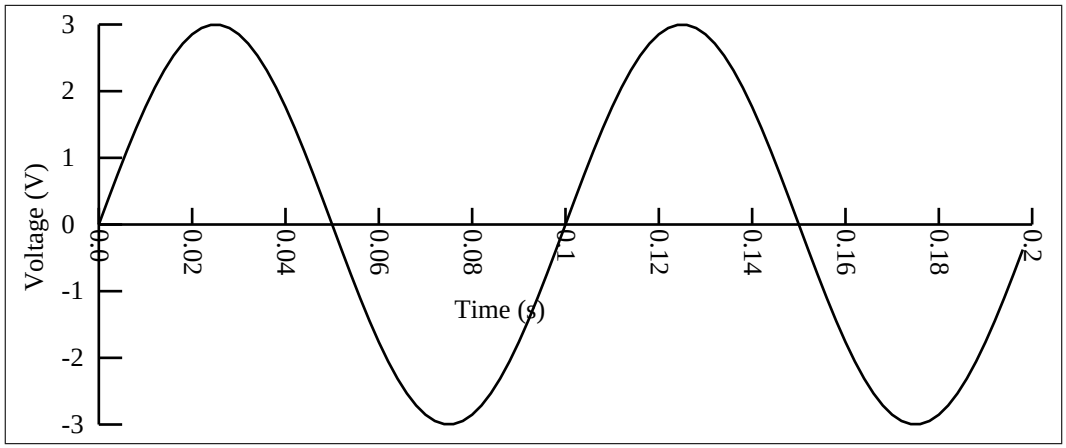


Figure 16.1: The time variation of an AC voltage.

If voltage is oscillating, will the current oscillate also?

Yes, and in this section I will provide the terminology we use to describe the oscillation in both voltage and current (the same terminology is used for both).

To represent how quickly the voltage oscillates, we use two terms: period (T) and frequency (f). Both can be used and you should become comfortable with both.

16.1.1 Period (T)

The **period**, T, is the time it takes to undergo one cycle. In the example shown in Figure 16.1, one cycle is represented by one up-down-up voltage variation. In this case, the cycle takes 0.1 seconds. In the 0.2 seconds that is shown in the figure, two cycles are completed.

Technically, the period has units of *time* but, for our purposes, we'll treat it as though it has units of *time per cycle*. In the example shown in Figure 16.1, it takes 0.1 seconds to complete one cycle, so we'll say the period is 0.1 s/cycle, even though technically the period is just 0.1 seconds.

16.1.2 Frequency (f)

The **frequency**, f, of the signal is the number of cycles that are completed in a given amount of time, usually one second. In the example shown in Figure 16.1, it takes 0.1 seconds to complete each cycle. That means it would get to complete ten cycles in one second. Consequently, the frequency of the voltage shown in the figure is 10 cycles/s.

Notice that the units of frequency are just the inverse of those for period: cycles per time, instead of time per cycle. In fact, there is a very simple relationship between frequency and period – they are inverses of one another:

$$f = \frac{1}{T} \qquad \text{and} \qquad T = \frac{1}{f}$$

The smaller the frequency, the slower the voltage oscillates up and down and the larger the period.

To express frequency, we tend to use the unit of "hertz"[iii] (abbreviated as Hz) which means a cycle per second. Consequently, 10 cycles per second is equivalent to 10 Hz.

Note that a zero frequency does not mean zero voltage amplitude (see next section). Rather, it means that the voltage value isn't varying – it is steady. Consequently an AC voltage of frequency zero is the same as a DC voltage (which is an unvarying voltage).

⁉ The frequency (measured in cycles per second) is the inverse of the period (measured in seconds per cycle).

⁉ One hertz is equivalent to a cycle per second.

✓ *Check Point 16.1: An AC voltage with peak value 10.0 V oscillates with a period of 1 ms, where "ms" stands for "millisecond." What is the frequency of the signal?*

16.2 Amplitude

The frequency of an AC voltage or current is like the heart rate – it represents how quickly the value oscillates. What we also need to do is measure the

[iii] This unit is named after Heinrich Hertz (1857-1894), a German physicist who studied electromagnetic phenomena.

strength of the AC voltage or current, much like how we measure the heart strength by determining how much blood it pumps for each beat.

To describe the strength of the voltage (or current), we use the same terminology we used in volume I when discussing oscillations, namely in terms of the amplitude. The **amplitude** is the maximum value *from the middle value*.

What is the middle value?

For oscillations like the type we are examining here, the middle value is zero.[iv]

⁝ The amplitude of an AC voltage represents how far the voltage gets from the middle value (zero for our cases).

In our example shown in Figure 16.1, the voltage varies between +3 V and −3 V, and the middle value is 0 V (directly between +3 V and −3 V). In this case, the voltage oscillates about zero, with the maximum above zero having the same absolute value as the minimum below zero (i.e., $V_{\text{min}} = -V_{\text{max}}$), with the maximum voltage being +3 V and the minimum voltage being −3 V.

If the middle value is zero, what is the amplitude?

Since the middle value is zero, the amplitude is simply V_{max} (i.e., 3 V in our example). Since we are assuming for our discussions that the middle value is zero, we will represent the amplitude as V_{max}.[v]

✓ *Check Point 16.2: An AC voltage with maximum value 10 V and minimum value of −10 V oscillates with a period of 1 ms. What is the amplitude of the signal?*

16.3 Signal generators

If we are going to study the effect of AC voltage on circuits, we need a reliable, adjustable source of AC voltage. The source should be able to provide a variety of frequencies and amplitudes.

If batteries produce DC voltage, how do we get AC voltage?

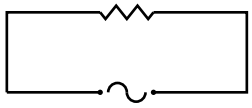
⁝ A signal generator generates an AC voltage.

[iv] In the cases we'll examine, zero is not only the middle value but also the average value. This need not be the case for AC but it makes the equations simpler.

[v] Sometimes this is written as V_{peak}.

We'll typically use a device called a **signal generator**[vi], which is basically just a very fancy battery that happens to provide a voltage that oscillates.

In a circuit schematic, an AC voltage source (such as that produced by a signal generator) is indicated by a wavy line (see schematic to the right, where the AC voltage source is connected to a resistor).



Unlike a battery, we use the signal generator to control both the *strength* of the voltage (e.g., the voltage amplitude) and the *frequency* of the voltage. These are controlled independently. One can change one while keeping the other constant.

✓ *Check Point 16.3: A signal generator is set to produce an AC voltage with a frequency of 100 Hz and an amplitude of 1 V. If the frequency is doubled to 200 Hz, what happens to the amplitude?*

16.4 Power in an AC circuit

Suppose you had two circuits, each with an identical bulb but one powered by an AC voltage and the other powered by a DC voltage that equals the amplitude value of the AC voltage. One thing you might notice is that the bulb in the AC circuit isn't as bright as the bulb in the DC circuit.

To understand why, recall that DC means steady and unvarying, so the current remains at a single value. That means the light bulb will have a steady brightness. AC, on the other hand, means the voltage and current are oscillating and not steady, with the current varying between some maximum value and zero. That means the bulb brightness also oscillates, between bright (when current is flowing) and dim (when it is not). If you stay at the maximum brightness (and current), the average brightness (and current) will be greater than if you varied between zero and that maximum.

[vi] There are many different types of signal generators available, ranging from those that only provide certain types of oscillations (e.g., sinusoidal) at only certain frequencies to those that allow you to specify any arbitrary signal. These latter generators usually go by the name **arbitrary waveform generators**. There are also special generators for tuning radios, and measuring various characteristics of circuits.

IF USE AC AT HOME, HOW COME I DON'T SEE THE LIGHT BULBS IN MY HOUSE FLICKER BETWEEN BRIGHT AND DIM?

If the period of the oscillation is long enough (i.e., long time to flip between bright and dim), we'll see the bulb flicker, oscillating between on and off, in time with the current, which is also oscillating. On the other hand, if the period is very short, say a small fraction of a second, the flicker can be so fast that our eyes cannot pick it up and we'll see the bulb emitting a steady brightness.[vii]

In fact, when the period is short an incandescent bulb won't even oscillate in brightness.[viii] This is because an incandescent bulb lights because it gets hot. If the oscillating current is only zero for a short time as it oscillates, the bulb doesn't cool enough to turn off completely and it stays on throughout the cycle.

Whether we have an incandescent bulb or not, the bulb brightness has more to do with the *average* current, which is necessarily less than what it would be if it stayed at its maximum value the entire time and didn't oscillate. Indeed, with AC the brightness we see is only about half of what it would be if the current remained at its maximum value the entire time and didn't oscillate.

Mathematically, we can write this as follows:

$$P_{\text{avg}} = \frac{P_{\text{max}}}{2}$$

Since the mathematical relationship between power, voltage and current is $P = IV$ (see equation 14.2), we can replace P_{max} by the product of the voltage and current amplitudes, $V_{\text{max}}I_{\text{max}}$ to get the following:[ix]

$$P_{\text{avg}} = \frac{I_{\text{max}}V_{\text{max}}}{2} \tag{16.1}$$

Even though we've used a bulb as an example, the same is true for any circuit, and any energy transformation that may occur. The power dissipated in the

[vii] That our eyes cannot distinguish between quick changes is called *vision persistence*.

[viii] LED and CFL bulbs don't work the same way but they have circuitry that make those bulbs mimic the way the brightness oscillates in an incandescent bulb.

[ix] Technically, this relationship only holds if the circuit is purely resistive and has no inductance or capacitance, properties that are discussed in chapter 17.

circuit will be half of what it would be if the voltage and current remained steady at their amplitude values rather than oscillating.

Example 16.1: The average power dissipated in particular circuit is 100 W. Assuming a typical AC voltage has an amplitude of 170 V (typical in the U.S.), find the amplitude of the AC current through the circuit.

Answer 16.1: Use $P_{\text{avg}} = I_{\max}V_{\max}/2$, with $P_{\text{avg}} = 100$ W and $V_{\max} = 170$ V. Solve to get a current amplitude of 1.2 A.

✓ *Check Point 16.4: When a 60-Hz voltage is applied to a bulb, the bulb seems to emit a steady light. If we want to find the average rate that energy is converted to heat and light by the bulb, why would half the product of* $I_{\max}$ *and* V_{ma} *be more appropriate than the product of* $I_{\max}$ *and* $V_{\max}$.

16.5 Root-Mean-Square (RMS)

In the previous section, it was mentioned how, if we had an oscillating voltage, the average brightness of a bulb would be equal to half of what it would be if the voltage remained constant at its amplitude value. Given that, suppose we had two circuits, one AC with amplitude $V_{\max}$ (oscilalting between zero and $V_{\max}$) and one DC with a steady voltage that we could set at whatever value we'd like. If we wanted both bulbs to have the same average brightness, what voltage value would we use for the DC one?

The answer is that we'd use a voltage value that is about 70% that of the amplitude.

Why is it 70% and not half?

Because if the voltage is 70% of the voltage amplitude value then the current is likewise 70% of the current amplitude value. The product of those two equals the power, and multiplying the two values, each 70% of the amplitude value, gives a product that is half of what it would be just using the amplitude values of each.

The 70% value is called the **RMS** value. In particular, the RMS voltage value is 70% of the voltage amplitude value, and the RMS current value is 70% of the current amplitude value.[x]

❢ For AC voltages, the average rate at which energy is dissipated in an element is equal to the product of the RMS current and RMS voltage.

$$\begin{aligned} V_{\text{rms}} &= 70\% \times V_{\text{max}} \\ I_{\text{rms}} &= 70\% \times I_{\text{max}} \end{aligned} \tag{16.2}$$

Consequently, the power equation ($P_{\text{avg}} = I_{\text{max}} V_{\text{max}}/2$) can be rewritten as follows:

$$P_{\text{avg}} = I_{\text{rms}} V_{\text{rms}} \tag{16.3}$$

It is the RMS value that everyone uses, not the amplitude value, because the RMS value is more closely tied to the actual energy dissipation in the circuit.

For example, the household outlets in the U.S. are rated as 120 V. The 120-V value is actually the RMS value of the voltage, not the amplitude. In addition, when used with AC, a typical multimeter provides the RMS value, not the amplitude value.[xi]

✍ When using a multimeter with AC, you must use the "AC" setting (usually indicated by the "V" or "A" with the wavy line on top, $\widetilde{\text{V}}$ or $\widetilde{\text{A}}$, for voltage and current, respectively). Doing so will give the RMS values. If you use the "DC" setting (indicated by the "V" and "A" with the straight line on top, $\overline{\text{V}}$ and $\overline{\text{A}}$), you'll get something close to zero when the voltage isn't steady because the meter is essentially trying to find the average voltage or current.[xii]

Example 16.2: The average power dissipated in particular circuit is 55 W. Assuming a typical AC RMS voltage of 120 V (typical in the U.S.), find the RMS value of the AC current through the circuit.

---

[x] Technically, the value is $1\sqrt{2}$, which I've approximated as 70%.

[xi] While a multimeter on AC setting will ideally provide the RMS value (as long as the amplitude is not changed), in reality it may be off, particularly at very low or very high frequencies. At low frequencies, the RMS mechanism may not be able to obtain the full signal and thus may give either an incorrect reading or a reading that varies. The reliability at high frequencies depends on the quality of the meter. For high quality, we need to use an **oscilloscope** (see the supplemental readings).

[xii] Even if the DC meter didn't average, the value would oscillate too quickly for us to see what the value is.

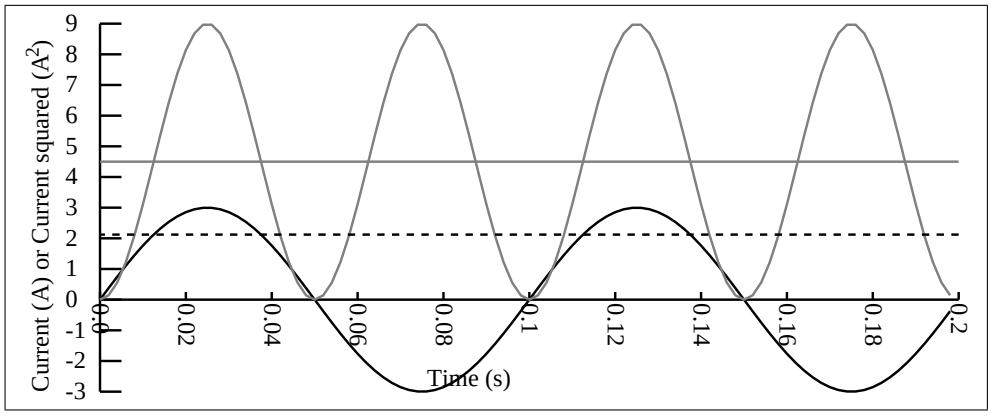


Figure 16.2: The time variation of an AC current.

Answer 16.2: Use $P_{\text{avg}} = I_{\text{rms}} V_{\text{rms}}$, with $P_{\text{avg}} = 55$ W and $V_{\text{rms}} = 120$ V. Solve to get an RMS current of 0.46 A.

Why is it called RMS?

The letters stand for Root-Mean-Square, which means that we take the square root of the average square value.

What is meant by "average square value"?

We know from before that the average value is zero because the values are positive as often as they are negative. However, if you square all of the values, you only get positive values. Consequently, the average of the squared values will not be zero.

The RMS is just the square root of the **average** of all of the **squared values**.

For example, consider the current that is illustrated in Figure 16.2. In this particular example, the current starts with a value of zero and oscillates between +3 A and −3 A (i.e., the amplitude is 3 A).

Keep in mind that the current is oscillating so that, at times, it is zero and at other times it is 3 A. That means the squared values vary between zero (when the current is zero) to 9 A^2 (when the current is +3 A or −3 A).

The squared values are given by the curved gray line in the figure. Notice how all the values on the gray curve are positive, since the square of a negative value is a positive value. Also notice how the maximum value of the gray curve is now 9 A^2 (i.e., the square of 3 A).

The average of all the values represented by the gray curve (values that vary from 0 to 9 A^2) is 4.5 A^2 (see horizontal gray line). The RMS value is then obtained by taking the square root of this result, which is 2.12 A in this case and is indicated by the horizontal dashed line in Figure 16.2. Notice how the RMS value is 70% of the original amplitude (3 A).

Example 16.3: What is the RMS value if the amplitude is 4 V?

Answer 16.3: Assuming it is sinusoidal, Since it is sinusoidal, multiply the amplitude (maximum value) by 70% to get the RMS value (i.e., 2.83 V).

✓ *Check Point 16.5: (a) What is the voltage amplitude for an AC outlet in the US, which has an RMS value of about 120 V? (b) Which is greater: the RMS value or the maximum value? Explain why your answer makes sense, given that "RMS" means "square root of the mean squared value".*

16.6 Voltage and current

For DC, the relationship between voltage and current was $V = IR$. What is the relationship for AC?

It is still $V = IR$.

In the expression, V represented the *strength* of the voltage and I represented the *strength* of the current. That is still the case with AC voltage.

Note that the frequency of the voltage and current does not appear in the expression. This is because frequency and amplitude are controlled separately and one doesn't impact the other.

For example, suppose you apply an AC voltage (via a signal generator) to a circuit that only contains a light bulb. The current in the light bulb oscillates with the same frequency as the voltage across the light bulb.

If the frequency is not large (say, less than 10 Hz or so), you'd likely notice the light blinking on and off (due to the oscillation in current). For the frequencies we'll be dealing with (60 Hz or more), you wouldn't notice the blinking and the bulb would appear to give off a steady light.

Once a steady light is achieved, the light wouldn't get any brighter or dimmer when the frequency is increased even more (e.g., to 1000 Hz) because the strength of the voltage is unaffected by the frequency. In other words, the strength of the current (as measured by the RMS or amplitude, for example) is the same, regardless of frequency as long as the strength of the voltage remains the same (as measured by the RMS or amplitude).

✍ If you consider our model of what is happening, this should make sense. After all, increasing the frequency doesn't change the maximum voltage reached during each cycle. It just reaches it more often each second. Consequently, the maximum current achieved during each cycle should remain the same as well (just achieving it more often during each second).

⁉ The relationship between the AC voltage across an element and the AC current through it is the same as it was for DC current and voltage.

COULD WE USE EITHER THE AMPLITUDE OR THE RMS FOR V AND I?

Yes, but whichever one you use for V, you must also use for I. For example, if you use the amplitude for voltage (V_{max}), you must also use the amplitude for current (I_{max}):

$$V_{\text{max}} = I_{\text{max}} R \tag{16.4}$$

Or, if you use the RMS value for voltage (V_{rms}), you must also use the RMS value for current (I_{rms}):

$$V_{\text{rms}} = I_{\text{rms}} R \tag{16.5}$$

✍ You can use equation 16.5 ($V_{\text{rms}} = I_{\text{rms}} R$) to replace I_{rms} or V_{rms} in equation 16.3 ($P_{\text{avg}} = V_{\text{rms}} I_{\text{rms}}$).

✓ *Check Point 16.6: (a) If we use $V = IR$ with the resistance and the RMS current, does V correspond to the RMS voltage or the voltage amplitude? (b) The RMS current through a 20-Ω resistor is 2.0 mA. What is the RMS voltage across the resistor?*

Summary

This chapter examined what is meant by AC voltage.

The main points of this chapter are as follows:

- An AC voltage is a voltage that oscillates in value.

- The frequency (measured in cycles per second) is the inverse of the period (measured in seconds per cycle).
- One hertz is equivalent to a cycle per second.
- The amplitude of an AC voltage represents how far the voltage gets from the middle value (zero for our cases).
- The RMS value represents a sort of "mean" positive value.
- A signal generator generates an AC voltage.
- The relationship between the AC voltage across an element and the AC current through it is the same as it was for DC current and voltage.
- For AC voltages, the rate at which energy is dissipated in an element is equal to the product of the RMS current and RMS voltage.

By now you should be able to do the following:

- Describe an oscillating voltage or current in terms of maximum voltage V_{max}, maximum current I_{max} and frequency f.
- Use a signal generator to generate an oscillating voltage.
- Describe the strength of an oscillating voltage and current in terms of the RMS (or root-mean-squared) value (e.g., $V_{\text{rms}} = V_{\text{max}}/\sqrt{2}$) and use it to calculate the average power consumed by a circuit (e.g., $P_{\text{avg}} = V_{\text{rms}} I_{\text{rms}}$).

Frequently asked questions

What does it mean to have a voltage with frequency equal to 0 Hz?

If the frequency is zero, that means the voltage doesn't change. That is the same as a steady, fixed voltage.

How do we measure AC voltage?

There are two ways. One way is to use the "AC" option of the multimeter, as described on page 292. The other way is to use an **oscilloscope** (see the supplemental readings).

What is an oscilloscope?

See the supplemental readings.

Is the power greater when the applied voltage has a higher frequency?

No. If all we are doing is changing the frequency, then the amplitude (and RMS) of the applied voltage will be the same and, from the relationship for power (equation 16.1), the power provided to the circuit depends only on the amplitudes (or RMS values) of the voltage and current. In this case, the amplitudes of the voltage and current are not changed and so the power dissipated by the circuit remains the same.[xiii]

What is meant by RMS?

See page 293.

Terminology introduced

AC voltage	Oscilloscope
Amplitude	Peak-to-peak
Arbitrary waveform generators	Period
Cycle	RMS
Frequency	Signal generator
Oscillates	Sinusoidal

Abbreviations introduced

Quantity	**SI unit**
frequency (f)	hertz (Hz)[xiv]
period (T)	second (s)

Additional problems

Problem 16.1: A signal generator is set to produce an AC voltage with a period of 1 s. If a bulb is connected to the signal generator, with what frequency does the bulb blink?

[xiii] One possible source of confusion is the fact that the energy of a photon (quantum of light) depends upon its frequency. That is not what is going on here. We are purposely keeping the applied voltage the same.

[xiv] A hertz is equal to a cycle per second (1/s).

Problem 16.2: The average power dissipated in a stereo speaker is 55 W. Assuming that the speaker can be treated as a resistor with 4.0-Ω resistance, find:
(a) The RMS value of the AC voltage applied to the speaker and the RMS value of the AC current through the speaker.
(b) The peak value of the AC voltage applied to the speaker and the peak value of the AC current through the speaker (i.e., find the amplitudes of the voltage and current).

17. Impedance

Puzzle #17: Is AC more dangerous to humans than DC?

Introduction

Now that we can describe AC voltage and current, we'll examine what happens when we apply an AC voltage to things.

As discussed in section 16.6, the relationship between current and voltage is the same with AC as with DC. However, that doesn't mean that everything responds the same way to both AC and DC. In other words, the response to AC can depend on the frequency. For example, the human body is affected more by AC voltage at a frequency of 60 Hz than zero Hz (DC) or higher frequencies (like 100 kHz).[i]

Part of the reason for the human response is physiological in nature, which is beyond the scope of the text. However, there are other reasons that are physical in nature, and those things we will discuss in this chapter.

17.1 Capacitors and inductors

So far, we've restricted our analysis to circuits containing resistors. Even a bulb is a type of resistor, albeit one whose resistance can vary since its temperature can be much hotter when current flows through it than when current is not. However, like a resistor, the amplitude of the current through a bulb (and thus its brightness) only depends upon the amplitude of the voltage across the bulb not upon the frequency of that voltage (assuming it is oscillating quick enough not to flicker).

[i]See, for example, https://iastate.pressbooks.pub/electriccircuits/chapter/chapter-1/.

However, there are some elements for which the amplitude of the current depends on *both* the amplitude of the voltage across the bulb *and* the frequency of that voltage. We will consider two such elements in this chapter.[ii]

- A **capacitor** allows more current to flow at *higher* frequencies (for the same voltage amplitude), allowing no current at zero frequency (DC).
- An **inductor** allows more current to flow at *lower* frequencies (for the same voltage amplitude), acting just like a wire at zero frequency (DC).

At this point you are probably wondering *why* the inductor and capacitor act this way. We will get to that shortly. First, we want to *describe* what happens with each, using what we learned in chapter 15. Once we are able to describe *what* happens, we can then explain *why* it happens.

The key point for now is that the capacitor and inductor, which we will explore in more detail in the next two sections, are elements that have opposite impacts on the circuit (more current at higher vs. lower frequency) but, unlike a resistor, share the property that the current depends on the voltage frequency, not just the voltage amplitude.[iii]

⁉ The resistance of a resistor is the same, regardless of the voltage frequency.

For example, let's revisit the circuit discussed in section 16.6 where we applied an AC voltage (via a signal generator) to a circuit that only contained a bulb. As mentioned in section 16.6, for the frequencies we'll be dealing with (60 Hz or more), the bulb would appear to give off a steady light, and the brightness of the bulb would be the same regardless of the voltage frequency (assuming the frequency is high enough that the bulb doesn't flicker).[iv]

Let's suppose we add a capacitor in line with the bulb (i.e., along the same path). The bulb would be dimmer, since now there is something else "in the way" but, unlike a resistor, the bulb would be even dimmer at lower frequencies because the capacitor allows less current to flow at lower frequencies

[ii] Note the phrase "for the same voltage." For example, suppose a voltage oscillates between +2 V and −2 V very slowly whereas a second voltage oscillates between +2 V and −2 V very quickly. They would have the same voltage amplitude but the second would have a higher voltage frequency.

[iii] As mentioned in the introduction, the human body is affected more by AC voltage at a frequency of 60 Hz than zero Hz (DC) or higher frequencies (like 100 kHz). It turns out that more current can flow at some middle frequency if both an inductor and a capacitor are present. So, the human body has characteristics of both inductors and capacitors.

[iv] This is because a resistor gets warm when current flows through it. It doesn't matter how quickly the current oscillates - as long as the average value of the current is the same (i.e., as long as the amplitude doesn't change) it will get just as warm and so the impedance would be the same (unless the current amplitude changes).

than at higher frequencies. The current would still be oscillating, and at the same frequency as the voltage. It is just that the current amplitude would be less (and so the bulb would be less bright) at lower frequencies than it would be at higher frequencies, even though the voltage amplitude is the same.

⁇ A capacitor allows more current to flow at higher frequencies.

✍ This is useful property, which is why capacitors are used. While you may not have a need for them in your future work, many objects (including the human body) have properties similar to capacitors, and those properties not only can be utilized for things like touchscreens but also predict the impact of electricity on the human body.

✓ *Check Point 17.1: A particular capacitor is used with an AC voltage of amplitude 10 V and frequency 1000 Hz. In which of the following two cases would the amplitude of the current be higher?*
(a) When the voltage is changing slowly (frequency ≪ 1000 Hz)
(b) When the voltage is changing rapidly (frequency ≫ 1000 Hz)

⁇ An inductor allows more current to flow at lower frequencies.

Now let's suppose we replace the capacitor with an inductor. It turns out that, once again, the bulb would be dimmer, since now there is something else "in the way" but, opposite the capacitor, the bulb would be even dimmer at *higher* frequencies because the inductor allows less current to flow at higher frequencies than at lower frequencies, even though the voltage *amplitude* remains unchanged.

✍ This is a useful property. Noise is essentially a high frequency sound wave. Similarly, we can have noise in an electronic signal, and that noise is typically characterized by high frequency oscillations. The inductor, then, can remove that noise. In a similar way, inductors can be used to prevent spikes and surges in voltages. In addition, many objects (like the human body) have properties similar to inductors and we can use inductors to illustrate those properties.

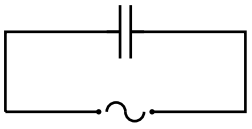
✓ *Check Point 17.2: A particular inductor is used with an AC voltage of amplitude 10 V and frequency 1000 Hz. In which of the following two cases would the amplitude of the current be higher?*
(a) When the voltage is changing slowly (frequency ≪ 1000 Hz)
(b) When the voltage is changing rapidly (frequency ≫ 1000 Hz)

17.1.1 Structure of a capacitor

WHY DOES THE CAPACITOR ACT LIKE A STRONGER RESISTOR AT LOWER FREQUENCIES?

To explain why the **capacitor** acts this way, allowing less current at low frequencies than at high frequencies (with almost no effect if the frequency is high enough), we first have to understand exactly what a capacitor is: essentially two tiny metal plates stuck together with some thin insulating material between them (so the two plates don't touch).

For this reason, in a circuit schematic a capacitor is indicated by two parallel lines with a small space between them (see figure to right, where a capacitor is drawn connected to an AC voltage source).



✍ The symbol for a capacitor is similar to the one used for a battery but for a battery the two lines are unequal in size. With a capacitor, the two lines are the same length.

This is why some capacitors look like little disks[v]. Sometimes the two plates are then "rolled" together to save space, in which case the capacitor has the shape of a cylinder, much like a resistor except without the color codes[vi].

Because an insulator separates the two plates, the capacitor is essentially a "gap" in the circuit. Current cannot flow through the gap. Consequently, when the frequency is zero (i.e., the applied voltage doesn't oscillate at all), no current flows at all.

THIS EXPLAINS WHY THE CAPACITOR DOESN'T ALLOW CURRENT TO FLOW WHEN THE FREQUENCY IS ZERO, BUT WHY WOULD CURRENT FLOW WHEN THE FREQUENCY IS NOT ZERO?

To explain how current can seem to flow through a capacitor, even though there is not a continuous conducting path through the capacitor, we need to modify our model somewhat.

The modification is that we will now allow for a limited amount of "bunching" up that we didn't allow before. In this case, the structure of the capacitor is such that, for a very short while, current can still flow through the wires

[v] Or little ticks, for you budding entomologists.

[vi] The color codes are colored stripes drawn on the resistor to indicate its resistance.

leading up to and from the capacitor. During this very short time, electrons are being pulled out of one side of the capacitor while electrons are being deposited onto the other side. As a result, one plate becomes positively charged (as electrons are removed) while the other becomes negatively charged (as electrons are deposited).

During this time, when positive charge is accumulating on one capacitor plate and leaving the other capacitor plate, current is flowing *into* the capacitor and flowing *out* of the capacitor, without any charge actually flowing *across* the gap between the two plates.

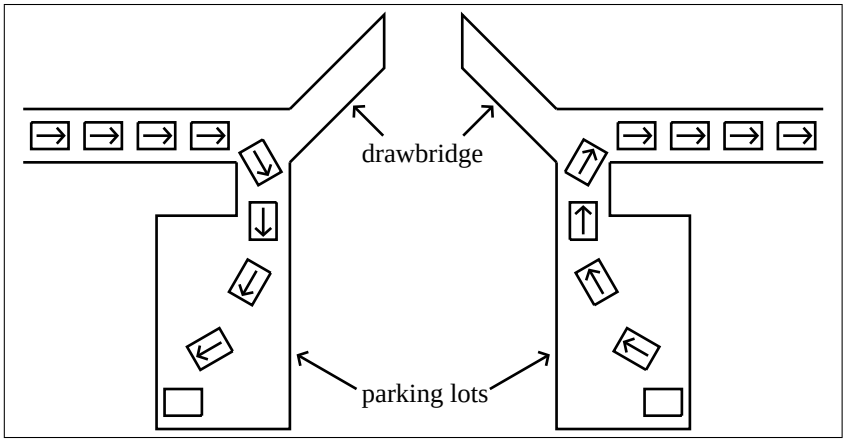
✍ The **plasma membrane** in an animal cell acts like a combined capacitor and resistor, since it impedes charges from flowing across (like a resistor) while, at the same time, it allows charges to build up on one side of the membrane (like a capacitor).

WHY WASN'T THIS "BUNCHING" OBSERVED BEFORE WHEN WE HAD A BREAK IN THE CIRCUIT?

Before now, we weren't able to see this effect because not enough bunching is present for a typical break. As far as we could tell, any break in the circuit immediately stopped the current. With a capacitor, however, the bunching can be significant.

DOES CHARGE FLOW ACROSS THE GAP FROM ONE PLATE TO THE OTHER?

No. As current flows into one plate and out the other plate, there is no current flowing *across* the gap. It is like a draw bridge that is up, so no cars can pass over the bridge. However, if there are parking lots on each side of the bridge, cars can drive into and out of the parking lots.



So current doesn't actually flow through the capacitor?

Technically, no. It just appears to do so.

Consider the draw bridge analogy again. Even when the draw bridge is up and no cars can get across the river, cars can continue to drive into the parking lot on one side and leave the parking lot on the other (until one parking lot is full and the other is empty). During this time, cars continue to move toward the bridge on one side and away from the bridge on the other, making it may appear as though cars are crossing the bridge. However, no cars actually cross the bridge.

⁉ A capacitor represents a break in the circuit, but can store charge, thereby giving the impression that current flows through it.

In a similar way, charges can continue to flow into one plate of a capacitor and out of the other plate. During this time, it may appear as though current is flowing *through* the capacitor, but it isn't.

Wouldn't the current eventually stop anyway?

Yes, the current stops eventually if the current remains in the same direction. The plates would get "full," so to speak, and the current would stop.

However, with an AC voltage, the current changes direction periodically. If the current changes direction before the plates get "full," the current never actually stops flowing into (or out of) the capacitor.

✍ Dry skin acts a little bit like a capacitor. Because of this, more current will flow through a person when the applied voltage has a higher frequency. This is one of the reasons why AC voltage is considered to be more dangerous than DC voltage (which has a zero frequency).

So each capacitor has a certain amount of charge that it can "store" before current stops?

Sort of. Unlike a parking lot, there essentially is no limit on the amount of charge we can deposit on the plates (considering how small the particles are). The only thing limiting the amount of charge is the voltage being applied. The harder the battery or power supply "pushes" the charges onto/off the plates, the more charge that can be placed on the plates.

In this way, our parking lot analogy breaks down. A better analogy might be to consider how many people you can pack into a subway car. With the help of "people pushers" (as in Tokyo), you can pack more people into the car. The people, like the charges on each plate of a capacitor, do not "want"

to be packed together. More people will be packed in with the help of people pushers, who act like an applied voltage to get more charge on each plate.[vii]

So, as the plates fill up, it gets harder and harder to put additional electrons on one and remove electrons from the other. In other words, the resistance to current increases as the plates fill up. When we refer to the impedance of the capacitor, we are actually referring to the average impedance as the current goes back and forth, and is somewhere between zero (the impedance when the plates are empty) and the value when the current switches direction (maximum charge on plates). The greater the maximum charge on the plates, the greater the average impedance, which is why the capacitor's impedance depends on how long the current flows before it switches directions (lower frequency means longer time before current switches directions).

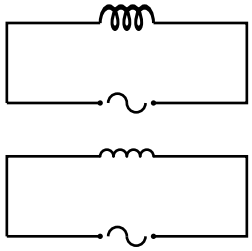
✓ *Check Point 17.3: (a) When high frequency is applied to a capacitor, a lot of current is allowed to flow. How is this possible if the two plates of the capacitor are actually separated by a thin insulator?*
(b) While current is flowing, does the capacitor remain neutral? Explain.

17.1.2 Structure of an inductor

WHY DOES THE INDUCTOR ALLOW MORE CURRENT TO FLOW AT LOWER FREQUENCIES THAN AT HIGHER FREQUENCIES?

To explain why the inductor has this dependency, opposite that of the capacitor, we need to examine the magnetic properties of the inductor.

An **inductor** is essentially a loop or winding of wire (of negligible resistance). On a circuit board, they may look like little coils of wire. In a schematic of a circuit, we use a series of curly lines (top schematic) or semi-circles (bottom schematic). In both circuits, the inductor is drawn connected to an AC voltage source.



⁉ An inductor is just a solenoid (coil of wire).

[vii] Another popular analogy utilizes the idea of a very tall dam. If no water is allowed to drain past the dam, the height behind the dam will continue to rise until the water level reaches the height of the stream at its source (assuming it can't go over the dam; see section 17.5.1). The higher the source, the higher the water will pile up behind the dam.

As we know from chapter 12, a coil of wire acts like an electromagnet when current is flowing through it.

DOES THAT MEAN THAT IF WE BRING A MAGNET NEAR THE INDUCTOR THE MAGNET WILL BE ATTRACTED OR REPELLED BY THE INDUCTOR?

Yes, but you'll only notice it if the inductor is connected to a DC voltage.[viii] If the inductor is connected to AC, the current is constantly flowing back and forth. Consequently, the inductor acts like a magnet that switches direction with a frequency equal to the frequency of the AC voltage that is applied. This may make it difficult to detect the magnetic properties.

SO HOW DOES THIS HELP EXPLAIN THE INDUCTOR'S PROPERTIES?

To answer this, let's compare the inductor to a resistor. If a resistor (like a bulb) is present in a circuit, it warms up as current flows through it, meaning that energy is being transferred to the environment (in the form of thermal energy[ix]). The inductor is also transferring energy, but to *magnetic* energy instead of thermal energy.

⁝ An inductor transfers energy between electric and magnetic energy.

The inductor is not a magnet (and has no magnetic energy) when no current flows but it does act like a magnet (and thus has magnetic energy) when current flows. This means it takes energy to set up an inductor with current just like it takes energy to send current through a resistor. However, there are two important differences. One difference is that the resistor continually transfers energy to thermal energy while current flows whereas the inductor transfers energy to magnetic energy only when the current *increases* (creating a stronger electromagnet). A second difference is that thermal energy is "lost" to the environment whereas magnetic energy remains with the inductor and is used to "drive" the current, like a battery, when the current *decreases* (using magnetic energy instead of chemical energy).

⁝ The inductor acts like a "status quo" device, helping to maintain the same current.

In this way, the inductor acts like a "status quo" device[x], helping to maintain the current. When current is increasing, the inductor acts like a resistor (to slow the increase in current). When the current is decreasing, the inductor acts like a battery (to slow the decrease in current).

This is why less current flows at higher frequencies. At higher frequencies, the current is oscillating more rapidly, and thus changing more rapidly. The

[viii] You also need to probe the inside of the coil, since the effect is significantly greater inside the coil than outside (as we know from chapter 5.

[ix] Also light energy but the light energy eventually transfers to thermal also.

[x] *Status quo* is Latin for 'state in which", meaning the present state of things.

faster the current changes, the more the inductor opposes those changes, leading to a lower current amplitude.

✍ Like with the capacitor, the impedance of the inductor actually refers to the average impedance as the current goes back and forth, and is somewhere between the zero (when the current is at its maximum value) and the value at the moment the current switches direction.

✓ *Check Point 17.4: According to our model, resistors transfer electric energy to thermal energy. What do inductors transfer electric energy to?*

17.2 Impedance vs. resistance

We use the term **impedance** to describe how a circuit responds to an AC voltage. Impedance, like resistance, is measured in ohms and represents how much an object impedes the flow of current.

IF RESISTANCE AND IMPEDANCE BOTH REPRESENT THE SAME THING, WHY DO WE NEED ANOTHER TERM FOR IT?

⁉ Impedance, like resistance, is measured in ohms and represents how much the element impedes the flow of current.

They don't represent *exactly* the same thing. Impedance is the general term and resistance is a particular type of impedance.

It is like the difference between fruits and oranges. Just as an orange is a type of fruit, resistance is a type of impedance. Resistors happen to have an impedance that is independent of frequency. A 100-Ω resistor has a resistance of 100 Ω, which is also its impedance.

✓ *Check Point 17.5: A resistor has an impedance of 100 Ω when an AC voltage of* $V_{\max} = 10.0$ *V and* $f = 1000$ *Hz is applied. In which of the following situations, if any, would the resistor's impedance be* $< 100\ \Omega$*?*
(a) When the voltage is changing slowly (frequency $\ll$ *1000 Hz)*
(b) When the voltage is changing rapidly (frequency $\gg$ *1000 Hz)*

Unlike the impedance of a resistor, which has an impedance that is the same regardless of the voltage frequency, the impedance of capacitors and inductors depend on the frequency. The capacitor's impedance is higher at

lower frequencies than at higher frequencies, consistent with the capacitor allowing less current to flow at lower frequencies than at higher frequencies. The inductor's impedance is the opposite and is impedance is higher at higher frequencies than at lower frequencies, consistent with the inductor allowing less current to flow at higher frequencies than at lower frequencies.

⁝ A capacitor's impedance is higher at lower frequencies, whereas the inductor's impedance is lower at lower frequencies.

Mathematically, we can write the relationships as follows:

$$Z_{\text{cap}} \propto \frac{1}{f} \tag{17.1}$$

$$Z_{\text{ind}} \propto f \tag{17.2}$$

where the impedance of each is indicated by the symbols Z_{cap} (for the capacitor) and Z_{ind} (for the inductor). Notice how the capacitor's impedance is inversely proportional to the frequency whereas the inductor's impedance is directly proportional to the frequency.

Example 17.1: When an AC voltage of frequency 10,000 Hz is applied to a circuit containing only a capacitor, the impedance is 100 Ω. What would the impedance be if the frequency is doubled to 20,000 Hz?

Answer 17.1: The impedance of a capacitor is inversely proportional to the frequency. Consequently, if the frequency doubles, the impedance is halved. So, the impedance would be halved to 50 Ω.

✓ *Check Point 17.6: A capacitor and an inductor each have an impedance of 100 Ω when an AC voltage of* $V_{\text{max}} = 10.0$ *V and* $f = 1000$ *Hz is applied. In which of the following situations, if any, would the capacitor's impedance be* < 100 *Ω? What about the inductor?*
(a) When the voltage is changing slowly (frequency $\ll$ *1000 Hz)*
(b) When the voltage is changing rapidly (frequency $\gg$ *1000 Hz)*

As mentioned before, impedance is the general term we use to describe how a circuit element impedes the current. Like resistance, impedance has units of ohms but the word "resistance" is only used for resistors.

Is impedance also represented by an R in equations?

No. We could but then it might be misinterpreted to only mean resistance, the impedance of a resistor. That is why I used the symbol Z for impedance. For example, we use Z_{cap} for the capacitor's impedance.

⁉ In equations, we use Z to represent the impedance.

WHY Z?

I don't know.[xi]

HOW DO WE DETERMINE THE IMPEDANCE OF AN ELEMENT?

We determine the *impedance* of an element the same way we determined the *resistance* of an element: by measuring the effect the element has on the circuit. In fact, impedance, like resistance, is defined as the ratio of the voltage across the element and the current through the element (compare to equation 15.1):

$$Z = \frac{V}{I} \tag{17.3}$$

Conversely, if you are given the impedance of an element and need to find the voltage across it (given the current through it) you can use

$$V = IZ \tag{17.4}$$

The expressions are the same as what we used for DC circuits except with Z instead of R to remind us that there may be a dependence on frequency.

IN THESE EXPRESSIONS, DOES IT MATTER IF WE USE THE MAXIMUM OR RMS VALUES (OF V AND I)?

No, as long as you are consistent. As before, if V is the voltage maximum then I is the current maximum. Similarly, if V is the RMS voltage then I is the RMS current.

✓ *Check Point 17.7: When an RMS voltage of 15 V is applied to a circuit, an RMS current of 5 A is produced. What is the impedance of the circuit?*

[xi] I have read that it is actually the Greek letter "zeta." However, I have also seen evidence that it was introduced along with a variable abbreviated as Y, which suggests that it was used because the letters at the beginning of the alphabet were already taken (see page 103).

17.3 Zero and infinite frequencies

Recall that the frequency represents how quickly something oscillates or repeats itself. A frequency of zero, then, means that there is no oscillation at all. A voltage with zero frequency essentially means it is a DC voltage, like the voltage from a battery. It doesn't oscillate at all.

Consider, now, what happens when we apply a zero-frequency voltage to a capacitor or an inductor. For the capacitor, the current stops, since the plates fill up, and current can only continue to flow if we alternately reverse the direction of the current. In such a situation (i.e., frequency is zero), the capacitor's impedance is infinity. This is consistent with equation 17.1 ($Z_{\text{cap}} \propto 1/f$), since one divided by zero is infinity.

The inductor, on the other hand, would have no impact since the inductor only has an impedance when the current is changing. In such a situation (i.e., frequency is zero), the inductor's impedance is zero.[xii] This is consistent with equation 17.2 ($Z_{\text{ind}} \propto f$).

Now let's consider what happens when we apply an infinite-frequency voltage to a capacitor. A frequency of infinity means that the voltage oscillates an infinite number of times every second. Of course this is not possible. However, that doesn't prevent us from consider what would happen under that circumstance.

With a capacitor, the current would flow unimpeded because there would be no time to fill up the capacitor plates. At infinite frequency, then, the current never has time to place any charge on the capacitor and the capacitor's impedance is zero (i.e., it doesn't impede the flow at all). This is consistent with equation 17.1 ($Z_{\text{cap}} \propto 1/f$), since one divided by infinity is zero.

The inductor, on the other hand, would have a really high impedance since the the current is changing so quickly. In such a situation (i.e., frequency being infinite), the inductor's impedance is also infinite, again consistent with equation 17.2 ($Z_{\text{ind}} \propto f$).

Example 17.2: When an AC voltage of frequency 10,000 Hz is applied to a circuit containing only an inductor, the impedance is 100 Ω. What would the impedance be if the frequency is doubled to 20,000 Hz?

[xii] This is ignoring the resistance that may be present with the wires themselves.

Answer 17.2: The impedance of an inductor is proportional to the frequency. Consequently, if the frequency doubles, the impedance also doubles (and the current is halved). So, the impedance would be doubled to 200 Ω.

✓ *Check Point 17.8: When an RMS voltage of 15 V is applied to a circuit containing only a capacitor, an RMS current of 5 A is produced.*
(a) What is the impedance of the capacitor?
(b) What would the impedance be if the frequency is doubled?
(c) What would the RMS current be if the frequency is doubled?
(d) What would the impedance of the capacitor be when the voltage doesn't oscillate at all (i.e., $f = 0$ Hz)?
(e) What would the impedance of the capacitor be when the frequency is infinitely high ($f = \infty$ Hz)?

17.4 Capacitance and inductance

Are all capacitors the same? Are all inductors the same?

While all capacitors have a larger impedance at lower frequencies, and all inductors have a larger impedance at higher frequencies, the actual impedance depends on the size of the capacitor and inductor.

Larger capacitors, with larger plates to store charge, have a lower impedance, all other things being equal. Meanwhile, larger inductors, which can create stronger magnets, have a higher impedance, all other things being equal.

The **capacitance** and **inductance** represent the sizes of the capacitor and inductor, respectively. So, a greater capacitance means the capacitor's impedance will be *lower* (all other things being equal) and a greater inductance means the inductor's impedance will be *higher* (all other things being equal).

⁉ A capacitor's capacitance represents its capacity to store charge.

✍ The capacitance depends not only on the size of the capacitor plates (larger plates = greater capacitance) but also how far apart the plates are (closer together = greater capacitance) and what is being used to separate the plates (more on this later). Similarly, the inductance depends not only on the size of the inductor but also the number of loops in the coil and what the coil is being wrapped around (more on this later).

Caution: The words "inductance" and "impedance" are very similar. They are so similar that many students confuse them. Remember that "inductance" describes the *inductor*: the greater the inductance, the stronger the magnet that a particular inductor creates (i.e., the more magnetic energy associated with a given current). The "impedance," on the other hand, represents the inductor's *effect on the circuit*. "Impedance" can be used with any element, not just inductors.

✓ *Check Point 17.9: For a given applied voltage and frequency, what should happen to the current if the inductance is increased?*

⁉ The SI unit of capacitance is the farad, F.

Although the impedance depends on the capacitance and inductance, capacitance and inductance are not the same thing as impedance. Consequently, the ohm is not the unit for capacitance and inductance. Instead, the SI unit for capacitance is the **farad**[xiii] (abbreviated as F) and the SI unit for inductance is the **henry**[xiv] (abbreviated as H).

⁉ The SI unit of inductance is the henry, H.

✍ According to the Electrostatic Discharge Association (ESDA), the human body has a capacitance[xv] of around 100 pF (picofarads) and an inductance between 0.4 and 2 μH. The capacitance of a cell membrane (for a 10-micron diameter cell) is about 3 pF. Most capacitors in electronic devices have a capacitance on the order of a micro-farad or less.

It turns out that the capacitor's impedance[xvi] is both inversely proportional to the voltage frequency and inversely proportional to the capacitor's capacitance.[xvii] We use C to represent the capacitance in equations so mathemat-

[xiii] The farad is named after English scientist Michael Faraday (1791-1867).

[xiv] The henry is named after Joseph Henry (1797-1878), an American physicist born in Albany, New York, who investigated inductance

[xv] The Electrostatic Discharge Association (ESDA) gives the capacitance as 100 pF, while *Electronic Noise and Interfering Signals - Principles and Applications* by G. Vasilescu (Springer-Verslag, 2005) suggests a range of 60 to 300 pF (see their page 349). The latter reference also provides a human body resistance between 330 and 10,000 Ω.

[xvi] The impedance of the capacitor is typically referred to as the **capacitive reactance** and is indicated by the variable abbreviation X_C. We'll use Z_{cap} not only so we don't have to add another variable abbreviation to remember but also because the word "reactance" can be easily confused with "reactants".

[xvii] It is possible to derive this relationship from $Z = V/I$ and the definition of capacitance, which is $C = Q/V$, where Q is the amount of charge stored on the capacitor for a given voltage V.

ically we can write the relationship as follows:[xviii]

$$Z_{\text{cap}} \propto \frac{1}{fC}$$

As mentioned earlier, the unit of impedance is the same as the unit of resistance: ohm (Ω).[xix]

In a similar way as with the capacitor, it turns out that the inductor's impedance[xx] is both proportional to the voltage frequency and proportional to the inductor's inductance.[xxi] We use L to represent the inductance in equations so mathematically we can Mathematically, we can write this as follows:

$$Z_{\text{ind}} \propto fL$$

As mentioned earlier, the unit of impedance is the same as the unit of resistance: ohm (Ω).[xxii]

WHY IS THE LETTER L USED FOR INDUCTANCE?

According to *The Physics Hypertextbook*, it was chosen to honor Emil Lenz.[xxiii]

✓ *Check Point 17.10: When an RMS voltage is applied to a circuit containing only a 10-nF capacitor, an RMS current of 4 A is produced. What RMS current would flow if the 10-nF capacitor was replaced by a 20-nF capacitor?*

[xviii] Don't confuse C, which represents the capacitance, with the SI unit C, which is the abbreviation for coulombs.

[xix] This relationship thus gives a relationship between ohm and farad: one ohm is equivalent to a second per farad.

[xx] The impedance of the inductor is typically referred to as the **inductive reactance** and is indicated by the variable abbreviation X_L. We'll use Z_{ind} for the same reasons we are using Z_{cap}.

[xxi] I haven't derived this. It can be derived using calculus.

[xxii] This relationship thus gives a relationship between ohm and henry: one ohm is equivalent to a henry second.

[xxiii] His full name was Heinrich Friedrich Emil Lenz, and he lived from 1804 to 1865. He formulated a relationship for inductance that we'll examine in chapter 18.

17.5 Dependence on structure

As mentioned earlier, larger capacitors have a larger capacitance because they can store a larger amount of charge for a given voltage before "filling up". Similarly, larger inductors have a larger inductance because they can store a larger amount of magnetic energy for a given current. What may not be so obvious, though, is how the capacitor's capacitance depends upon the material used to separate the plates (what we call the dielectric), and how the inductor's inductance depends upon the material we wrap the wire around (what we call the core).

17.5.1 Dielectrics

YOU DESCRIBED A CAPACITOR AS TWO PARALLEL PLATES SEPARATED BY A SMALL GAP. WHAT IS IN THE GAP?

In our idealistic capacitor, there was nothing in the gap. It was just a vacuum.

While a vacuum acts like an insulator (see section 10.3.2), in a real capacitor there needs to be some physical insulator in the middle to keep the two sides from touching. After all, if the two sides touch then charges will flow from one side to the other and we no longer have a capacitor.

IF WE PUT SOMETHING BETWEEN THE TWO PLATES, DOES THAT AFFECT THE CAPACITANCE OF THE CAPACITOR?

Yes, but probably not how you might expect. Putting the insulator between the plates *increases* the capacitance.

The two plates of a capacitor are separated by an insulator, which increases the capacitance by effectively decreasing the separation distance between the plates.

WHY DOES IT INCREASE THE CAPACITANCE? I WOULD HAVE THOUGHT PUTTING SOMETHING IN THE WAY WOULD DECREASE THE CAPACITANCE.

It increases the capacitance because it essentially *decreases* the space between the two plates, increasing the attraction of charges on opposite plates and leading to more charge (a higher capacity) for a given applied voltage.

To see why this is, let's first consider a capacitor with nothing in it (see figure 17.1.a). We'll indicate the capacitance of the capacitor with nothing in it (i.e., a vacuum) by C_0. We can then see how the presence of the insulator affects the capacitance by comparing the capacitance C with C_0.

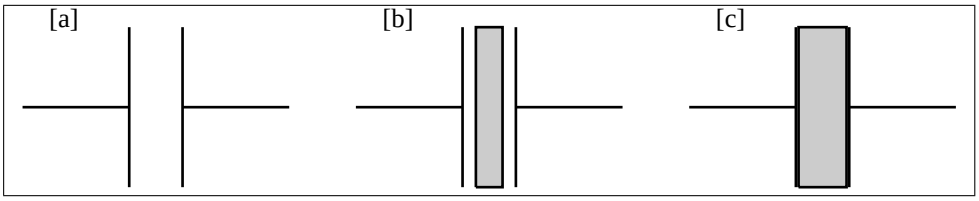


Figure 17.1: [a] A schematic of a capacitor with nothing in the gap. [b] A schematic of a capacitor with half of the gap filled by a conductor. [c] A schematic of a capacitor with the gap filled by an insulator.

Keep in mind that the vacuum doesn't prevent the charges on one side from "feeling the presence" of the charges on the other side. The charges on one side have a charge that is opposite the charges on the other. They *want* to cross over the gap but it takes too much energy to leave the conductor.

Let's suppose we now replace half of the gap with a conductor (see figure 17.1.b). What we've done is essentially halved the width of the gap. As we noted above, decreasing the gap makes the capacitance increase. In much the same way, whenever the gap is filled, the capacitance increases. Mathematically, this is expressed as $C \geq C_0$.

Unfortunately, putting a conductor in the gap provides a path for the charge to flow across the gap. We don't want that. So, instead, we fill the gap with an insulator (see figure 17.1.c). The increase in capacitance won't be as great as if we had filled it with a conductor but at least there won't be a path available for the charge to cross the gap.

When an insulator is used in this way, it is called a **dielectric** and its impact on the capacitance is called its **dielectric constant** (see table 17.1). A larger dielectric constant means the capacitance would be larger with the dielectric than with nothing occupying the gap. Using a dielectric with a dielectric constant of two would mean the capacitance is twice as much as what it would be with nothing occupying the gap.

WOULDN'T THE INSULATOR PREVENT THE CHARGES ON ONE PLATE FROM BEING ATTRACTED TO THE OTHER PLATE?

No. An insulator prevents charges from flowing through it but it doesn't act like a shield, preventing the charges on one side from "feeling the presence" of the charges on the other side. The charges on one side still *want* to cross over the gap but can't because it takes too much energy to leave the conductor.

Material	Dielectric Constant
Vacuum	1
Air (dry, °C)	1.000536
Paper (dry)	2.0
Rubber (hard)	2.8
Glass (silica)	3.8
Water (20°C)	80.4
Water (0°C)	88.4

Table 17.1: Dielectric constant for various materials (ASI Instruments, Inc.)

✍ Perhaps I shouldn't say there "won't be a path" but rather there is "less likelihood". As mentioned in section 6.2, an insulator can break down if the electric field is too great (as the negative and positive particles are pulled apart). The maximum electric field an insulator can withstand before breaking down is called the **dielectric strength**.

✓ *Check Point 17.11: A capacitor has a capacitance of 3 μF when used with a dielectric with dielectric constant equal to 2. What is its capacitance if it is used with a dielectric with dielectric constant equal to 4?*

17.5.2 Ferromagnetic cores

What happens if we place a metal bar into the inductor?

If the bar is ferromagnetic, the inductor's impedance will be greater with the bar inside it because, as mentioned on page 208 in chapter 12, a ferromagnetic core increases the overall magnetic strength of the electromagnet[xxiv], thus increasing the overall magnetic energy and increasing the inductor's impact on the circuit.

✓ *Check Point 17.12: According to our model, why does the impedance of an inductor depend on whether a metal bar is placed within the inductor?*

[xxiv] This is because the electromagnet aligns the little magnets in the bar, making it into a magnet (as discussed in section 4.5), assuming the metal bar is not already a magnet.

Metal detectors utilize this property. Suppose we had an inductor so big you could walk through it. The inductor's impedance is larger when you walk through it with ferromagnetic metal, allowing us to determine if you are carrying any ferromagnetic metal objects simply by measuring its impedance.[xxv]

A small-scale metal detector can be constructed with a simple circuit. For example, you can construct a circuit with an inductor such that a bulb goes off or on depending on whether metal is placed within the inductor.

How is this done?

One way would be to place the inductor along the same path as a bulb. Use a frequency of applied voltage such that the impedance of the inductor is small enough to just light the bulb.

If you then place a piece of metal inside the inductor, the impedance of the inductor goes up. If you are careful about it, you can arrange it so that the increase in impedance is just enough to make the bulb turn off.

✓ *Check Point 17.13: As described above, the bulb will turn off when metal is placed within the inductor. Describe the configuration and explain why adding the metal turns the bulb off.*

Summary

This chapter examined the relationship between the structure and properties of capacitors and inductors.

The main points of this chapter are as follows:

- The resistance of a resistor is independent of voltage frequency.
- A capacitor allows more current to flow at higher frequencies.
- An inductor allows more current to flow at lower frequencies.
- A capacitor represents a break in the circuit, but can store charge, thereby giving the impression that current flows through it.

[xxv] Non-ferromagnetic metals can also be detected, using a property called induction, which is explained in chapter 18.

- An inductor is a solenoid (coil of wire) and transfers energy between electric and magnetic energy, allowing it to act like a "status quo" device, helping to maintain the same current.
- Impedance, like resistance, is measured in ohms and represents how much the element impedes the flow of current.
- A capacitor's impedance is higher at lower frequencies, whereas the inductor's impedance is lower at lower frequencies.
- In equations, we use Z to represent the impedance.
- A capacitor's capacitance represents its capacity to store charge.
- The SI unit of capacitance is the farad, F.
- The SI unit of inductance is the henry, H.
- The two plates of a capacitor are separated by an insulator, which increases the capacitance by effectively decreasing the separation distance between the plates.

By now you should be able to explain how the structure of an inductor and a capacitor impacts how they effect the circuit (in terms of its impedance and the dependence on frequency).

Frequently asked questions

What does it mean to have a voltage with frequency equal to ∞ Hz (in checkpoint 17.8)?

The symbol ∞ means infinity. A voltage with infinite frequency means the voltage changes directions so quickly that no charge gets deposited onto the capacitor plates. We never actually have a voltage of infinite frequency. I just use that as an extreme example.

How can current flow if the two sides of the capacitor are not connected?

Because of the gap between the plates of capacitor, current cannot flow for an extended period of time. However, it *can* oscillate back and forth if it oscillates quick enough. In a sense, it like rubbing your hands to warm them up. To do so, you need to move your hands back and forth. If you don't oscillate your hands, you'll "run out" of space on your hand and your hands won't warm up.

When current flows "through" the capacitor, does the capacitor as a whole get charged?

No. The capacitor as a whole remains neutral in that the negative charge on one plate is balanced out by the positive charge on the other plate.

What is the difference between R and Z?

They are really the same thing, in that both are measured in ohms and both represent the extent to which the element impedes the current. We just use Z (impedance) as the general term and restrict R (resistance) to just the impedance that does not depend upon frequency. For a resistor, its impedance is equal to its resistance.

Does C stand for coulombs or does it stand for capacitance?

It is unfortunate that we use the same letter for the unit of coulombs *and* the variable of capacitance. However, remember that units are written in Roman font and variables are written in *Italic* font. The unit of charge (Q) is coulomb (C). The unit of capacitance (C) is farad (F).

Is the impedance of the inductor due to the resistance of the wire that makes it up?

No. We usually assume that the resistance of the inductor is zero (i.e., it is made up of a very low resistance wire). In other words, the impedance of the inductor is *not* due to the resistance of the wire. Indeed, if it were, its impedance would not depend upon frequency.

When discussing the energy "stored" in an inductor, do we also include the kinetic energy of the electrons?

No. The kinetic energy of the electrons is tiny since the electrons have such a small mass and drift velocity. The kinetic energy associated with the movement of electrons through a typical inductor is about 10^{-12} J (see footnote[xxvi]), much smaller than the typical magnetic energy associated with an inductor.[xxvii]

[xxvi] In copper there are $\approx 8.5 \times 10^{28}$ free electrons/m^3 (assuming one free electron per atom). Multiply that by the wire's cross-sectional area (≈ 0.50 mm^2) and the electron mass (9.11×10^{-31} kg) to get the linear mass density of free electrons (3.87×10^{-8} kg/m). For a typical inductor of 500 loops of radius 2 cm, the total length is 62.8 m and so the total mass of moving electrons is 2.43×10^{-6} kg. The kinetic energy is obtained by using the kinetic energy definition ($\frac{1}{2}mv^2$) and a typical drift speed (≈ 1 mm/s).

[xxvii] The magnetic energy associated with a typical inductor of 10 mH and a current of 0.1 A is around 0.05 mJ. The actually equation is $^1/_2LI^2$.

Terminology introduced

Capacitance	Dielectric	Inductive reactance
Capacitive reactance	Farad	Inductor
Capacitor	Henry	Plasma membrane
Dielectric constant	Impedance	
Dielectric strength	Inductance	

Abbreviations introduced

Quantity	**SI unit**
capacitance (C)	farad (F)[xxviii]
impedance (Z)	ohm (Ω)
inductance (L)	henry (H)[xxix]

Additional problems

Problem 17.1: When an RMS voltage of 15 V and 10,000 Hz is applied to a circuit containing only a capacitor, an RMS current of 3.7 A is produced. What is the impedance of the capacitor?

Problem 17.2: When a RMS voltage of 15 V is applied to a circuit containing only an inductor, an RMS current of 5 A is produced.
(a) What is the impedance of the inductor?
(b) What would the impedance be if the frequency is doubled?
(c) What would the RMS current be if the frequency is doubled?
(d) What would the impedance of the inductor be when the voltage doesn't oscillate at all (i.e., $f = 0$ Hz)?
(e) What would the hypothetical impedance of the inductor be when the frequency is infinitely high ($f = \infty$ Hz)?

[xxviii] A farad is equal to a second per ohm (s/Ω).
[xxix] A henry is equal to a ohm per second (Ω/s).

18. Magnetic Induction

Puzzle #18: If an inductor is just a solenoid (coil of wire), why do we call it an inductor instead of a solenoid?

Introduction

I'll address the puzzle right away. The reason we call an inductor an inductor, rather than just a coil of wire or solenoid, is because of a property called **magnetic induction**, which lies at the heart of what an inductor does. In this chapter, we'll explore magnetic induction.

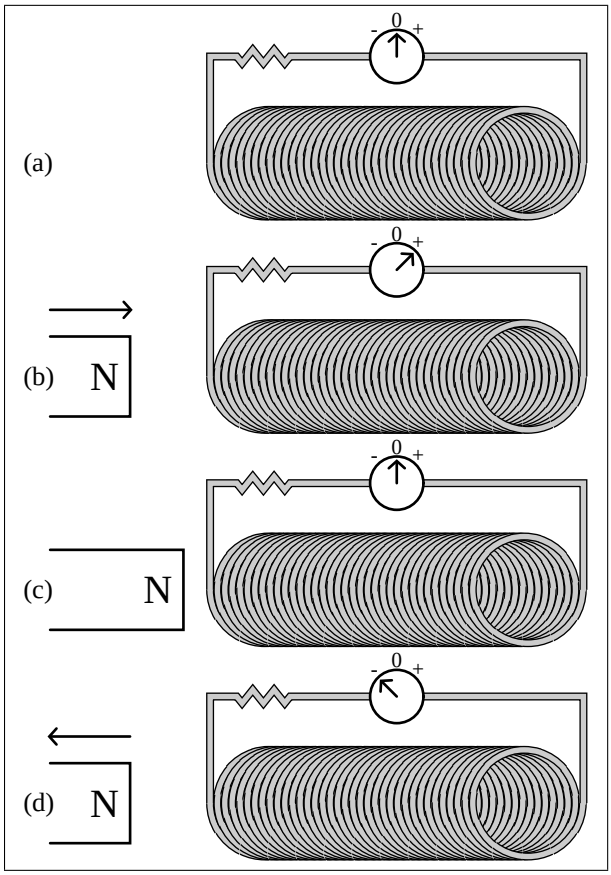
As we've shown in chapter 17, it isn't necessary to understand magnetic induction to explain how an inductor works. Still, understanding magnetic induction allows us to not only gain insight into how an inductor works and why it is called an inductor, but it also allows us to understand how electric generators and transformers work (we'll explore each as part of this chapter).

18.1 Current induction

To see magnetic induction in action, we can probe a solenoid with a magnet. Something very interesting happens when you do this.

Suppose we connect a solenoid to an ammeter (circle with arrow) as illustrated in part (a) of the figure. If nothing else is in the circuit, there is no current flowing through the solenoid and the ammeter reads zero. This is exactly what we would expect (based on our model so far).

One observes, however, that the ammeter gives a *non-zero* reading when a magnet is moved toward the solenoid or away from it, as illustrated in figures (b) and (d), with the current flowing one way as the magnet is brought into the coil and the current flowing the opposite way as the magnet is brought out of the coil.



Curiously, if you repeat the process but with the south end of the magnet instead of the north end, the current direction in each case is flipped (leftward through the ammeter as the magnet is brought into the coil and rightward as the magnet is brought out). The direction of the current also depends on how the wires are coiled in the solenoid.

Yet, regardless of which end of the magnet is used, or how the solenoid is coiled, simply leaving the magnet nearby (without moving it) has no effect. The current is zero. This is illustrated in figure (c).

This is not something that we can explain with our model so far. After all, the magnet is electrically neutral. It should have no impact at all on the electrons inside the wires that make up the solenoid.

To explain this, we need to modify our model.

In our new model, a coil is a *magnetic field change sensor* in that it can sense

when the magnetic field inside the coil changes. As we know, each magnet has a magnetic field, which is strongest close to the magnet and weaker farther away from the magnet. Consequently, when you move a magnet further away from the coil, the coil is able to sense the decrease in magnetic field at the coil's location. Conversely, when you move a magnet toward the coil, the coil is able to sense the increase in magnetic field at the coil's location.

What is really interesting is what the coil does when it senses a change in the magnetic field – current is *induced* to flow through the coil – a process called magnetic induction.*induction*[i].

❣ Changing the magnetic field inside a solenoid induces current to flow in the solenoid.

What do you mean by "induce"?

Here the term **induce** is used in much the same way as it was used in section 10.2.3. We are "coaxing" current to flow (or stop) without applying a voltage directly or touching the solenoid with the magnet.

It is important to note that it isn't the *strength* of the magnetic field that is important but rather whether the magnetic field is *changing*.

Leaving the magnet close by, as in part (c) of the figure, does *not* induce current to flow. Only by changing the magnetic field (by moving the magnet closer or farther away) will induce current to flow.

✍ It turns out that the *quicker* one changes the magnetic field, the *more* current that will be induced to flow. This property will be used in the next section.

✓ *Check Point 18.1: How do the observations illustrated in the figure on the previous page show that it isn't the presence of the magnet that induces the current but rather the* change *in the magnet that does?*

18.2 Transformers

To demonstrate that the motion of the magnet is unnecessary and that it is really the changing magnetic field inside the solenoid that is important, let's

[i] As mentioned in chapter 10, we are using the word "induction" in much the same way one would use it for placing someone into elected office. In both cases, the word "induct" has to do with leading into, and comes from the Latin *ducere*, which means "to lead."

consider how a transformer works.

A **transformer** is a device that outputs a voltage that has a different amplitude (but same frequency) than the input voltage. In other words, it "transforms" the voltage amplitude to a new value, either higher or lower.

This type of device is used when connecting a house to the power grid. Electric energy is distributed via high voltage lines. The voltage on those lines is much higher than what a typical house uses.[ii] So, a transformer is used to "step down" the high voltage to a lower voltage.[iii]

It turns out we can explain how a transformer works by using our knowledge of magnetic induction. It basically involves two solenoids. When an AC voltage is applied to one solenoid, current is *induced* to flow in the other.

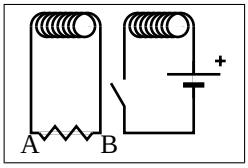
To understand the process, let's review a few basic ideas.

We know that moving a magnet toward or away from a solenoid will induce current to flow in the solenoid because the magnetic field is changing within the cavity of the solenoid.

As mentioned in the previous section, the *quicker* one changes the magnetic field, the *more* current that will be induced to flow. In the previous section this could be done by moving the magnet more quickly.

In this case, we are replacing the moving magnet with a stationary electromagnet. To change the magnetic field inside the solenoid, we just need to "turn on" and "turn off" the electromagnet (instead of moving the electromagnet toward and away from the solenoid).

For example, consider the two solenoids side by side as indicated to the right. With the switch open, current is not flowing through either.



What happens when the switch is closed?

[ii] The neighborhood distribution lines are at 13,000 V, whereas the voltage in a house is 120 V (or 240 V for some devices, like an oven). Transmission lines from the power generation station are typically at 100,000 V.

[iii] The outside lines are at a higher voltage in order to reduce frictional heating. To understand why, consider that the power delivered to the home is IV (from equation 14.2). By making the voltage high in the outside wire, that means the current can be very small in the outside wire. The smaller the current, the less frictional heating in the wire. The high outside voltage allows us to reduce frictional heating loss while still providing the necessary power to the community.

When the switch is closed, current starts to flow through the right solenoid. That means that the right solenoid becomes an electromagnet.

It is as though we just brought a magnet toward the left solenoid. As a consequence, current is induced in the *left* solenoid.

Not only will current be induced when the switch is closed and current *starts* to flow through the right solenoid, but current will also be induced when the switch is opened and current *stops* flowing through the right solenoid. That would be like moving a magnet away from the left solenoid.

Rather than opening and closing a switch, we can turn the current on and off just by applying an AC voltage.

For example, if a 60-Hz AC voltage is applied to a solenoid, this produces a magnetic field through the solenoid that is likewise changing at 60 Hz. If we want to increase the induced current, simply increase the frequency (rather than increase the strength of the field).

⁉ Oscillating the current in one solenoid will induce an oscillating current to flow in the other solenoid.

✍ The higher the frequency, the greater the effect, since the magnetic field is changing at a higher frequency.

A transformer, then, is essentially two solenoids, with one solenoid inducing current to flow in the other.

To enhance the effect, one can physically insert one solenoid into the other (i.e., "wrap" one solenoid around the other). Another way would be insert a ferromagnetic rod that goes through both solenoids. Since the entire piece of metal acts like a magnet, which not only helps to ensure that both solenoids experience similar magnetic field values (since the same bar is in both solenoids) but would would also increase the overall magnetic field.

✍ The induced voltage depends upon the number of coils in each solenoid. The coil with the higher number of loops will end up being at a higher voltage.

If one is inside the other, what if the two solenoids switch positions so that solenoid 2 is on the outside rather than the inside?

It turns out that it doesn't matter. The same expression applies since they both experience the same magnetic field within them.

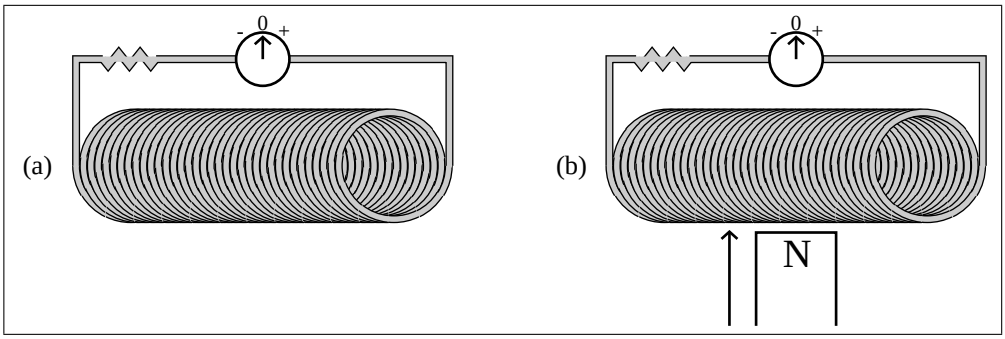


Figure 18.1: (a) A solenoid connected to an ammeter and resistor. The ammeter shows no current flowing through the solenoid. (b) As the north pole of a magnet is brought near the solenoid, current still doesn't flow because the magnetic field associated with the magnet is not parallel to the solenoid axis.

✓ *Check Point 18.2: A bulb is connected to a solenoid with nothing else. We try the following methods to light the bulb:*
(1) Inserting a magnet into the solenoid.
(2) Leaving a magnet inside the solenoid.
(3) Leaving a second solenoid inside the first solenoid, where the second solenoid is connected to an AC voltage source oscillating at 1000 Hz.
We find that only method (3) lights the bulb. Why?

18.3 Electric generators

So far we've seen two ways that we can induce current to flow in a solenoid. In section 18.1, we moved a magnet toward and away from the solenoid. In section 18.2, we turned on and off an electromagnet. In both cases, the key is that we are changing the magnetic field inside a solenoid, and that changing magnetic field induces current to flow in the solenoid.

It turns out that solenoid is only sensitive to changes in the magnetic field *parallel* to its axis. This property can be used to generate electricity.

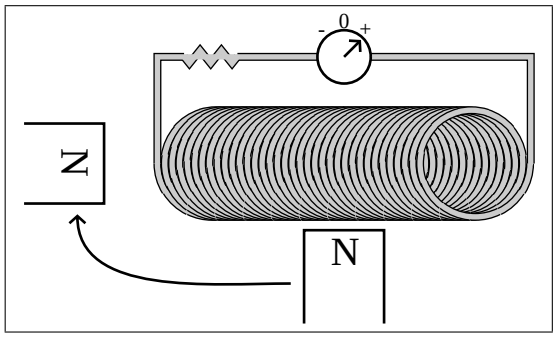
⁝ The solenoid is only "sensitive" to changes in the magnetic field that is parallel to the solenoid axis.

To understand how this sensitivity can be used to generate electricity, consider the process illustrated in Figure 18.1, where a magnet is oriented per-

pendicular to the solenoid axis. In that situation, nothing happens. In other words, one does not induce a voltage if the magnet's field is perpendicular to the solenoid axis, regardless of whether the magnet moves or not.

On the other hand, since the solenoid is sensitive to the magnetic field parallel to its axis, one can induce current simply by re-orienting the magnetic field, even if the *strength* of the magnetic field remains the same. All we need to do is change the part parallel to the solenoid axis.

This is illustrated to the right. If one moves the magnet as shown, current flows through the meter because moving the magnet in this way introduces a magnetic field *parallel* to the axis that wasn't there before. The key, then, is the change in the magnetic field *parallel* to the axis.



✓ *Check Point 18.3: A solenoid of cross-sectional area 0.0160 m^2 is placed in a uniform magnetic field of magnitude 0.35 T directed toward the* **north**. *For each of the following situations, answer "yes" or "no" and explain your choice.*
(a) If the axis of the solenoid is also directed toward the **north**, *will current be induced when the field increases in strength to 0.70 T in 5 seconds?*
(b) If the axis of the solenoid is also directed toward the **north**, *will current be induced when the magnitude of the field is held steady but the direction of the field changes toward the east?*
(c) If the axis of the solenoid is directed vertically **downward** *(instead of north), will current be induced when the field increases in strength to 0.70 T in 5 seconds?*
(d) If the axis of the solenoid is directed vertically **downward**, *will current be induced when the magnitude of the field is held steady but the direction of the field changes toward the east?*

To generate electricity, we have to use the property that the solenoid is only sensitive to changes in the magnetic field parallel to its axis.

One way to do change the magnetic field parallel to the axis is to reorient

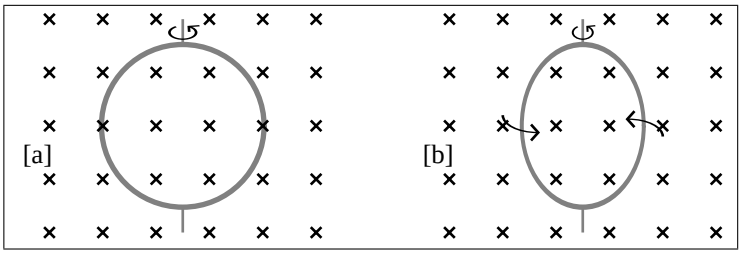


Figure 18.2: Two pictures of a loop that is spinning, like a top, in a region where the magnetic field is directed into the page. [a] At this moment, the axis of the loop is perpendicular to the page and so is the magnetic field. [b] At this moment, the loop has rotated slightly (counter-clockwise as seen from above). The axis of the loop is no longer parallel to the magnetic field.

the magnet. Another way is to reoriented the *solenoid*.[iv]

So, rather than moving the magnet, we could instead rotate the solenoid.

For example, consider the loop in Figure 18.2, which is placed in an externally-applied magnetic field directed into the page. Here I use a single loop instead of a solenoid only to make the picture clearer.[v]

One finds that as the loop is rotated in the presence of the fixed magnetic field, current is induced in the loop.[vi] Simply moving the loop is not sufficient. It must *rotate*.

⁝ One can induce current to flow by either changing the orientation of the magnetic field through the solenoid or by changing the orientation of the solenoid itself.

What if the entire loop is moved parallel to the magnetic field?

Then nothing happens (assuming the strength of externally-applied magnetic field is the same everywhere – if the external field is provided by a bar magnet, this will not be true). For example, if the loop in Figure 18.2a is moved toward you and the strength of the externally-applied magnetic field (parallel to the loop axis) remains unchanged, then no current is induced in the loop.

What if the entire loop is moved perpendicular to the magnetic field?

[iv] In a sense, this is a result of one of nature's symmetry laws.

[v] The same effect will be observed in both a single loop and a solenoid except that the solenoid, being constructed of many loops, will amplify the effect.

[vi] We are essentially creating an a voltage or **emf** by moving the wire within the magnetic field (see footnote about emf on page x). For this reason, this particular phenomenon is sometimes called a **motional emf**.

Again, nothing happens (assuming the strength of magnetic field is the same everywhere). For example, if the loop in Figure 18.2a is moved toward the left and the strength of the externally-applied magnetic field (parallel to the loop axis) remains unchanged then no current is induced in the loop.

WHY DO WE HAVE TO ASSUME THAT THE STRENGTH OF THE MAGNETIC FIELD IS THE SAME EVERYWHERE?

Because if you move the loop into a region that has a different magnetic field, that will induce a current in the loop.

IS THAT A GOOD ASSUMPTION TO MAKE?

Not if the external magnetic field is being provided by a small magnet. However, if we consider the magnetic field produced by Earth, it is a pretty good assumption. We could also produce a magnet that is large compared with the motion of our loop.

SO WHAT DOES THIS HAVE TO DO WITH THE GENERATION OF ELECTRICITY?

We can induce current to flow by rotating a solenoid. All we need is a magnet nearby and a way to rotate the solenoid.

⁉ In practice, rotating a solenoid is how electricity is generated.

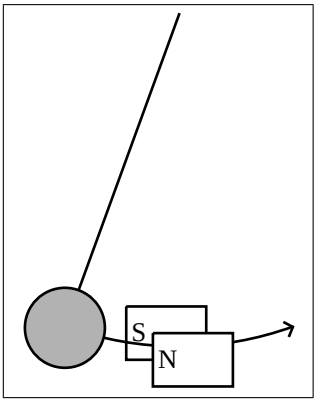
✍ This method of electricity generation produces alternating current. As the solenoid rotates, the magnetic field parallel to its axis continually switches directions. This induces current that also changes directions. This is discussed further in section 18.4.

✓ *Check Point 18.4: A circular loop of cross-sectional area 0.0160 m^2 is placed in a uniform magnetic field of magnitude 0.35 T directed toward the north (i.e., it has the same strength and direction everywhere). For each of the following situations, answer "yes" or "no" and explain your choice.*
(a) If the loop axis is also directed toward the north, will current be induced when the loop is moved northward at 10 m/s?
(b) If the loop axis is also directed toward the north, will current be induced when the loop is moved eastward at 10 m/s?
(c) If the loop axis is also directed toward the north, will current be induced when the loop axis is reoriented toward the east?

18.4 Magnetic brakes

A third application of induction is **magnetic braking**.

Magnetic braking can be observed by passing an aluminum plate between the poles of a large magnet, as illustrated to the right, where a circular aluminum plate (in gray), hanging from a string, passes between the poles of a magnet as the plate swings back and forth.



Aluminum is not ferromagnetic and so it is not attracted to the magnet. However, one finds that when the plate enters the region of the magnet, there is a force on the plate acting to slow it down.

This last point is important – the aluminum plate is not ferromagnetic so it is not forced *toward* the magnet as a ferromagnetic material would. Rather, the force is opposite the plate's *motion*.

⁉ In magnetic braking, there is a magnetic force due to the induced current, directed opposite the motion of the conductor.

It turns out that magnetic induction is causing the magnetic braking. Aluminum is a conductor so the plate acts like solenoid in the sense that current can flow through it. When the aluminum plate is moved into the region between the magnets, the plate experiences a change in the magnetic field and, consistent with magnetic induction, that change induces a current in the plate. Similarly, when the plate is moved out of the magnetic region, the plate again experiences a change in the magnetic field and so current is again induced in the plate.

✍ Note that the braking effect only occurs when the plate is entering or leaving the magnetic region (so part of the plate is within the magnetic region and part of the plate is outside) because that is when the plate is experiencing a changing magnetic field. There is no braking regardless of any motion when the plate is fully within the magnetic region.

✓ *Check Point 18.5: Consider the situation shown on page 330 where a small aluminum plate is passed through the poles of a large magnet. As the plate is moving rightward and leaving the magnetic region, is there any force on the plate? If so, in which direction is it?*

WHY DOES THIS RESULT IN BRAKING?

It turns out that magnetic braking is a consequence of the "opposing" nature of magnetic induction.

To see this opposing nature, go back and re-examine the figure on page 322 that showed a solenoid connected to an ammeter. That is the figure that showed how current is induced in the solenoid when a magnet is brought toward the solenoid (part b of the figure) or away (part d), but no current is induced when the magnet is stationary (part c).

If you examine the direction of the current in part (b) carefully, you'll find that the direction of the solenoid's magnetic field is *opposite* the magnet's magnetic field. In other words, it is as though the current is produced in such a way as to maintain the magnetic field that was there prior to the introduction of the magnet. In particular, the magnet in part (b) introduces a *rightward*-pointing magnetic field (away from the magnet's north pole) and, in response, the solenoid sets up a clockwise current (as seen by someone to the right) which, according to our right-hand rule, produces a *leftward*-pointing magnetic field.

⁉ The induced magnetic field counters the change that is introduced (Lenz's law).

In other words, introducing the magnet has upset the status quo (no magnetic field prior to the introduction of the magnet) and, in response, current is set up in the solenoid that produces a magnetic field counter to that of the magnet. This relationship is known as **Lenz's law** (i.e., the induced magnetic field of the solenoid counters the introduced magnetic field).

IN PART (C) OF THE FIGURE, THERE IS NO CURRENT FLOWING EVEN THOUGH THE MAGNET IS REALLY CLOSE. WHY NOT?

In part (c), the magnet is no longer moving. Its contribution to the magnetic field inside the solenoid is no longer changing. Since it is no longer changing, the solenoid no longer needs to provide a counter magnetic field.

WHAT HAPPENED TO THE CURRENT THAT WAS FLOWING THROUGH THE SOLENOID?

It very quickly decreased to zero because of the small resistance in the wires.

WHAT HAPPENS WHEN THE MAGNET IS REMOVED?

As the magnet is removed, the magnetic field inside the solenoid changes. In particular, the rightward-pointing magnetic field that was there is no longer present. Again, the status quo has been upset. This time, the current set up

in the solenoid (in response to the change) runs counter-clockwise through the solenoid (as seen by someone to the right). Such a current, according to our right-hand rule, produces a rightward-pointing magnetic field (helping to maintain what was there before).

IN THE FIGURE, THE CURRENT THROUGH THE AMMETER IS ALWAYS IN THE SAME DIRECTION AS THE MOTION OF THE MAGNET. IS THIS ALWAYS GOING TO BE THE CASE?

No. It just so happens to be that way here. It depends on where you measure the current, which way the solenoid is wound and which pole is being brought near the solenoid.

✍ In chapter 6, we found that moving charges (within a wire) have magnetic properties. Magnetic induction is just the reverse: moving magnets have electric properties. This is just a consequence of velocity being relative: there is no difference between the object moving relative to the magnet vs. the magnet moving relative to the object.

✓ *Check Point 18.6: How would the current through the ammeter be different in the four parts of the figure on page 322 if a* south *pole of a magnet (rather than a north pole) was brought toward the solenoid? Assume the same motion from the left.*

We can use Lenz's law not only to explain magnetic braking but also to explain why an inductor is called an inductor.

The inductor is just a solenoid. And, as we've discussed, a solenoid senses change in the magnetic field within it. And, if there is a change, current flows through the solenoid in a way that creates a magnetic field of its own such to counter the introduced magnetic field. In other words, the inductor uses magnetic induction to "maintains" a constant magnetic field within it (or at least tries to). This is why it acts as a as a "status quo" device – always trying to maintain the same current through it. It is basically sensing the changing magnetic field within it and using magnetic induction to modify the current and minimize the change.

We can also use Lenz's Law to determine the direction of the induced current in the transformer and the generator.

For example, on page 324, there is an illustration of two solenoids that are

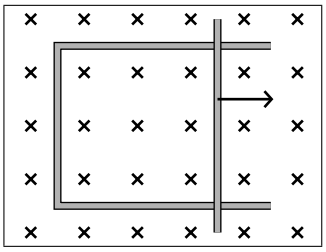


Figure 18.3: A square made of two pieces of wire: a C-shaped piece that is held fixed and a straight piece of wire that is allowed to move on top of the two extensions of the C-shaped piece. If the straight wire is moved to the right, the size of the loop increases.

side-by-side. Looking from the right, the current starts flowing counter-clockwise through the right solenoid when the switch is closed. This produces a magnetic field directed toward the right (using the right-hand rule). Consequently, closing the switch introduces a rightward-directed magnetic field in the *left* solenoid. In an attempt to maintain the status quo, the left solenoid then produces a magnetic field directed toward the left. This means current must flow through the left solenoid clockwise (as seen from the right). Use your finger to indicate this in the figure and you'll find that this means current flows from B to A through the resistor.

The electric generator is illustrated in Figure 18.2 on page 328. In Figure 18.2a, the magnetic field (applied from some outside source) is directed parallel to the loop axis. In Figure 18.2b, the magnetic field is no longer parallel to the loop axis (it is only partly parallel). Consequently, the component parallel to the axis has *decreased*. To maintain the magnetic field parallel to the axis, current must flow clockwise through the loop (according to the right-hand rule described on page 211).

As the loop continues to rotate, it will eventually return to its original position. Along the way, the induced current will need to switch directions (as seen from the original perspective). In this way, AC current is produced.

As one final investigation of Lenz's Law, one can show that rather than *rotating* the loop, one can also just decrease or increase the *size* of the loop. An example is shown in figure 18.3.

In the figure, there is a square-shaped loop made up of a C-shaped wire and a straight wire. As the straight wire moves rightward, the square-shaped loop increases in size. Since the magnetic field is pointed into the page, this is essentially *increasing* the "amount" of the magnetic field through that loop. To maintain the magnetic field that was there prior to the loop increasing in size, current is induced to flow counter-clockwise around the loop. Using the right-hand rule, one can see that a counter-clockwise current produces a

magnetic field *out* of the page, maintaining the "amount" of magnetic field that was present before.

It turns out that, just as with magnetic braking, there is a force on the straight wire, opposing its rightward motion.

✓ *Check Point 18.7: Suppose the straight wire piece in Figure 18.3 was moving leftward. Does current flow through the square-shaped loop? If so, in what direction?*

Summary

This chapter examined how current is induced to flow when a conductor (like a coil of wire) experiences a changing magnetic field. We call this phenomenon *magnetic induction*.

The main points of this chapter are as follows:

- An inductor acts to maintain a constant magnetic field.
- Changing the magnetic field inside a solenoid induces current to flow in the solenoid.
- The solenoid is only "sensitive" to changes in the magnetic field that is parallel to the solenoid axis.
- One can induce current to flow by either changing the orientation of the magnetic field through the solenoid or by changing the orientation of the solenoid itself.
- Oscillating the current in one solenoid will induce an oscillating current to flow in the other solenoid.
- In practice, rotating a solenoid is how electricity is generated.
- In magnetic braking, there is a magnetic force due to the induced current, directed opposite the motion of the conductor.
- The induced magnetic field counters the change that is introduced (Lenz's law).

By now you should be able to describe magnetic induction and how it is used in inductors, electricity generation, magnetic braking and transformers.

Frequently asked questions

Does the magnet exert a force on the charges in a wire?

A magnet doesn't exert a magnetic force on a charge unless the charge is already moving (and thus has magnetic properties, like with electric current through a wire). In comparison, magnetic induction is a force that occurs when the magnetic field is *changing*. It is not the same as the magnetic force (which occurs whenever the field is present, whether it is changing or not).

Terminology introduced

Induce	Magnetic induction
Lenz's law	Motional emf
Magnetic braking	Transformer

Additional problems

Problem 18.1: A circular loop of cross-sectional area 0.0160 m^2, oriented with its loop axis directed north-south, is placed in a uniform magnetic field of magnitude 0.35 T directed toward the north.
(a) If the loop is moved toward the east at 10 m/s, in what direction is the magnetic force exerted on the protons in the loop?
(b) Is any current induced in the loop? Why or why not?
(c) Will current be induced if the cross-sectional area of the loop is decreased to 0.0080 m^2 in 5 seconds? If so, in what direction (as seen from the from the north)? If not, why not?

Part E

Waves

19. Sound

Puzzle #19: How does a speaker, like that in a phone or radio, work to make sound?

Introduction

The way sound is produced, as with a speaker, and gets from the sound "producer" to us, is fascinating yet relatively simple. It involves waves, the focus of the rest of this book. The language we use to describe waves is similar to the language we used to describe the current and voltage oscillations in an AC circuit, like frequency and amplitude. There are lots of phenomena that involve waves, including water waves, musical instruments, and light. We start our examination of waves in this chapter by focusing on sound.

19.1 Describing sound

We'll start our examination of sound by considering how it is produced.

Basically, to make a sound, we have to make something vibrate. The vibrating object could be our vocal chords, a bell, a string, or even a paper plate. By changing the way the object vibrates, we can create different sounds.

Sounds vary in two basic ways: **pitch** and **loudness**. Examples of high pitch sounds are the squeak of a mouse and the sound made by a piccolo (a small flute), whereas low pitch sounds include the moo of a cow and the sound made by a tuba. High pitch sounds can be loud or soft, just as low pitch sounds can be loud or soft.

19.1.1 Pitch and frequency

To create a high pitch sound, the object needs to be vibrating very quickly. We know from our investigation of AC voltage (see section 16.1) that something oscillating quickly has a high frequency[i]. It is beyond the scope of this course to explain how the ear is able to interpret a high frequency sound as having a high pitch. All we need to recognize is that high-pitch sounds are high-frequency sounds. Similarly, low-pitch sounds are low-frequency sounds.

Note that frequency only controls the pitch, not the loudness. So, high-frequency (high-pitch) sounds can be loud or soft, just as low-frequency (low-pitch) sounds can be loud or soft.

To illustrate how this works, consider a speaker. A speaker is really just a paper or plastic plate that is attached to something that makes the plate vibrate. For example, one can make a primitive speaker by placing a magnet in a solenoid, attaching a paper plate to the magnet, and then making the magnet vibrate back and forth by oscillating the current through the solenoid.[ii] By varying the frequency of the back and forth motion of the magnet, one can use this setup to create sounds with varying pitches.

✍ Unless the frequency is really low, the back and forth motion is so quick that it is very difficult to tell that the magnet/plate assembly is actually moving back and forth.

⁉ The period of a sound is the time it takes to complete one cycle of the oscillation. The frequency is the inverse of the period.

As before, the **frequency**, f, is the rate at which cycles are completed, where one cycle is one back and forth motion of the object. It is the inverse of the **period** (T, the time it takes to complete one cycle) and so is typically given in **hertz** (cycles per second). Mathematically, the relationship between the two is as follows:

$$f = \frac{1}{T} \qquad \text{and} \qquad T = \frac{1}{f}$$

The smaller the frequency, the longer it takes for the object to complete one cycle of its back and forth motion.

[i] Recall from chapter 16 that frequency has to do with how quickly something is oscillating back and forth. The frequency of your heart beat, for example, might be 60 beats per minute when you are resting but is higher when you are exercising.

[ii] For the solenoid to exert a force on the magnet, part of the magnet has to be outside the solenoid or near the end, where the strength of the solenoid's magnetic field is less. In that way, one end is pulled into the solenoid more than the other end is pushed away.

For example, consider an object that is vibrating such that its position with time is indicated by the graph in Figure 19.1(a). The object is seen to vibrate, moving from 0.1 mm one way to 0.1 mm the other way, taking 1.0 ms (milliseconds) to complete one cycle. That means its period is 1 ms. Since there are one thousand milliseconds in one second (1 ms = 0.001 s), and the object completes one cycle every millisecond (T = 0.001 s), it completes a thousand cycles in one second. For that reason, its frequency is 1000 Hz (equal to one cycle divided by 0.001 s).

✓ *Check Point 19.1: Suppose an object makes a sound with frequency equal to 300 Hz. According to our model, how long does it take the object to complete one cycle of its vibration?*

19.1.2 Frequency vs. amplitude

Does changing the frequency of the sound also affect how loud the sound is?

As mentioned earlier, the frequency of vibration governs the pitch, not the loudness. After all, we can change the pitch of a note without changing how loud it is.[iii]

Whereas the pitch is governed by how quickly the object moves back and forth (the frequency), the loudness is governed by "how much" the object moves back and forth (the amplitude).

To understand what amplitude is, compare the (a) and (b) graphs in Figure 19.1. Both represent a back and forth motion that has a period of 1 ms, so they represent sounds of equal frequencies (1000 Hz) and thus the same pitch. However, object (b) moves back and forth only 0.05 mm each way, half of what object (a) does. As a result, the sound produced by object (b) is softer, not as loud as the sound produced by object (a). It would be like comparing two flutes and a single flute, if all three flutes were playing the same note the same way. The two flutes together would be louder than a

[iii] As we'll see in chapter 23, just as we can change the pitch of a note without changing how loud it is, we can likewise change the color of light without changing how bright it is (since frequency impacts the color and amplitude impacts the brightness).

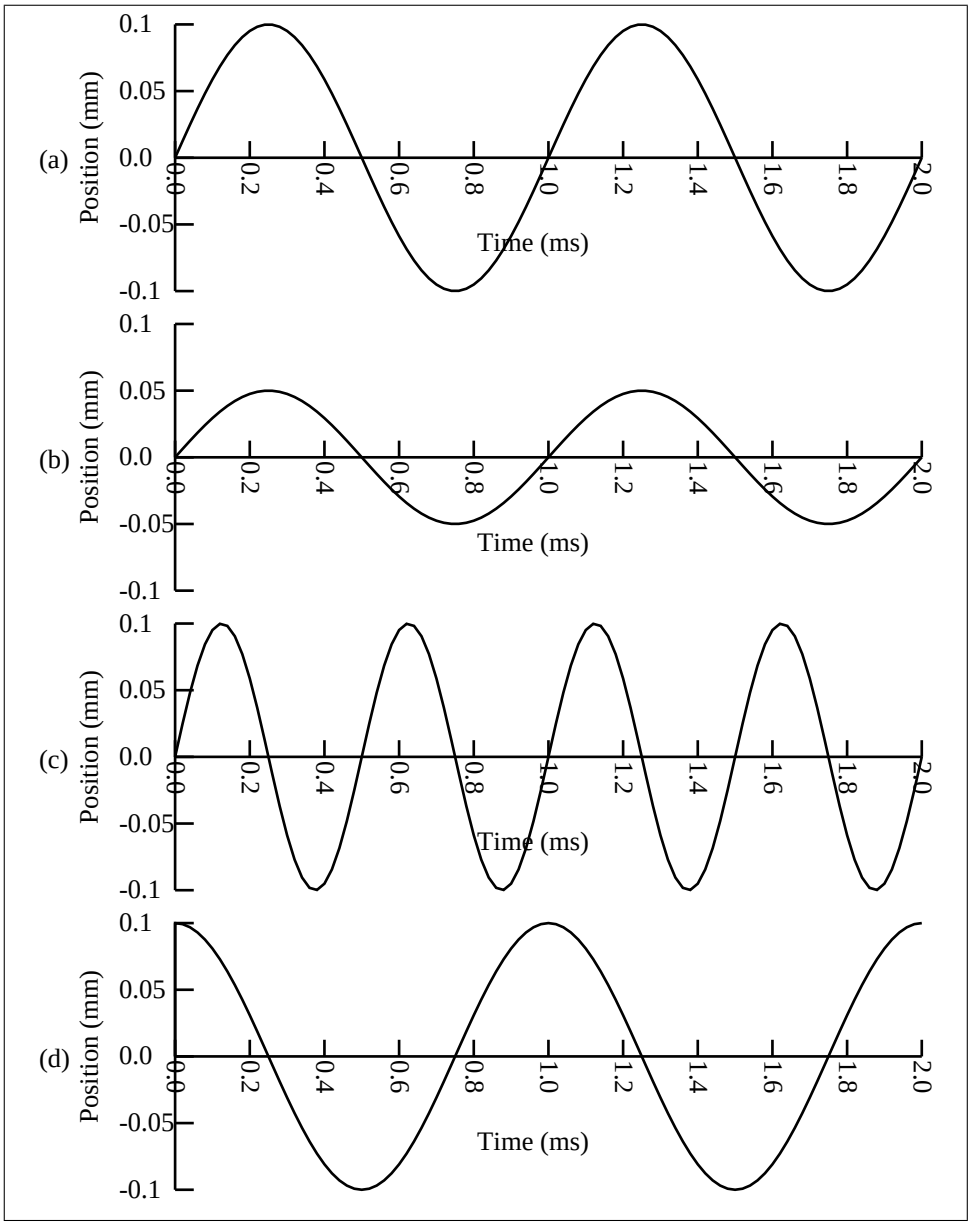


Figure 19.1: Position vs. time graphs for four vibrating objects. Graph (b) has the same frequency as that in (a) but with half the amplitude. Graph (c) has the same amplitude as that in (a) but with twice the frequency. Graph (d) has the same amplitude and frequency as that in (a) but at a different time.

single flute but the pitch we'd hear with two flutes together would be the same as the pitch we'd hear with just the single flute.

The maximum extent of the movement each way is called the **amplitude**.[iv] So, object (a) in Figure 19.1 is vibrating with an amplitude of 0.1 mm, whereas object (b) is vibrating with an amplitude of 0.05 mm.

In comparison, consider graph (c) in Figure 19.1. The amplitude of the motion is 0.1 mm, the same as object (a)'s motion, so it represents a sound that is just as loud as the one produced by object (a). However, object (c)'s vibration has a period that is half that of object (a) and thus has twice the frequency (2000 Hz vs. 1000 Hz). As a result, the sound produced by object (c) is just as loud but has a higher pitch than the sound produced by object (a).

Keep in mind that the objects represented by these graphs are vibrating very quickly, 1000 or 2000 times every second. What we hear is a steady tone that is unvarying. Indeed, as long as the amplitude remains the same, the sound doesn't get louder or softer. It stays the same loudness. And, as long as the frequency remains the same, the sound doesn't rise or fall in pitch. It stays the same pitch.

Also note that we can't tell, by what we hear, *where* the object happens to be at any given movement. What we hear is not a result of its *position* but rather how it is *vibrating*. More is said on this in the next section but to illustrate what I mean, consider the vibration illustrated in graph (d). Since the back and forth motion has the same amplitude and frequency as that of object (a), the sound is exactly the same in terms of loudness and pitch.

✓ *Check Point 19.2: For the sounds represented by the four graphs in Figure 19.1, which one(s) represent a sound with the highest pitch, and why? Is that sound also the loudest? Why or why not?*

19.1.3 Amplitude and intensity

The **intensity** of a sound basically represents how loud a sound is. The larger the amplitude, the greater the sound intensity and the louder the

[iv] You can think of the volume knob on a radio as controlling the amplitude of the sound.

sound.[v] The intensity is equal to the rate at which energy (energy per time; SI unit of W) is transferred per area (SI unit of m^2) and so it has SI units of W/m^2.

Is a W/m^2 the same as a decibel?

Both are used to describe the intensity of sound but they are not the same thing. Since most people use **decibels** instead of W/m^2 to describe the intensity of sound, it makes sense to explain the difference.

The reason why people prefer decibels is because the human ear is sensitive to a wide range of intensities (i.e., over 12 orders of magnitude). Decibels allow us to convey the wide range of intensities without using large numbers.

It does this by conveying how many **orders of magnitude** a sound is relative to the threshold of human hearing (10^{-12} W/m^2; softest one can hear). For example, a sound with intensity 10^{-5} W/m^2 is said to be seven orders of magnitude greater than the threshold of human hearing (i.e., compare the exponents: -5 is seven more than -12).

The threshold of pain (1 W/m^2; the loudest one can hear without pain) is then 12 orders of magnitude greater than the threshold of human hearing (since 1 is the same as 10^0, and zero is 12 more than -12).

✍ For consistency, we use the same threshold intensity for all notes, even though technically the softest one can hear depends on the person and the frequency of the note.

This way of indicating the intensity is equivalent to expressing the intensity in units of **bels** (named after Alexander Graham Bell), which is abbreviated as B. Consequently, the threshold of hearing would have an intensity of 0 B and the threshold of pain would have an intensity of 12 B.

What is the difference between a bel and a decibel?

⁇ A decibel is a unit used to describe the intensity of sound.

The prefix "deci" means "tenth" (see the supplemental readings for metric prefixes). This means that 12 bels is equivalent to 120 decibels. The abbreviation for decibel is "dB", which means that the threshold of pain would have an intensity of 120 dB.

[v] Technically, intensity is proportional to both the amplitude and the frequency (actually, the square of each), which is why, for the same intensity, a speaker cone moves with a smaller amplitude at higher frequencies. However, it is the amplitude dependence, and its influence on loudness that is of importance to us here.

WHY USE DECIBEL INSTEAD OF BEL?

One decibel (or 1/10 of a bel) is considered to be the smallest change in intensity that the human ear can distinguish.

IF THE INTENSITY IS 0 dB DOES THAT MEAN THERE IS NO SOUND?

No. It only means that the intensity of the sound is at the threshold of hearing. The sound is still there but the human ear cannot hear it.

CAN THE INTENSITY BE LESS THAN 0 dB?

Yes. A -10 dB sound, for example, has an intensity that is one order of magnitude less than the threshold of hearing. Mathematically, that would mean the intensity was 10^{-13} W/m^2. We just can't hear it.

✓ *Check Point 19.3: A typical office or classroom has a sound intensity that is 5 orders of magnitude (i.e., 100,000 times) greater than the threshold of hearing. Indicate the intensity of the sound in W/m^2, bels, and decibels.*

19.1.4 Audible range

⁉ Our ears are more sensitive to some frequencies of sound than others.

As mentioned earlier, a low frequency sound is not necessarily louder or softer than a high frequency sound. However, it turns out that our ear is more sensitive to certain frequencies and so certain frequencies may appear louder to us than others, even for the same intensity. A human ear is most sensitive, it turns out, to frequencies between 1000 and 5000 Hz. Consequently, sounds with those frequencies will appear louder than frequencies outside the range with the same intensity.

In addition, the human ear can't detect sounds with frequencies less than 20 Hz or greater than 20,000 Hz. Such sounds can be very intense and the human ear would still be unable to detect them.

The frequency range from 20 Hz to 20,000 Hz is called the **audible range**. It varies with age, with the high end being lower for older people.[vi]

IS THE TYPICAL HUMAN EAR EQUALLY SENSITIVE TO ALL FREQUENCIES BETWEEN 20 HZ AND 20,000 HZ?

[vi] This has to do with the cilia hair cells in the ear stiffening up with age and not being able to vibrate at the higher frequencies.

If the ear was perfect, sounds of identical intensities would be equally loud for all frequencies. However, as mentioned above, the ear is not perfect. The typical human ear is more sensitive[vii] to frequencies around 1,000 to 5,000 Hz than to other frequencies. As we increase the frequency from 0 Hz up to 1,000 Hz, while keeping the intensity the same, the sound appears to get progressively louder as the ear becomes more sensitive. As we increase the frequency above 5,000 Hz, the sound appears to get softer (assuming the intensity remains the same) because the ear is less sensitive to higher frequencies (though most people find such high frequencies to be annoying).

In general, we won't be considering changes in frequency that are so great that we need to consider the frequency-dependence of our hearing ability. Thus, it will be sufficient to assume that the sound gets neither louder nor softer as we change the frequency. However, it is still important for you to recognize that the ear does have this dependence.

✓ *Check Point 19.4: As mentioned above, the ear is more sensitive to 2000 Hz than to 20 Hz. If a 2000 Hz tone appears louder than a 20 Hz tone, does that necessarily mean the 2000-Hz tone is more intense? Why or why not?*

19.2 Sound propagation

Sound is an example of a **wave**. A wave is a traveling series of pulses, where a **pulse** is a single back-and-forth motion of something. To create a traveling series of pulses, one must periodically create pulses that then travel away from where the pulses are generated.

How does the sound get to our ear from the vibrating object?

When an object vibrates, it alternately compresses and decompresses the air next to it. Compressed air expands, compressing the surrounding air. That air, in turn, expands, leading to another region being compressed. In this way, the compression moves from the vibrating object to your ear.

[vii] By "more sensitive", I mean that the ear requires a higher intensity to hear lower-frequency (less than 1,000 Hz) and higher-frequency (more than 5,000 Hz) sound. For example, the threshold of hearing is 10^{-11} W/ ^{2} for sound at 500 Hz or 10,000 Hz (compared to 10^{-12} W/m^2 for frequencies between 1,000 and 5,000 Hz).

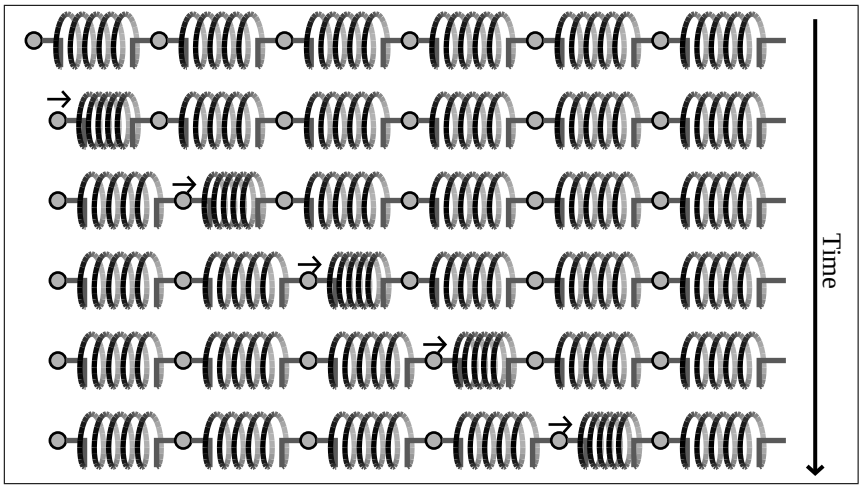


Figure 19.2: Six snapshots of a series of balls connected by springs. The sequence starts at the top, where the balls are at rest. In the second snapshot, the left-most ball is moved slightly toward the right. Due to the connections, the "compression" moves to the right. To see this sequence in motion, along with the other wave motions discussed in this chapter, go to http://www.acs.psu.edu/drussell/demos/waves/wavemotion.html.

To illustrate what I mean, we can treat air molecules as a bunch of balls connected by springs, as shown in Figure 19.2 (top snapshot). Because the balls are connected by the springs, moving one ball forward (see second snapshot in Figure 19.2) compresses the spring that is next to it. That spring then does two things as it tries to expand back to how it was: it pushes the first ball back to where it was initially, and it pushes the second ball forward. The motion of the second ball, in turn, compresses the spring to *its* right.

This is repeated along down the line. The left spring compresses and then expands, but in expanding it compresses the spring next to it. That next spring expands but in so doing it compresses the spring next to it. In this way, the "compression" moves along the line.

WHY DON'T ALL THE BALLS MOVE TO THE RIGHT AT THE SAME TIME?

The compression doesn't happen instantaneously. It takes some time for the "information" about the compression to be conveyed to the next spring in line. If the springs are very stiff, the information can be conveyed very quickly and there is hardly any delay between when one ball moves and the next ball moves. The softer the springs, the slower the information is conveyed and

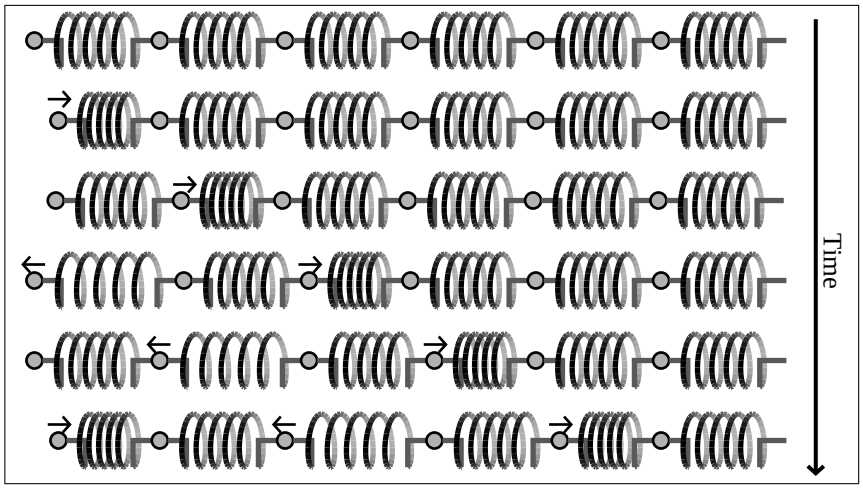


Figure 19.3: Six snapshots of balls connected by springs, as in Figure 19.2, except that the left-most ball is moved back and forth (with slight pauses between each movement). As a result, a series of "compressions" and "rarefactions" move to the right.

the larger the time delay between when the first ball is moved and when the last ball moves.

Now let's apply this to the situation where we have a vibrating object. Let's suppose the vibrating object is on the left side of the group of balls connected by springs. This is illustrated in Figure 19.3, where the first ball is moved slightly toward the right and then, after a short pause, it moves back toward the left. Then, after another short pause, it starts over again. This back and forth motion is an **oscillation**.

Because of the connections, the other balls also undergo an oscillation. The springs between the balls alternate between **compression** (balls closer together) and **rarefaction** (balls farther apart).

How is this related to sound?

With sound, the particles are air molecules, not balls. And the mechanism for pushing the molecules back toward equilibrium is air pressure, not springs.[viii]

[viii] Keep in mind that the oscillation occurs rather quickly. For a 2000-Hz sound, it oscillates 2000 times each second. Also the variation in pressure between the compression and rarefaction areas is small, about twenty-billionths of the regular atmospheric pressure. Given that, it is amazing how sensitive the ear is. Even for a very intense sound,

However, other than that, the process is the same.

⁝ Sound is a series of compressions and rarefactions in a material.

How do the compressions and rarefactions make a sound?

Eventually the series of compressions and rarefactions reaches your ear. Each time a compression hits your ear drum, the ear drum is pushed in. Each time a rarefaction hits your ear drum, the ear drum is pulled out. This results in a back and forth motion of your ear drum that is interpreted as sound.[ix]

What is really interesting, and important to recognize, is that the brain interprets the back and forth motion as a steady tone, rather than as a sound that goes on and off as the ear drum goes back and forth.

⁝ The material doesn't get carried along with the wave.

Where does the air go when it hits your ear?

It doesn't go anywhere. It is important to distinguish between the propagation of the sound and the back and forth motion of the particles. For example, with the balls and springs (see Figure 19.3), the oscillation travels toward the right. Each ball, however, just moves back and forth slightly as the wave passes by. The "material" as a whole just stays where it is.

The same thing holds for sound except that we have air molecules instead of balls. As the sound travels through the air, the air molecules move back and forth slightly. The air as a whole just stays where it is.

✓ *Check Point 19.5: Which of the following correctly describes what happens as a sound travels through the air?*
(A) The air molecules are carried along with the sound.
(B) The air molecules move back and forth a small distance.
(C) The air molecules are unaffected, even as the sound passes by.

19.3 Other types of waves

As mentioned in the introduction, sound is not the only phenomenon to involve waves. There are light waves, water waves, and others. To have a

like a sound at the threshold of pain, the pressure difference is only about 0.03% of the atmospheric pressure.

[ix] The back and forth motion of your ear drum produces a similar motion in the cochlea, a fluid-filled cavity in the inner ear, and hair cells in the cochlea sense this motion and submit nerve impulses to the brain, which interprets the nerve impulses as sound.

wave, we need two things to happen.

The first is we need to periodically create pulses. Basically, that means that we need to make something oscillate back and forth. An *oscillation*[x] refers to the periodic back and forth motion of something, whether it is the up and down motion of a washing machine during the agitate cycle or the back and forth motion of a pendulum.

The second requirement is that the object undergoing the back and forth motion must be "connected" to other objects such that each object[xi] acts to force its neighbors back toward the "equilibrium" position (the position it had prior to oscillating). For the balls in Figure 19.3 the restoring force is provided by the springs.

⁉ A wave is an oscillation that moves through space because of the connections between the oscillators.

This is why we refer to sound as a wave. It is essentially a series of compressions and rarefactions of air that propagate through the air due to the air pressure pushing the air molecules back toward equilibrium.

✓ *Check Point 19.6: Are any of the following waves? If so, which ones and why? If not, why not?*
(A) Ten chairs placed in a row
(B) The oscillation of a pendulum
(C) The movement of the second hand on a clock

A sound wave is an example of a wave where the direction the pulses travel (rightward in Figure 19.3) is parallel to the back and forth motion of the objects (air molecules) that are oscillating (right and left in Figure 19.3).[xii] Not all waves involve oscillations parallel to the pulse motion. Figure 19.4 illustrates a wave (on a string) where the pulse motion (rightward in the figure) is *perpendicular* to the motion of the oscillating object (the string, which is oscillating up and down in the figure).[xiii] And some waves have characteristics of both, like water waves, where the water itself moves in a

[x] You can think of an oscillation like **vibration** or shaking, although an oscillation tends to be more periodic.

[xi] The technical term for such connected objects is **coupled oscillators**.

[xii] This type of wave is called a **longitudinal** wave. The word "longitudinal" comes from a word that is related to "lengthwise".

[xiii] This type of wave is called a **transverse** wave, from a Latin word meaning "to turn" (similar to "cross-wise").

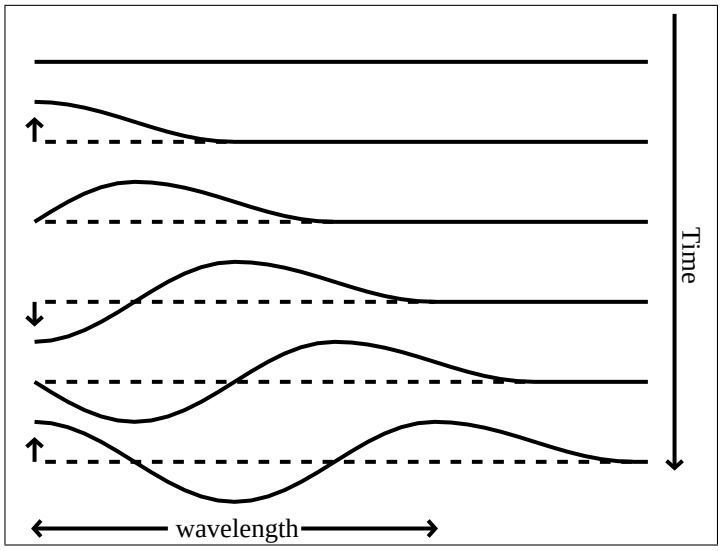


Figure 19.4: Six snapshots of a string. Initially (see snapshot on top) the string is flat. Then the left end of the string is oscillated upward and downward. This causes the string next to the end to move upward and downward. The series of up and down oscillations moves rightward.

circular manner, as illustrated in Figure 19.5.[xiv] In particular, as a crest of water passes by, a cork floating on the water first travels toward the crest and then up and with the crest as the crest overtakes it. As the crest leaves, the cork travels down and back to where it started. Notice how, just like with other types of waves, the water molecules don't get transported along with the wave, but rather just oscillate about some equilibrium position (which would be sea level for waves on the sea).

As mentioned before, to get a wave we need oscillators (in Figure 19.3 that would be the balls) that experience a restoring force pushing them back toward some equilibrium position.

With the string, the equilibrium position is represented by the flat line in the top snapshot in Figure 19.4 and the dashed line in the other snapshots. If the left end of the string is above the equilibrium position, the portion of the string right next to it pulls the left end back toward equilibrium. In doing so,

[xiv] Keep in mind that, in the language we are using here, a wave is a traveling series of oscillations. Consequently, a water wave would be the entire series of oscillations in the water height, not a single wave that a surfer might ride.

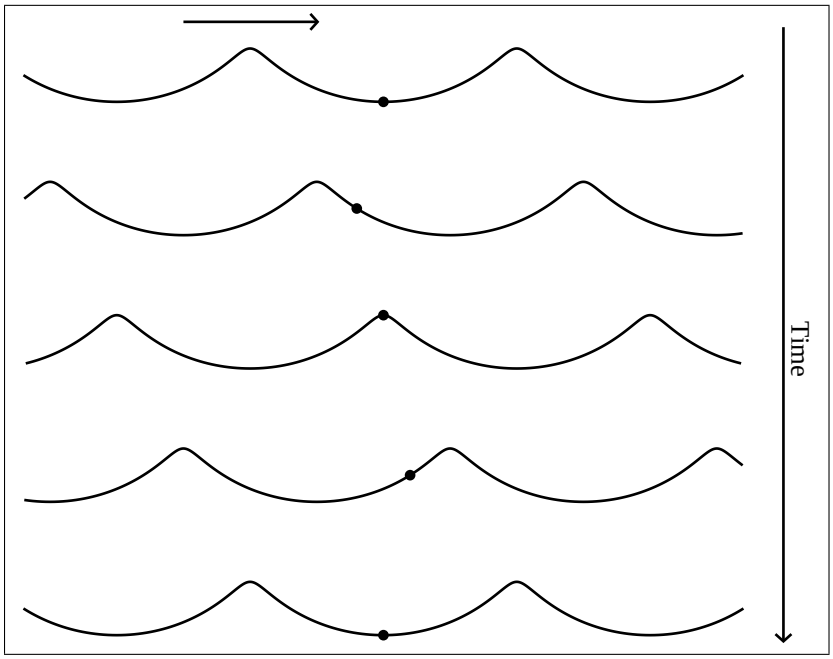


Figure 19.5: Five snapshots of a water wave. Time progresses downward, with the wave moving to the right. The dot indicates the motion of an object floating on the water. To see this sequence in motion, go to http://www.acs.psu.edu/drussell/demos/waves/wavemotion.html.

the string right next to the left end is pulled above the equilibrium position. Via this mechanism, the pulse moves toward the right while the string itself moves up and down.

With the water wave, gravity exerts a vertical force on the water (moving it up and down) while varying water pressure along the wave (due to different heights of water) leads to a horizontal force on the water (moving it left and right). The end result is that the water moves in a circular path as the wave moves across the water surface.

✓ *Check Point 19.7: For each of the following, identify the material is oscillating parallel to the wave motion, perpendicular to the wave motion or both (i.e., the particles go in a circular motion): (a) sound waves, (b) water waves; (c) a wave on a string.*

19.4 Wave speed

The generator of the wave is the vibrating object that "starts" everything off, alternately compressing and decompressing the air next to it, and the compression pulses move away from the generator. The speed at which the pulses move away from the generator is called the **wave speed**.

We have already noted that the pulses travels away from the generating object at a different speed than the back and forth motion of the individual particles. The **wave speed** is the speed that the pulses travels away from the generating object, not the speed of the individual particles as they move back and forth. Be careful not to confuse these two speeds.

⁉ The wave speed is the speed at which the oscillations travel through the material.

What does the wave speed depend on?

First let me state what it *doesn't* depend on. It doesn't depend on the amplitude of the wave, and it doesn't depend on the frequency of the wave. In other words, the speed of sound (how quickly the sound travels away from the object generating the sound) is the same for a loud sound and a soft sound, and it is the same for a high pitch sound and a low pitch sound.

Indeed, the wave speed doesn't depend upon the wave itself at all, but rather the *material* through which the wave travels. In particular, the wave speed depends on the magnitude of the restoring force between the particles. For example, in Figure 19.3 (page 348), stronger springs cause the wave to travel more quickly (and, similarly, the wave speed is faster for tighter strings).

The wave speed also depends on the mass (or density) of the materials. For example, heavier balls cause the wave to travel more slowly down the line in Figure 19.3 (because there is more inertia to overcome).

With air, the same pattern holds but it is less obvious. For air, the speed of sound is dependent on the temperature[xv] (see, for example, Table 19.1[xvi]. Consequently, when doing problems with sound in air, one usually needs

[xv] If all other things are kept the same, increasing the air pressure leads to an increased speed of sound. However, in practice, increasing the pressure also leads to increasing the density and the two effects cancel out, leading to the same wave speed. Since the temperature is proportional to the ratio of the two (ideal gas law), it turns out that the speed of sound is more closely related to the temperature.

[xvi] Notice that there is no value listed for a vacuum. Sound can't be transmitted through a vacuum since there are no molecules that can act as coupled oscillators.

Table 19.1: The speed of sound in different media. Source: Handbook of Chemistry and Physics (http://www.hbcpnetbase.com/).

Substance	Speed (m/s)
Air (dry; 0°C)	331.5
Air (dry; 20°C)	343.4
Air (100%; 20°C)	344.7
Helium (0°C)	965
Hydrogen (27°C)	1310
Water (20°C)	1483.3
Glass (Pyrex; 20°C)	5640
Steel (1% C; 20°C)	5940

to specify the temperature of the air[xvii] For our purposes, we'll assume the temperature is 20°C. The speed of sound in 20°C air is 343 m/s.

✓ *Check Point 19.8: Suppose we assume the speed of sound is 343 m/s. What relative error would that introduce when the temperature is 0° C. In other words, what is the percent difference between 343 m/s and the actual speed of sound at 0° C?*

19.5 Wavelength

Figures 19.4 and 19.5 illustrate the state of a wave (string and water, respectively) at different times. If we take a single snapshot of each, we'll get a picture like what I show in Figure 19.6 (string wave on top, water wave on bottom).

A single snapshot doesn't convey how each part of the material is oscillating or even which way the pulses are traveling (although we can add an arrow, as in Figure 19.6 to indicate the direction) but it does show the spacing of the crests, which I've highlighted in Figure 19.6 via the vertical dashed lines. In fact, to simplify the illustration, it is common to show the position of the crests (or compressions, for sound) at a particular moment in time, as in Figure 19.7.

[xvii] The dependence on temperature leads to a phenomenon called refraction (see chapter 24), which is responsible for how one can hear sounds farther away at night time.

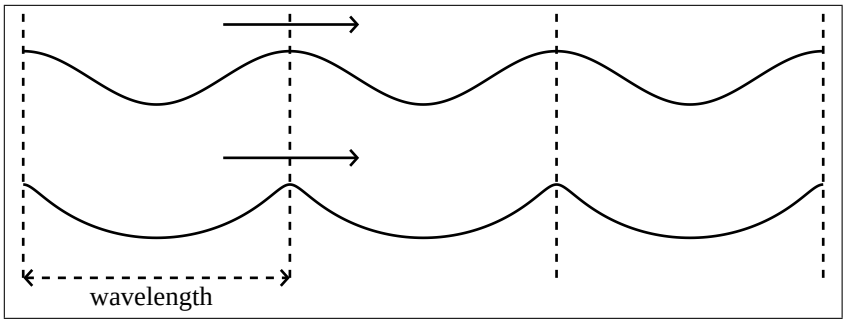


Figure 19.6: Two snapshots of a wave (top: wave on a string; bottom: water wave) with positions of crests indicated by via the vertical dashed lines.

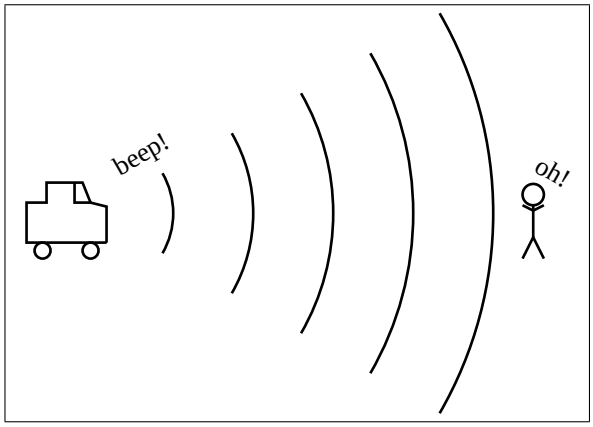


Figure 19.7: An illustration of the sound associated with a car horn.

The vertical dashed lines in Figure 19.6 and the arcs in Figure 19.7 are called **wavefronts**. The wavefronts are separated by a distance known as the **wavelength**, as shown in Figure 19.6. The closer together the crests, the smaller the wavefront spacing and the smaller the wavelength.

⁝ The wavelength is the distance between oscillators at the same stage of their oscillation.

Notice that the wavefronts don't need to correspond to the positions of the crests. They could just as easily correspond to the position of the valleys. The spacing would be the same. The only requirement is that they indicate the location where the oscillators are at the same stage of the oscillation.[xviii]

[xviii] With sound, we could will graph the air pressure (or air density) at a particular time, in which case the graph would look like the snapshot of the wave on a string, with the peaks and valleys corresponding to high and low air pressure (due to compression and rarefaction).

How will the wavelength be indicated in equations?

A lower-case Greek letter lambda (λ) is used indicate the wavelength.

Why λ?

This is the convention. Many times, physicists will use Greek letters because the corresponding Roman letter has already been used for something similar. In this case, a Roman lower-case "L" is also difficult to distinguish (i.e., it might be interpreted as a one).

✓ *Check Point 19.9: For the situation illustrated in Figure 19.7, suppose the total distance is 11 m from the front of the car to the pedestrian. What is the wavelength of the wave being represented?*

Does the intensity of a sound wave depend on how close you are to the source?

Yes. This is because the sound spreads out over three dimensions (i.e., a volume rather than a line like the string). In fact, one limitation of the illustrations in Figures 19.6 and 19.7 is that they do not adequately convey how the intensity decreases the farther one is from the source. As the wave moves away from the center, the energy spreads out, which lowers the amplitude of the wave. This is why a sound wave becomes softer as the wave moves away from the source.

To illustrate how the intensity decreases the farther one is from the source, we could using shading, as in Figure 19.8a, where the darker and lighter areas represent where the air is compressed vs. expanded. Notice how the contrast between dark and light is greater near the center than farther away. Drawing the wavefronts, as in Figure 19.8b, is simpler but doesn't provide a sense of how the amplitude decreases with distance.

Both methods fail to indicate *direction* the pulses are moving, though. We can address this by adding arrows, as in Figure 19.8c. These arrows are called **rays**.[xix] The rays indicate the direction the pulses travel but fail to illustrate the wavelenth or how the wavelength (distance between adjacent wavefronts) remains the same, regardless of how far one is from the center (where the source is located).

[xix] In a similar way, arrows can be used to indicate the migration path of animals.

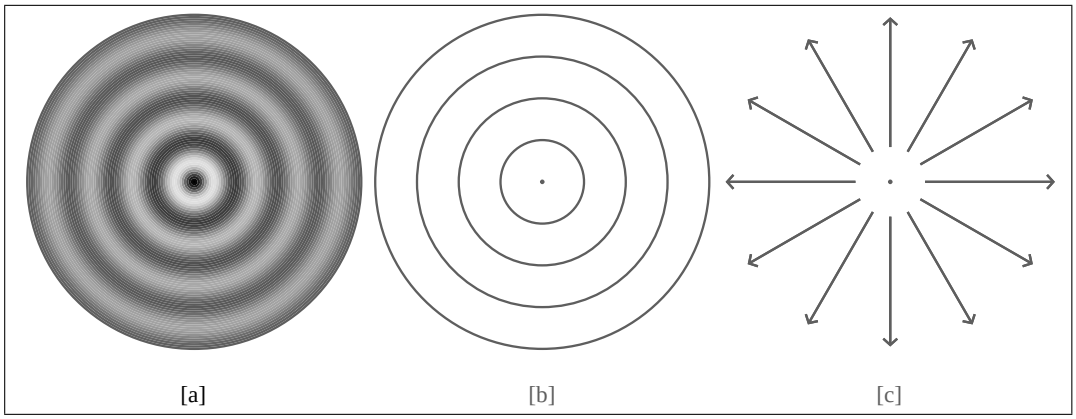


Figure 19.8: Three pictures of a sound wave moving away from a central location. [a] air density is shaded; dark and light shades correspond to compressed and expanded areas, respectively. [b] Compressed areas are indicated by wavefronts. [c] Direction of motion is indicated by rays.

✍ The wave always moves in a direction perpendicular to the wavefront. Consequently, if you superimpose the rays upon the wave fronts, the rays will always be oriented perpendicular to the wave fronts.

Each method has its advantages and disadvantages. Which method we use depends on the aspect we want to illustrate.

✓ *Check Point 19.10: Suppose the outermost circle in Figure 19.8[b] has a radius of 3.0 meters. What is the wavelength of the wave being represented by the wavefronts?*

19.6 The wave equation

The three wave characteristics discussed so far (wave speed v, wavelength λ and frequency f) are related, and rather simply it turns out:

$$v = f\lambda \tag{19.1}$$

This is known as the **wave equation**.

⁝ The wave equation relates the three aspects of the wave: the wave speed, wavelength and frequency.

Where does this equation come from?

By definition, an object's speed is equal to the distance the object travels divided by the time it takes to travel that distance. In this case, our "object" is a pulse within the wave. If we wait a time equal to the period, T, then the pulse travel a distance equal to the wavelength, λ. Since speed is defined as the distance divided by the time, we then have that our wave speed v must equal to the wavelength λ divided by the period T:

$$v = \frac{\lambda}{T}$$

Since the frequency is the inverse of the period ($f = 1/T$) we can replace $1/T$ with f to get the wave equation.

How do you know that after one period the information has traveled a distance equal to the wavelength?

Let's say that the pulse travels a distance equal to Δx in an amount of time equal to Δt. If, at that time, we introduce another compression, the two compressions will be separated by a distance equal to the wavelength λ. Since the time between compressions is what we call the period T, that must be the time that corresponds to $\Delta x = \lambda$.

Example 19.1: A sound wave travels at a speed of 343 m/s in air. If the wavelength is 3 m, what is the frequency of the sound wave?

Answer 19.1: From equation 19.1, $v = f\lambda$. Plugging in 343 m/s for v and 3 m for λ, we find that the frequency is 114 cycles per second (or 114 Hz).

✍ Usually the wave speed v is determined by the material (like the air density and temperature) and the frequency f is determined by the object that is vibrating (like your vocal chords). The wavelength λ, then, is dependent on the wave speed and frequency, and can be determined using the wave equation.[xx]

✓ *Check Point 19.11: A wave with a frequency of 3.0 Hz takes 2 s to travel the length of a 3-m Slinky*[xxi]*. Determine the wavelength of the wave.*

[xx] In chapter 22 we'll explore situations where we control the wavelength, with the frequency then determined by the wave equation.

[xxi] A Slinky is just a long spring, much like that shown in Figure 19.3 but without the balls.

Summary

This chapter examined the characteristics of waves and how they are related.

The main points of this chapter are as follows:

- The period of a sound is the time it takes to complete one cycle of the oscillation. The frequency is the inverse of the period.
- A decibel is a unit used to describe the intensity of sound.
- Our ears are more sensitive to some frequencies of sound than others.
- Sound is a series of compressions and rarefactions in a material.
- The material doesn't get carried along with the wave.
- A wave is an oscillation that moves through space because of the connections between the oscillators.
- The wave speed is the speed at which the oscillations travel through the material.
- The wavelength is the distance between oscillators at the same stage of their oscillation.
- The wave equation relates the three aspects of the wave: the wave speed, wavelength and frequency.

By now you should be able to describe sound (with appropriate units) in terms of the frequency, period, wavelength, wave speed, amplitude and intensity (in both W/m^2 and decibels) of air molecule vibrations (compressions and rarefactions), and the audible frequency range, and relate the wavelength, wavelength, and frequency (or period), via the wave equation. You should also be able to utilize the difference between the motion of the wave and the motion of the material within which the wave travels and distinguish between transverse waves, longitudinal waves, and those waves that are both transverse and longitudinal.

Frequently asked questions

Can sound exist if there is no air?

If there is really nothing (i.e., a vacuum), there is nothing to vibrate. Consequently, sound cannot exist if there is no air.

How, then, can you hear explosions in space?

You can't.[xxii]

IS A HIGH-FREQUENCY SOUND NECESSARILY LOUDER?

No. Frequency just influences the *pitch* of the sound (i.e., higher note or lower note). However, we may be more sensitive to certain frequencies (see section 19.1.3). For example, a note at 50 Hz may be barely audible just because our ears are not as sensitive to frequencies that low. A similar thing happens for very high frequencies, like those at 20,000 Hz or above.

HOW DO WE KNOW THAT THE v IN EQUATION 19.1 STANDS FOR THE WAVE SPEED AND NOT THE SPEED OF THE PARTICLES AS THEY MOVE BACK AND FORTH?

Since I will only be writing equations for the wave speed, not the speed of the particles, you can be safe in assuming that v corresponds to the wave speed.

WHY IS EQUATION 19.1 WRITTEN IN TERMS OF $f\lambda$ INSTEAD OF λ/T?

Since the period is just the inverse of the frequency (i.e., $T = 1/f$), you can write the wave equation as either $v = \lambda/T$ or $v = f\lambda$, but the convention is to use frequency when the period is smaller than one second per cycle.

Terminology introduced

Amplitude	Longitudinal	Rays
Audible range	Loudness	Transverse
Bels	Orders of magnitude	Vibration
Compression	Oscillation	Wave
Coupled oscillators	Period	Wave equation
Decibels	Pitch	Wave speed
Frequency	Pulse	Wavefronts
Hertz	Rarefaction	Wavelength
Intensity		

[xxii] Unless, of course, you are a movie director and you think it is important for audiences to hear the explosions as they would hear them on Earth.

Abbreviations introduced

Quantity	**SI unit**
intensity (I)	watt per square meter (W/m^2)
wave speed (v)	meter per second (m/s)
wavelength (λ)	meter (m)

Quantity	**non-SI unit**
intensity (I)	decibel (dB)

20. Doppler Effect

Puzzle #20: You are standing alongside a road when a car drives past with its horn sounding. Why does the pitch you *hear* (of the horn) appear to change *as the car drives past you* but remain the same *while the car drives toward you* (even as the sound gets louder and louder)?

Introduction

To strengthen your understanding of waves, particularly the difference between intensity, frequency and wavelength, in this chapter we examine the particularly strange phenomenon described in the puzzle. While we'll focus on sound, the phenomenon occurs with all types of waves. We'll just focus on sound waves because the phenomenon is easier to observe with sound waves.

20.1 Describing the Doppler effect

When a car drives past you with the horn sounding, the sound you hear changes in two ways:

- The sound you hear gets *louder and louder* as the car gets closer, being loudest when it is closest. Then, after the car passes and drives away, the sound gets *softer and softer*.
- The sound you hear has *one pitch* as the car approaches and a *lower pitch* when the car moves away.

I want to emphasize that the horn is producing a steady, unchanging sound, not getting louder or softer, or changing pitch. In fact, to someone in the

car, that is exactly how the horn sounds – with a steady, unchanging sound, not getting louder or softer, or changing pitch.

I also need to be clear that the pitch you hear does not change *as the car approaches* or *as the car moves away* – only the intensity (loudness) changes (getting louder and louder as the car approaches, and getting softer and softer as the car moves away). The pitch changes only at the moment the car passes by[i], an effect that is called the **Doppler effect**. In this chapter, we'll examine why the pitch you hear does this (despite the fact that the sound being produced does not change in pitch).

✍ We'll also examine why the intensity changes as the car approaches and leaves, but the Doppler effect refers to how the *pitch* changes, not the intensity.

✓ *Check Point 20.1: While the car approaches with the horn sounding, what happens to the pitch you hear: does it continually get higher and higher, or does it remain the same the entire time the car approaches?*

20.2 Observer vs. source

Before we explain why the pitch does what it does, we first need to distinguish between the *observer* and the *sound source*.

The **observer** is the one *receiving* the wave while the **source** is the one *producing* the wave. In our puzzle, the wave is sound, the source is the car's horn and the observer is you.

The Doppler effect occurs when the source and observer, being two different objects, are moving toward or away from each other. In the puzzle, the source is moving and the observer is stationary. In this chapter, we will not only examine this type of situation but we will also examine the situation where the source is stationary and the observer is moving, and when both the source and the observer are moving.

[i] How quickly the pitch changes depends on how close to the observer the car gets as it passes. Assuming it passes relatively close, the change is somewhat rapid because the "passing by" period is rather short, but it isn't instantaneous (since the car doesn't go "through" the observer).

✓ *Check Point 20.2: In the puzzle (see page 363),*
(a) What object is the source? Is it moving or is it stationary?
(b) What object is the observer? Is it moving or is it stationary?

20.3 Moving observers

To begin our exploration, we examine the case where the observer is moving and the source is stationary. This isn't the case in the puzzle but it turns out it is a little simpler to understand.

Once we understand the case of moving observers, we'll examine the case of moving sources. After that, we will tackle the more difficult task where both the observer and the source are moving. In each case, we'll first *describe* what happens and then *explain* why it happens.

Suppose, then, that we observe a stationary car with its horn sounding. Suppose further that if we are likewise stationary, we hear a sound with frequency 400 Hz ($f_{\text{emitted}} = 400$ Hz).

Even though the frequency of the horn does not change and continues to produce a 400-Hz sound, it turns out that we hear a higher frequency ($f_{\text{obs}} > 400$ Hz) when we are moving toward the car, and we'd continue to hear that higher frequency the entire time we are moving toward the car (as long as we maintain a constant speed). Conversely, we hear a lower frequency ($f_{\text{obs}} < 400$ Hz) when we are moving away from the car, and we'd continue to hear that lower frequency the entire time we are moving away from the car (as long as we maintain a constant speed).

For example, it might appear to be 440 Hz to us when we are moving toward the car and 370 Hz when we are moving away from the car, even though the car is still producing the sound at 400 Hz, and someone at rest would hear it at 400 Hz. This means that if we move toward the car and then pass it, the frequency we'd hear would go from the higher frequency of 440 Hz down to the lower frequency of 370 Hz at the moment we pass the car.

If we stop, the sound returns to 400 Hz, the same as that heard by someone in the car. It doesn't matter where we stop. We could be on the way toward

⁉ The observed frequency is *higher* if the observer is moving *toward* the source and *lower* if the observer is moving *away from* the source.

the car when we stop, or we could have already passed it. When we stop, the sound returns to 400 Hz.

AS WE GET CLOSER, THE SOUND GETS LOUDER. DOESN'T THE FREQUENCY INCREASE ALSO?

? As we move toward the source, the sound we hear gets louder but the observed frequency doesn't change.

No. Remember, frequency is not the same as **intensity**. The intensity (loudness) of the sound we hear increases as we get closer and closer. The frequency we hear is not related to how close we are but rather whether we are moving toward or away.

In general, the observed frequency is *higher* if the observer is moving *toward* the source and *lower* if the observer is moving *away from* the source.

Example 20.1: A police car is parked with its siren on. Assume the siren has a constant frequency and your speed is constant.
(a) As you approach the police car (in your own car), does the frequency you hear increase, decrease or stay the same?
(b) As you pass the police car, does the frequency you hear increase, decrease or stay the same?
(c) As you move away from the police car (after passing the police car), does the frequency you hear increase, decrease or stay the same?

Answer 20.1: As noted above, the observed frequency is *higher* if the observer is moving *toward* the source and *lower* if the observer is moving *away from* the source. Consequently, the observed frequency of the siren decreases as the car passes by (part b). However, as you approach (part a), the observed frequency stays the same (i.e., remains higher than the emitted frequency). Conversely, as you leave (part c), the observed frequency stays the same (i.e., remains lower than the emitted frequency).

WHY DOES THE OBSERVED FREQUENCY DEPEND ON WHETHER WE ARE MOVING AWAY FROM THE SOURCE OR TOWARD IT?

To understand why this happens, consider the sound wave illustrated in Figure 20.1. Here the sound wave is produced by some stationary object (indicated by the dot at the center of the circles).

The sound produced by the source has a constant frequency, illustrated by equally-spaced circular wavefronts in Figure 20.1 moving away from the

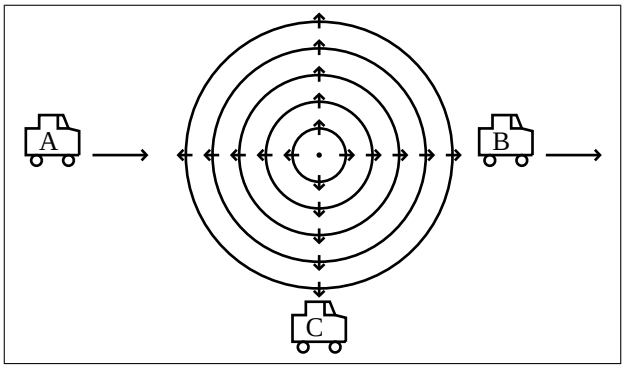


Figure 20.1: A snapshot of a sound wave produced by a stationary object (indicated by the dot at the center of the circles). The arrows indicate that the sound waves are moving away from the source. Car A moves toward the sound source, car B moves away from the sound source and car C is stationary.

source (with the direction of motion indicated by the arrows on each wavefront), similar to what you'd see when someone drops a rock into a pool of water, producing concentric water waves. Car C, being stationary, observes a frequency equal to that which is emitted.

Car A, on the other hand, is moving toward the source and hears a frequency that is higher than that emitted because car A *encounters* the pulses more often. Basically, each pulse doesn't have to travel as far to reach car A as the previous pulse did. There is less time between when consecutive pulses are received by car A vs. by car C, so car A receives them at a higher frequency.

Conversely, car B, moving away from the source, hears a frequency lower than that emitted because each pulse has to travel farther than the previous pulse as it tries to "catch up"' to the car. This means the period between receiving successive pulses is greater for car B vs. car C, so car B receives them at a lower frequency.

✓ *Check Point 20.3: A police car is parked with its siren on. The frequency of the siren is 1000 Hz.*
(a) As you approach the police car (in your own car), is the frequency you hear higher, lower or equal to 1000 Hz?
(b) After you pass the police car and are moving away from it, is the frequency you hear higher, lower or equal to 1000 Hz?

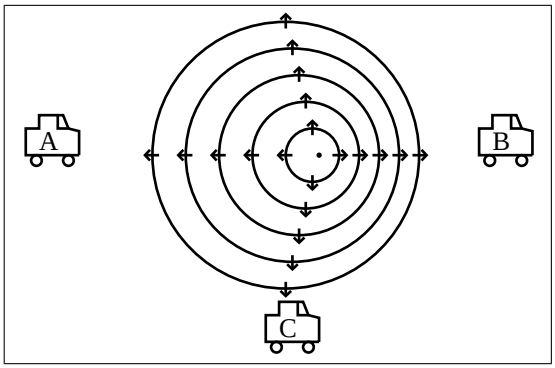


Figure 20.2: A snapshot of a sound wave produced by a rightward moving object (indicated by the dot at the center of the circles). The arrows indicate that the sound waves are moving away from the source. Observers at points A, B and C all hear different frequencies.

20.4 Moving sources

What happens if the source is moving?

Let's suppose the observer is stationary and the source is moving either toward the observer or away from it.

⁝ For a stationary observer (and moving source), the effect is similar to what happens with a stationary source (and moving observer).

In that case, one finds that the observed frequency is *higher* (than the emitted frequency) when the source is moving *toward* the observer and the observed frequency is *lower* (than the emitted frequency) when the source is moving *away from* the observer. Notice the similarity with what happens when the *observer* is moving instead of the source: the observed frequency is *higher* (than the emitted frequency) when the observer is moving *toward* the source and the observed frequency is *lower* (than the emitted frequency) when the observer is moving *away from* the source.

To see why this is the case, consider the sound wave illustrated in Figure 20.2. Here the sound wave is produced by a rightward-moving object (indicated by the dot at the center of the circles). As before, I use arrows on each wavefront to indicate the motion of the sound waves away from the source. Since the circular wavefronts move away from the location where they are *produced*, and that location is changing (as the source moves rightward), the center of each circular wavefront is shifted from the one before.

The end result is that the wavefronts at A (to the left) are farther apart than the wavefronts in other areas (B and C). At all locations, each compression moves one wavelength during one period, but for someone at point A each compression is initiated from a location further to the right (further from point A) than the previous compression.

Someone at position A, by being behind the source, experiences a wave that is more spread out than it would otherwise be, and thus with a longer wavelength and lower frequency. Someone at position B, by being ahead of the source, experiences a wave that is more compressed that it would otherwise be, and thus with a shorter wavelength and higher frequency. Someone at position C, in comparison, experiences a wave with wavelength equal to what would be received with a stationary source.

In general, the observed frequency is *higher* if the source is moving *toward* the observer and *lower* if the source is moving *away from* the observer.

Example 20.2: A police car is driving at 17 m/s with its siren on. You are standing on the side of the road as the police car passes by. As the car passes by, does the observed frequency of the siren increase, decrease or stay the same?

Answer 20.2: As noted above, the observed frequency is *higher* if the source is moving *toward* the observer and *lower* if the source is moving *away from* the observer. Consequently, the observed frequency of the siren decreases as the car passes by.

✓ *Check Point 20.4: A police car moving to the east with its siren on. The frequency of the siren would be 1000 Hz if the police car were stationary.*
(a) If we were east of the police car, so the car was driving toward us, would the frequency we hear be higher, lower or equal to 1000 Hz?
(b) If we were west of the police car, so the car was driving away from us, would the frequency we hear be higher, lower or equal to 1000 Hz?

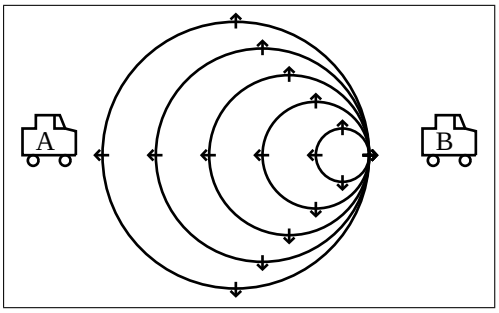


Figure 20.3: A snapshot of a sound wave produced by an object moving rightward at the same speed as the wave.

20.5 Moving at the speed of the wave

To get additional insight into the Doppler effect, let's examine some extreme examples. What happens, for example, if the observer is moving at the same speed as the wave?

In Figure 20.1 (on page 367), for example, a wave is illustrated moving away from a stationary source. Consider, then, what would happen if car B was moving away at a speed equal to that of the wave. In such a case, the wavefronts would move with the observer and so the observed frequency should be zero ($f_{\text{obs}} = 0$). The observer would hear nothing, as the pressure at their location wouldn't be changing.

What if the observer is moving toward the source at the wave speed?

That means the observer will be encountering the wavefronts more quickly than if stationary. It turns out the observed frequency is twice that emitted ($f_{\text{obs}} = 2f_{\text{emitted}}$).

To see why this is, consider what would happen if you and friend run toward each other at the same speed. You'd find that the two of you would meet midway between your initial locations. That halves the time between when you receive the pulses, which means you are doubling the frequency at which you encounter the pulses;

What if the source is moving instead of the observer?

Such a situation is illustrated in Figure 20.2 (on page 368). However, if the source is moving at the same speed as the wave, the wavefronts bunch up even further on one side, as seen in Figure 20.3.

The poor observer who is in the way will receive all of the wavefronts at once. Perhaps you have heard the "sonic boom" associated with when a jet "breaks" the "sound barrier". The "sonic boom" is due to all of the sound waves "piling up" on each other when the plane is traveling at the same speed as the sound wave. Mathematically, since all of the wavefronts are received at once, the observed frequency is infinite ($f_{\text{obs}} = \infty$).

⁉ A sonic boom is created as a source moves at the speed of sound.

WHAT IF THE SOURCE IS MOVING AWAY FROM THE OBSERVER AT THE WAVE SPEED?

That means the observer will be encountering the wavefronts more infrequently than if stationary. It turns out the observed frequency is half that emitted ($f_{\text{obs}} = \frac{1}{2} f_{\text{emitted}}$). To see why this is, imagine two friends who are both ten meters from you. If the one friend walks toward you while the other friend walks away from you then at the moment the first friend reaches you the second friend would be twenty meters from you. The distance between you and the second friend would be double what it was initially. In the same way, the effective wavelength would be doubled when the source moves away from you at the speed of sound, halving the frequency.

As you can see, there is a big difference between an observer moving toward the source at the wave speed ($f_{\text{obs}} = 2f_{\text{emitted}}$) and the source moving toward the observer at the wave speed ($f_{\text{obs}} = \infty$). Conversely, there is a big difference between an observer moving away from the source at the wave speed ($f_{\text{obs}} = 0$) and the source moving away from the observer at the wave speed ($f_{\text{obs}} = \frac{1}{2} f_{\text{emitted}}$).

In both cases, the observed frequency is higher if the observer and source are coming together and the observed frequency is lower if the observer and source are moving apart. However, the actual observed frequency depends on which is moving. For speeds much less than the wave speed, the difference is small. However, when the observer or source are moving quickly, the difference can be significant.

✓ *Check Point 20.5: Does the observed frequency depend on which object is moving – observer or source – even if the relative motion is the same? To answer, it helps to consider the extreme case (with one or the other moving at the speed of sound).*

Summary

This chapter examined how the observed frequency of a sound wave depends on the relative motion between the source and the observer.

The main points of this chapter are as follows:

- The observed frequency is *higher* if the observer is moving *toward* the source and *lower* if the observer is moving *away from* the source.
- As we move toward the source, the sound we hear gets louder but the observed frequency doesn't change.
- For a stationary observer (and moving source), the effect is similar to what happens with a stationary source (and moving observer).
- A sonic boom is created as a source moves at the speed of sound.

By now you should be able to relate the emitted and observed frequency when the transmitter and/or receiver are moving (Doppler Effect).

Frequently asked questions

Is the Doppler effect impacted by the wind?

Technically, the motion of the source and observer are measured relative to the medium. In other words, if the medium is moving (e.g., for a sound wave, if the wind is blowing), you need to add that wind speed (if moving against the wind) or subtract that wind speed (if moving with the wind) from the motion.

Does the frequency increase as a sound source approaches because, as it gets closer, the sound gets louder?

No. Although the sound does get louder as the source approaches (and gets softer as it leaves), frequency is not the same as intensity. The frequency is not related to how close we are but rather whether we are *moving* toward or away.

Terminology introduced

Doppler Effect
Intensity
Observer
Source

Additional problems

Problem 20.1: Suppose a car on an interstate is approaching a police car that is parked on the side of the road. The police car has its sirens on (frequency = 1000 Hz). For each questions (a) through (d), state whether the answer is $>$ 1000 Hz, $<$ 1000 Hz or = 1000 Hz. Assume the car is traveling at a constant speed.
(a) What siren frequency is heard by the driver of the car?
(b) Suppose the car honks his horn which has a tone at 1000 Hz also. What is the frequency of the horn as heard by the stationary police officer?
(c) Suppose the car passes the police car (so that it is now moving away from the police car at 50 mph instead of toward), what siren frequency is heard by the driver of the car?
(d) Suppose after passing the police car (so that it is now moving away from the police car at 50 mph instead of toward), the driver then honks his horn (1.000×10^3 Hz). What is the frequency of the horn as heard by the (still) stationary police officer?
(e) Of the four frequencies observed in the situations described in (a) through (d), which frequency is lowest?
(f) Of the four frequencies observed in the situations described in (a) through (d), which frequency is highest?

Problem 20.2: Repeat the questions in Problem 20.1 but with the car traveling at the speed of sound. Give numerical values with each answer.

Problem 20.3: Suppose a car is traveling west on interstate 80 at the speed limit (50 mph) and a police car comes up from behind (also traveling west) at 70 mph. The police car has its sirens on (frequency = 1.0×10^3 Hz). What is the siren frequency as heard by the driver of the car: $>$ 1000 Hz, $<$ 1000 Hz or = 1000 Hz?

21. Interference

Puzzle #21: When a radio is hooked up to two speakers, there may be "dead" spots in the room where some frequencies are softer than they would be with only one speaker. What causes such dead spots?

Introduction

The puzzle refers to a phenomenon called **interference**. We'll focus on sound but keep in mind that interference occurs with all waves.

21.1 Interference with single pulses

Wave interference occurs when we have two separate waves that are incident upon the same location. In the puzzle, for example, there are two speakers. Each speaker produces a sound wave and so an observer will receive two sets of waves: one from each speaker. It is the *combined* wave of the two waves added together that the observer hears.

Does the observer hear a louder sound?

This is where things get a little strange. You might expect to hear that you'd a louder sound than what you'd hear with either speaker by itself. While the combined sound wave *may* be louder, sometimes the combined wave is *softer* than each individual wave, resulting in dead spots, as mentioned in the puzzle.

✍ We'll assume that both speakers are directed right at the observer, so the softer sound is not simply a result of one or both speakers being turned away from the observer.

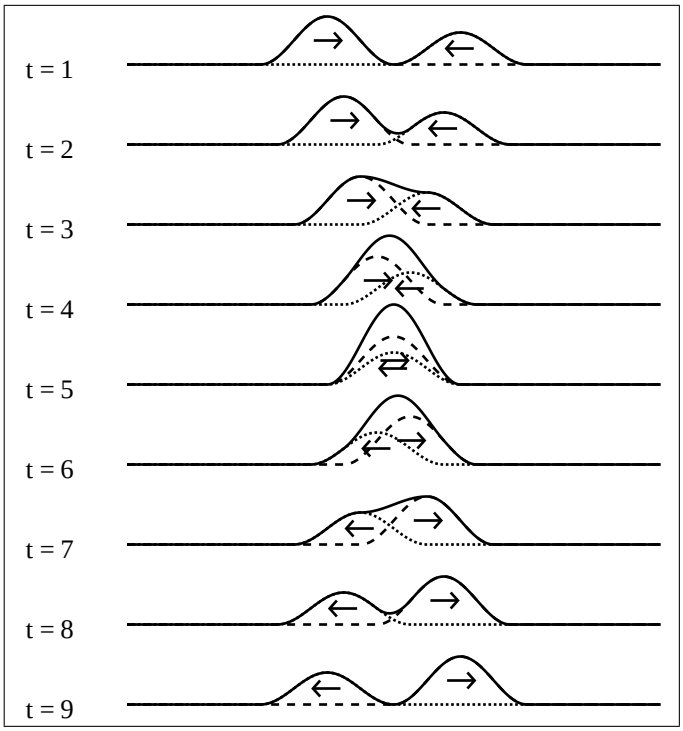


Figure 21.1: Nine "snapshots" of two pulses, traveling in opposite directions. The snapshots start at the top.

To illustrate how the combination of two sounds can produce a sound that is *softer* than the two sounds that make it up, we'll first consider the simpler situation where we have two single pulses rather than two sounds. As you may recall, sound is a wave, and a wave is made up of a series of pulses. By consider only one pulse, we can more easily see what is happening when waves interact.

To visualize what happens, consider the series of nine illustrations shown in Figure 21.1. Each illustration shows two separate pulses, traveling in opposite directions, a larger pulse traveling from left to right, and a smaller pulse traveling from right to left. One can see that when the pulses reach the same spot, there is a moment when they add together, creating an even bigger pulse.

Once they pass, they continue on their merry way as if nothing has happened. The larger pulse continues moving rightward, unchanged from what

it was before the interaction. Likewise, the smaller continues moving leftward, unchanged from what it was before the interaction. In the same way that two peaks combine to (temporarily) form a bigger peak, two troughs will combine to form a bigger trough (not shown) and then continue on their way, unchanged from what they were before the interaction.

In contrast to how two peaks or two troughs combine, a peak and trough will combine to form a *smaller* pulse, not a larger one. To visualize this, consider the series of illustrations shown in Figure 21.2. Notice how the two pulses "cancel" somewhat when they meet. Someone sitting at that spot would experience a smaller pulse than either of the two pulses by themselves. Also notice that, as before, the two pulses continue on their merry way after passing as though nothing has happened.

The two cases illustrate the essential nature of wave **interference**. Basically, when two ridges or two troughs happen upon the same location at the same time, the result is an even bigger ridge or trough. However, if a ridge happens upon a trough, the result is something smaller. In fact, if the ridges and troughs are the same size then combining a trough and ridge would result in nothing at all.[i]

✓ *Check Point 21.1: Why does snapshot 5 of Figure 21.2 show a combined pulse that is smaller than the individual waves but snapshot 5 of Figure 21.1 shows a larger pulse?*

21.2 Phase

In the previous section we examined a single pulse, either a ridge or a trough, interfering with a second pulse, either a ridge or trough. Since a wave is a

A **rogue wave** is a particularly tall wave that can occur in the open ocean. Though many people used to think rogue waves were just the stuff of folklore, nowadays scientists recognize that such waves form when two separate water peaks reach a particular location at the same time, combining together to create a bigger peak that seems to appear out of nowhere and then disappear (and thus the name "rogue wave"). Note that a rogue wave is *not* a tsunami or tidal wave, which is basically just a really wide wave due to a large amount of water being displaced due to an earthquake.

[i] If two ridges are the same size then combining them would result in a ridge twice as big. A similar thing would happen with two troughs that are the same size.

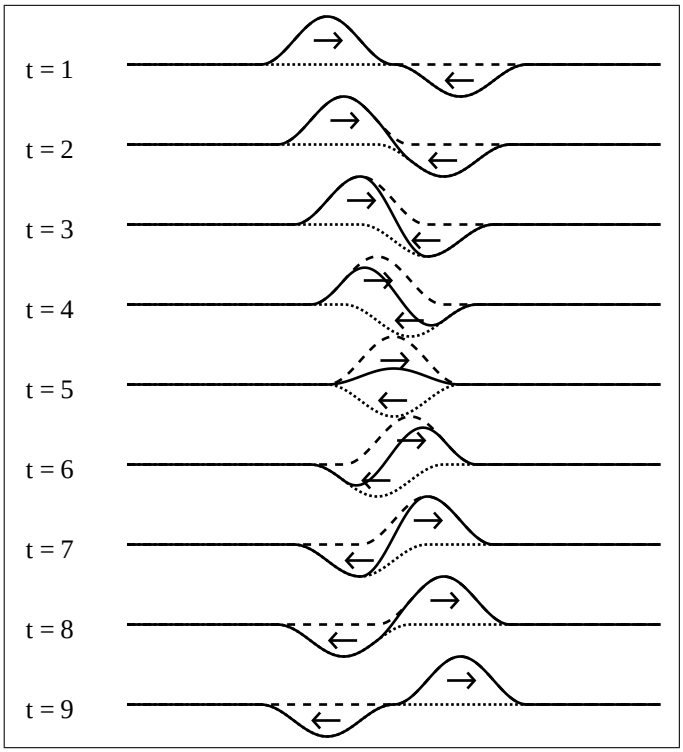


Figure 21.2: Nine "snapshots" of a single ridge encountering a single trough traveling in the opposite direction. The snapshots start at the top.

traveling series of ridges and troughs, what we want to know is how a *series* of ridges and troughs will interfere with each other.

As mentioned before, if two ridges are present at a given location at the same time, they combine to form a bigger ridge. In a wave, each ridge is followed by a trough, which combine to form a even bigger trough. The two troughs are then followed by ridges, forming a big ridge again. This will continue on and on as the ridges and troughs follow one another. And, just like single pulses, waves pass through each other without affecting one another. It is only at the location where both waves are present that we see the combination of the two.

With sound waves, the wave is made up of compressions and rarefactions instead of ridges and troughs. However, the idea is the same. When two compressions (from two different sources) are present at a given location at the same time, they combine to form a bigger compression. Since each com-

pression is followed by a rarefaction, that same location will then experience a bigger rarefaction than either individual rarefaction. This will continue on and on as the compressions and rarefactions follow one another. Such a situation would correspond to two sound waves combining to form a *louder* sound.

However, let's suppose we are at a location where a compression from one source reaches us at the same time as a rarefaction from the second source. If the two waves have identical amplitudes then the result is a complete cancellation and we'd experience neither a compression nor a rarefaction. If the amplitudes are not identical, there would still be some cancellation but not a total cancellation.

As the wave continues by, we again receive a compression and a rarefaction, but this time the compression is from the second source and the rarefaction is from the first source. Again, the result is a cancellation. This will continue on and on as the compressions and rarefactions follow one another. Such a situation would correspond to two sound waves combining to form a *softer* sound, as in the puzzle.

✍ As with single pulses, even when cancellation occurs at the location where the two waves combine, the waves still pass through that location unchanged from what they were prior to combining at that location.

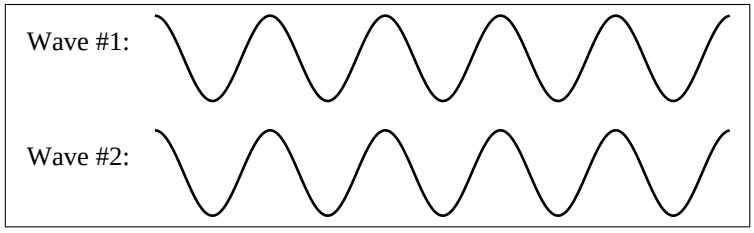
Predicting whether two sounds form a louder sound or a softer sound, then, is basically a matter of identifying whether we receive the compressions at the same time or at opposite times.

To help in predicting which occurs, I'll first introduce a language for describing and distinguishing between the two situations. The language will be based on something called **phase**.

When the compressions are received at the same time, we say that the two waves are **in phase**, meaning that they are in sync. When a compression of one wave is received at the same time as a rarefaction from a second wave, we say that the two waves are **out of phase**, meaning that they are out of sync.

✍ If two wave sources are in phase, that means that each source creates a compression at the same time. However, if the two sources are not the same distance away from the observer, the observer may not *receive* the compressions at the same time. In that case, the two waves may not be in phase at the observer.

I've illustrated two waves in phase below. Notice that the peaks coincide with each other and, similarly, the valleys coincide.



If the two waves were sound waves, the peaks and valleys would represent the compressions and rarefactions. If the two waves were water waves, the peaks and valleys would represent the ridges and troughs of the water.

In the drawing, the idea is that the observer experiences the peaks at the same time, and the valleys at the same time, which represents the two waves being in phase.

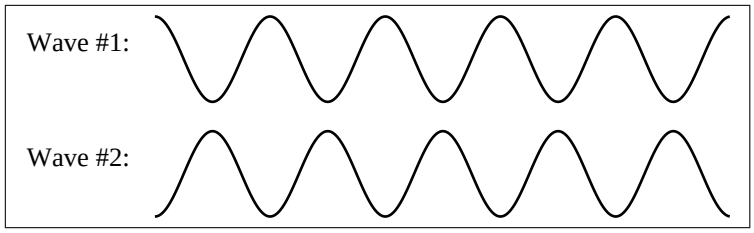
⁝ When you add two waves that are in phase, you get a wave with a bigger amplitude.

When you add two waves that are in phase, the peaks and valleys reinforce each other and you get a wave with a bigger amplitude. Mathematically, you get a wave with an amplitude equal to the *sum* of the individual amplitudes.

For example, if two waves are in phase and they have amplitudes A_1 and A_2 then the combination of the two will be a wave with an amplitude equal to $A_1 + A_2$.

What happens when the two waves are out of phase?

This is illustrated in the figure below. The peak of one wave coincides with the valley of the other (and visa-versa).



⁝ When you add two waves that out of phase, you get a wave with a smaller amplitude.

When you add two waves that are out of phase, the peaks of one wave tend to cancel out the valleys of the other. The result is a wave with an amplitude equal to the *difference* in the individual amplitudes.

For example, if two waves are out of phase and they have amplitudes A_1 and A_2 then the combination of the two will be a wave with an amplitude equal to $A_1 - A_2$.

✍ When you add two *identical* waves (same amplitude) that are out of phase, you get nothing (i.e., an amplitude of zero).

⁇ We use angle values to indicate how much two waves are out of phase.

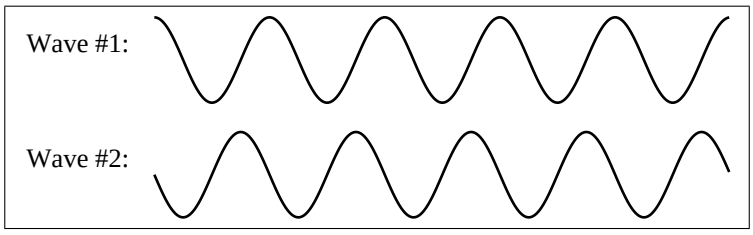
Technically, we say that two out of phase waves are 180° out of phase. This wording is based upon the convention of using 360° (or 2π radians) to represent an entire wavelength. Consequently, a phase difference of 180° means the wave is shifted by half a wavelength. In a similar way, waves in phase are 0° out of phase.

✍ It isn't that the waves are at an angle with each other but rather that we indicate our position along the up and down in terms of a number out of 360 that happens to have degrees as a unit. It is sort of like setting the temperature to be between 0 at freezing and 100 at boiling and using degrees to indicate where we are between freezing and boiling but there is no angle.

What if two waves are between 0° and 180° out of phase?

If two waves are between 0° and 180° out of phase, the combination will have an amplitude that is greater than the amplitude difference (i.e., what it would be if 180° out of phase) but less than the amplitude sum (i.e., what it would be if 0° out of phase). Mathematically, if the two waves had amplitudes A_1 and A_2, the combination would have an amplitude greater than $A_1 - A_2$ but less than $A_1 + A_2$.

For example, the two waves illustrated below are 90° out of phase (i.e., one-quarter of a cycle). The combination of these two waves will have an amplitude equal to the *square root of the sum of the squares* of the original two amplitudes ($\sqrt{A_1^2 + A_2^2}$).



✍ Just as two vectors of amplitude A combine to produce a sum of amplitude A when the two vectors are 120° apart (which you can show with a little trigonometry), it turns out that two waves of amplitude A will combine to produce a wave of amplitude A when the phase difference is 120° (i.e., one-third of a cycle).

Do all waves combine in this way?

Yes. This effect is true for all waves. We will be focusing on sound only because the effect is easier to demonstrate with sound.

✓ *Check Point 21.2: Suppose two waves have the same amplitude A but are 45° out of phase. Based on the discussion in the text, should the two waves produce a wave with an amplitude greater than A or less than A? If greater than A, does it equal $2A$? If less than A, does it equal zero? Explain.*

21.3 Constructive vs. destructive

It can be awkward to refer to the combination of two waves as the "sum" of two waves since the combined wave doesn't necessarily have an amplitude equal to the sum of the individual amplitudes. In fact, as seen in the previous section, if the two waves are completely out of phase, the combined wave has an amplitude equal to the *difference* of the two amplitudes.

For this reason, we refer to the combination of two waves as "interference". When the two waves are in phase, we call it **constructive interference** and when the two waves are out of phase we call it **destructive interference**.

What if two waves are between 0° and 180° out of phase?

For our purposes, we'll refer to that as *partial* constructive interference (or partial destructive interference; they mean the same thing).

✓ *Check Point 21.3: Suppose two waves of amplitudes A and B experience destructive interference. What is the amplitude of the combined wave?*

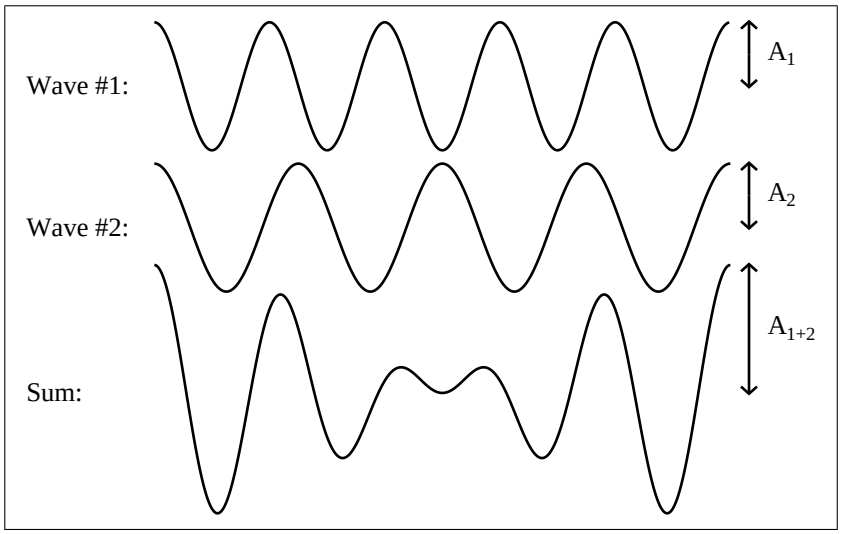
21.4 Beats

In the previous section, we examined how two waves combine. The assumption was that the two waves had exactly the same frequency (and, similarly, the same period). That way, if two peaks are coincident at some time then the valleys will necessarily be coincident a short time later, as will all the following peaks and valleys.

What happens if the two waves don't have the same frequency?

If the frequencies are not identical, the periods won't be identical either. That means if the peaks are coincident at one time, the next peaks won't be coincident. This results in a phenomenon called **beating**, where the sound alternates between loud (beat) and soft, like the beating of a drum.

To illustrate what happens, consider the figure below. Waves #1 and #2 have the same amplitude but different periods (wave #1 has five oscillations while wave #2 has four).



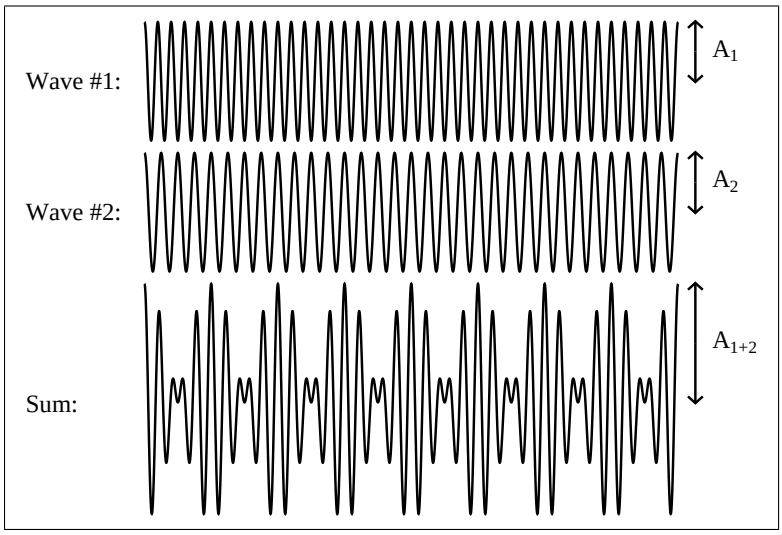
Notice that waves #1 and #2 are in phase at the beginning (left) and end (right) but out of phase in the middle. Consequently, the sum of the two waves (illustrated by the bottom wave in the figure) has a large amplitude at the ends (when the two waves are in phase) but a zero amplitude in the middle (when the two waves are out of phase). This results in beating, with

⁝ If the frequencies are similar but not identical, beats occur as the interference alternates between constructive and destructive.

the sound alternating between louder and softer as the interference alternates between constructive and destructive.

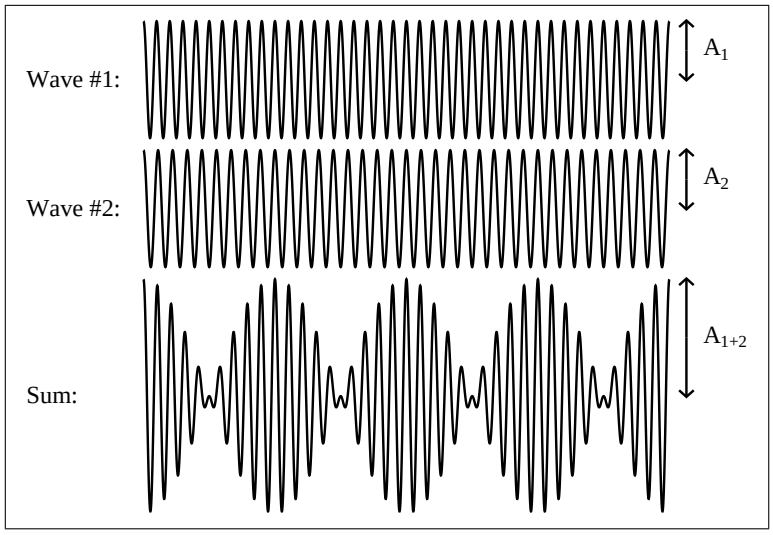
Does each peak correspond to a loud sound?

No. Remember that sound is due to the oscillation itself so a loud sound corresponds to where the oscillation amplitude is large (at the beginning and end in the case illustrated), not at each peak. This can be seen more clearly in the figure below, where wave #1 has 40 oscillations and wave #2 has 32.



In the illustration, wave #1 represents a sound with a *steady* unchanging loudness, as does wave #2. If you could hear both sounds, each sound would be equally as loud since they have the same amplitude. They'd only differ in pitch, although the pitch would be very similar (since their frequencies are similar). The sound you'd hear when listening to the *combined* wave (sum), however, would *not* have a steady unchanging loudness. Instead, you'd hear it being loud when the oscillation back and forth is large (nine times in the figure), and soft when the oscillation back and forth is small (eight times in the figure).

The closer the two original frequencies, the less often the two waves will switch from being in phase (combined wave has larger amplitude) to out of phase (combined wave has smaller amplitude). This can be seen in the figure below, where the two frequencies are even more similar (36 and 32 oscillations). Notice that when the frequencies are closer (as in the second case) there are *less* beats.



The time it takes from one in-phase moment (a beat) to the next in-phase moment (another beat) is called the **beat period**. Alternatively, the number of beats per time is known as the **beat frequency**.

For example, in the first case (with 40 and 32 oscillations) the combined wave exhibits eight full beats.[ii] If the total time is one second, the beat frequency would be 8 Hz. In comparison, in second case (with 36 and 32 oscillations), the combined wave exhibits four full beats. If the total time is one second, the beat frequency would be 4 Hz. The beat frequency in the second case is lower because the two waves in that case have frequencies that are more similar than the two waves in the first case.

It turns out that the beat frequency is equal to the *difference* between the two original frequencies:

$$f_{\text{beat}} = |f_2 - f_1| \tag{21.1}$$

For example, in the first case (with 40 and 32 oscillations), the difference is eight, which is the number of beats (during the time interval shown). In comparison, in the second case (with 36 and 32 oscillations), the difference is only four, which is the number of beats (during the time interval shown). This is in keeping with the expectation that the smaller the difference in the two frequencies, the smaller the beat frequency (and the larger the beat period) of the combined wave.

[ii] Notice that there is a half of a beat at the beginning and another half of a beat at the end.

✍ If the two frequencies are identical, the beat frequency is zero (i.e., the phase difference is constant and there are no beats).

IS THE BEAT FREQUENCY THE SAME AS THE FREQUENCY OF THE COMBINED WAVE?

No. The "beat frequency" refers to the pulsing of the combined wave as the original two waves move in and out of phase. The combined wave actually oscillates up and down with a frequency that is greater than beat frequency.

Indeed, the frequency of the combined wave is the **average** of the two original frequencies:

$$f_{\text{average}} = \frac{f_1 + f_2}{2} \tag{21.2}$$

Distinguishing between these two may be confusing. For sound, the beat frequency describes how quickly the tone is beating (i.e., how quickly the tone is oscillating between loud and soft) while the average frequency describes the pitch of the combined wave, which is very close to the pitch of the two original waves.

✓ *Check Point 21.4: (a) When a guitar string is sounded along with a 440-Hz tuning fork, a beat frequency of 5 Hz is heard. What two frequencies could the string have?*
(b) When the same string is sounded along with a 436-Hz tuning fork, the beat frequency is 9 Hz. Which of the two frequencies found in (a) must the string have?

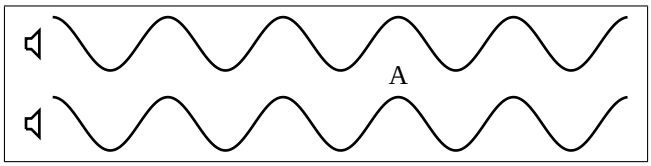
21.5 Two point sources

To explain the puzzle, we have to recognize that each speaker produces sound waves and that these sound waves can combine in ways to produce louder or softer sounds, depending on their phase difference.

Let's consider the situation illustrated below. In this case, the illustration shows two speakers right next to each other. Each speaker produces waves of identical wavelength[iii] and in phase, represented by the wavy lines. At a

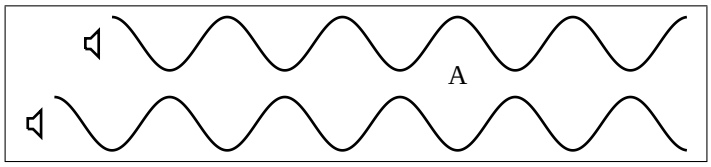
[iii] For stereo speakers, each speaker can be producing different sounds and so the frequencies (and wavelengths) may not be the same.

particular moment in time, we take a snapshot of the two waves, which is what is represented in the figure.



Since the speakers are in phase (i.e., when one speaker cone is compressing the air, so is the other), the two sound waves are also in phase. For example, at the location indicated as A, the peak of each wave is present. That produces a bigger peak at A (combined wave not shown). A short time later, as the waves continue to travel to the right, two valleys will be present at location A, producing a deeper valley.

In fact, since the speakers are in phase and located at the same place, they interfere constructively at all points, not just at location A. However, suppose one speaker is in front of the other, as illustrated below.



The two speakers are still in phase (as evidenced by the fact that each wave *starts* at the same part of the wave) but the sound waves are *no longer* in phase at location A because one has a "head start", so to speak. At the moment the picture was taken, each speaker has produced six peaks. The third of the six produced by the top speaker has reached location A at the moment the picture was taken. However, the third of the six produced by the bottom speaker has not. Instead, at the moment the third peak from the top speaker is at location A, a valley from the bottom speaker is there. These cancel at location A (as well as at every other location).

In other words, the two waves are 180° out of phase and interfere destructively.

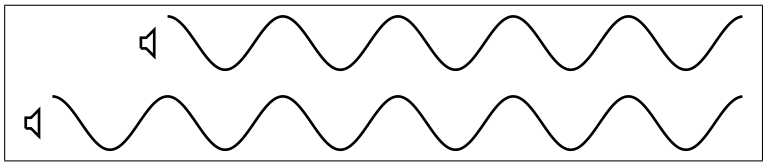
In this case, the two speakers are separated by a distance equal to one-half of a wavelength. When two speakers are arranged in this way, the sound we

hear (i.e., the *combined* wave) is softer than it would otherwise be due to destructive interference.[iv]

For a "dead spot" to occur, then, the two waves must be *out of phase* at that location. And, in order to keep the sound soft, with no beating, the two waves must have the same frequency.[v]

✍ The two *speakers* are in phase with each other but the *sound waves* are not, since the speakers are not at the same location.

If we continue to shift the position of one speaker relative to the other, we'll find that at some point the two waves are once again in phase. To do that, they have to be separated by a distance equal to one wavelength of the sound wave. This is illustrated below.



It turns out that the sound waves will be in phase whenever the two sources are separated by an integer multiple of the wavelength distance (i.e., $0, 1\lambda, 2\lambda$, etc.) and they will be 180° out of phase when they are a half-integer wavelength different (i.e., $\frac{1}{2}\lambda$, $\frac{3}{2}\lambda$, $\frac{5}{2}\lambda$, etc.).

✓ *Check Point 21.5: A speaker is produce a pure tone of 680 Hz, which has a wavelength in air of 0.5 m (since the speed of sound is 340 m/s). How far behind this speaker would we need to place a second speaker, producing the same tone and in phase with the first speaker, such that they interfere destructively and produce the softest sound?*

[iv] In "real life", the two sound waves likely don't cancel each other totally because the closer speaker is going to have a slightly larger amplitude at your location, as it is closer.

[v] This is called **coherency**. If the two frequencies are exactly the same, then we necessarily have coherency. However, if they deviate even by a fraction, we have lost coherency. In other words, they may be out of phase at one particular moment but they won't stay out of phase. Rather, they will alternate between being in phase and out of phase (which we call beating).

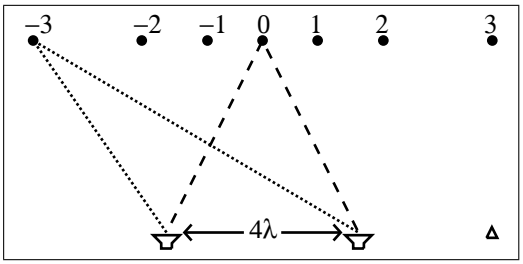


Figure 21.3: Two speakers that are separated by a distance equal to 4 wavelengths. The meaning of the dots and triangle are discussed in the text.

21.6 Interference in two dimensions

The puzzle mentions the existence of dead spots. Why would it be "dead" only in one spot? Wouldn't it be dead everywhere?

Just because the two waves are out of phase at one location does not mean the two waves would be out of phase at *every* location.

The reason has to do with the fact that we are dealing with a two-dimensional situation (unlike the situations discussed up to now, which were in one dimension).

Two sources of identical frequency but separated in space will create regions of constructive interference and regions of destructive interference.

To illustrate what I mean, again consider two speakers that are *in phase*. However, rather than placing one in front of the other, let's consider the situation where there are side-by-side, as illustrated in Figure 21.3.

Each speaker produces an identical sound. Let's suppose the two speakers are separated by a distance that is four times the wavelength of the sound. As we know from the last section, someone positioned directly to the right of the right speaker (see triangle in the figure) will hear a loud tone because, from that location, the right speaker is an integer-number of wavelengths (i.e., 4) closer to the observer than the left speaker is and thus constructive interference occurs at the observer's location.

Now consider an observer located at the dot labeled "0". At that position, the observer is equidistant to each speaker. Again, as discussed in the last section, someone positioned at that location will hear a loud tone because, from that location, each speaker is the same distance away (see dashed lines) and thus constructive interference occurs at that location also.

What about the other locations indicated in the figure?

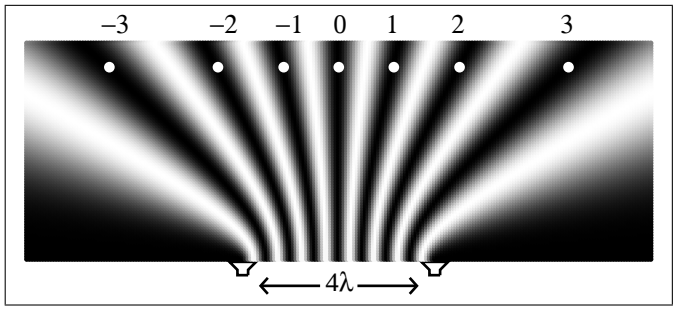


Figure 21.4: Two speakers are separated by a distance equal to four wavelengths as in Figure 21.3. White areas indicate regions of destructive interference (soft sound) and black areas indicate regions of constructive interference (loud sound). The white dots represent the locations of the black dots in Figure 21.3.

It can be difficult to tell without further information. It turns out that the dots indicate locations of constructive interference, with the number above each dot indicating much closer (in terms of number of wavelengths) the dot is to one speaker than the other speaker. For example, the dot labeled "−3" happens to be three wavelengths closer to the left speaker than the right speaker.[vi]

To better visualize where the areas of constructive and destructive interference are, we can use shading, as in Figure 21.4. Notice how the dots, reproduced from Figure 21.3 but now white instead of black, lie in regions where there is constructive interference (black areas; loud sound), with regions of destructive interference at locations midway between the dots (white areas; soft sound). However, there are lots of locations that experience constructive interference and lots of locations that experience destructive interference. All locations directly to the left or right of the speakers experience constructive interference (black shading in the figure) but between the speakers the type of interference depends on where the observer is (alternating between white and black).

To understand why Figure 21.4 looks the way it does, consider the illustration in Figure 21.5. Notice that all of the black areas (constructive interference; loud sound) in Figure 21.4 lie along the dotted lines in Figure 21.5.

[vi]The difference in distance must be less than four wavelengths, since it is not along the line of the speakers, as with the location indicated by the triangle.

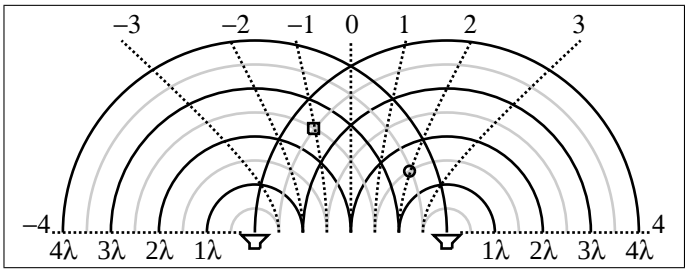


Figure 21.5: Two speakers are separated by a distance equal to four wavelengths as in Figure 21.3. The vertical dotted line 0 marks all of the points that are equidistant between the two speakers. The other dotted lines mark all of the points that are one, two or three wavelengths closer to one speaker than the other.

To see why constructive interference occurs along the dotted lines, I've drawn solid black semicircles so you can see how far each location is from each speaker (the number of wavelengths is marked on the bottom). In between the solid black semicircles are lighter semicircles that are halfway between the others.

For example, the right speaker lies along the semicircular labeled as 4λ (see bottom left of figure), which means the right speaker is four wavelengths from the left speaker, consistent with the speakers being four wavelengths apart.

As another example, consider the position indicated by the open circle, which is drawn along the dotted line labeled "2". That location is 1.5 wavelengths away from the *right* speaker (as evidenced by the fact that it sits on a light gray semicircle between the 1λ semicircle and the 2λ semicircle). That location is *also* 3.5 wavelengths away from the *left* speaker (as evidenced by the fact that it sits on a light gray semicircle between the 3λ semicircle and the 4λ semicircle). The *difference* between those distances is two wavelengths, as indicated by the dotted line, which means that constructive interference is experienced at that location.

In fact, all positions along the dotted line labeled "2" are two wavelengths closer to the right speaker than the left speaker. Similarly, all positions along the dotted line labeled "−2" are two wavelengths closer to the left speaker than the right speaker. This means that the two waves interfere constructively there, resulting in a louder sound than each speaker individually.

⁉ Constructive interference will occur at locations that are an integer-number of wavelengths closer to one source than the other.

✍ Since the two sources are four wavelengths apart, there are no points that are five or more wavelengths closer to one source than the other.

Mathematically, if we use $\Delta\ell$ as the difference between how far that location is from each source, then $\Delta\ell$ divided by λ, the wavelength, will be equal to the difference in distance in terms of the number of wavelengths.

We can then state that constructive interference will occur whenever the difference (in terms of number of wavelengths) is whole number:

$$\Delta\ell = (\text{whole number}) \times \lambda \qquad \text{[constructive]} \qquad (21.3)$$

For the dotted lines labeled "-1" and "1", $\Delta\ell/\lambda$ equals 1, which is a whole number and corresponds to constructive interference.

What would happen if the two speakers were separated by a different number of wavelengths?

If the two speakers were moved further apart, or if the wavelength of the sound was shortened (i.e., higher frequency), there would be more lines of constructive interference (i.e., more dotted lines in Figure 21.5).

✓ *Check Point 21.6: Examine the open square in Figure 21.5 that lies along the line labeled "-1".*
(a) How far is it (in number of wavelengths) from the left source?
(b) How far is it (in number of wavelengths) from the right source?
(c) What is the difference in distance between (a) and (b)?
(d) Is your finding consistent with constructive interference occurring at that location?

What happens if the difference in distance is not an integer multiple of the wavelength?

Then we don't have total constructive interference.

For example, destructive interference occurs at locations halfway between the dotted lines in Figure 21.5.

⁉ Destructive interference will occur at locations that are an half-integer number of wavelengths closer to one source than the other.

Does destructive interference occur when the location is half a wavelength closer to one source than the other?

Yes, and destructive interference not only occurs at locations that are half a wavelength closer to one source than the other but also at locations that are

1.5 wavelengths closer to one source than the other, and locations that are 2.5 wavelengths closer to one source than the other, and so on.

In fact, destructive interference will occur whenever the difference is a "half-integer", like 0.5, 1.5, 2.5, etc. Mathematically, we can represent this as follows:

$$\Delta\ell = (\text{half integer}) \times \lambda \qquad \text{[destructive]} \qquad (21.4)$$

✍ As with constructive interference, the distances to each source do not have to be half or whole integers. For example, if a location is 1.23 wavelengths from one source, destructive interference will occur if the distance to the other source is 0.73 wavelengths, 1.73 wavelengths. 2.73 wavelengths, 3.73 wavelengths, etc.

✓ *Check Point 21.7: Two sources are in phase and emit waves that have a wavelength of 0.44 m. For each of the following pairs of distances, determine whether constructive or destructive interference occurs at a point whose distances from the two sources correspond to the numbers given.*
(a) 1.32 m away from one source and 3.08 m away from the other source
(b) 2.67 m away from one source and 3.33 m away from the other source
(c) 2.20 m away from one source and 3.74 m away from the other source
(d) 1.10 m away from one source and 4.18 m away from the other source

---

Summary

This chapter examined how waves interfere with one another.

The main points of this chapter are as follows:

- When you add two waves that are in phase, you get a wave with a bigger amplitude.
- When you add two waves that out of phase, you get a wave with a smaller amplitude.
- We use angle values to indicate how much two waves are out of phase.
- If the frequencies are similar but not identical, beats occur as the interference alternates between constructive and destructive.

- Two sources of identical frequency but separated in space will create regions of constructive interference and regions of destructive interference.
- Constructive interference will occur at locations that are an integer-number of wavelengths closer to one source than the other.
- Destructive interference will occur at locations that are an half-integer-number of wavelengths closer to one source than the other.

By now you should be able to predict when and where two waves will interfere constructively or destructively.

Frequently asked questions

By "interference", do you mean that the combined wave is noisy?

No. Our use of the term "interference" may be a little different from the one you are used to. When people outside physics use the term (as in "interference in a phone line"), the implication is that the result is noisy.

In physics, interference simply refers to the combination of two waves that are incident upon the same point. The two waves can interfere "constructively" or "destructively". In constructive interference, the two waves are in phase and the resultant wave has an amplitude equal to the sum of the two individual ones.

Note that the result may be noisy but then again it may be not noisy at all. For example, in our usage, when two pure tones interfere we will simply get a louder tone or softer tone. We will not get a noisy tone.

Does constructive interference have to occur an integer-multiple of wavelengths from a source?

No. For example, if a location is 1.23 wavelengths from each source, constructive interference will occur. Or, the location could be 1.23 from one source and 0.23 wavelengths from the other source. Or, 2.23 wavelengths from the other source. Or, 3.23 wavelengths from the source. Or, well, hopefully you get the idea that it is the *difference* in the distance to each source that matters.

Suppose we are equidistant from the source of two water waves, with each source having the same frequency and in phase. Does that mean the water is higher where we are?

No. It means that the water is oscillating with a larger amplitude than it would if only a single source was present.

Are there always the same number of constructive interference lines regardless of how close the sources are to each other?

No.

If the two sources are separated by less than one wavelength, then it is impossible to be more than one wavelength further from one than from the other. Thus, there is only one line of constructive interference (the central line). If the two sources are separated by many wavelengths, then it is possible to be several wavelengths further from one than from the other. Thus, there are many more lines of constructive interference.

Terminology introduced

Beat frequency	Coherency	In phase
Beat period	Constructive interference	Out of phase
Beating	Destructive interference	Phase

Additional problems

Problem 21.1: Why is it difficult to obtain "dead areas" or locations that always have destructive interference if the two source have different frequencies?

22. Standing Waves

Puzzle #22: Why is the sound produced by a violin different from the sound produced by a trumpet, even when they play the same note?

Introduction

In chapter 21 we investigated the phenomenon of interference and showed that waves can combine either constructively or destructively. At first glance, it may not appear that musical instruments (like the violin and trumpet) create sound by interference, but they do. In this chapter we explore how each instrument does that, how each instrument selects the note to be played, and why the sound produced by different instruments sound different, even when they play the same note. While the context of our exploration will be mostly with sound and musical instruments, keep in mind that we can apply these ideas to any wave.

22.1 Visualizing a vibrating string

Plucking a guitar string makes it vibrate. Figure 22.1 illustrates what the vibrating string looks like by indicating the position of the string (black curve) at eight different instants in time.[i] The first instant is at the top, with the string drawn in black. In each succeeding image until time 5 the string has moved slightly downward, with the previous positions drawn in gray. The resulting time-lapse (multiple exposure) representation of the string (in gray) looks something like an elongated onion.[ii]

[i] While an actual string might not look exactly like what is shown in Figure 22.1, it is close enough to help us analyze what is going on and start our path toward answering the puzzle.

[ii] At least it does to me.

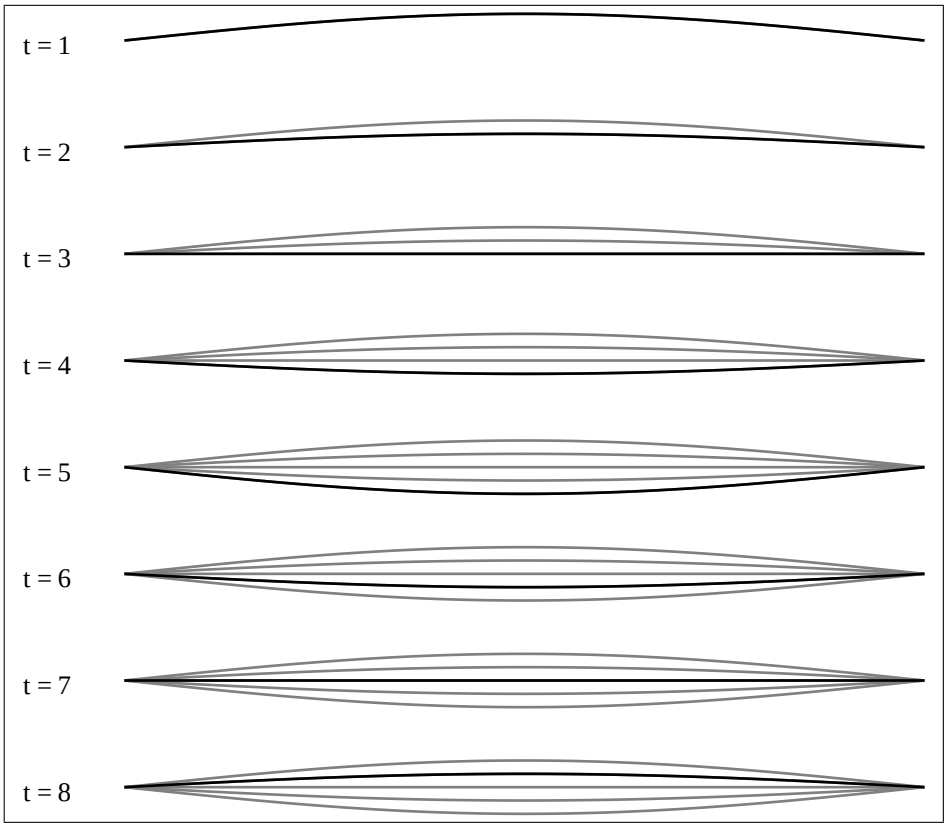


Figure 22.1: Eight "snapshots" of a vibrating string (black curves) superimposed on the time lapse representation (in gray). The snapshots start at the top.

The ends of the string are fixed to the guitar so the right and left ends of the string can't move. The greatest motion is in the middle, a region called the **antinode**, to contrast it with the ends, each called a **node**, where the string is stationary and does not move.

✍ Note that when the string is flat (at $t = 3$ and $t = 7$), the string is still moving up (at $t = 7$) or down (at $t = 3$) except for at the ends.

✓ *Check Point 22.1: How many nodes are present for the string in Figure 22.1?*

22.2 How a standing wave is generated

A vibrating string, as illustrated in Figure 22.1 is example of a **standing wave**. To explain why it is called a standing wave, we need to explain how it comes about.

Basically, plucking the string produces a pulse that splits in two, with one half traveling one way along the string and the other half traveling the other way.[iii] The pulses then reflect off the ends and travel back toward the middle, where they pass through each other and continue to reflect off the ends, moving back and forth along the string but in opposite directions. What we see on the string is just a combination of the two oppositely traveling pulses.

By visualizing two identical waves traveling in opposite directions, as in Figure 22.2, we can observe how this results in the standing wave illustrated in Figure 22.1. Each wave, after all, is a continuous series of peaks and valleys. As discussed in section 21.1, two peaks at the same location result in a peak that is twice as big, while a peak and a valley at the same location results in a cancellation. As the rightward-traveling wave interferes with the leftward-traveling wave, sometimes a peak of one coincides with the valley of the other (total destructive interference; see snapshots 4, 6 and 8) and sometimes the two peaks or two valleys coincide (total constructive interference; see snapshots 5, 7 and 9).

How does this produce what is illustrated in Figure 22.1?

To see why the result is what we see in Figure 22.1, it helps to extend our examination to times beyond what is shown in Figure 22.2. Figure 22.3, for example, illustrates the result for snapshots 11 through 16.

Notice that when the two waves (illustrated in gray) are out of phase, the string (black line) is flat.[iv] This is illustrated in snapshots 12, 14 and 16.

In comparison, when the two waves (illustrated in gray) are in phase, the string (black line) exhibits the peaks and valleys of a regular wave but with an amplitude twice as big as the original two waves. This is illustrated in snapshots 11, 13 and 15.

⁉ Standing waves can be set up with fixed ends, with the result being that there is a node at each end.

[iii] The tension in the string brings the string back toward equilibrium much like gravity brings a pendulum ball back toward equilibrium. This produces a pulse that travels along the string.

[iv] In general, during this moment the material matches the equilibrium position. That happens to be a flat string for the situation involving a string.

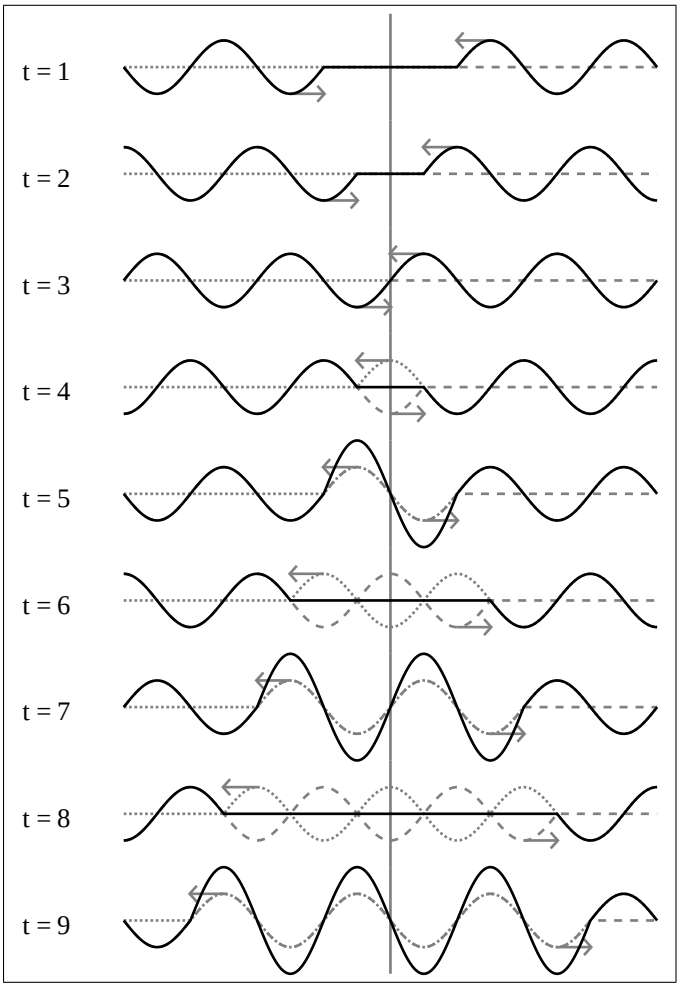


Figure 22.2: Nine "snapshots" of a wave traveling toward the right encountering a wave (of the same wavelength and frequency) traveling toward the left. The snapshots start at the top. The motion of the two waves are indicated by the gray arrows. The gray dashed and dotted lines indicate what the two waves would look like if they didn't interfere with each other.

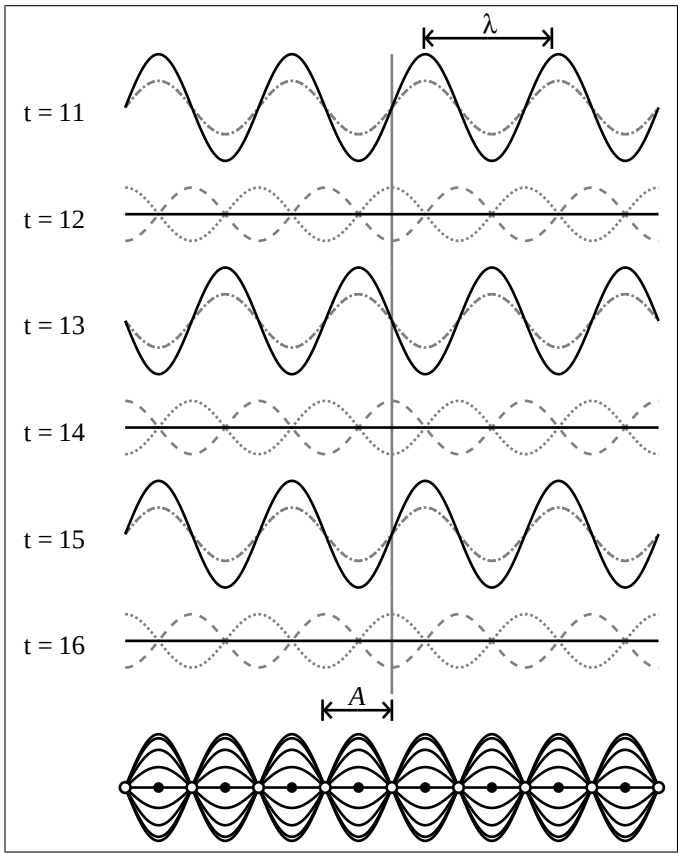


Figure 22.3: Six "snapshots" of a two identical but opposite-traveling waves (each traveling the entire length shown). The gray dashed and dotted lines indicate what the two waves would look like if they didn't interfere with each other. The black line is what we observe.

In practice, the string oscillates so rapidly that the string appears as a blur as it transitions through all the different stages (represented by the snapshots in Figure 22.3). A time-lapse (multiple exposure) image, where images at several instants are superimposed on top of one another, produces a series of elongated onions. The vibrating string in Figure 22.1 is just the portion in Figure 22.3 indicated by the double arrow with label "A".

The result is called a standing wave because it appears to remain stationary – it doesn't "travel" along the string like a wave. While the terminology may be a little misleading, since the standing wave is actually a combination of two individual waves traveling in opposite directions, it reveals an important thing about the vibrating string, namely that *the wave equation applies*. Recall that the wave equation relates the frequency of the vibrating string with the wavelength and wave speed of the traveling waves along the string. That means we can use the wave equation to predict the frequency of the vibrating string and thus the pitch of the sound (since the sound produced by the vibrating string has the same frequency as the vibrating string).

This is exactly what we'll do (use the wave equation) in the next section. To do that, though, we first have to recognize what we mean by the wavelength. Remember, it is the wavelength (or pulse separation) of the wave traveling along the string. For the case illustrated in Figure 22.3, the wavelength is indicated by the double arrow with label "λ".

Notice that the wavelength is *twice* the length of the distance between two nodes (see double arrow with label "A"). In other words, if we are to use the wave equation to determine the frequency of a vibrating guitar string, the wavelength value will be *twice* the length of the vibrating string.

✓ *Check Point 22.2: Suppose the length of the string in Figure 22.1 is 18 cm. What is the wavelength of the standing wave illustrated in that figure?*

22.3 Controlling the pitch of the string

There are two main ways one can select the desired note on a guitar – plucking a different string and placing one's finger along the string. A third way is

to tighten or loosen the string. To see why these things impact the note, we need to return to the wave equation ($v = f\lambda$, equation 19.1).

First, as mentioned in the previous section, we need to recognize that the pitch (frequency) of the sound produced by the vibrating string is equal to the frequency at which the string itself vibrates.[v] According to the wave equation, that frequency (f) is equal to the wave speed (v) divided by the wavelength (λ). To change the frequency (pitch), then, we need to either change the wave speed or change the wavelength, as described below.

22.3.1 Changing the wave speed

The frequency of a standing wave is proportional to the wave speed.

From the wave equation, the frequency is proportional to the wave speed. Increase the wave speed and the frequency must likewise increase (if the wavelength is kept the same). But what is the wave speed?

For a standing wave on a string, the wave speed is the speed a pulse travels along the string. There are two ways to change that wave speed. One way is to is to change the string tension. Tightening the string results in a faster wave speed. So, to change the note one can tighten or loosen the string. Tighter strings correspond to higher notes.

Another way to change the speed of a wave is to change the string mass per length. The heavier the string (per length), the slower the wave speed. So, to change a note, one simply has to pluck a string with a different weight. This is why stringed instruments like a guitar and violin have strings of different masses, with heavier strings corresponding to lower notes.

✍ Note that the way we pluck a guitar string (or bow[vi] a violin string) can change how *loud* the sound is (by changing the amplitude of the resulting standing wave) but does not change the wave speed and thus has no (or little) impact on the note produced.

[v] Technically, the body of the instrument vibrates at the same rate as the string, and the body acts like a speaker cone to compress and rarefy the air, which then produces the sound wave (see discussion about pitch in section 19.2).

[vi] The string on a violin is set to vibrate by taking a bow and sliding the bow across the string. As it slides across the string, it continually grabs the string and releases it.

✓ *Check Point 22.3: The "A" string on a violin is 66 cm long and produces a sound of frequency 440 Hz. What must be the wave speed?*

22.3.2 Wavelength

❣ The frequency of a standing wave is inversely proportional to its wavelength.

The instrumentalist usually just fiddles with the tension once, before playing a particular piece, just to get the desired notes for each string, and then leaves the tension alone during the piece. After all, it would take too long to continually change the tension to keep up with the different notes that are needed. Instead, to change the note, the instrumentalist plucks a different string. However, a guitar only has six strings (usually). How does the instrumentalist play more than six notes?

The main way one selects different notes is by changing the wavelength. From the wave equation, the frequency is inversely proportional to the wavelength. Increase the wavelength and the frequency must likewise decrease (if the wave speed is kept the same). Recall that the wavelength of the standing wave is twice the length of the string. To change the wavelength, then, we just have to change the length of the string. More precisely, we just have to change the length of the string that is *vibrating*. On a violin or guitar, the length of the vibrating portion of the string can be shortened by pressing one's finger against the string. This forces the traveling waves to reflect off the finger location rather than where the string is fixed to the instrument. With a shorter wavelength, the frequency is higher, producing a higher pitched note.

Of course, with a string, there is a limit to where you can put your finger and shorten the string. A beginning violinist tends to keep their hand in one place and use their different fingers to get different notes. With that fixed hand position, the pinkie finger can be placed at a location such that the resulting pitch is about four notes higher than what is achieved without placing any fingers on the string.[vii] To get higher notes, the violinist simply switches to one of the lighter strings on the instrument.

[vii] By shifting where the hand is placed, one can get a much larger range of notes from a single string. A really good violinist can get multiple octaves out of a single string.

✍ In chapter 19, it was mentioned that we usually control the frequency (by controlling the speaker cone oscillation, for example), which then impacts the wavelength (since the wave speed is determined by the material). For standing waves, however, we usually control the wavelength, with frequency determined by the wavelength, rather than the other way around.

✓ *Check Point 22.4: The "A" string on a violin is 66 cm long and produces a sound of frequency 440 Hz. What is the frequency of the note produced when you place your finger in such a way to make the vibrating portion only 64 cm long? You can assume that the tension hasn't changed (meaning that the wave speed is the same as that determined in the previous checkpoint).*

22.4 Wind instruments

Whereas string instruments (like the violin and guitar) use a vibrating string to produce sound, wind instruments (like the flute and clarinet) vibrate the air directly (the air being inside the instrument). However, the same ideas apply, namely that a standing wave is set up within the instrument, and the frequency of the standing wave is related to the wave speed and the wavelength via the wave equation. Since the vibrating medium is the air itself, for a wind instrument the wave speed is just the speed of sound in air.

Is the wavelength twice the length of the instrument, as with a string instrument?

It depends. For some instruments, like the flute and oboe, the wavelength is twice the length of the instrument. For other instruments, like the clarinet, the wavelength is *four* times the length of the instrument. And in others, like with brass instruments, it is a bit more complicated. We'll explain why in this section.

22.4.1 Open pipes

To explain the variation, let's first start with the flute and oboe, where the standing wave has a wavelength equal to twice the length of the instrument,

just like with string instruments. To understand why, you first have to recognize that it is the *air* inside the instrument that is vibrating, not a string, and unlike a string there is nothing to hold the air in place at each end of the instrument. The flute, like the oboe, is considered to be an **open pipe**, meaning it is a tube open at each end. Since the ends are open, the air at each end is free to oscillate. That means there is an *antinode* at each end of the instrument, not a node, as illustrated below.

⁝ Standing waves can be set up with free ends, with the result being that there is an antinode at each end.

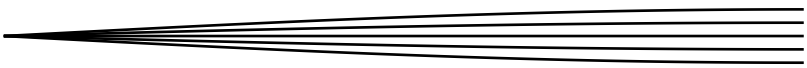
Notice that there is still a node, where the air doesn't oscillate, but it is at the center, not the ends. And the resulting standing still has the size of one whole onion, so to speak, because there are two half-onions. Consequently, just like with the string fixed on both ends where the standing wave has a wavelength equal to twice the length, a tube open on both ends allows for a standing wave with wavelength equal to twice the length.

✓ *Check Point 22.5: What frequency is produced on a flute that is 66 cm long if it acts like an open pipe? The speed of sound in air is 343 m/s (at 20° C).*

22.4.2 Closed pipes

⁝ Standing waves can be set up with mixed ends, with the result being that there is a node at the fixed end and an antinode at the free end.

Let's now consider the clarinet, where the wavelength is *four* times the length of the instrument. Whereas a flute acts like an open pipe, or a tube that is open on each end, a clarinet acts like a **closed pipe**, or a tube that is open on one end and closed on the other.[viii] At the open end, the air is free to oscillate, forcing an *antinode* to be present there. At the closed end, the air is not free to oscillate and so, just like a fixed string, a *node* is forced there. The resulting standing wave is illustrated below.



[viii] Why it acts like this has to do with its shape and how each end is structured.

Notice that the standing wave is just half of an onion. Since the wavelength is twice as long as an onion, that means the wavelength is four times as long as a half an onion. Thus, for the clarinet, being a closed pipe, the wavelength is four times as long as the length of the clarinet. Since the clarinet is roughly the same length as the oboe and flute, that means the clarinet can play a note that has twice the wavelength and thus half the frequency (an octave lower).

✍ Brass instruments have characteristics of both open and closed pipes and thus are too complicated to be considered as just one or the other.

✓ *Check Point 22.6: What frequency is produced on a clarinet that is 66 cm long if it acts like a closed pipe? The speed of sound in air is 343 m/s (at 20° C).*

22.5 Normal modes

How do you create different notes with a wind instrument?

Just like with a string instrument, you can change the note by changing the length of the tube. For a flute and oboe, the length of the standing wave can be shortened by placing a hole along the pipe, thus forcing the "open end" to be at some other place, and forcing the antinode to be where the hole is rather than at the end of the instrument. By shortening the tube, the wavelength is smaller and thus a higher frequency (for the same wave speed) is produced. Each instrument has holes (and keys that be used to open or close holes) that essentially change the length of the standing wave that is produced inside the instrument. The same approach can be used with instruments like clarinets. Placing a hole along the clarinet shortens the length of the standing wave inside the instrument, and thus produces a higher frequency note.

Brass instruments can also change the length of the instrument but instead of holes they "extend" the tube. A trumpet, for example, has valves, with each valve allowing the instrumentalist to add some tubing to the instrument, thus lengthening the standing wave that is set up.

The problem with both of these methods is that there is a limit to how high or low you can change the frequency since there is a limit to how short (for flutes and oboes, for example) or how long (for brass instruments) you can make the tube. Shortening a pipe to half as long gives a note that is an octave higher (since the wavelength is half as long) and lengthening a pipe to twice as long gives a note that is an octave lower (since the wavelength is twice as long). A string instrument achieves the wider range of notes by utilizing multiple strings, each with a different wave speed. A wind instrument, on the other hand, cannot change the wave speed. Since the air itself is the medium that is vibrating, the wave speed for a wind instrument is just the speed of sound in air, and that doesn't change.

So how do wind instruments achieve such a wide range of notes?

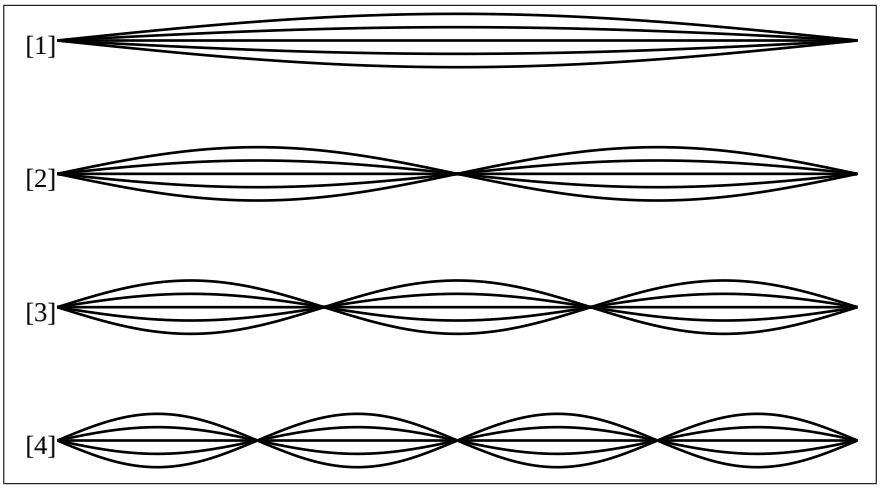
Many standing waves are typically set up (called normal modes) with the condition that they meet the boundary conditions.

To produce a larger range of notes, the wind instrument uses a characteristic of standing waves called normal modes. Basically, the standing waves we've described so far are just one of many possible standing waves that can be set up. *Any* standing wave can be set up *as long as* it meets the boundary conditions.

What does this mean?

For example, consider a string instrument, which has a string with two fixed ends. Since the ends are fixed there must be a node at each end. However, there are multiple standing waves that are possible that meet that boundary condition. Figure 22.5 shows four possibilities. Each possibility that can exist is called a **normal mode**[ix], and we refer to each normal mode by a number. The simplest normal mode is the first normal mode, which has only two nodes (and one antinode), one at each end. The second simplest, called the second normal mode, has an additional node in the center (with two antinodes).

[ix] The word "normal" here has more to do with how the word is used with forces, where the normal force is the force exerted by a hard surface, than how it is used statistics (like a normal distribution). With normal forces, "normal" means perpendicular. In statistics, "normal" means common. At first glance, you may be hard pressed to see the relationship between normal forces and a vibrating string, but every vibrating string can be thought of as consisting of some combination of the normal modes, just as any coordinate in space can be thought of as consisting of a set of perpendicular coordinate values.



Notice how each has a different onion size but the onions that make up each normal mode have the *same* width. This must be the case since any single wave must have a single wavelength value (and thus a single onion size value).

Since each normal mode has a different onion size, that means they each have a different wavelength and thus a different frequency. In fact, when a string is plucked or bowed, *all* of the possibilities are set up on the string and what we hear is the *combination* of all of those frequencies. Usually the first normal mode (indicated as #1 in the figure) is the strongest and thus the loudest, which is why our discussion earlier is mostly accurate.

Since each normal mode has a different wavelength (but same wave speed), each normal mode must also have a different frequency.

✍ Although many standing waves are set up, only certain onion sizes (and thus only certain wavelengths) work with the boundary conditions. Since only certain wavelengths respond, that means only certain frequencies respond. So, while many standing waves are possible, they correspond to only certain frequencies.

Notice how the second normal mode has an onion size that is half that of the first normal mode. That means it has a wavelength that is half as much, which corresponds to a frequency that is twice as much.[x] The third normal mode has an onion size that is one-third that of the first normal mode, and the fourth normal mode has an onion size that is one-fourth that of the first

[x]This assumes the wave speed is the same for all of the normal modes. Since all normal modes share the same string (for a string instrument) and same air (for a wind instrument), that is a pretty safe assumption though in real life there may be very small changes to the wave speed depending on the normal mode.

normal mode. Correspondingly, the third normal mode has a frequency that is three times that of the first normal mode, and the fourth normal mode has a frequency that is four times that of the first normal mode.[xi]

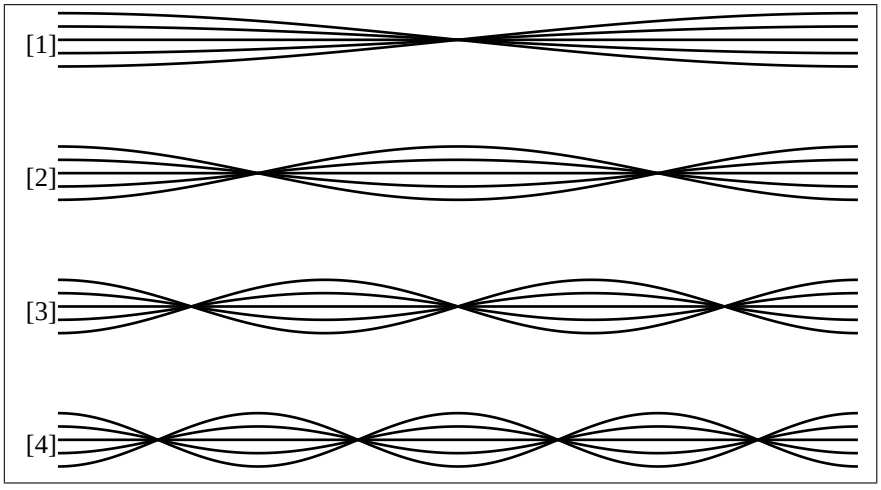
IF ALL OF THE NORMAL MODES RESPOND WHEN A VIOLIN STRING IS PLAYED, WHY DO WE HEAR A SINGLE NOTE RATHER THAN MANY FREQUENCIES AT ONCE?

Actually, you do hear many frequencies, but since the frequency associated with the first normal mode is *loudest*, that is the frequency you associate with the note being played.

✓ *Check Point 22.7: The "A" string on a violin is 66 cm long. The frequency of the first normal mode on that string is 440 Hz. What is the wavelength, wave speed, and frequency of the second normal mode?*

WHAT ABOUT AN OPEN PIPE WITH FREE ENDS?

As with a string instrument, *any* standing wave in an open pipe can be set up *as long as* it meets the boundary conditions, namely that the two ends are free and have antinodes. Figure 22.5 shows four possibilities (normal modes).



[xi] In music, the first normal mode is called the **fundamental**, with each succeeding mode called an **overtone**, such that the second normal mode is the first overtone, the third normal mode is the second overtone, and so on.

As with a string, when a flute or oboe is played, *all* of the possibilities are set up and what we hear is the *combination* of all of those frequencies. Again, usually the first normal mode (indicated as #1 in the figure) is the strongest and thus the loudest, which is why our discussion earlier is mostly accurate.

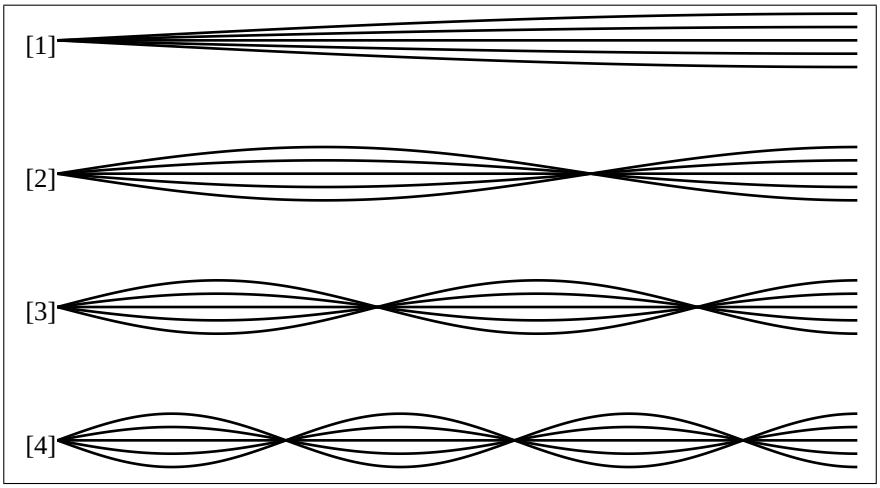
IF A FLUTE, OBOE AND STRING ALL PRODUCE THE SAME NORMAL MODES, WHY DO THEY SOUND DIFFERENT?

The relative strengths of the various normal modes depends on the instrument, and it is the particular combination of normal modes that give the sound a particular "flavor" or "substance" that distinguishes that sound from the sound played on another instrument. A single frequency, by itself, like that heard from a tuning fork, doesn't sound as rich as the same note played on a violin string, for example.

✓ *Check Point 22.8: Sketch the vibration pattern associated with the fifth normal mode of a standing wave free at both ends.*

WHAT ABOUT A CLOSED PIPE WITH ONE FREE END AND ONE FIXED END?

Again, *any* can be set up *as long as* it meets the boundary conditions. For a closed pipe, with one free end and one fixed end, that means one end must be an antinode and the other end must be a node. Figure 22.5 shows four possibilities (normal modes).



Unlike the open pipe, with a closed pipe the second normal mode has an onion size that is one-third that of the first normal mode. That means it has a wavelength that is one-third as much, which corresponds to a frequency that is three times greater. The third normal mode has an onion size that is one-fifth that of the first normal mode, and the fourth normal mode has an onion size that is one-seventh that of the first normal mode. Correspondingly, the third normal mode has a frequency that is five times that of the first normal mode, and the fourth normal mode has a frequency that is seven times that of the first normal mode. Notice how the closed pipe doesn't have the "even" multiples that the open pipe has.[xii]

✓ *Check Point 22.9: Suppose we have an closed pipe of length 63 cm. What is the wavelength of the fourth normal mode?*

22.6 Resonance

In the previous section it was pointed out that there are many different standing waves, called normal modes, that are set up on a string or in a pipe but usually the lowest normal mode is the loudest and so that is the one that we've been focusing on. However, by utilizing **resonance**, we can "select" the normal mode that we want to be loudest and thus change the note that is being played, without changing anything about the pipe or string.

To understand resonance, consider a playground swing. Suppose you are helping a young child to swing. To do this, you can stand behind the swing and push periodically, in time with the swing. Basically, the swing moves back and forth with a particular frequency, called its *natural* frequency, and, to keep the swing going, we push with the same frequency as the swing's natural frequency. Resonance occurs when the natural frequency of something is amplified by applying a force that matches that natural frequency.

Now let's consider a rope that is stretched between you and a wall, where you hold onto one end and the other end of the rope is attached to the wall. Like a string on a violin or guitar, the rope has several natural frequencies,

[xii] In music, integer multiples of the fundamental or first normal mode frequency are called **harmonics**. Standing waves with mixed ends only exhibit the odd harmonics.

each corresponding to a normal mode, and you can get the rope to oscillate at one of those frequencies if you oscillate your end of the rope at one of those natural frequencies. When your applied frequency (called the **driving frequency**) matches one of the natural frequencies of the rope, you'll get the rope to look like one of the figures illustrated earlier for a string fixed at both ends. When it doesn't match any of the normal mode frequencies, you won't get anything and the rope won't vibrate much at all.

❣ A standing wave is produced only when the driving frequency matches one of the normal mode frequencies.

✍ Instead of a rope, you can use water in a bathtub. Slosh the water back and forth with a frequency equal to one of its natural sloshing frequencies and you'll get the water to match one of the figures illustrated earlier for a pipe open at both ends. When your driving frequency doesn't match any of the natural sloshing frequencies, you don't get much of a response at all.

Like the rope, the string of a violin or guitar has several natural frequencies, each corresponding to a normal mode. When an instrumentalist bows a violin string or plucks a guitar string, they are essentially forcing all possible frequencies, not just one, but only some of those frequencies correspond to the frequencies associated with the normal modes. The other frequencies don't get any response at all. Consequently, we hear something soothing, not noise.

With a wind instrument, the forcing is done by blowing into the instrument, but the basic idea is the same. For example, a trumpet player buzzes their lips as they blow into the trumpet, and the instrument resonates with the frequencies associated with the normal modes. However, whereas plucking or bowing a string reinforces all frequencies and there isn't much[xiii] we can do about that, a trumpet player can change the way they buzz their lips when blowing into the trumpet. Buzzing their lips with a higher frequency will amplify a higher-frequency normal mode and thus create a higher note. Of course, not every possible note can be create this way – only those notes that correspond to the normal modes. However, with the addition of the valves, which can lengthen the tube, the trumpet player can be play a wider range

[xiii] One can modify the response on a string instrument by shifting *where* along the string you pluck or bow the string. In addition, by touching the string with one hand while plucking or bowing with the other, one can dampen out certain normal modes while allowing others to respond. This is done by touching the string at a location where the undesired normal modes have an antinode.

of notes by "selecting" which normal mode they want to amplify.[xiv]

✓ *Check Point 22.10: Consider a rope fixed at both ends, like a rope stretched between you and a wall, where you hold onto one end and the other end of the rope is attached to the wall. If the natural frequency associated with the rope's first normal mode is 2 Hz, would the rope respond (by oscillating) when you apply a driving frequency of 1 Hz (by rapidly move your hand back and forth)? What about 4 Hz? Why or why not?*

Summary

This chapter examined how standing waves can be set up when a wave reflects back upon itself.

The main points of this chapter are as follows:

- Standing waves can be set up with fixed ends, with the result being that there is a node at each end.
- Standing waves can be set up with free ends, with the result being that there is an antinode at each end.
- Standing waves can be set up with mixed ends, with the result being that there is a node at the fixed end and an antinode at the free end.
- Consistent with the wave equation, the frequency of a standing wave is proportional to the wave speed and inversely proportional to its wavelength.
- Many standing waves are typically set up (called normal modes) with the condition that they meet the boundary conditions.
- Since each normal mode has a different wavelength (but same wave speed), each normal mode must also have a different frequency.
- A standing wave is produced only when the driving frequency matches one of the normal mode frequencies.

By now you should be able to do the following:

- Describe a standing wave according to nodes and antinodes.

[xiv] A bugle is like a trumpet but without any valves. Consequently, a bugle can only play the notes that correspond to the normal modes, not any other notes.

- Explain how a standing wave is formed.
- For a standing wave with fixed ends (forced node at each end), one end free (forced node at one end point; forced antinode at the other), or both ends free (forced antinode at each end), sketch the vibration patterns of the first five normal modes.
- For a standing wave, given the length of the n-th normal mode, determine the wavelength, and visa-versa, and from that predict the frequency and/or wave speed.

Frequently asked questions

How come I don't see nodes and antinodes when a violin is played?

There are several reasons why you don't see the nodes when a string vibrates. One reason is that a string typically vibrates with a small amplitude. Consequently, it is hard to tell the difference between a node (where the string doesn't vibrate) and an antinode (where the string does). Another reason is that the fundamental (or first normal mode) is typically the dominant mode, and that normal mode does not have any nodes except for the two at the ends.

Terminology introduced

Antinodes	Node	Resonance
Closed pipe	Normal mode	Standing wave
Driving frequency	Open pipe	

Part F

Optics

23. Light as a Wave

Puzzle #23: In part E, we discussed various wave phenomena like sound waves, water waves and waves on a string. Here we will examine the wave-like properties of light. What evidence is there that light is a wave?

Introduction

This part of the textbook is entitled "optics," which refers to the behavior of light. In this chapter, we'll examine the wave properties of light.[i]

Since we are treating light as a wave, the language and mathematics of chapter 19 can still be used, and section 23.2 will examine how that language is used with light. We'll then examine how light, like other waves, can exhibit the Doppler effect and interference.

23.1 Do waves require a medium?

Before describing the wave characteristics of light, we need to address one significant difference between light and the other waves we've examined up to now: light, unlike other waves, doesn't require a medium to travel in. Indeed, it travels fastest[ii] in a **vacuum**, which is empty space, with no air, water or any other material.

How do we know that light can travel in a vacuum?

[i] Light exhibits characteristics of both waves and particles, as do electrons. However, we have been able to treat electrons as particles and, for the phenomena we'll be investigating, we can treat light as a wave.

[ii] Faster light doesn't make it more dangerous.

Outer space is essentially a vacuum. Yet, we see light from the sun, which must travel through millions of miles of emptiness to get to us.

Since waves are made of *coupled oscillators* (see section 19.2) or vibrating "things" such as strings, water, or air molecules that are connected in some way, it is reasonable to ask how a wave can exist where there is nothing. However, it was never stated that waves require a *physical* material that is oscillating. Rather, what is necessary is a "restoring force" of some kind that tries to bring the material (string, air, water, etc.) back to the "status quo" (or equilibrium state) that existed before the material was disturbed. This restoring force disturbs the neighboring areas which leads to a "domino" effect in which the disturbance (from equilibrium) travels from one place to another through the material. A wave is just a series of such traveling disturbances or pulses.

The difference with light is that what is oscillating is not a physical thing but rather the electric and magnetic fields. As we know from chapter 18, there appears to be a "status quo" law for electric and magnetic fields, where a changing magnetic field induces an electric field in such a way that nature tries to maintain the magnetic field that was present before the change.[iii]

This means that disturbances in the electric and magnetic fields (usually combined together in one term called the **electromagnetic field**) will travel – an **electromagnetic wave** – and that is what light is.

⁝ Light is an electromagnetic wave and, as such, can travel in a vacuum.

✓ *Check Point 23.1: (a) Can sound exist in a vacuum? Why or why not? (b) Can light exist in a vacuum? Why or why not?*

[iii] In a similar way, although we haven't shown it, a changing electric field induces a magnetic field. We can show this by considering a capacitor. As the capacitor charges or discharges, current must flow in the wires leading to and from the capacitor. Since a magnetic field is generated around a wire with current, one might ask if the magnetic field simply "stops" at the location of the charging/discharging capacitor. It turns out that the magnetic field continues across the capacitor as though there was no gap at all. Apparently, current by itself is not necessary to create the magnetic field, since there is no current flowing *between* the capacitor plates. As the capacitor charges, the electric field between the plates increases. It turns out that this changing electric field will produce a magnetic field just as if there was a current flowing between the two plates.

23.2 Electromagnetic spectrum

Since electromagnetic waves (of which light is one type) are waves in the electromagnetic field, they can be created by oscillating a charge. When the electromagnetic field is made to oscillate at a frequency between 4.0×10^{14} Hz and 7.9×10^{14} Hz, **visible light** is produced.

Don't we plug in our lamps to a 60-Hz AC outlet? Shouldn't the light have a frequency of 60 Hz?

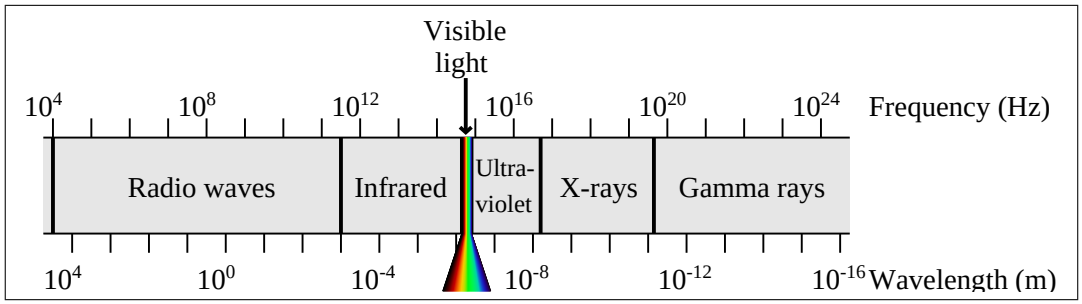
The frequency of the AC voltage just tells us how the current is changing through the light bulb. When we say that the light frequency is 4.0×10^{14} Hz, for example, we don't mean that the light is turning on and off at that frequency (as with an AC voltage). Rather, we mean that the electromagnetic field that makes up the light itself is oscillating at that frequency.

That is why visible light has a frequency that is very, very high compared to the frequency of the AC voltage that was applied (60 Hz), why we can light a light bulb using 0-Hz voltage (i.e., a battery), and why light isn't emitted from the wires in our circuit. The current in the wires are oscillating with frequencies similar to that produced by the signal generator, which is at most 10^6 Hz.

What happens when the electromagnetic field is made to oscillate at a frequency less than 4.0×10^{14} Hz or greater than 7.9×10^{14} Hz?

Electromagnetic waves will still be created at those frequencies – it is just that the eye cannot see them. Just as the ear can only sense certain frequencies of sound, the eye can only sense certain frequencies of electromagnetic oscillations. As shown below, the visible light portion (between 4.0×10^{14} Hz and 7.9×10^{14} Hz) is just a small portion of the entire electromagnetic spectrum. The hotter an incandescent light bulb filament gets hot, the higher the frequency of the electromagnetic waves generated. If the filament is too cool, it may only emit infrared light, which we can't see.

Visible light is just a small part of the electromagnetic spectrum.



WHERE DO SOUND WAVES FIT IN?

Sound waves are not electromagnetic in nature. After all, sound waves cannot exist in a vacuum whereas electromagnetic waves can.

THEN WHY ARE RADIO WAVES LISTED IN THE DIAGRAM?

Radio broadcasts convert sound waves into electromagnetic waves of a particular frequency (called **radio waves**). We can not "see" or "hear" radio waves. If we could, it would be pretty noisy and blinding with all of the radio signals being broadcast all over the place. A radio is a device that is able to interpret a particular frequency of electromagnetic waves, converting those waves back to sound waves so we can hear them.

Radio waves are not the only electromagnetic waves invisible to the eye. Other frequencies invisible to the eye include microwaves (like those used in microwave ovens, and with cell phones and satellite communication) and X-rays. The term "light" can actually be applied to all of these waves but usually different names are used depending upon the wavelength/frequency of the wave. For example, in the diagram, electromagnetic waves have been broken down into radio waves, **infrared** waves (i.e., "frequencies below red"), visible light, **ultraviolet** light (i.e., "frequencies beyond violet"), **X-rays** and **gamma rays**.

All are due to the oscillation of charges and the radiation of the electric field away from the charge. To avoid confusion, we use the term "electromagnetic waves" as the general term for all frequencies and "visible light" to describe that subset of frequencies that we can see. Usually, if "light" is used, we assume it to mean "visible light".

To detect the non-visible frequencies (like radio waves), we need special instruments. A radio is basically just an instrument designed to pick up a specific range of frequencies that are lower than those of visible light. Whereas

visible light has frequencies around 500 trillion hertz, FM radio waves are around 100 million hertz (100 MHz) and AM radio waves are around 1000 thousand hertz (1000 kHz). Those frequencies should be familiar to you, as they are the numbers shown on the radio dial.[iv]

✓ *Check Point 23.2: Identify each statement as true or false:*
(a) Visible light is a form of electromagnetic waves.
(b) Microwaves are a form of electromagnetic waves.
(c) Radio waves are a form of electromagnetic waves.

23.2.1 The speed of light

One special feature of light is that it travels very, very fast. Indeed, the time it takes between turning on a flashlight and the light beam hitting the wall is very small!

It turns out electromagnetic waves travel at about[v] 3×10^8 m/s in a vacuum, and visible light, being just one type of electromagnetic wave, likewise travels at a speed of about 3×10^8 m/s in a vacuum.

The speed of light is usually represented by a c in equations (e.g., $E = mc^2$). Consequently, in a vacuum, c has the following value:

$$c \approx 3 \times 10^8 \text{ m/s}$$

✓ *Check Point 23.3: (a) Which travels faster: the speed of light in a vacuum or the speed of sound in air?*
(b) Which travels faster: the speed of microwaves in a vacuum or the speed of visible light in a vacuum?

[iv] FM stands for frequency modulation, which means that the frequency of the sound is represented by modulations in the signal's frequency. For this reason, FM signals need a buffer of about 0.2 MHz on either side of the station frequency. This is why FM stations are assigned no closer than 0.2 MHz apart. AM stands for amplitude modulation. They don't need as much of a buffer and, as such, can be separated by only 10 kHz.

[v] See the supplemental readings for a more precise value.

Substance	Speed of light (10^8 m/s)
Vacuum	2.998
Air	2.997
Ice (0°C)	2.290
Water	2.249
Ethyl alcohol	2.201
Gasoline	2.148
Wesson™ Oil	2.034
Pyrex Glass	2.034
Glass, crown	1.968
Diamond	1.239

Table 23.1: The speed of light in a sample of materials (all at 20°C unless otherwise noted).

23.2.2 Speed of light in different media

Previously, it was mentioned that light doesn't need any physical material through which to travel. This doesn't mean that light *can't* travel through physical material. It can. For example, we all know that light can go through glass (e.g., a window or glasses) and certainly it can travel through air.

However, if something is present, the speed of the light will be slower. Indeed, it has been found that it travels 25% slower in water than it does in vacuum, and about 33% slower in glass than it does in a vacuum.[vi] That is still really, really fast, yet slower than what it is in a vacuum. The speed of light in sample materials are listed in Table 23.1. Notice how the speed is fastest for a vacuum and all other values are lower, but still pretty fast.

✓ *Check Point 23.4: If light is found to travel at about 73.4% the speed in a vacuum then, based on Table 23.1, in what material is the light traveling?*

[vi] Light doesn't slow down because of friction. After all, it speeds right back up after passing through a material and returning to a vacuum. The effect is more like how you can ride your bike from pavement onto sand and then back to pavement. Peddling at the same intensity, you'll slow down while on the sand then speed back up when you return to pavement. After all, even water and glass are mostly empty space – it is just the interaction of light with the material that makes the overall speed less.

Note that light travels slower in air than in a vacuum but the difference between the speed of light in a vacuum and the speed of light in air is very small. In fact, they differ by less than 0.03%. Thus, we typically use the speed of light in a vacuum even if we are dealing with light traveling through air.

How do we know light is slower in air?

Because light is so fast, it is very difficult to measure its speed. Even in a material like glass, the speed of light is still pretty fast (about 2×10^8 m/s in glass). Rather than measure the speed directly, we instead observe how the way light behaves when it encounters a different material. The process is described in chapter 24.

✓ *Check Point 23.5: (a) Describe an observation you can make (or have made) that suggests that light can travel through a vacuum.*
(b) Describe an observation you can make (or have made) that suggests that light can travel through glass?

23.2.3 Color

Since light acts like a wave and travels at a finite speed, we can use the same terminology and techniques to describe light that we use to describe other waves. For example, we can relate the frequency, wavelength, and speed of the light wave via equation (19.1), $v = \lambda f$, and we can use this equation to convert the frequency range for visible light (4.0×10^{14} Hz to 7.9×10^{14} Hz) to a wavelength range. For example, we can rewrite the equation as $\lambda = v/f$ (by dividing both sides by f) then, to get the wavelength range, we just divide the speed of light by the minimum and maximum frequencies. This produces a range of 7.5×10^{-7} m to 3.8×10^{-7} m (380 to 750 nm).

Our eyes interpret the higher frequencies (shorter wavelengths) as blue and the lower frequencies (longer wavelengths) as red.

What frequency is white light?

White light is what our eye "sees" when it receives frequencies from across the visible light range. There are three types of sensors in our eyes.[vii] Each

⁉ White light is made up of light of many different frequencies.

[vii] The three sensors, called cones, roughly correspond to red, green and blue. The blue

type is sensitive to a different range of frequencies. If all three types are stimulated, we perceive the light to be white.[viii] Otherwise, we see colors.

What happens if you view light in a different material?

The color doesn't change. And, it turns out, the frequency doesn't change either[ix]. However, the wavelength *does* change[x], which means that color depends on frequency, not wavelength.

How can the wavelength change but not the frequency?

We have three quantities in the wave equation, v, f and λ. Just because one changes (v in this case) does not mean that *both* of the others change as well.[xi] In this case, only the wavelength (λ) changes when light changes speed upon entering a new material, like from air to water.

✓ *Check Point 23.6: Use the wave equation to find the frequency for blue light of wavelength 440 nm in a vacuum.*

23.2.4 Transmission

Given the discussion so far, you might be wondering what happens to light when it passes through different materials, like with a window, where it travels through air then glass then air again. This is called **transmission**.

When light enters a new material, we'll assume its color (and thus its frequency) remains the same. However, the speed may not be. As mentioned in section 23.2.2, light travels at different speeds in different materials because of the way the light interacts with the material. In a vacuum (no air, water

---

cone seems to be most sensitive to frequencies on the violet side of blue and the red cone most sensitive to the yellow side of red, so there is a greater overlap between the red and green than green and blue, allowing us to better distinguish between red and green objects (with yellow and orange in between).

[viii] This is essentially how a color TV or computer monitor produces "white" light. It simply produces red, green and blue light at the same time.

[ix] The field can't oscillate at two different frequencies at the same location.

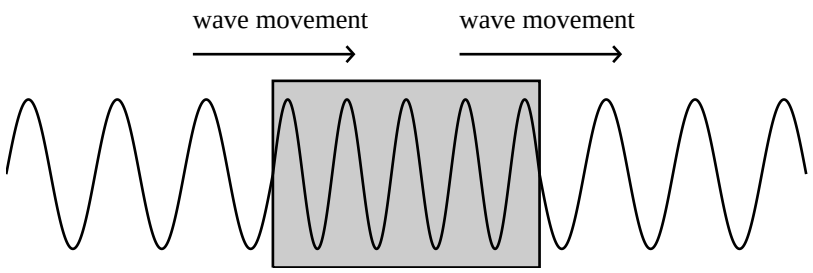
[x] In a similar way, cars slow down and bunch together (less spacing between them) when they encounter a construction zone with a lower speed limit.

[xi] In a similar way, just because force is equal to the product of mass and acceleration, changing the force on an object doesn't change its mass, just its acceleration.

or any other material), the speed of light is about 3×10^8 m/s (see section 23.2.1) and it is slower than that in all other materials.

It is important to recognize that the wave travels at a constant, albeit slower, speed in the material. In other words, it is *not* slowing down in the material (i.e., it is not getting slower and slower). And, once the wave leaves the material, it reverts back to the same speed it had prior to entering the material.

From the wave equation ($v = f\lambda$), if the speed v is different but the frequency f is not then the wavelength λ must be different to keep the equation balanced. A slower speed corresponds to a smaller wavelength (for the same frequency). We can therefore illustrate the change in speed by plotting the wavelength, as shown below. In the case illustrated in the figure, the wave is moving to the right. After entering the slower material (shaded in the figure), the wave is still moving to the right but it has a smaller wavelength. Upon re-entering the faster material, the wave continues to the right but the wavelength returns to the size it was prior to entering the material.

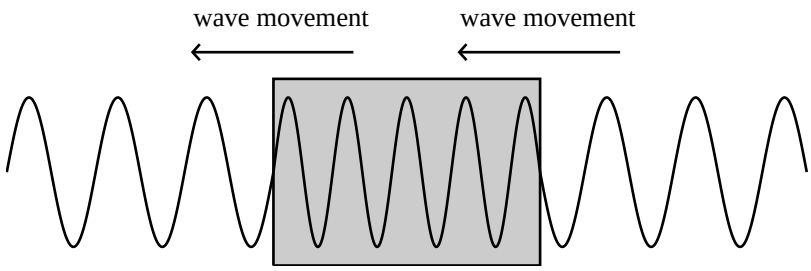


Notice how the wavelength shortens when going from a faster material to a slower material, and then lengthens when going from a slower material to a faster material. This is because light, like all waves, is not a physical object that requires an external force to speed up or slow down. In other words, light doesn't slow down because of friction. If it did, it would get slower and slower the further it got into the material. Instead, the speed has the new value immediately upon entering the new material and it maintains that new speed until it exits the material.[xii]

[xii] In a related note, the slowing isn't due to the material having a higher density or more "stuff" in the way. For example, water is more dense than vegetable oil, gasoline and ethyl alcohol, but the speed of light is faster in water than in any of those other liquids. In fact, the speed of light is faster in water than in Pyrex, a solid, in which the speed is similar to for vegetable oil.

What if the wave was moving from right to left instead of from left to right?

The illustration would look exactly the same, as in the following figure. In both cases, the wavelength shortens when going from a faster material to a slower material, and then lengthens when going from a slower material to a faster material.



The more significant the change in speed, the greater the difference in wavelength. For example, as shown in Table 23.1 in section 23.2.4, the speed of light in a vacuum is roughly the same as the speed of light in air so the wavelength of the light in a vacuum will be roughly the same as the wavelength of that light in air. The speed in Pyrex glass, however, is about two-thirds that in a vacuum, so the wavelength in Pyrex glass is about two-thirds that in a vacuum.

✓ *Check Point 23.7: Light in ethyl alcohol travels at a speed that is about 87% the speed in a vacuum. If light of wavelength 500 nm in a vacuum encounters ethyl alcohol, what would be the wavelength of that light in the ethyl alcohol?*

23.3 Doppler effect with light

Does light experience a Doppler Effect?

Yes, although we won't notice it unless the source or observer is moving very quickly. After all, the speed of electromagnetic waves is much, much greater than the speed of sound waves (300,000,000 m/s compared to 340 m/s). Consequently, to see the effect without special equipment, the source (or

observer) has to be moving very quickly. Examples with special equipment include Doppler radar, which meteorologists use to examine the potential for tornadoes and police use to determine if cars are speeding. Both involve the Doppler effect with electromagnetic waves and, as with sound, the observed frequency is higher when the source and observer are moving toward each other.

✓ *Check Point 23.8: When an object is moving toward you very quickly, should the color of the object be shifted toward longer wavelengths (lower frequencies, like red) or shorter wavelengths (higher frequencies, like blue)? Why?*

23.4 Interference with light

Can light exhibit constructive and destructive interference?

Yes. Light is a wave and, as discussed in chapter 21, two superimposed waves can exhibit constructive and destructive interference, meaning that the combination of the two waves can have an amplitude greater than or less than the amplitudes of each wave individually.

If light can exhibit destructive interference, how come two superimposed light beams are always brighter than each one individually?

This is because we need two sources with *exactly* the same frequency to observe interference. A standard incandescent bulb, for example, emits many different frequencies. If you had two such bulbs shining on a screen, some of the frequencies would interfere constructively while others interfere destructively, and the destructive interference would be overwhelmed by the huge number that doesn't experience destructive interference there. The end result would be that the screen would always be brighter with two bulbs than one by itself.

Compared to light waves, it is relatively simple to create two sources of sound or water waves that have the same frequency. For sound, all one needs to do is to set up two speakers that are run by the same controllable generator.

For water, all one needs to do is poke two sticks in the water at the same constant frequency.

How, then, can we observe interference with light?

⁉ A laser allows us to produce light of only a single frequency.

We can use the laser. The reason why laser light works for observing interference is because lasers produce a single frequency (or close to it).[xiii] For example, in the last checkpoint, it mentions that the light from a helium-neon **laser** has a wavelength of 632.8 nm.

Can we use two lasers?

No. Although two of the same type of lasers produce approximately the same wavelength, even a slight difference will cause a high-frequency beating.

Why?

As mentioned in section 21.4, two waves of frequencies f_1 and f_2 will interfere in such a way as to cause a wave that "beats" with frequency equal to the difference $f_2 - f_1$. In the case of light, we could have two helium-neon lasers, each providing red light (with frequency equal to 4.7×10^{14} Hz or so; see section 23.2). However, even if they differ by only 0.01%, that means the frequency difference is 4.7×10^{10} Hz, which produces a beating that is way too quick to be noticeable.

Then how can we observe interference with light?

We can create two "sources" of exactly the same frequency by shining the light from a single laser onto *two* slits. Each slit, then, will act as a source of light.[xiv] And, since the same laser light is passing through each slit, the slits will have the same frequency.

✓ *Check Point 23.9: Suppose two individual helium-neon lasers are aimed toward the same point. Why don't we get total destructive interference at the point and thus no light there? Try it out.*

[xiii] Not only is the laser light just one frequency but it also has a property called coherency, due to the atoms in the laser being excited in phase with the incident light upon them.

[xiv] Chapter 24 discusses why each slit acts like a source of light.

23.5 Polarization

In section 19.3, it was mentioned that for a wave on a string the string oscillates in a direction that is perpendicular to the direction that the pulses travel (along the string) whereas with sound the air oscillates in a direction that is parallel to the direction pulses travel. It turns out that light is like waves on a string, and even though we are unable to see the vibrations the make up the light waves, we can demonstrate that this is case because of its ability to be **polarized**.[xv]

What does it mean for a wave to be polarized?

To answer that, let's examine a wave on a string. If you could tag one piece of the string, you'd find that it is moving in a direction that is *perpendicular* to the motion of the wave. In other words, the string itself may be vibrating up and down, side to side, or both up and down and side to side. However, it isn't vibrating back and forth *along* the length of the string.

If the string is constrained to move in only one direction (e.g., up/down or side/side), then we say it is *polarized*. A string vibrating up/down is vertically polarized and a string vibrating side to side is horizontally polarized. A string that is vibrating sometimes up/down, other times side to side, and other times in other directions, is said to be *unpolarized*.

It isn't that the wave change directions but rather that the oscillation direction is in *all* directions rather than along a single direction. Since it is difficult to visualize something vibrating in more than one direction *at the same time*, my examples will consider situations where the oscillation is in a single direction at one moment and then a *different* direction at another moment.

A string can be polarized and the direction of polarization can be determined by sending it through a filter, such as a slot in a barrier, as shown in Figure 23.1a. If the filter is oriented the *same* way, the polarized wave will pass through *unaffected*. On the other hand, if the filter is oriented *perpendicular*

[xv] The meaning of the word "polarized" is similar to when it was used with electric dipoles, in that there is a specific orientation. With light polarization, we are referring to the orientation of the field oscillations that make up the light wave. With electric polarization, in comparison, we are referring to the orientation of the positive vs. negative sides of an object.

to the orientation of the wave, the wave energy is *absorbed* and the wave doesn't pass through (see Figure 23.1b).

The wave isn't destroyed if the filter is oriented at an *angle*. Rather, two things happen: the amplitude decreases (but not to zero) and the orientation of the polarization changes to the orientation of the filter (see Figure 23.1c).

One indication that light is similar to a wave on a string, then, is that filters[xvi] can be created that produce the same effect on light. These are called **polarizing filters**.

Usually[xvii] light is unpolarized. Consequently, if light is sent through a polarizing filter, it will become polarized. It will also become less intense, since those waves not parallel to the filter axis will be diminished. In general, when completely unpolarized light is sent through a polarizing filter, its intensity decreases by 50%.

✍ Consider a flashlight that is pointed straight up, as though you were trying to illuminate something on the ceiling. The electric field associated with that beam of light would be directed horizontally, perpendicular to the direction the light travels (upward). If the light from the flashlight was polarized, the electric field in the beam would be oscillating in just one direction, like north-south or east-west. However, unpolarized light is oscillating in all the horizontal directions (i.e., north-south *and* east-west at the same time).

How do we tell the difference between polarized light and unpolarized light?

Our eyes can't tell the difference. For example, light from the sun is unpolarized but light from LCD flat screens tends to be polarized (depending on the color). Light reflecting off of surfaces can be polarized also. However, the light in all those cases looks the same to us.

If they all look the same to us, how can we tell when light is polarized or not?

[xvi] An actual polarizing filter essentially has conducting bars that absorb the electric field that is parallel to it. Consequently, the string model described in the text is somewhat opposite the model used for electromagnetic waves.

[xvii] This is because light from, say, an incandescent light is produced by electrons oscillating in random directions and each electron produces a field in a particular direction, based upon the orientation of the oscillation.

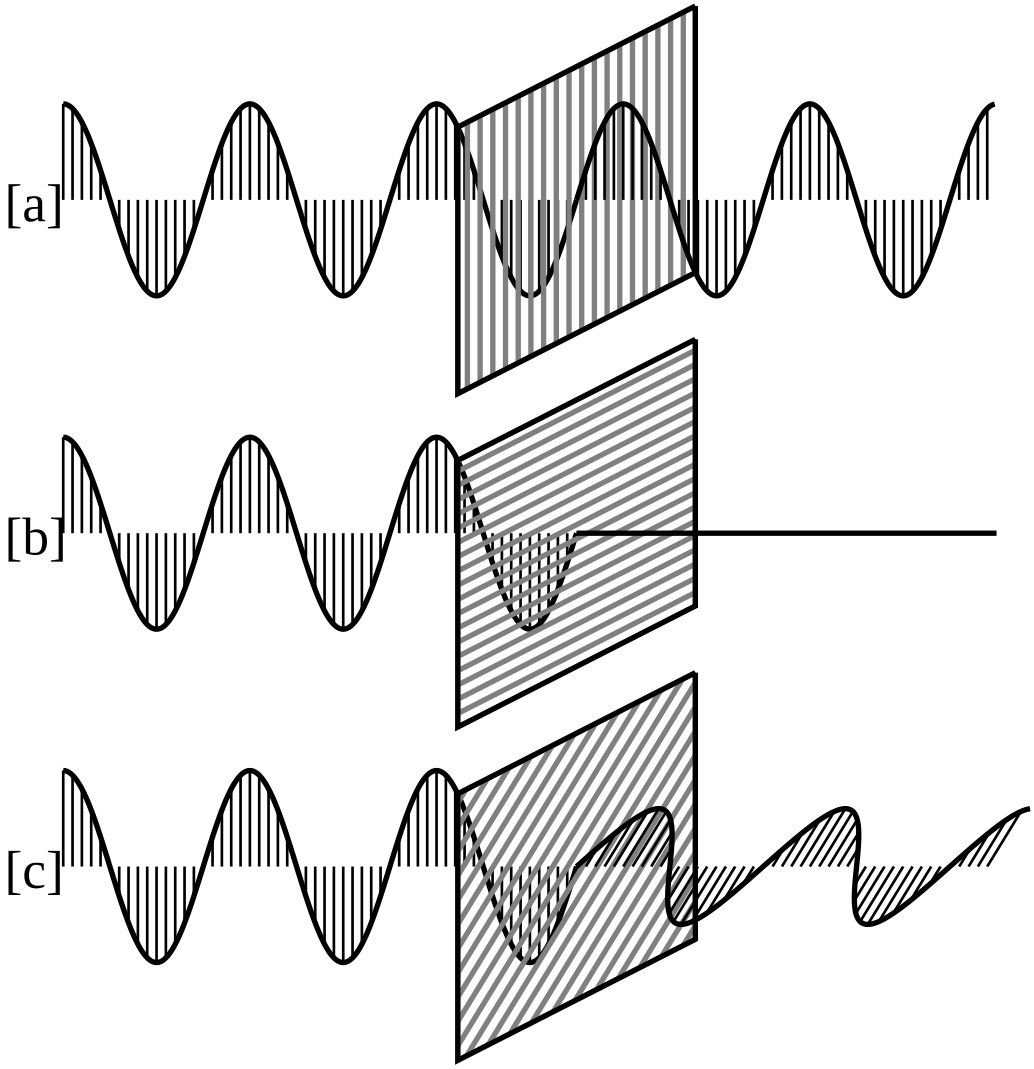


Figure 23.1: A vertically-oriented wave [a] passes through a vertically-oriented polarizer unchanged, [b] is *totally* absorbed by a horizontally-oriented polarizer (idealized), and [c] is *partially* absorbed by a polarizer oriented at an angle and also reoriented to match the orientation of the polarizer.

We can use a polarizing filter to reveal whether a particular light source is polarized or not.

When unpolarized light is sent through a single polarizing filter, the intensity decreases by 50% regardless of the orientation of the filter. In comparison, with polarized light, the resulting intensity (after passing through the filter) could be anywhere from 0% to 100% of the intensity prior to the filter, depending on the orientation of the filter.

To understand why, consider that there is a decrease in intensity only if some of the light is being absorbed by the filter. There will be no decrease if the light polarization direction and the filter polarization direction are oriented the same way. There will be 100% decrease if the light polarization direction and the filter polarization direction are oriented perpendicular to each other. Unpolarized light consists of both types of light: that which is oriented the same way as the filter and that which is oriented perpendicular to the filter. Consequently, half will always be absorbed (leading to 50% decrease in intensity), regardless of the filter's orientation.

Example 23.1: When polarized light is sent through a polarizing filter, the intensity is found to decrease by 50%. What does this imply about the orientation of the polarizing filter?

Answer 23.1: Since the intensity decreases, the orientation is not parallel to the light polarization. Since the intensity did not decrease to zero, the orientation is not perpendicular to the light polarization. Consequently, the orientation must be at some other angle.

Example 23.2: Suppose unpolarized light is sent through two perpendicular polarizing filters and the intensity is found to decrease to zero. What happens when a third filter is inserted between the two?

Answer 23.2: It depends on the orientation of the inserted filter. If it is parallel to the first or second filter, it has no effect and the intensity still decreases to zero. However, if it is inserted at an angle, then the orientation of the polarization is changed and so some light will get through the third filter. Sounds strange - you need to try it out to believe it.

Is there any value to having polarized vs. unpolarized light?

Both look the same to our eyes, so there is no value in that sense. However, we can take advantage of the fact that the impact of a polarizing filter on polarized light depends on the orientation of the filter. For example, light that reflects off water tends to be polarized. Consequently, wearing polarizing filters as glasses can, if oriented properly, remove the glare reflecting off the water while letting the wearer see the rest of the light (though diminished).

LCD screens also take advantage of this. The screen itself is essentially just a whole set of tiny polarizing filters (one for each pixel). Each filter can be rotated independently, controlling whether the polarized light (behind the screen at that pixel location) shines through that pixel area.

✓ *Check Point 23.10: Suppose unpolarized light is sent through a polarizing filter and we find that the intensity decreases by 50%. Do we know what will happen if the polarized light is then sent through a second polarizing filter? If so, what? If not, why not?*

Summary

This chapter examined the wave characteristics of light.

The main points of this chapter are as follows:

- Light is an electromagnetic wave and, as such, can travel in a vacuum.
- Visible light is just a small part of the electromagnetic spectrum.
- White light is made up of light of many different frequencies.
- A laser allows us to produce light of only a single frequency.

By now you should be able to do the following:

- Recall the speed of light in a vacuum.
- Identify light as an example of electromagnetic radiation or electromagnetic waves that, unlike the other waves mentioned so far, doesn't require a material through which it must travel.
- Recall that light travels at different speeds in different media, with light traveling fastest in a vacuum, and be able to explain why we typically for air use the speed of light in a vacuum.

- Describe the polarization of waves and be able to use the concept of polarization to identify the nature of light.

Frequently asked questions

What is electromagnetic radiation?

To scientists, **electromagnetic radiation** means the same as "electromagnetic waves." To the scientist, the term ""radiation is a generic term for light waves, and includes visible light and radio waves, as well as infrared radiation and ultraviolet radiation. The light coming from the sun actually includes these other frequencies as well as the frequencies in visible light. All travel at the same speed in a vacuum.

What may be confusing is that non-scientist typically consider "radiation" to be very high-frequency light such as X-rays and Gamma Rays, which can be dangerous.

Terminology introduced

Electromagnetic field	Laser	Ultraviolet
Electromagnetic radiation	Polarized	Vacuum
Electromagnetic wave	Polarizing filters	Visible light
Gamma rays	Radio waves	X-rays
Infrared	Transmission	

Abbreviations introduced

Quantity	**SI unit**
speed of light (c)	meter per second (m/s)

Additional problems

Problem 23.1: Suppose a call is made from East Stroudsburg, PA (41°N, 75°W), to Lima, Peru (12°S, 75°W) via a communications satellite in geosynchronous orbit directly above the equator at 75°W, about 3.6×10^7 m above

Earth.[xviii]
(a) How long would it take for an electromagnetic wave to travel this distance? Note that the electromagnetic wave must first travel up to the satellite and then back down.
(b) Assuming a typical speed of sound equal to 340 m/s, how long would it take for a sound wave to travel this distance?
(c) Which of these two represents the time delay between when one person says something and the other person hears it?

Problem 23.2: (a) What would the time delay be in problem 23.1 if the speed of light in air was used instead of the speed of light in a vacuum?
(b) What is the difference in time between that calculated in part (a) and that calculated in problem 23.1? Be careful not to round your answers. Otherwise, you'll get no difference. You might get better results just calculating what is 0.03% of the time.
(c) How far does light travel in a vacuum in the time given in part (b)?

[xviii]To understand where this distance comes from, consider that a satellite in geosynchronous orbit orbits Earth once a day. This is possible because its centripetal acceleration ($4\pi R/T^2$) at that height is equal to Earth's gravitational field at that height (GM_e/r^2). Since $v = 2\pi r/T$, we can solve for r, from which we can get the distance the signal must travel from East Stroudsburg to Lima.

24. Bending of Light

Puzzle #24: What happens to a wave when it encounters a different material, like the glass of a window? Is it affected at all?

Introduction

In this chapter, we'll examine ways to bend light. Everything we discuss can also be applied to other types of waves but we'll focus on light because it is easier to see the bending that occurs.

There are three phenomena that we'll examine: diffraction, reflection and refraction. We've already mentioned reflection in chapter 22, since standing waves are set up when a wave bounces off a boundary and back upon itself, which is a reflection. However, we'll now examine it in more detail.

24.1 Diffraction

Suppose you are talking to someone who is about five feet in front of you. If you hold up a book in front of you, that will block you from *seeing* the other person but you will still be able to *hear* them because sound can bend around the book. That bending is called **diffraction**.

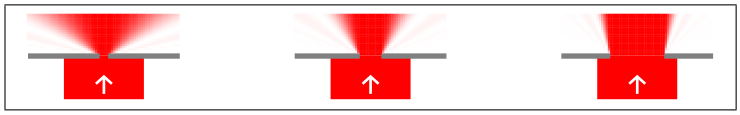
All waves exhibit diffraction, in that they can bend around corners and go around barriers and obstacles. And if light is a wave, it should also exhibit diffraction. This presents an apparent problem, though. We are used to light traveling in straight lines. For example, if you direct the light of a laser pointer or a flashlight at something, you expect the light to travel in a straight line from the laser pointer or flashlight to where you are pointing it.

So if light is a wave, why doesn't it bend around obstacles?

It does bend, but the bending is very, very slight because its wavelength is so small. For a wave to exhibit the "bending" characteristic, its wavelength must be about the size of the obstacle or larger. This means that for a 20-cm-wide book, the wavelength needs to be at least 20 cm or so. The human voice generates sound waves of wavelength about 13 cm to 150 cm, which allows for the bending.[i] Visible light, on the other hand, has wavelengths between 380 to 740 nanometers (see section 23.2.3), much smaller than 20 cm (a nanometer is a billionth of a meter), which is why visible light doesn't noticeably bend around the book.

The same relationship holds for openings as well as obstacles,[ii] namely that the amount of bending depends on how big the wavelength is compared to the obstacle or opening size, with more bending the smaller the opening (compared to the wavelength) and less bending the larger the opening. This is why you can hear people on the other side of a 20-cm-wide opening in a wall even if you can't see them.

The figure below illustrates how the bending depends on the size of the opening. Three openings are shown and a beam of light (or other wave, moving up the page as indicated by the arrow) passes through the opening. The wavelength isn't shown in the figure but happens to be the same in all three illustrations and equal to the width of the opening in the *middle* illustration (if this was indeed an illustration with a light beam then the width of the opening would be too small to see). The wave "spread" is wider when it passes through the narrower opening (left) than the wider opening (right).



[i] This assumes a frequency range of 230 Hz to 2600 Hz, and a speed of sound equal to 340 m/s.

[ii] The bending can be explained by first recognizing that each point within a wave acts as its own point source, in the sense that each point in space acts to push the neighboring points "out of balance" so to speak. A wave front is really just the result of constructive interference from all those point sources within the wave. If some of the points are missing, because they are blocked by an obstacle, then the waves spreading out from the remaining points produce the bending around the obstacle. The same is true for an opening, where the points within the opening produce a wave that spreads as it passes through the opening.

If we went to the extreme and made the opening really large then there would be no spread at all. On the other extreme, if we made the opening really tiny then the spread would be similar to what we'd get with a "point source," where the wave travels away from the point in all directions equally (see, for example, Figure 19.8b on page 357).

Note that the initial wave is shown as a beam, like a beam of light, and the wall blocks the part of the beam that isn't able to pass through the opening. That means that, in reality, the intensity of the wave (after passing through the opening) would be significantly less for the case on the left than for the cast on the right. However, in the figure the same intensity is used for both just so you can more clearly see the spread.

So is it possible to get light to exhibit diffraction?

⁉ Light can bend around corners but the opening has to be very tiny since the wavelength of light is so small.

Yes, but to observe such diffraction of light, we must either use a wavelength that is much larger (say, radio waves) than the opening (or obstacle) or use an opening (or obstacle) that is much smaller. To see the effect with visible light, which has a very tiny wavelength (380 to 740 nm), we need a very narrow slit or a very small obstacle. An example of a small obstacle is something like DNA. In fact, scientists used the diffraction of X-rays by DNA to determine the structure of DNA.

We can observe the diffraction if we send the light through a very narrow opening, since the smaller the opening the greater the diffraction. However, the smaller the opening the amount of light that gets through, which can make it harder to see. Consider, for example, a narrow opening made by painting one side of a piece of glass and then using a razor blade to make a thin scratch in it. A razor blade is about 0.1 mm thick, so the scratch made by the razor blade would be an opening about 0.1 mm wide. If you shined light from a laser onto the slit, a tiny amount of light would pass through. You could probably still see the light that passes through the slit but you'd have to turn off all the other lights in the room. In addition, an opening 0.1 mm wide would still be *two hundred times* wider than the wavelength of light, so the diffraction would be very slight. You'd have to observe the beam width several meters *after* passing through the slit to see the difference.

Figure 24.1 tries to illustrate this by considering what the beam width would be on a wall five meters after passing through the slit. In the bottom portion of Figure 24.1, I show a 1-mm wide slit (shown actual size). Above the slit is an illustration of how wide the beam would be (after passing through the slit)

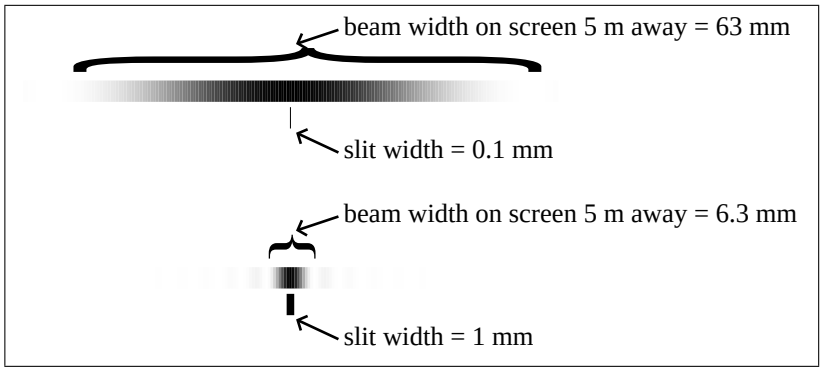


Figure 24.1: An illustration of the pattern observed on a wall after laser light is sent through a narrow 0.1-mm wide slit (top pattern) and wider 1.0-mm wide slit (bottom pattern). The actual size of the slit is shown directly below the pattern observed on the wall. Both assume light of wavelength 632.8 nm (red) and a wall 5 m from the slit.

on a screen five meters from the slit (here the shaded areas represent areas where the light beam is). Keep in mind that the beam is only as wide as the slit (1 mm in this case) at the moment it passes through the slit. However, due to diffraction, the beam "spreads" as it "bends" around the obstacle. The spread is not much but, five meters away, even a slight spreading results in a significant widening.

✍ The slit has a height and a width. The width is the narrow portion. In the figure, the slit height is 4 mm. The slit width, on the other hand, is 0.1 mm in the top case and 1 mm in the bottom case. There will also be spreading vertically but much less than the spread horizontally because the slit height is so much greater than the slit width.

How bright would the band be on the screen?

It wouldn't be very bright. After all, not much light can get through such a narrow slit. And then, after spreading apart due to diffraction, the beam would be even dimmer. However, if you are in a darkened room, you should easily be able to make out the band on the screen.

What happens if the slit is made narrower?

If the slit is narrower, the beam spreads even more. This is illustrated in the top portion of Figure 24.1 with a 0.1-mm wide slit. The actual beam would

be even dimmer since even less light would be able to get through a 0.1-mm wide slit.

IF LESS LIGHT GETS THROUGH THE 0.1-MM SLIT, WHY IS THE SHADING JUST AS DARK?

I created the figure as though the same amount of light gets through both slits but, in reality, much more light will get through the wider slit and so the band produced by that slit would be brighter.[iii]

⁝ The narrower the opening, the greater the spread of light that occurs.

The important thing is that, as illustrated in Figure 24.1, the narrower the slit, the wider the spread. This is why the spread isn't normally noticed. For normal openings, the width of the opening is so large compared to the wavelength that the spread is not noticeable.

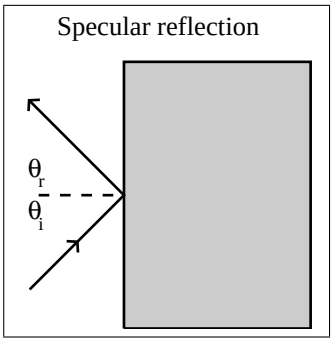
✍ It probably isn't noticeable in Figure 24.1 but, in addition to the spreading, there can also appear "bands" of light and dark areas outside the central band. These bands, called **fringes**, are due to interference effects, which are explained in section 23.4.

✓ *Check Point 24.1: A helium-neon laser emits light with a wavelength of 632.8 nm. In which of the following situations would you more likely find a greater spread of the beam after passing through the slit? Explain.*
(a) When the beam is incident upon a single slit of width 0.04 mm.
(b) When the beam is incident upon a single slit of width 4.5 mm.

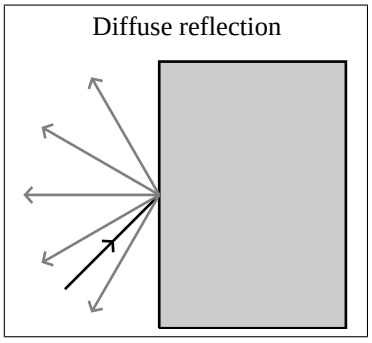
24.2 Reflection

Another way to get light to bend is to reflect it off the surface of an object, like a mirror. In general, there are two types of reflections: **specular reflection** and **diffuse reflection**. We'll focus on specular reflection in this chapter but it is important that you understand both.

[iii] If I made the top portion with one-tenth the ink of the bottom portion (to represent one-tenth of the light getting through the slit), you'd hardly be able to see it on the printed version of this textbook.



With specular reflection, the light only goes in one direction upon bouncing off the surface. This is illustrated in the figure to the right. The arrow represents a beam of light, like from a laser, initially directed upward and rightward then directed upward and leftward after bending at the surface of the object (shaded region).[iv] This is what happens when light bounces off an extremely smooth surface like a mirror.[v]



In comparison, diffuse reflection refers to the type of reflection that occurs when the surface isn't smooth. With diffuse reflection the light spreads out in all directions (see figure), not just a single direction.

Most surfaces are not smooth, even if they appear flat. For example, regular writing paper may be flat but the tiny bumps and crevices in the paper will result in diffuse reflection, not specular reflection.

One can distinguish between the two types by shining a laser on a wall vs. a mirror. When the laser beam is incident upon a wall (diffuse reflection), one sees a dot of light at the location where the light hits the wall, and it doesn't matter where you happen to be observing the wall – all observers see the dot (as long as they have an unobstructed view of the wall). In comparison, when one shines the laser onto a mirror (specular reflection), most of the people in the room are *not* be able to see the dot.[vi]

To understand why this is, one needs to first recognize that our eye only sees the dot if the reflected light is directed toward our eye. With diffuse reflection

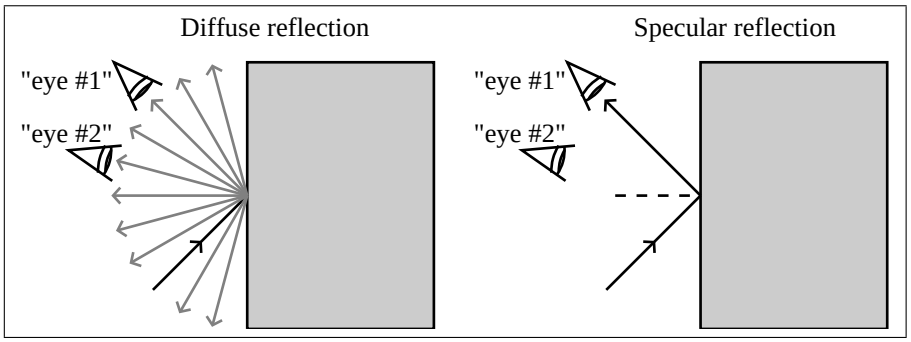
[iv] This approach is similar to what we used in section 21.6, when we examined interference in two dimensions. However, rather than drawing the wave as a squiggly line (which illustrates the oscillations within the wave), here we are just drawing the *direction* that the wave (light in this case) is traveling.

[v] In fact, it is called specular reflection because the word "specular" means "mirror-like" (from the same root as spectator and speculate).

[vi] This is true only for a perfect mirror. For a real mirror, there will be a little diffuse reflection at the mirror, which allows everyone to see a dim dot of light where the light hits the mirror.

off the wall, the light reflects at all angles, meaning that some of the light is directed toward your eye no matter where you happen to be (assuming you are looking at that point on the wall).

This is illustrated in the left drawing below, illustrating diffuse reflection (off the wall). Two eyes are drawn, indicating the location of two people who are observing the wall. Both the top person (eye #1) and the bottom person (eye #2) receive reflected light off the wall and thus both see the dot.



Compare that with the drawing on the right, illustrating specular reflection (off the mirror). With specular reflection off the mirror, the light only reflects toward one particular direction. Consequently, only the top person (eye #1) receives the reflected light off the mirror and sees the dot.[vii]

✍ Unless the object is a perfect mirror, not all of the light reflects off the object. Some light is "absorbed" into the material (possibly making the object warmer). For example, when we shine light onto a black material, most of the light is absorbed[viii] , whereas when we shine light onto a mirror, most of the light is reflected. To simplify things, we'll examine situations where there is little or no **absorption**.

[vii] This is not recommended. For someone at position #1 looking into the mirror it would be equivalent to looking straight into the laser, as that person would be receiving "all" of the light from the laser. In comparison, if the laser was pointed toward a wall, the diffuse reflection would send the light in all directions, and someone looking at the dot would only be receiving a portion of the light, which would not be as bright as looking straight into the laser.

[viii] A black object is black because the object absorbs the light, not because of interference.

✓ *Check Point 24.2: Suppose some takes a flashlight and shines it against the wall at an angle, as in the illustrations shown above. In which case, specular or diffuse reflection, would the person holding the flashlight be able to see the light being reflecting off the wall?*

24.3 Law of reflection

Most reflections are diffuse in nature, and light from some source (like a light bulb or the sun) reflects in a diffuse manner off the object. Diffuse reflection is how we see things and each other, since light reflects off of us (as opposed to us emitting light), and people around us can then detect that reflected light with their eyes.

However, in some cases we want to make the light travel in a particular direction and in those cases we need to use specular reflection because there will be a single reflection angle. To determine that reflection angle, we'll use the **law of reflection**. The law of reflection states that the reflected angle θ_r is equal to the incident angle θ_i. Mathematically, this is written as follows:

For specular reflection, the incident angle equals the reflected angle.

$$\theta_i = \theta_r \tag{24.1}$$

Since reflection occurs at the *surface* of the object, we'll use the convention where those angles are measured at the object's surface (where the bending of light occurs) with zero degrees being the direction *perpendicular* to the surface. This is illustrated in the specular reflection figure from before, with the dashed line being the perpendicular direction and θ_r and θ_i marked in the figure. One can see that the reflection angle (θ_r) is equal to the initial angle (θ_i).

The relationship is neither surprising nor complicated if you think about it. As such, it is usually pretty easy to apply. It can get a little tricky, though, when we have multiple beams of light reflecting off the surface. Each beam may hit the surface at a different angle but, for each beam of light, its reflected angle equals its incident angle. This is illustrated in the Figure 24.2.

In part (a), four arrows are drawn from the source, which is indicated by the small circle labeled "O". Each arrow represents a beam of light or **light ray**.

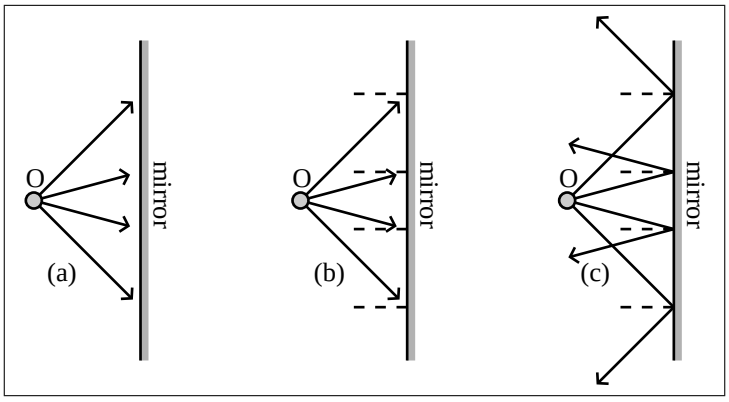


Figure 24.2

These rays could be four of the many rays emitted from a bulb, for example. The process for predicting where each ray will go after hitting the mirror is illustrated in parts (b) and (c).

In part (b), a dashed line is drawn at the point where each ray hits the mirror. That dashed line is perpendicular to the mirror *at the point where the particular ray hits the mirror*. The reason for drawing this line is because it serves as a reference for applying the law of reflection. Recall that the law of reflection states that the reflected angle equals the incident angle. These angles are measured relative to the direction perpendicular to the mirror. By drawing the perpendicular direction at each point on the mirror, it is easier to identify what the angles are.

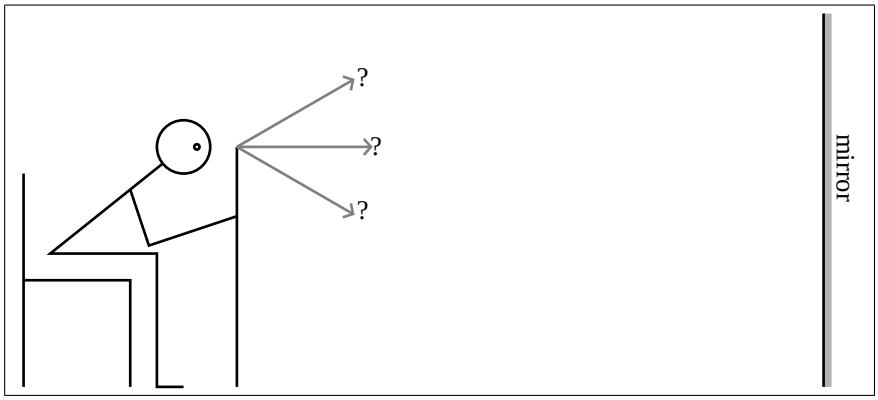
In part (c), then, each reflected ray is drawn on the same side of the mirror as the incident rays but on opposite sides of the dashed reference line such that the angle of the reflected ray is equal to, but opposite, the incident angle for that particular ray. In this way, you can predict where light will go when it hits a mirror.

✍ It is common to refer to the perpendicular direction as the **normal** direction, because the word "normal" means "perpendicular" in the mathematical sense.[ix] Thus, each dashed line corresponds to the normal direction (perpendicular to surface) at the point where the light ray hits the mirror.

We can also use the law of reflection to work backwards. For example,

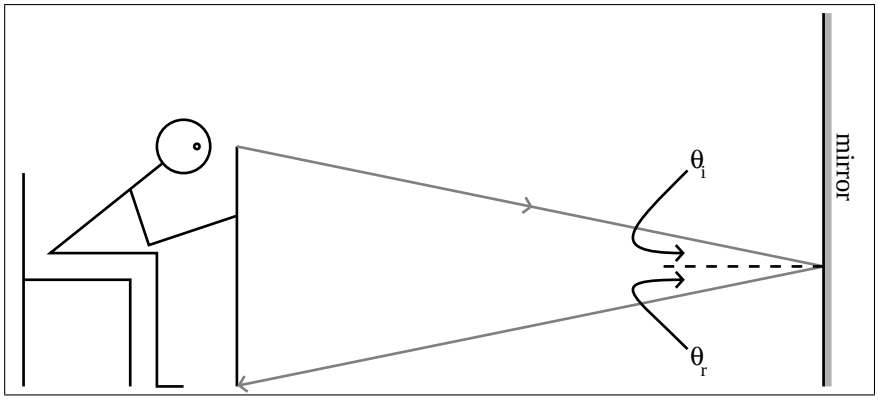
[ix] This usage of the term is the same as how one uses it when referring to the surface repulsion force (see Volume I) as the normal force, and is similar to how we used it when referring to the normal modes of standing waves.

consider the situation illustrated in the figure below, where a person is seated and holds a meter stick standing on end (so that the top of the meter stick is one meter above the floor). About two and half meters away, a mirror is mounted on the wall.



Suppose a laser is placed at the top of the meter stick and aimed at the mirror. Where does the laser have to hit the mirror to make the reflected light hit the bottom of the meter stick?

To answer this, we use the geometry of the law of reflection. We know that the incident angle must equal the reflected angle. Since it is symmetric (equal on both sides), it must hit the mirror half a meter above the floor (half the height of the meter stick). This is illustrated below.



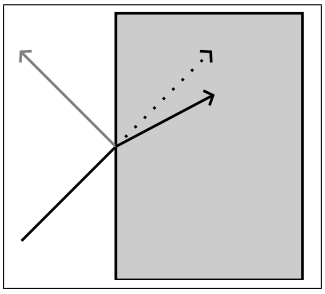
✍ Notice that in the example it doesn't matter how close the meter stick is to the mirror. Basically, no matter how far it is from the mirror, the laser must hit the mirror half a meter above the floor.[x]

✓ *Check Point 24.3: Consider the situation described earlier, with a laser placed at the top of a meter stick and aimed at a mirror. For each of the following questions, support your answer with a picture of the light ray.*
(a) How high above the floor does the laser light have to hit the mirror for the reflected light to hit the top of the meter stick (i.e., eye level)?
(b) How high above the floor does the laser light have to hit the middle of the meter stick (i.e., 50 cm above the ground)?

24.4 Refraction

In chapter 23, it mentioned that light can travel at different speeds in different materials. Unfortunately, we can't see it slowing down – it is simply too fast. Given that, you might be wondering how we know the speed is actually different. In turns out that we know this because of how light bends when it enters a faster or slower material *at an angle*.

For example, the figure to the right illustrates what one would observe when a beam of light (like from a laser) is sent into water or a glass block (indicated by the shaded region). The solid black arrow in the figure indicates the direction the light travels within the material. For reference, the gray arrow indicates the reflected light beam.



✍ A laser is nice because it produces a narrow beam, so it is easy to see the path the light is taking.

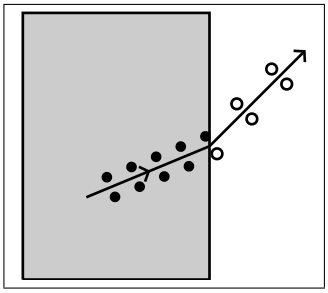
Notice how the light experiences an abrupt change in direction at the interface (compare to the dotted arrow, which indicates the direction the light would travel if the shaded material wasn't there). The change in direction upon entering a new material is called **refraction**.

⁉ Unless the wave is directed perpendicular to a boundary, it will change its direction upon entering the new material.

[x]For the same reason, if you are looking into a mirror, you need a mirror half your height to see your entire body.

✓ *Check Point 24.4: In your own words, describe refraction.*

Why does the light bend?



To explain why the light bends, consider what we'd see if we were looking down on two people marching shoulder to shoulder. The two people are illustrated by dots in the figure. Each pair of dots represents the two people, one on each side of the arrow. The pair starts off on the lower left side of the figure, marching toward the boundary between the shaded area and the non-shaded area.

The spacing between each pair of dots is smaller in the shaded region (solid dots) than in the non-shaded region (open circles). This is because the pair is moving faster in the non-shaded region. As long as both people march at the same speed, though, they remain together and march in a straight line.

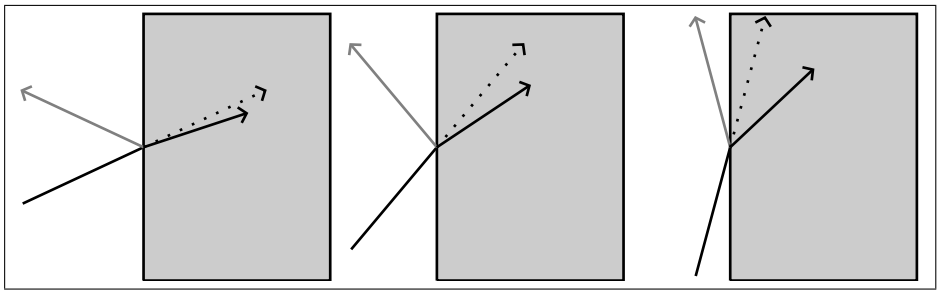
However, consider what happens when they encounter the boundary between the two materials. Notice how the pair now consists of a solid dot and an open circle. The person on the right (open circle) is in the "new" material while the person on the left (solid dot) is still in the "old" material. Consequently, for a short period of time, the person on the right is moving faster than the person on the left. That "rotates" their direction around the slower person (on the left). By the time the person on the left has entered the faster material, the orientation of their line has changed.

The direction of the bend is the same as what happens with light (or any other wave) when it encounters a material in which it travels faster. The direction of the bend goes the reverse way when the wave encounters a slower material.

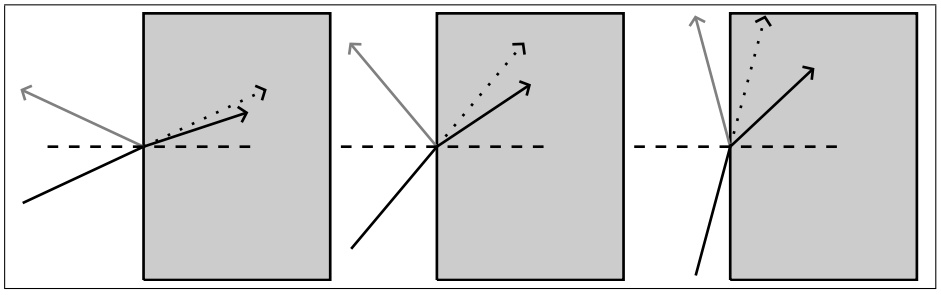
✓ *Check Point 24.5: In the analogy with the marchers, which marcher is traveling further during the transition when one marcher is in the faster material and the other marcher is in the slower material? Is this consistent with the way the marchers turning during that time period? Explain.*

Based on the analogy with the marchers, you might expect that how far the beam refracts (or bends), and which way it refracts, should depend on the

angle at which the wave encounters the interface and speed of the wave in the two materials. That is exactly what happens. Experimenting with a laser and a block of glass, one can easily observe that the amount of bending depends on the angle at which the light encounters the boundary. This is illustrated below, where the block of glass is shaded and the light ray comes in from the left.



Notice that the further off of head-on (perpendicular to the boundary), the greater the bending that occurs (the right panel shows more bending than the left panel). Remember to focus on the solid black arrow (the transmitted ray) rather than the solid gray arrow (the reflected ray). Also notice that, no matter what angle the incident light is relative to the boundary, the transmitted ray, like the incident ray, is directed upward and rightward. This may be a little easier to see in the figure below, which is the same as the previous figure but with dashed lines indicating the direction perpendicular to the boundary at the point where each ray is incident upon the surface. In each case, the incident ray and the transmitted ray are on opposite sides of the boundary (one ray in one material and the other ray in the other material) and on opposite sides of the dashed line. Another way of looking at it is that the solid and dotted arrows are in the same quadrant, bounded by the material surface and the perpendicular to that surface (dashed line).



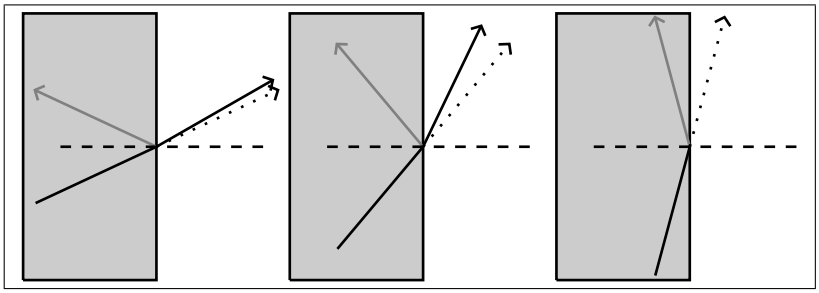
✓ *Check Point 24.6: Suppose light is sent into a material and is observed to change directions slightly when it enters the new material. Does the amount of bending depend on the angle at which the light encounters the boundary?*

Which way light bends depends on whether the speed is greater or less in the new material.

Which way does the light bend?

In each case illustrated so far, the bending is such that the transmitted ray is *closer* to the dashed line than the unchanged direction. However, it is only *closer* to the dashed line when the wave is traveling from a faster material into a *slower* material (like from air into glass). If the wave is traveling from a slower material into a *faster* material (like from glass into air) then the ray will be bend in the other direction – the transmitted ray will be *farther* from the perpendicular direction than the unchanged direction is.

For example, if you take the previous illustrations and leave them all the same except put the arrow head on the opposite end of the light ray, you'd have what happens when the light travels from a slower material into a faster material. This "reverse" bending is illustrated the figures below, where the light rays are now starting inside the slower material (shaded area).[xi] Notice that the transmitted direction (solid arrow) is always further from the perpendicular direction (dashed line) than the unchanged direction (dotted arrow), and after bending the light ray continues to move upward and rightward, as it did before it entered the material.



Why isn't there any transmitted ray (solid black arrow) in the rightmost illustration?

[xi] The speed of light is less in water than in air (see section 23.2.4) so the figure illustrates what happens to light when it enters air from water.

Table 24.1: A comparison of the transmitted angles for various incident angles to the air/water interface.

Refraction of light			
from air into water		from water into air	
θ_i	θ_t	θ_i	θ_t
0°	0°	0°	0°
10°	7.5°	10°	13.4°
20°	14.9°	20°	27.1°
30°	22.0°	30°	41.8°
40°	28.8°	40°	59.0°
50°	35.1°	50°	none
60°	40.5°	60°	none
70°	44.8°	70°	none
80°	47.6°	80°	none

When light (or any wave) enters a *faster* material, there becomes a point when there isn't enough "room", so to speak, for the transmitted ray. The result is no transmitted ray at all (there is still a reflected ray, though).

To illustrate the problem, let's suppose the two materials are air and water. As we know, the light will bend upon entering the air or water. The actual bending for each incident angle is provided in Table 24.4 (where the incident and transmitted angles, θ_i and θ_t, are measured from the direction perpendicular to the air/water boundary).

Consistent with what was discussed earlier, when the light is incident at an angle there is a bending, as can be seen by the incident angle not equaling the transmitted angle, which greater bending at greater incident angles. However, for light going from water into air, above a certain incident angle, the bending is so great that there is no way for the light to bend that much and still be transmitted into the material.

What happens to the light then?

There is no transmitted light at all. When there is no transmission, then 100% is reflected. In fact, we call this **total internal reflection**. We add the word "internal" because it only happens when the light (or wave) is moving from the slower material to the faster material (i.e., from within the slower material).

⁉ Total internal reflection occurs when the light has a large incident angle and is hitting the boundary from within a slower material.

✍ If you ever swam underwater (you can also see this if you visit an aquarium), you may notice that at certain angles the surface looks like a mirror. This is why. Total internal reflection is also what keeps the light inside an **optical fiber**.

At what angle does this happen?

As you can see in Table 24.4, total internal reflection for the air/water interface starts at an incident angle between 40° and 50° (for light moving from water into air). The actual angle is 48.6°. This value would be the **critical angle** for light moving from water into air. The actual value of the critical angle will depend on the speed of light in the two materials.

✍ Mirages are a consequence of the same effect, as the speed of light is slight faster in hot air than in cold air. Consequently, light coming from the sky can appear to reflect off the ground when the ground is very hot. In actuality, it is experiencing total internal reflection off the hot air near the ground.

✓ *Check Point 24.7: Is it possible to get total internal reflection when light goes from a material in which it is faster into a material in which it is slower? Why or why not?*

Why is the transmitted angle zero when the incident angle is zero?

This is because light doesn't bend when the incident angle is zero (light directed perpendicular to boundary), which corresponds to the ray being oriented perpendicular to the boundary (head-on).

✓ *Check Point 24.8: Suppose light is sent into a material and is observed to change directions slightly when it enters the new material. Does the light change direction when incident upon the interface in a direction perpendicular to the interface?*

What happens if the speed is the same in both materials?

If the speed is the same in both materials then there is no bending – the transmitted direction will be the same as the unchanged direction – regardless of the incident angle.

✍ As indicated in Table 23.1 in section 23.2.4, the speed of light in Wesson™ Oil is the same as the speed of light in Pyrex Glass. It is for this reason that immersion oils are used with microscope slides, so that the light doesn't bend as it passes into or out of any air pockets between the microscope slides.

✓ *Check Point 24.9: If the light is traveling from a slower material to a faster material, is it possible to have the light bend so much that the incident and transmitted rays are on the same side of the surface boundary and/or the same side of the perpendicular to the surface boundary (at the point where the ray is incident)? What about if the light is traveling from a faster material to a slower material?*

24.5 Index of refraction

It is common to describe materials in terms of their **index of refraction**, which is related to the amount of refraction (bending) that occurs when light enters the material from air or a vacuum. The greater the index of refraction, the greater the amount of refraction.

Indices of refraction from sample materials are listed in Table 24.2. Notice how the index of refraction is equal to one for a vacuum (no bending) and is greater than one for all other materials.

✍ Since the speed of light in air is approximately the speed of light in a vacuum, we typically take the index of refraction of air to be 1.

You may have noticed that the order of the materials in Table 24.2 is the same as that in the list of speeds indicated in Table 23.1. This is not a coincidence. The reason for the correspondence is because the index of refraction is actually defined in terms of the speed of light. However, the index of refraction is *larger* when the speed of light is *slower*. This is a consequence of how the index of refraction is defined.

⁝ The index of refraction represents how much slower light travels in a material, compared to the speed in a vacuum.

By definition, the index of refraction (usually indicated by the letter n) is defined as the ratio c/v_{material}, where c is the speed of light in a vacuum and v_{material} is the speed of light in the material. For example, an index of 2

Substance	Index of Refraction
Vacuum	1 (exact)
Air	1.0003
Ice (0°C)	1.309
Water	1.333
Ethyl alcohol	1.362
Gasoline	1.396
Wesson™ Oil	1.474
Pyrex Glass	1.474
Glass, crown	1.523
Diamond	2.419

Table 24.2: The index of refraction of a sample of materials (all at 20°C unless otherwise noted).

means the speed is half that in a vacuum. Mathematically, the relationship is written as follows:

$$n_{\text{material}} = \frac{c}{v_{\text{material}}} \tag{24.2}$$

Example 24.1: If the speed of light in a material is 2×10^8 m/s, what is the material's index of refraction?

Answer 24.1: The material's index of refraction is the ratio c/v_{material}. Since the speed of light in a vacuum is 3×10^8 m/s, the material's index of refraction is 1.5.

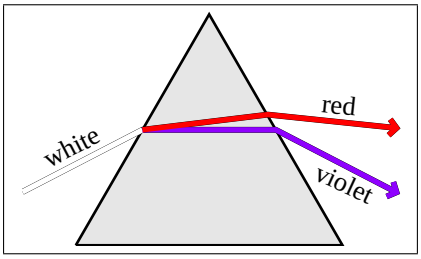
✓ *Check Point 24.10: In which material does light travel faster: water (n equal to 1.333) or crown glass (n equal to 1.523)?*

Technically, the index of refraction depends on the frequency of the light. Table 24.3 shows the dependence for various sample materials. You may noticed that there is no light frequency in Table 24.3 that corresponds to "white". As mentioned on page 425, white light is what our eye "sees" when it receives frequencies from across the visible light range.

Color (approx)	Wavelength in vacuum (nm)	ice	water (15°C)	crown glass	diamond
Red	700	1.307	1.329	1.520	2.410
Red	660		1.331	1.520	2.410
Orange	610			1.522	2.415
Orange-Yellow	600	1.310	1.333	1.522	2.415
Yellow	580			1.523	2.417
Green	550			1.526	2.426
Green-Blue	500	1.314	1.339	1.526	2.426
Blue	470		1.340	1.531	2.444
Violet	410			1.538	2.458
Violet	400	1.320	1.345	1.538	2.458

Table 24.3: Indices of refraction of selected materials at various wavelengths (from various sources).

Since the colors travel at different speeds, each color bends a slightly different amount, leading to a separation of the colors in white light, an effect called **dispersion**. This is what produces the colors in a **rainbow**.



✓ *Check Point 24.11: Consider a ray of 700-nm red light that is incident upon a boundary at an angle of 40°. Which should bend more: when encountering water from air, or encountering diamond from air?*

Summary

This chapter examined diffusion, reflection and refraction. The main points of this chapter are as follows:

- Light can bend around corners but the opening has to be very tiny since the wavelength of light is so small.
- The narrower the opening, the greater the spread of light that occurs.

- For specular reflection, the incident angle equals the reflected angle.
- Unless the wave is directed perpendicular to a boundary, it will change its direction upon entering the new material.
- Which way light bends depends on whether the speed is greater or less in the new material.
- The index of refraction represents how much slower light travels in a material, compared to the speed in a vacuum.
- Total internal reflection occurs when the light has a large incident angle and is hitting the boundary from within a slower material.
- The law of refraction describes the relationship between the angles in the two materials.
- Total internal reflection occurs when the light has a large incident angle and is hitting the boundary from within a slower material.

By now you should be able to predict how light will bend in various situations.

Terminology introduced

Absorption	Fringes	Optical fiber
Critical angle	Incident angle	Optics
Diffraction	Law of reflection	Reflection
Diffuse reflection	Light ray	Specular reflection
Dispersion	Normal direction	Total internal reflection
Index of refraction	Rainbow	

Abbreviations introduced

Quantity	**non-SI unit**
angle (θ)	degrees ($^\circ$)

25. Lenses and Mirrors

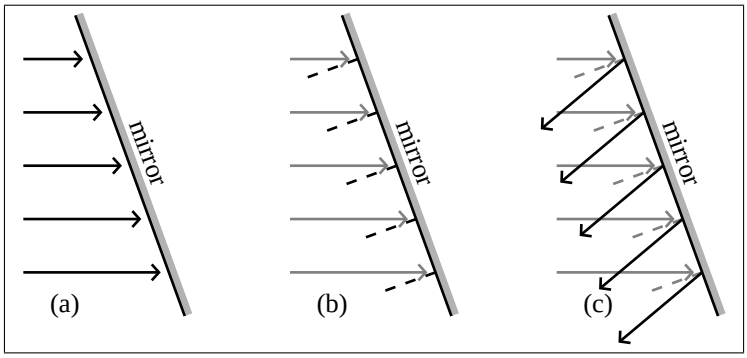
Puzzle #25: One can use a magnifying glass with the sun to start a fire. Why does that work?

Introduction

In this chapter, we utilize our understanding of reflection and refraction to explore how mirrors and lenses can be used to focus light, like using a magnifying[i] glass with the sun to start a fire.

25.1 Diverging and converging mirrors

To understand how a mirror or lens can focus light, we need to recognize that a beam of light, like that from the sun, consists of multiple light rays, which hit the mirror at different points. This is illustrated in part (a) of the figure below, where the rays come in from the left (as though the sun is off to the left) and hit a mirror that is drawn at an angle.



[i]In chapter 26, we'll explore why it is called a *magnifying* glass.

Each light ray undergoes reflection, following the law of reflection, where the reflected angle is equal to the incident angle (with the angles measured from a direction perpendicular to the surface). The process for determining the reflected direction is illustrated in parts (b) and (c) of the figure.

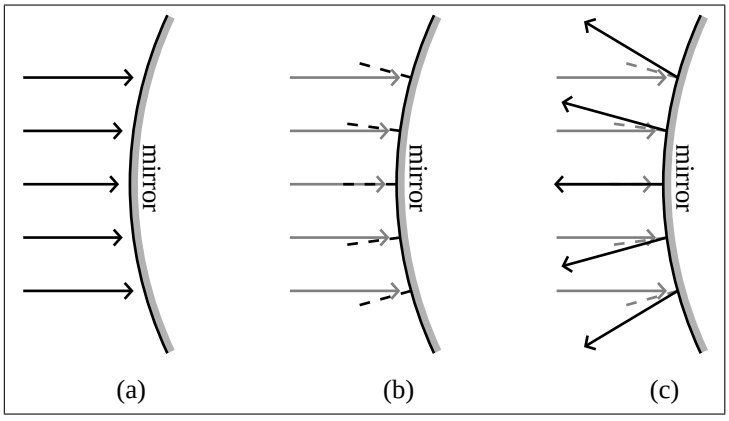
In part (b), a dashed line is drawn at the point where each ray hits the mirror. The dashed line is drawn perpendicular to the mirror to allow us to determine the incident and reflected angles. Notice that separate dashed lines are drawn at *each point where the particular ray hits the mirror*. The reflected angle, then, is equal to the incident angle but the direction of the reflected ray is on the opposite side of the dashed line (see part c of figure). In this way, you can predict where light will go when it hits a mirror.

❣ For mirrors, the reflected angle equals the incident angle at the point where the ray hits the mirror.

What happens if the mirror is curved?

The process is the same, in that the reflected angle equals the incident angle at the point where the light ray hits the mirror. However, the reflected directions will be different for each ray because the orientation of the mirror will be depend on where each ray happens to hit the mirror.

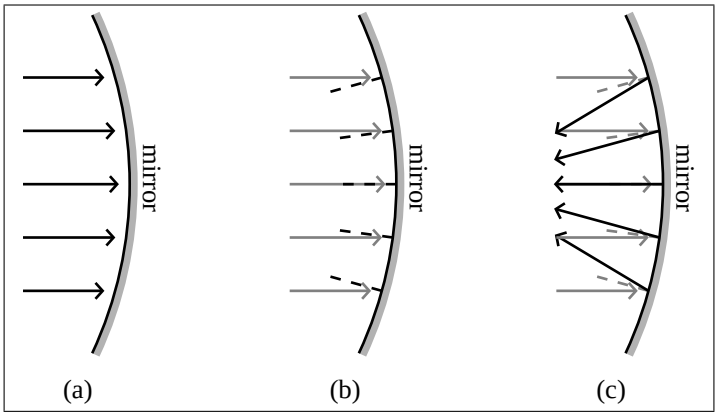
The process, though, is the same and is illustrated in the figure below.



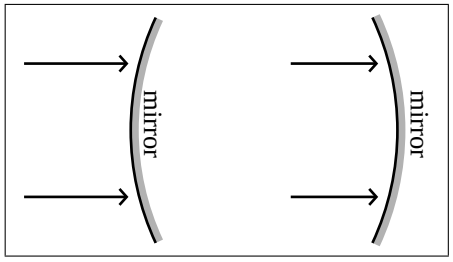
As before, five parallel rays are drawn, representing the light coming from the sun or some other source. In part (a), only the five incident rays are drawn. In part (b), a dashed line is drawn at the point where each ray hits the mirror. As before, that dashed line is perpendicular (normal) to the mirror *at the point where the particular ray hits the mirror*. However, unlike with the flat mirror, the dashed lines are not parallel *to each other* because the mirror is curved.

The reason we draw a separate dashed line for each ray is because the law of reflection states that the reflected angle equals the incident angle *at the point where each ray hits the mirror*. Each ray has a different incident angle, measured from the dashed line at the location where the ray hits the mirror surface. The reflected ray, then, is on the opposite side of its *own* dashed line (see part c of figure). In this way, you can predict where light will go when it hits a curved mirror.

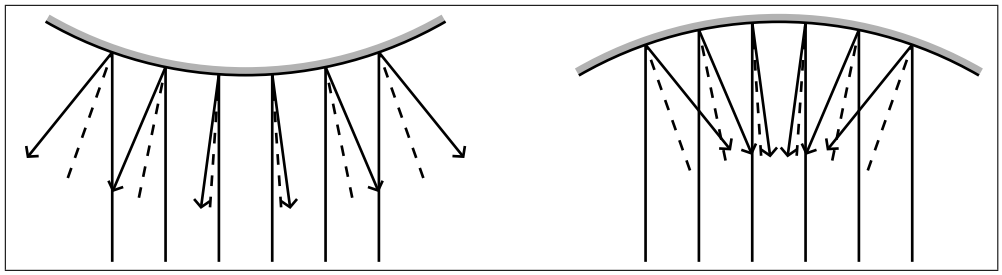
The process is the same for mirrors curved the opposite way, as illustrated in the figure below. At each location where a ray hits the mirror, I've drawn a dashed line that indicates the perpendicular (normal) to the mirror at that location. I then draw the reflected rays such that the reflected angle at each location equals the incident angle at that location.



✓ *Check Point 25.1: For the two cases illustrated to the right, redraw the figures and then add in the perpendicular line at the location where the incident ray hits the surface and the reflected ray from that location.*



You may have noticed that with the flat mirror, the reflected rays are parallel to each other, like the incident rays. However, for the curved mirror, the reflected rays are *not* parallel to each other. In the first curved mirror, the reflected rays are angled *away from* each other, and in the second curved mirror the reflected rays are angled *toward* each other. The difference in the two curved mirrors is illustrated on the next page.



To simplify our analysis of curved mirrors, we can classify curved mirrors into these two types, depending on what the mirror does to parallel incident rays. The two types are called converging and diverging mirrors.

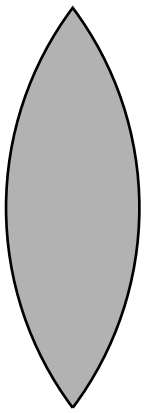
With a **diverging** mirror, like the one shown on the left in the illustration above, incident parallel rays are reflected such that the reflected rays diverge from one another.

In comparison, with a **converging** mirror, like the one shown on the right in the illustration above, incident parallel rays are reflected such that the reflected rays converge toward one another. Only a converging mirror can be used to focus light. That is why converging mirrors are used for reflecting telescopes (and why satellite dishes are curved inward like a converging mirror).

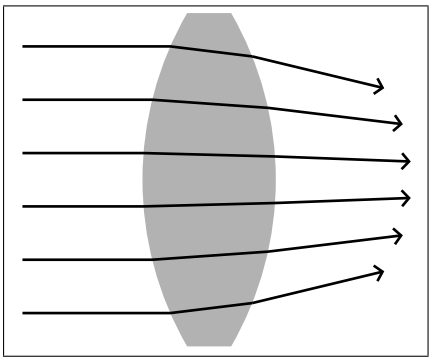
✍ An alternate way of describing the two types of mirrors is based on their shape. The left mirror is a **convex** mirror (because it is the outer surface of a sphere) and the right mirror is a **concave** mirror (because it is like the inside of a cave). I prefer converging and diverging because the words concave and convex just describe the *shape* of the mirror – they do not describe what the mirror does to parallel incident rays.

✓ *Check Point 25.2: Suppose the two mirrors illustrated above were mirrored on the top surface (instead of the bottom surface) and the light was incident upon the top surface (instead of the bottom surface). What kind of mirror would the left mirror be: a diverging mirror or a converging mirror?*

25.2 Converging and diverging lenses



Just as there are converging and diverging *mirrors*, there are also converging and diverging *lenses*, which are transparent pieces of material like glass or plastic. The difference is that light rays *reflect off* a mirror whereas the light rays *refract through* a lens. An example of a converging lens is the **magnifying glass**, which can be used to focus the light from the sun to start a fire. A magnifying lens is thicker in the center than along edges, as illustrated in the cross-section shape shown. To understand how this shape uses refraction to focus the light rays, we need to first review a little bit about refraction.



As before, we'll treat the beam of sunlight as a set of parallel light rays. As we know from chapter 24, light undergoes a refraction when it enters the lens because the lens has an index of refraction (around 1.4 for the eye) that is different than that of surrounding material (air). In this case, the light undergoes refraction *twice*: once upon entering the lens and then again upon leaving it (see illustration).

The end result of those two refractions is a converging of the light rays, as shown in the figure.

WHY DOES THE LIGHT REFRACT IN THAT PARTICULAR WAY?

To see why the lens does this, let's examine what happens at each boundary separately and then we'll combine the two.

The left figure on the next page illustrates what happens when parallel light rays first hit the lens (shaded). Since the speed of light is slower within the lens, each incident ray is bent toward the perpendicular line (dashed)at whatever location the incident ray hits the lens. Notice how that results in a converging of the light rays.

The right figure illustrates what happens when parallel light rays *leave* the lens (again, the shaded part is the lens). Since the speed of light is faster outside the lens, each incident ray is bent away from the perpendicular line

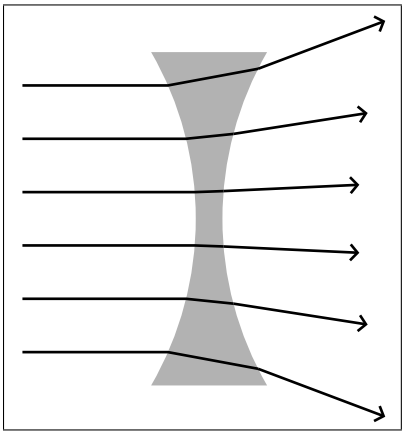
(dashed) at whatever location the incident ray hits the lens. Notice how that *also* results in a converging of the light rays.

In a real lens, the rays hitting the right edge would be converging, not parallel, because of the refraction at the left edge. However, the end result is the same – both sides contribute to the convergence of the light rays. In fact, it turns out that *any* lens thicker in the center than on the edges acts to converge the light rays.

⁝ Converging lenses are thicker in the center than the edges.

✍ As with mirrors, we can instead name lenses based upon their shape but I am not doing so because it makes more sense to name them based on what they do to initially parallel rays. However, the lenses above have surfaces that are convex (either on one side or both sides). Note that a converging lens is convex whereas a converging mirror is concave.

Is it possible to have a diverging lens?



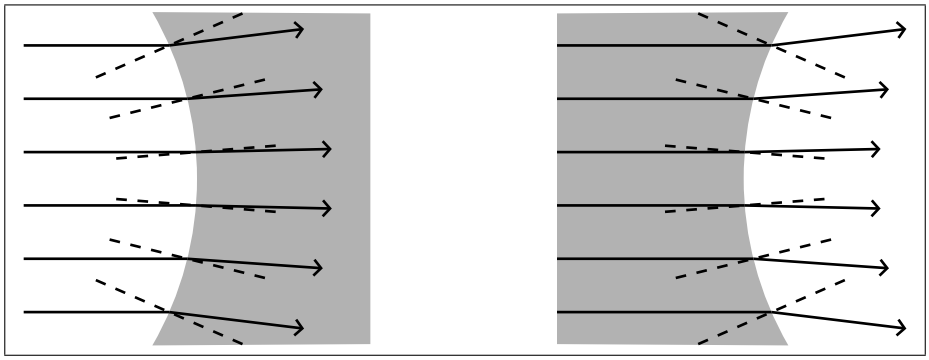
Yes, it turns out a diverging lens is thinner in the center than at the ends.[ii] An example is shown to the right.

⁝ Diverging lenses are thinner in the center than the edges.

To see why this is a diverging lens, we'll do the same exercise as before, applying the ideas of chapter 24 to each side of the lens, but with the two sides of the diverging lens.

[ii] The lens shown is concave on each side.

The left figure below illustrates what happens when parallel light rays first hit the lens (shaded). Since the speed of light is slower within the lens, each incident ray is bent toward the perpendicular line (dashed) at whatever location the incident ray hits the lens, as before, but because of the curvature of the lens surface it now results in a diverging of the light rays.



The right figure illustrates what happens when parallel light rays *leave* the lens. Since the speed of light is faster outside the lens, each incident ray is bent away from the perpendicular line (dashed) at whatever location the incident ray hits the lens. Notice how that *also* results in a diverging of the light rays.

Are diverging lenses of any use?

Diverging lenses are used in eyeglasses when someone is nearsighted as well as peepholes.[iii] Chapter 26 explores why this is.

✓ *Check Point 25.3: Figure 25.1 illustrates six types of lenses.*[iv] *Which lenses are diverging lenses?*

[iii] Diverging lenses are also used to correct chromatic aberration (dispersion due to different frequencies refracting a slightly different amount), as with binoculars, which otherwise would just use converging lenses.

[iv] The word **plane** is used in the mathematical sense, as in a flat or level surface, or carpentry sense, as in smoothing or finishing.

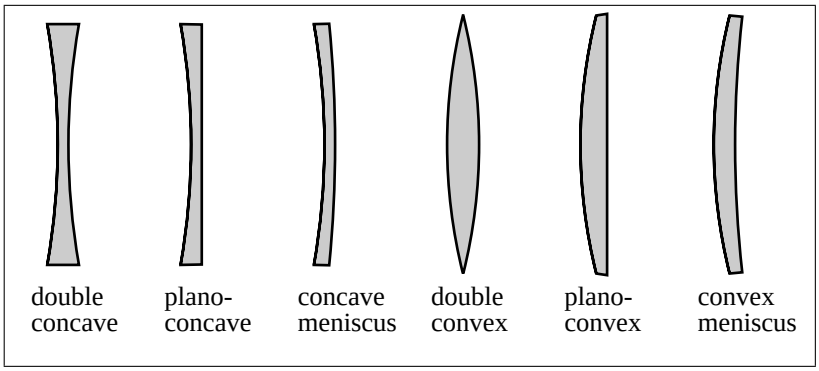


Figure 25.1: Cross-sections of six types of lenses.

25.3 Initially non-parallel rays

As noted earlier, we can consider a beam of sunlight as consisting of parallel light rays. Converging lenses make parallel light rays converge, whereas diverging lenses make parallel light rays diverge.

It turns out that *any* source of light, not just sunlight, can consist of parallel light rays, but only if the source of light is very far away (like the sun). To see why, consider the illustrations in Figure 25.2.

In each case the "O" indicates the location from where the light rays are coming. Notice the light rays that are incident upon the lens in each case. As the source is placed further away from the lens, the rays that are incident to the lens are more and more parallel. In the bottom illustration, the incident rays are parallel because the source is so far away (and too far away to be included in he illustration).

It is important to keep in mind that the light rays diverge *at the source* in the same way in each case. The rays incident *upon the lens* are more parallel as the source is farther away only because I draw a narrower and narrower range of the light rays.

How do lenses impact the light rays when the source isn't very far away?

Let's consider converging lenses first. To show how non-parallel light rays are impacted by a converging lens, again consider the four illustrations in Figure 25.2.

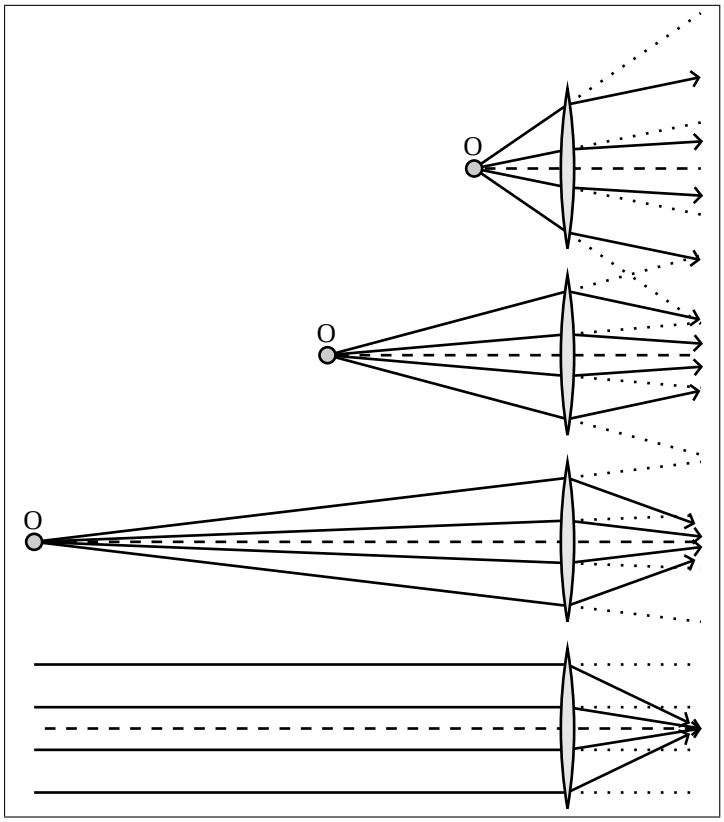


Figure 25.2

To help you see the bending in each case, I've drawn dotted lines that indicate what the rays would have done without the lens in place (i.e., the rays wouldn't bend). Notice how the bend direction is the same all four case (compare the angle between the dotted line and solid arrow) but only in the bottom three cases do the rays actually converge after passing through the lens.

In the bottom case, with the source really far away, the light rays end up converging to a point not far from the lens. As the source becomes closer, the light rays converge to a point farther and farther from the lens. In the top case, the source is so close that the lens isn't able to bend the rays enough to actually converge at all.

In other words, a converging lens always acts to bend the rays such that they are either converging (if the source is far enough way) or less diverging (if

⁉ A converging lens always acts to bend the rays to be more toward the optical axis.

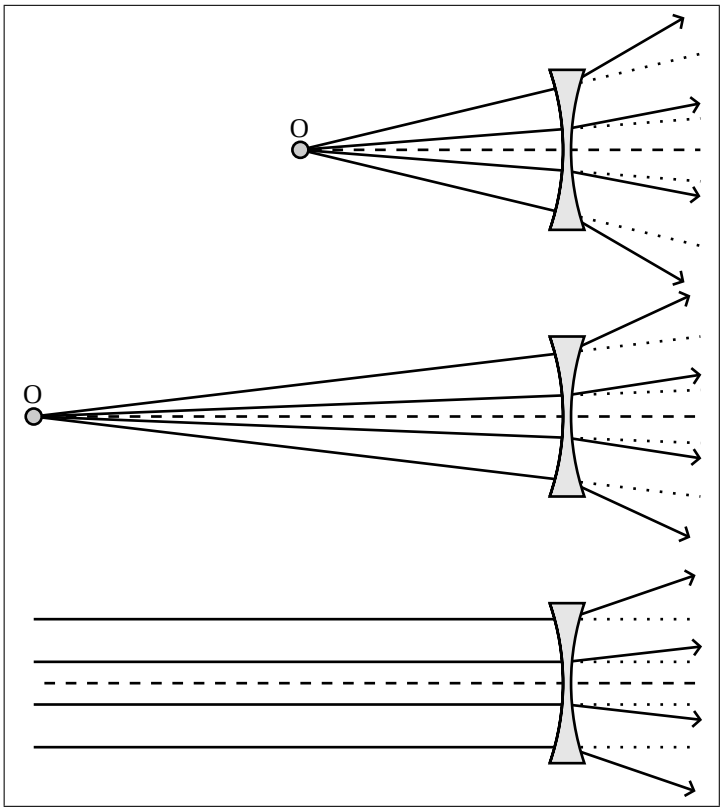


Figure 25.3

the source is close).[v]

The same is true for converging mirrors (not shown) in that a converging mirror always acts to bend the rays so that they are either converging (if the source is far enough way) or *less diverging* (if the source is close).

What about a diverging lens or mirror?

Figure 25.3 shows the impact of a diverging lens with sources at different distances (again, in the bottom case the source is really far away).

⁉ A diverging lens always acts to bend the rays to be more away the optical axis.

Notice how the diverging lens makes the rays bend in each case but, opposite to the converging lens, the diverging lens bends the rays to be more *away* from the optical axis (dashed line). In addition, since the rays from the

[v] In a sense, the converging lens always "converges" the rays but not always enough to make the rays converge. It is like cooling down a hot bowl of soup but not enough to make it cool (so that it remains hot).

source are diverging even when the source is close, the rays remain diverging after passing through the diverging lens, just more so.

The same is true for diverging mirrors (not shown) in that a diverging mirror always acts to bend the rays so that the reflected rays diverge more than the incident rays.

✓ *Check Point 25.4: After passing through a converging lens, do the rays always converge back to a point? After passing through a diverging lens, do the rays always diverge away from each other?*

25.4 Focal Length

In the previous section, it was mentioned how converging lenses and mirrors always act to bend the rays to be more toward (or less away from) the optical axis. However, just because they bend toward the optical axis does not mean they bend enough to *converge*. The rays only converge if the source isn't too close. If the source is too close the rays are still diverging after passing through the lens.

⁉ The strength of lenses and mirrors are described in terms of the focal length.

The transition occurs when the source is a certain distance from the lens or mirror. That distance is called the **focal length** and is typically represented by a lower-case letter f. Strong lenses and mirrors have short focal distances, meaning that the lens or mirror is so strong that the source has to be very close to the lens or mirror before the lens or mirror is unable to converge the rays. Weak lenses and mirrors, on the other hand, have long focal distances, meaning that the lens or mirror is only able to converge the rays if the source is far away.

✓ *Check Point 25.5: Is the focal length equal to the distance from the lens or mirror to where the light rays converge or appear to diverge from?*

Notice how stronger converging lenses, with their greater "converging" power, have a *smaller* focal length. Since this is an inverse relationship, opticians typically use the *inverse* of the focal length, called the **diopter strength** of

the lens (i.e., the diopter strength = $1/f$). In this way, the "stronger" the lens, the higher the diopter strength. The units of the diopter strength is inverse meters or **diopters**.

✍ The diopter strength does not indicate the durability of the lens or its inability to break but rather the effect it has on the bending of the light that passes through it.

✓ *Check Point 25.6: A particular lens has a focal length of 20 cm. What is the diopter strength of the lens?*

Do diverging lenses and mirrors also have a focal length?

Since diverging lenses and mirrors can't converge rays, regardless of how far the source is from the lens or mirror, they do not have focal lengths in the same sense as converging lenses and mirrors. However, they still have a focal length, just negative.

What does a negative focal length mean?

To understand what a negative focal length means for diverging lenses and mirrors, let's revisit what the focal length means for converging lenses and mirrors. As stated before, the focal length represents the transition between when the converging lens or mirror is able to converge the light rays and when it can't, depending on where the source is. Consider, then, what happens when the source is at a location that is *equal* to the focal length. In that case, the source is right at the transition point, and the light rays, after the refraction or refraction, neither converges nor diverges – the light rays become parallel.

⁝ The focal length is the distance from the lens or mirror to where the rays converge or appear to diverge from when the source of light rays are very far away.

This is just the reverse of what happens if the light rays are initially parallel – the lens or mirrors makes the rays converge to a point that is a distance from the lens or mirror *equal* to the focal length. This is illustrated by the three lenses on the right in Figure 25.4. In each case, I've used a double arrow marked with an f to indicate the focal length. Notice that, in each case, the focal length happens to be equal to the distance from the lens to the location where the light rays converge (since each case uses initially parallel rays). The top lens, being thicker and stronger, converges the rays more than the bottom lens. In each case, however, the distance from the lens to the convergence point is equal to the focal length of that lens.

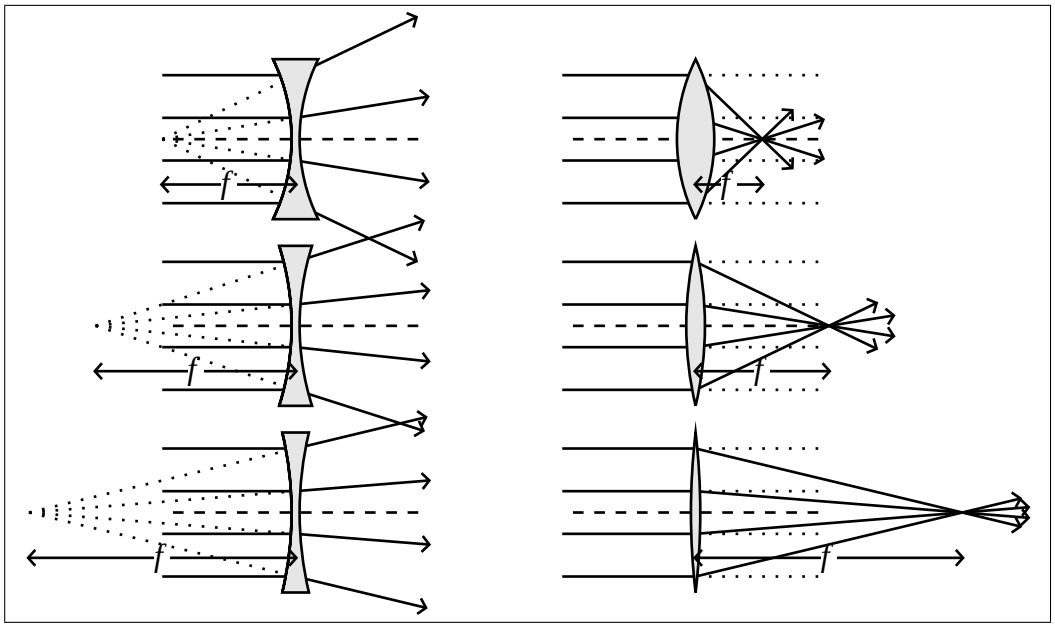


Figure 25.4

✍ Note that the focal length does not indicate where the light focuses. Indeed, that location depends on where the source is and only happens to equal the focal length when the incident rays are parallel.

Given that finding, we can use the same approach to define the focal length of diverging lenses and mirrors, as in the three lenses on the left side of Figure 25.4. The rays end up diverging, not converging, but they "appear" to diverge from a point, and that point is closer to the lens for the stronger lens, and farther from the lens for the weaker lens.

To distinguish between the meaning of the focal length for converging lenses and mirrors, where the light actually converges when the initial rays are parallel, we indicate the lengths as *negative* focal lengths.

Keep in mind that the focal length only depends on the strength of the lens or mirror (with a large focal length corresponding to a weak lens or mirror). The focal length does *not* depend on where the source is or what the incident rays are like.

⁝ Diverging lenses and mirrors have negative focal lengths. Converging lenses and mirrors have positive focal lengths.

✓ *Check Point 25.7: A particular lens has a focal length of +20 cm. What*

kind of lens is it – converging or diverging?

25.5 Ray diagrams

Knowing where the source is relative to the focal length, one can figure out where the light rays will converge to or diverge from by using **ray diagrams**, a graphical technique similar to what I've shown in illustrations.

Before showing the technique, I want to point out that in every case so far, the source was placed on the dashed line that passes through the center of the lens or mirror. That line is called the **optical axis**, and represents an axis of symmetry (where the portion of the lens on one side of he axis looks the same as the portion on the other). With the ray diagram technique, the source need not be on the optical axis, though, so I'll be using examples where the source is off the optical axis.

To illustrate the method, I'll apply it to three different situations. First, I'll apply it to the diverging lens and diverging mirror. Then I'll apply it to the converging lens and mirror, first with the source closer than the focal length then farther.[vi]

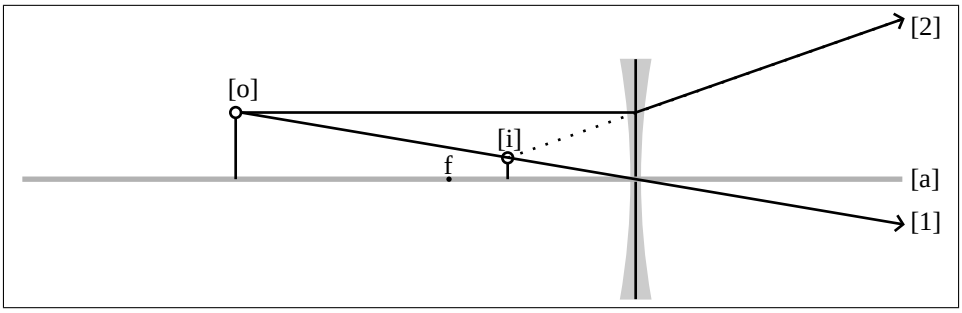
In every case, we'll draw two rays from the source, and our task is to identify where those two rays converge to (or diverge from) after passing through the lens or reflecting off the mirror.

We know there are an infinite number of rays from the source but we only need two to figure out this out, as every other ray will also converge to (or diverge from) the same place.

25.5.1 Diverging lenses and mirrors

The process involves six steps that I describe below and illustrate in the figure with a diverging lens.

[vi] Since refraction depends slightly on the frequency of the light, one can get dispersion with a lens (with different colors bending slightly different amounts depending on their frequency, as with rainbows; see chapter 24) whereas dispersion doesn't occur with mirrors because the angle of reflection is independent of frequency. For this reason, mirrors may be preferred in certain circumstances.



1. Draw the optical axis, which is the axis of symmetry that goes through the center of the lens. In the figure the optical axis is indicated as [a].
2. Indicate the position of the lens with a long vertical line. The method will only really work for very thin lenses so the long vertical line will act as our lens. I superimposed an outline of the lens only to indicate the type of lens, and the outline won't be used.
3. Draw a dot on the optical axis that represents a distance from the lens equal to the focal length. Label this as "f". The actual distance doesn't matter – it just serves as a reference, as you'll see in step 5.
4. Draw a ray from the source that is directed toward the point on the lens that lies on the optical axis (see ray [1]). This particular ray continues straight through the lens, as will be explained later.
5. Draw a ray that travels parallel to the optical axis until it reaches the lens (see ray [2]). From the definition of the focal length (see section 25.4), that particular ray bends away from the optical axis in a way that it is directed away from the point you've indicated as "f".
6. With the two rays drawn, identify where the rays are diverging from.

WHY DID YOU CHOOSE THESE TWO RAYS TO DRAW?

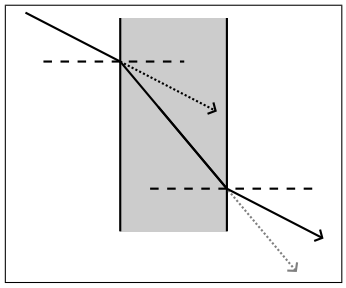
There are, of course, many rays coming from the source. We have chosen these two rays because we know how they will change as they hit the lens. Although two is a small number, it turns out we only really need to know two since all rays will diverge from the same point anyway.

⁝ While many rays can be used for the ray tracing, only two are needed to identify where the rays converge to or appear to diverge from.

WHY DOESN'T RAY 1 BEND AT THE LENS?

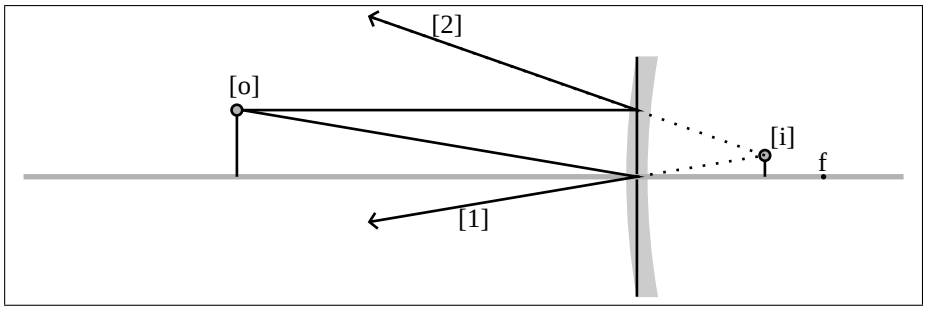
That ray, and only that ray, is not bent because the two sides of the lens at that point are parallel to each other. That means that the ray undergoes a refraction on the way into the lens but then an equal and opposite refraction on the way out, leading to no change in direction. This is illustrated in the figure below, which provides a magnified view of what is happening to the ray

as it passes through the center of the lens and how the original ray direction (dark dotted arrow) is parallel to the final ray direction (solid arrow).



Basically, at the center of the lens, the lens acts like a window, and rays passing through undergo a small "offset" that is proportional to the thickness of the lens at that point. For thin lenses, the offset should be negligible, which is why we are drawing ray 1 without any bend at all. That is also why I chose that ray out of the billions that are emitted from the source.

Below I illustrate the method with a diverging mirror. The only difference between this and the diverging lens is that the light reflects off the mirror rather than passing through it. For that reason, I draw the dot for the focal length on the side *opposite* the source of the rays.



The first ray hits the mirror right on the optical axis. Consequently, it reflects in the same way as if it were reflecting off of a plane (flat) mirror.

Being a diverging mirror, the second ray reflects away from a point on the optical axis that is a focal length away from the mirror (indicated as "f" in the figure).

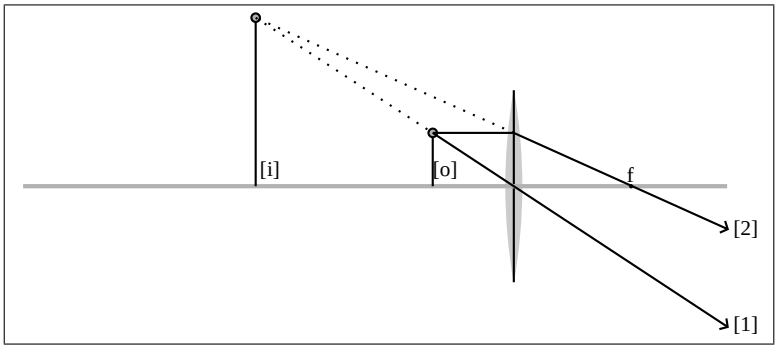
The end result is similar to what was obtained with the diverging lens, except that the reflected rays appear to diverge from a location on the *opposite* side of the mirror.

✓ *Check Point 25.8: Why does ray 2 bend when passing through the lens but ray 1 doesn't?*

25.5.2 Converging lenses and mirrors

The same ray tracing technique can be used for a converging lens. First I'll consider a source that is closer than the focal length and then I'll consider a source that is further.

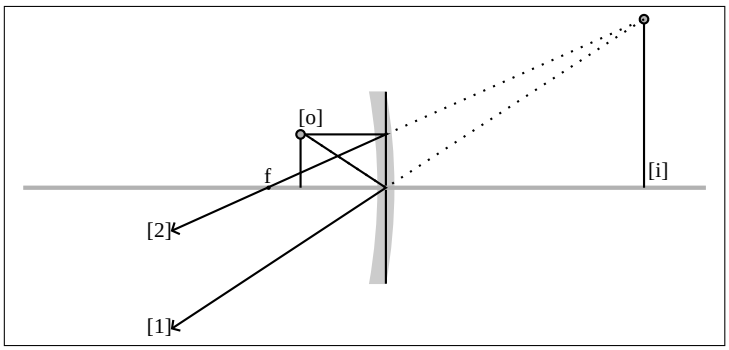
The diagram below shows the ray diagram with the source is closer than the focal length. Again, I drew a horizontal line for the optical axis (indicated as [a] in the figure) along with a vertical line to indicate the lens. I also included two dots: one dot to indicate the point a focal length away from the axis (see point "f") and a second dot, offset from the optical axis, to indicate the source location (see point "o").



This time, I drew the dot for the focal length on the *opposite* side of the lens from the source. I did this because it allows me to figure out how the parallel ray (indicated as [2] in the figure) is bent *toward* the optical axis (whereas for diverging rays the parallel ray is bent *away* from the optical axis). The other ray (indicated as [1]) is the one that goes through the center of the

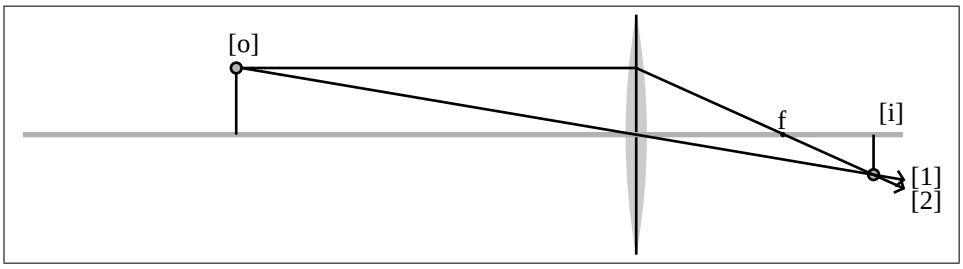
lens and is unbent. With the two rays drawn, we can see that the two rays appear to diverge from the location indicated as [i] in the figure.

The process for a converging mirror is the same, and is illustrated below. Notice how the source is again closer to the mirror than the focal length and the rays continue to diverge after reflecting off the converging mirror, in the same way that the rays continue to diverge after passing through the converging lens shown earlier.



✓ *Check Point 25.9: In the last ray diagram, with the converging mirror, do the rays converge to a point after reflecting off the mirror?*

For a converging lens or mirror to actually make the rays converge after passing through the lens or reflecting off the mirror, the source of the rays must be farther from the lens or mirror than the focal length. The process for a converging lens is illustrated in the ray diagram below.

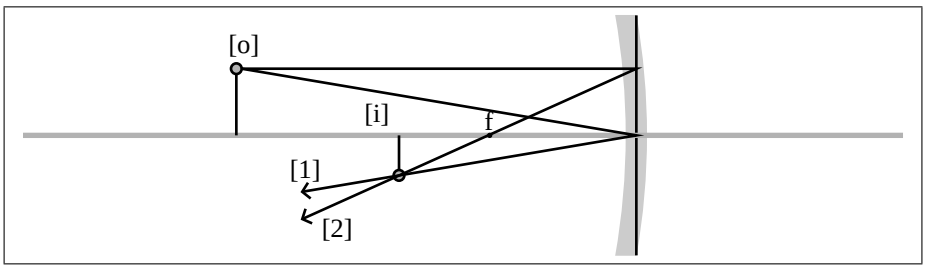


Again, the first ray passes through the center of the lens and is unchanged in direction while the second ray travels parallel to the optical axis until it

reaches the lens (see ray [2]) and then ray refracts toward the point on the optical axis that is a focal length away from the lens (see point f).

With the two rays drawn, we can see that the rays now converge to a point on the other side of the lens.

The process for a converging mirror and a source farther from the mirror than the focal length is illustrated below. Again, with the two rays drawn, we can see that the rays now converge to a point, but on the same side of the mirror as the source.



✓ *Check Point 25.10: Suppose the source of light rays was placed a distance from a converging lens or mirror equal to the focal length. After passing through the lens (or reflecting off the mirror), will the rays be (a) converging, (b) diverging or (c) parallel?*

Summary

This chapter examined mirrors and lenses. The main points of this chapter are as follows:

- For mirrors, the reflected angle equals the incident angle at the point where the ray hits the mirror.
- Converging lenses are thicker in the center than the edges.
- Diverging lenses are thinner in the center than the edges.
- A converging lens always acts to bend the rays to be more toward the optical axis.
- A diverging lens always acts to bend the rays to be more away the optical axis.

- The strength of lenses and mirrors are described in terms of the focal length.
- Diverging lenses and mirrors have negative focal lengths. Converging lenses and mirrors have positive focal lengths.
- The focal length is the distance from the lens or mirror to where the rays converge or appear to diverge from when the source of light rays are very far away.
- While many rays can be used for the ray tracing, only two are needed to identify where the rays converge to or appear to diverge from.

By now you should be able to predict what will happen to light rays that impinge upon a lens.

Frequently asked questions

In the figures, does point O represent where my eye is?

No. That is the source of the light (either where it is emitted or where it is reflected, as in the diffuse reflection shown on page 444). Your eye does not emit any light. The light must initially come from some source like the sun or a light bulb.

In the figures, the source is just a point. What if the source is bigger than just a point?

A larger source of light is essentially made up of lots of individual sources of light. To be complete, I could draw lots of sources but that would overly complicate the picture.

Terminology introduced

Concave	Diopters	Optical axis
Converging	Diverging	Plane
Convex	Focal length	Ray diagrams
Diopter strength	Magnifying glass	

Abbreviations introduced

Quantity	**SI unit**
diopter strength ($1/f$)	inverse meter (m^{-1})
focal length (f)	meter (m)

Additional problems

Problem 25.1: A particular lens is concave on one side (like the inside of a cave) and convex on the other (like the outside of a ball). Can one tell if it is converging or diverging? If so, which is it? If not, what additional information about the shape would you need?

Problem 25.2: Why do we say that sources infinitely far away produce rays that are parallel?

26. Objects and Images

Puzzle #26: How does a magnifying glass magnify?

Introduction

To answer the question in the puzzle, we need to distinguish between images (what we "see" as being there when we look into the lens or mirror) and objects (what is actually there when the lens or mirror isn't present). When we look through a lens, what we see is an image of the object, and the image may differ from the object, both in terms of where it is and how big it is. In this chapter, we'll examine how lenses and mirrors create images.

26.1 Magnification

To explain why a magnifying lens magnifies, we first have to define what we mean by **magnification**. To do that, consider the three illustrations in Figure 26.1, where each circle represents a lens that is placed in front of a piece of paper with the phrase "Physics is interesting!" on the paper. The result depends on the type of lens used (converging vs. diverging) and, for converging lenses, also whether the paper is placed closer than the focal length (center illustration) or farther than the focal length (right illustration). In later sections we'll explore why the different lenses produce the different images seen through the lenses. For now, we just want to describe what we see in each case.

⁉ The magnification represents the size of the image compared to the size of the object.

In the left illustration, the words appear smaller through the lens than without the lens. In comparison, in the center illustration, the words appear larger through the lens than without the lens. By convention, the magnification indicates of the ratio of the size of what is seen through the lens compared to what is seen without the lens. Consequently, the magnification

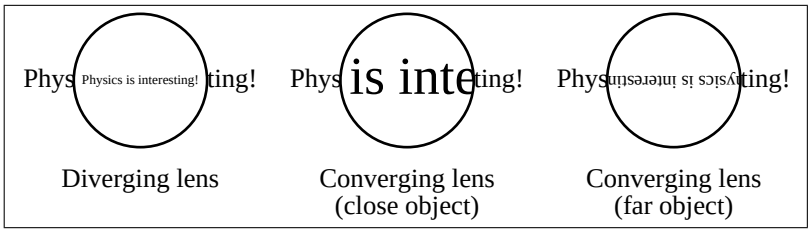


Figure 26.1: Illustration of what one might see when three different lenses are placed in front of the text "Physics is interesting!".

is smaller than 1 when the image is smaller, and the magnification is larger than 1 when the image is larger. A magnification of 1 means that the image seen through the lens is the same size as what is seen without the lens.

In the case of the illustrations in Figure 26.1, I drew the left image with a magnification of 0.5, which means the image seen through the lens is half the size as what is seen without the lens. In comparison, the center image is drawn with a magnification of 2, which means the image seen through the lens is twice the size as what is seen without the lens.

WHAT ABOUT THE IMAGE IN THE RIGHT ILLUSTRATION?

⁝ Positive and negative magnifications refer to whether the image is upright or inverted.

That image is about 70% the size of what is seen without the lens but it is *also* upside-down. By convention, we indicate a "flipped" image by using a *negative* magnification. Consequently, the right image is drawn with a magnification of -0.7, which means the image seen through the lens is inverted and 70% the size of what is seen without the lens.

Mathematically, the magnification, m, is defined as follows, where h_o and h_i are the heights of the object and image, respectively:

$$m = \frac{h_\text{i}}{h_\text{o}} \tag{26.1}$$

DOES THE MAGNIFICATION HAVE ANY UNITS?

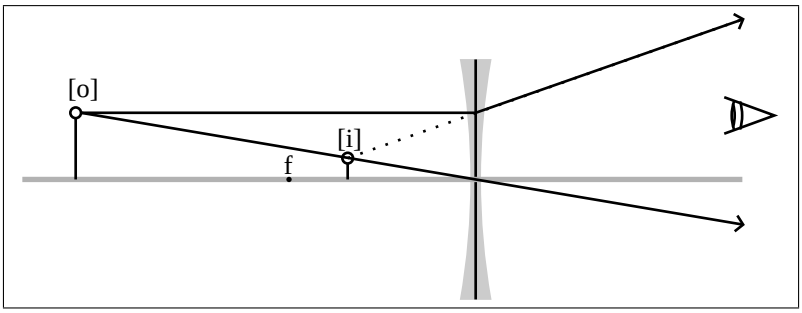
No. It is unitless. It represents how much you have to multiply the object size by to get the image size.

✓ *Check Point 26.1: Suppose we have a magnification of* -3*. (a) Which is bigger: the object or the image? (b) Is the image inverted or upright?*

26.2 Explaining magnification

In the previous section, it was mentioned that what we see through a lens can be different than what we see without the lens in place. For diverging lenses, we see an image that is smaller than what we'd see without the lens. The same is true for diverging mirrors. For example, if you could look at yourself in a diverging mirror[i], you'd see an image of yourself that is smaller than what you'd see in a regular flat mirror.

To explain why this is, let's revisit the ray diagrams introduced in chapter 25. Consider, for example, the ray diagram on page 473 and reproduced below illustrating the rays that pass through a diverging lens. The **object** in this case is the original source of the light rays (at [o]).



✍ The object need not be *emitting* light (like a bulb). For example, the object could be the phrase "Physics is interesting!" written on a piece of paper, since the paper reflects light in all directions (what we call *diffuse reflection*) and thus can be considered to be equivalent to a source of light for our purposes.

The illustration includes an "eye," representing where we'd be positioned (on the right) as we look leftward *through* the lens. The rays observed by the eye have been modified by the lens. Notice that those rays, if you trace them *straight* back, do not appear to diverge from [o] (where the object is). The rays instead appear to diverge from the position marked [i]. Indeed, from our vantage point on the right of the lens, we don't see anything at position [o]. Instead, we see it being at position [i]. We refer to what we see at [i] as the **image**.

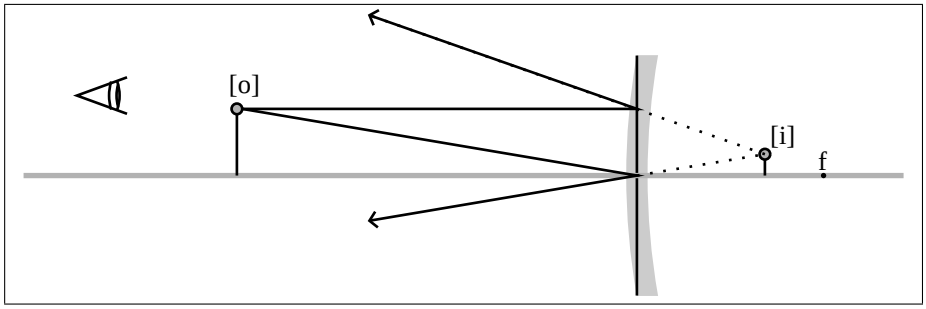
[i] An example of a diverging mirror is a shiny Christmas ball ornament.

✍ If the object (at [o]) was the phrase "Physics is interesting!" on a piece of paper with the lens between you (on the right) and the paper (on the left), you'd see the paper as being at position [i], not at [o]. The only way you'd see the paper at [o] is if the lens was removed, so that the two rays would reach you without being bent by the lens.

There are two important things to notice.

1. The image appears to be where the rays appear to *come from* not where the rays are *going*. Just as the object is where the rays originate *from*, the image is where the rays *appear* to originate *from*.
2. The image at [i] in this case is *smaller* than the object at [o], consistent with how we see something smaller when we look through a diverging lens, as illustrated in Figure 26.1.

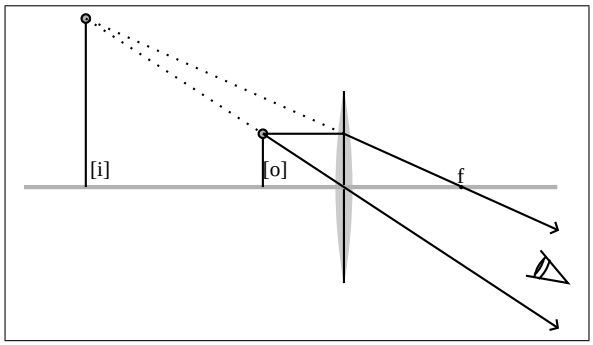
The same holds true for a diverging mirror, as illustrated by the ray diagram on page 474 and reproduced below, with the addition of an "eye," representing where we'd be positioned (on the left) as we look *into* the mirror at the reflection of the light from the object at [o]. Again notice (1) that the image is at [i], where the rays received by the eye *appear* to originate *from* and (2) how the image (at [i]) is smaller than the object (at [o]). You can experience this for yourself by looking at yourself using the "bottom" of a shiny spoon.



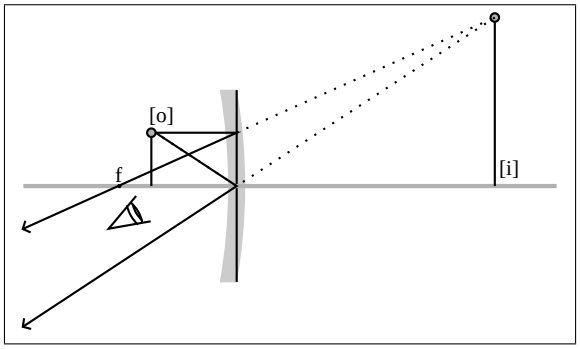
✓ *Check Point 26.2: In the ray diagram with the diverging mirror, would you still be able to see the image (located at [i] in the illustration) if you were positioned to the right of the image looking leftward?*

Now let's consider what happens with a converging lens, where we see an enlarged image when the object is closer than the focal length (center illustration in Figure 26.1). To see why, consider the ray diagram on page 474

and reproduced below, illustrating the rays that pass through a converging lens with the source closer than the focal length, with the addition of an "eye" on the right looking leftward through the lens.[ii] Notice how the image at [i] is now *bigger* than the object at [o], consistent with what was illustrated in Figure 26.1.



The same is true when we look into a converging mirror, with the object closer to the mirror than the mirror's focal length. This is illustrated by the ray diagram on page 476 and reproduced below, with the "eye" added to show how we'd be on the left looking rightward into the mirror. Again notice how the image at [i] is larger than the object at [o]. This is why a make-up mirror, which is a converging mirror, enlarges our reflection when we happen to be close to the mirror when we look into it. If you don't have a make-up mirror to experiment with, you can try using looking into the "cup" part of a shiny spoon (but hold it close to your face so that the object, your face, is closer than the mirror's focal length).

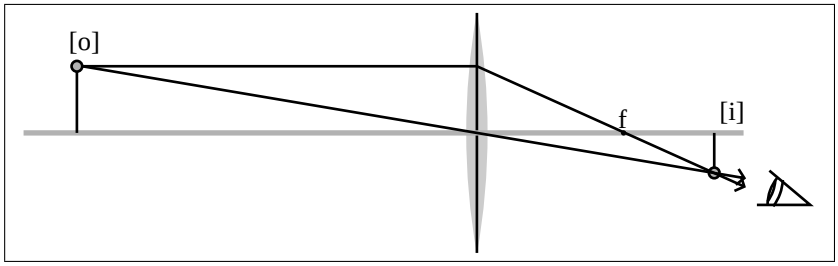


[ii]The reason I place the "eye" on the right is because, to see the image, we need to be in a position to receive rays from the source *after* the rays have passed through the lens. Without the lens present, we'd see the object at [o].

✓ *Check Point 26.3: In the ray diagram on the previous page with the converging mirror, would you be able to see the image if you were on the right side of the mirror instead of its left side?*

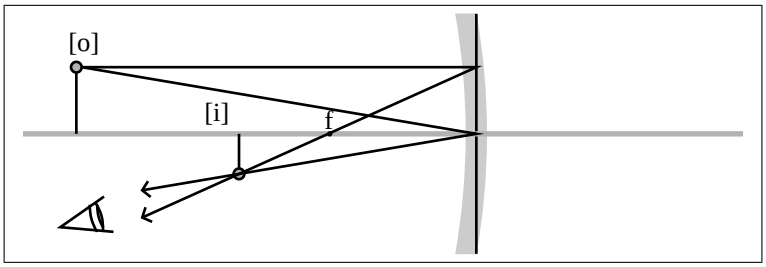
As our final analysis, let's examine what happens with a converging lens and mirror when the object is more than the focal length away from the lens or mirror. As illustrated in Figure 26.1, we see an inverted image.

To see why, consider the ray diagram on page 474 and reproduced below, illustrating the rays that pass through a converging lens with the source *farther* than the focal length, with the "eye" added to show how we'd be on the right looking leftward through the lens.[iii] Notice how the image at [i] is now on the *opposite* side of the optical axis as the object [o]. This is consistent with the inverted image illustrated in Figure 26.1.



The same is true when we look into a converging mirror, with our face farther from the mirror than the mirror's focal length. This is illustrated by the ray diagram on page 477 and reproduced on the next page,with the "eye" added to show how we'd be on the left looking rightward into the mirror. Again notice how the image at [i] is inverted from the object at [o]. This is what you'd see with a make-up mirror if you held it far from your face (which most people don't do). If you don't have a make-up mirror to experiment with, you can try using looking into the "cup" part of a shiny spoon (but hold it far from your face).

[iii] We have to position ourselves on the opposite side of the lens as the paper, which is to the right of the lens in the illustration, and looking leftward, so that we would be receiving the rays *after* they have passed through the lens. Otherwise, we wouldn't see the image.



✓ *Check Point 26.4: In the ray diagram with the converging mirror and inverted image, where would you have to be, and which way would you have to be facing, to see the image produced by the mirror (located at [i] in the illustration)?*

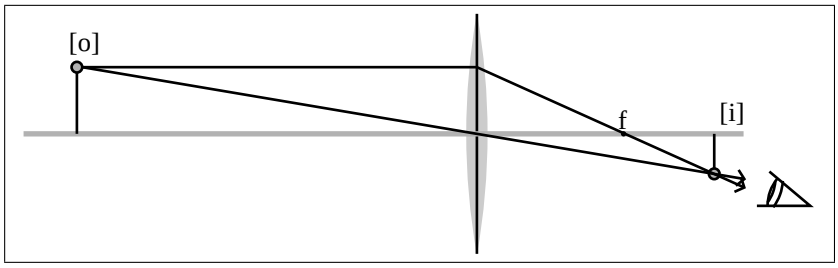
26.3 Real vs. virtual images

If you tried to view the inverted image, either by using a converging lens with the object far from the lens or by using a converging mirror with yourself far from the mirror, you may have noticed that it was hard to see the inverted image clearly. The reason for this has to do with *where* the image is.

You see, the image not only has a particular size but a particular *location* as well. The position depends on the lens and where the lens is placed but in no case is the image on the lens itself, where many people incorrectly assume the image should be. And, if the image is not where you expect it to be you won't see it clearly.

Let's revisit the ray diagram, reproduced on the next page, for the converging lens and inverted image. The image is at [i]. Notice that is not on the lens. Instead, it is to the right of the lens. For you to see that image, you have to position yourself to the right of that image, looking to the left. Only then would the rays be reaching you from point [i].[iv]

[iv] In addition, you'd have to be far enough away from the image to be able to focus on it clearly, as our eyes can't focus on things that are too close. And if you were between the image and the lens or mirror, you wouldn't see the image at all.



Is the image just floating in midair?

Yes, in the sense that the image is not "on" anything. That is what makes it hard to see clearly. One way to clearly see the image, though, is to put something at that location to coincide with the image location. For example, if the object is bright enough, you can place a screen at the image location and, voilà, you'll see the image on the screen. This is because the light undergoes diffuse reflection upon hitting the screen, allowing you to see the image from any location (as long as you are on the appropriate side of the screen) and, being as the screen is at the same location as the image, it makes it easier to see the image clearly.

✍ This is essentially the way movie theaters project images onto a screen. In those cases, the screen is placed at the exact location of the image, allowing the audience to see it.

Note that the screen technique can only be used for the case with the converging lens or mirror and object farther than the focal length (and if the object is bright enough). Only in that case do the light rays, after passing through the lens or reflecting off the mirror, actually converge to a point. For the other cases, the light rays do not converge after passing through the lens or reflecting off the mirror. You can still see the image, but only by looking *through* the lens or *into* the mirror because the rays appear to diverge from a location on a side of the lens or mirror that is opposite you.

⁉ The words "real" and "virtual" describe whether the light actually diverges from the image location or not.

To distinguish between the two types of images – those that you can use a screen to observe and those you cannot – we call the first type **real** images and the second type **virtual** images. Don't get hung up on the names – both types are real enough to the people who see them.

The only way to see a virtual image is to look *through* the lens or *into* the mirror. To see a real image, you can also look through the lens or into the

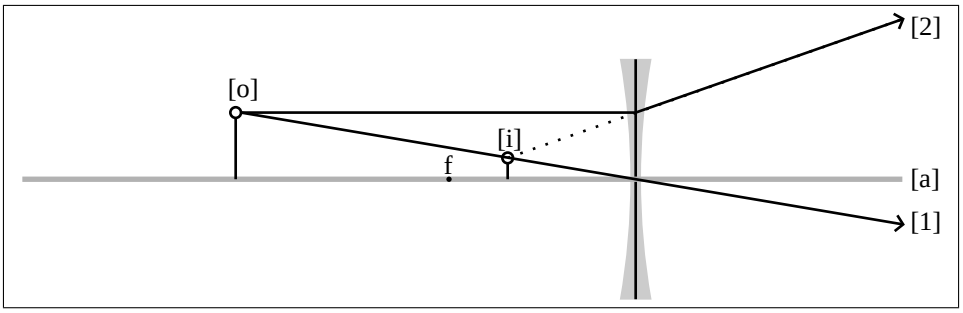
mirror (if you are not too close to the lens or mirror[v]) but you can *also* use a screen (at the location of the image).

✍ Remember that both real and virtual images are, well, images. We just tend to use the terms "real" and "virtual" to distinguish between those that form by the rays that *actually* diverge from a point (after converging to a point) and those that only *appear* to do so.

✓ *Check Point 26.5: When you look at a movie in a movie theater, an image is produced by a lens in the movie projector. The screen then reflects the image back to the audience. If the screen was removed, where would the image be?*

Can we tell where the image is just by looking at it?

It is actually pretty hard to tell where an image just by looking at it. For example, consider the image produced by the diverging lens, as illustrated in Figure 26.1. We know from the ray diagram of the situation, reproduced below, that the image is not only smaller than the object but it is also *closer*.



Go back and look at the left illustration in Figure 26.1. We can tell it is smaller, but can you tell that it is also *closer*? Chances are you assumed it was farther, not closer, since farther objects tend to appear smaller to you. However, the image in this case is actually smaller and closer, not smaller and farther.[vi]

⁉ If an image is smaller than the object that doesn't mean the image is farther away; conversely a larger image isn't closer than the object.

[v] And you know ahead of time where the image is supposed to be, since you probably won't be expecting it to be between you and the lens.

[vi] In the United States (and some other countries), the phrase "objects in mirror are closer than they appear" must be stated on the passenger side mirrors of cars. The reason

Conversely, consider the center illustration in Figure 26.1, showing the magnification that occurs with a converging lens and an object closer than the focal length. Chances are you assumed it was closer, not farther, since closer objects tend to appear bigger. However, the image in that case is actually larger and farther, not larger and closer.

Why is the image not the same size as the object?

Everything has a size. We can tell its size by examining where the top is vs. the bottom. The size of the image is determined by where the light rays appear to come from – the rays from the top appear to diverge from one place and the rays from the bottom appear to diverge from another place. Notice that it isn't the *spread* of the rays that tell us the size. Rather, it is *where* the top is relative to the bottom.

It turns out that for *every* image (created by a lens or mirror) the size is proportional to its distance from the lens or mirror. For magnifications equal to one, the image is the same distance from the lens or mirror as the object. For magnifications greater than one (image larger than object), the image is farther than the object. And for magnification less than one (image smaller than object), the image is closer than the object.

⁉ Images that are closer than the object are also smaller than the object, and images that are farther than the object are also larger than the object.

Mathematically, we can write this relationship as follows, where h_o and h_i are the heights of the object and image, respectively, and d_o and d_i are the distances from the lens or mirror to the object and image, respectively:

$$\frac{h_i}{h_o} = -\frac{d_i}{d_o} \tag{26.2}$$

✍ Note that the image and object distances are measured *from the lens or mirror*, not from each other.

Why is there a minus sign?

The minus sign is a consequence of our sign convention. Recall that the magnification is negative when the image is inverted. The negative sign is there to indicate the inversion.

---

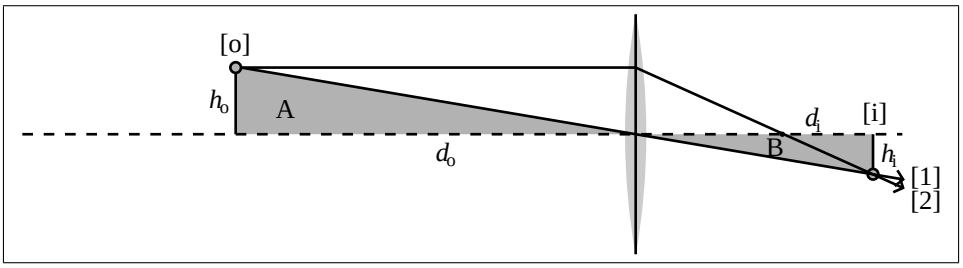
for this is the mirrors are diverging mirrors, which are used to provide a wider field of view, but produce images that are small. People naturally assume that smaller means farther away and may, as a result, over-estimate how far back the vehicles are (that are seen in the mirror).

Does this work for images that aren't inverted as well?

Yes. However, to get a positive magnification for such images, we'll follow the convention that the image distance is negative. In other words, the image distance is negative for virtual images (not inverted) and positive for real images (inverted).

How do we know that the sizes are proportional to their distances from the lens or mirror?

Certainly it seems consistent with the ray diagrams we've drawn so far, but one can also show that is true by using a little geometry with any of the ray diagrams we've used so far. For example, consider the ray diagram for the converging lens and an object farther than the focal length, reproduced below with two extra triangles added.



The two triangles, shaded in the figure, are similar in the sense that the interior angles of triangle A is the same as the interior angles of triangle B. The only difference is the size. That means the ratio of triangle A's sides (h_o to d_o) is the same as the ratio of triangle B's sides (h_i to d_i), which is what equation 26.2 represents.

✓ *Check Point 26.6: Suppose we place an object 20 cm from a diverging lens such that the image is 10 cm from the lens.*
(a) What is the object distance and what is the image distance?
(b) What is the magnification?
(c) Given the magnification, what would the image height be if the object height was 5 cm?

26.4 Image angular size

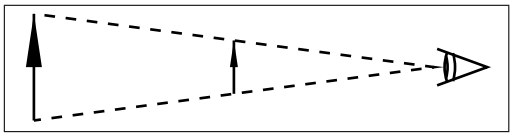
Even if an image is larger, shouldn't it appear smaller if it is farther away?

That depends on how far away it is. For example, the moon is much bigger than your house but it appears smaller (in that you can easily cover it with your hand) because it is so much farther away. However, if the moon was only twice as far from you as your house, it would certainly appear larger.

For this reason, as long as your eye is not right up against the lens, larger images will actually appear larger.

What happens if your eye is right up against the lens?

In that case, the image will have the same *angular* size as the object from your perspective. To illustrate what I mean, consider the illustration below. Notice how the left arrow and the right arrow (both pointing up) are seen as having the same *angular* size from the perspective of the "eye" (the eye, looking left, is that thing I drew on the right side of the illustration).

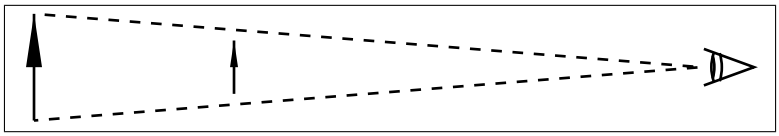


Since image and object sizes are proportional to their distances *from the lens*, that means they'll have the same *angular* size as seen from the perspective of someone viewing them *from the lens*. In the illustration, the angular size is indicated by the dashed lines.

✍ This is why people wearing glasses see things having the same size as those who don't wear glasses.

Why then, does a magnifying glass magnify?

It wouldn't appear to magnify anything if you placed the magnifying glass right up against your eye. The image would only appear to be larger if the magnifying glass was away from your eye. This is illustrated below, with the same two arrows as before but with the "eye" moved farther away. Notice how the angular size of the larger arrow (dashed lines) is larger than what it is for the smaller arrow (dashed lines for smaller arrow not shown so as to not crowd the picture).



✓ *Check Point 26.7: A diverging lens is held 22 cm from your eye as you look at far-away mountains through the lens. Which is closer: the mountains or the image created by the lens? If the image is closer, which is closer to you: the image or the lens?*

26.5 Multiple lenses

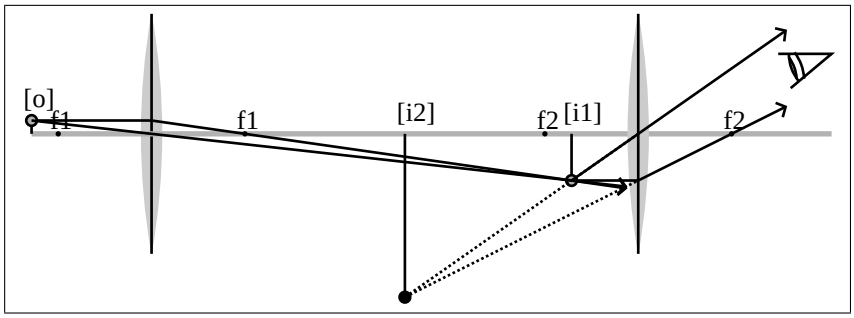
Whereas a magnifying glass uses a single lens, many optical instruments, like microscopes and telescopes, don't just use a single lens. There are two reasons.

First, for a telescope, the object (like a planet) is very far away, much farther than the focal length, so the image isn't magnified at all. The image is only a focal length away from the lens, so it is much closer than the object, but it is also very tiny (since its size is proportional to how far it is from the lens).

Since it is closer, though, we can then use a *second* lens to examine that image. Basically, we use the image from the first lens as the object for the second lens. In a sense, the second lens is the magnifying glass, examining the details of the small, but close, image created by the first lens.

A **compound microscope** does something similar. In a compound microscope, the object is placed slightly *farther* from the lens than the focal length. This produces a real image that is magnified and real. The observer can then use *another* converging lens to look at *that* image in the same way as the magnifying glass. The end result is an image that has a much larger angular size than the object.

This is illustrated in the following diagram.



In the diagram, an object [o] is placed on the left of a converging lens (i.e., the left lens in the diagram), a little bit farther than the focal length of the lens on the left (see f1 in the diagram). That produces a real image at [i1] that is larger (and inverted).

The lens on the right is placed such that the image at [i1] is slightly closer than the focal length of that lens (see f2 in diagram). That produces a virtual image at [i2] that is even larger.

The magnification of the second lens doesn't produce an angular magnification since our eye is right at the second lens. However, it moves it further away, which makes it easier to focus on.

Where is our eye in the illustration?

To see the images, our eye has to receive the light rays. That means we must be on the right looking toward the left.

✓ *Check Point 26.8: When you look through a compound microscope or telescope, the image is inverted. Why?*

26.6 Vision correction

The eye is essentially a converging lens[vii] and a "screen" (called the retina), and your brain interprets the images that are produced on the screen.

[vii]The actual shape of the lens in the eye is a bit more like an oval than pointy like my drawings but the key thing is that it is thicker in the middle than the edge, which means it is converging lens.

How does the eye deal with different object distances if the retina is always the same distance from the lens?

It does this by changing the shape of the lens in the eye. The more the lens is stretched, the "thinner" it gets. That effects the curvature of its surface, and we know that the amount of bending depends on the curvature. The greater the curvature, the greater the bending and the stronger the lens. For a closer object, a stronger lens is needed (top case), whereas a weaker lens is needed for a farther object (bottom case).[viii]

Why do people wear glasses?

⁝ Converging and diverging lenses can be used to correct for eyes that are too weak or too strong.

When people get older, the lens in the eye becomes less flexible. If the lens gets stuck in a shape that is too strong, then it won't be able to see clearly things that are far away. This means the person is **nearsighted** (they can see near but not far). To correct their vision, they wear diverging lenses. A diverging lens not only counters the overly converging nature of the own lens in the eye, but the diverging lens also creates an image that is closer to the wearer (as discussed earlier), allowing the lens in the eye to "see" it clearly.

⁝ Nearsighted people wear diverging lenses whereas farsighted people wear converging lenses.

⁝ Corrective lenses produce images at a location where the wearer can see them clearly.

Conversely, if the lens gets stuck in a shape that is too weak, then it won't be able to see clearly things that are close. This means the person is **farsighted** (they can see far but not near). To correct their vision, they wear converging lenses. A converging lens not only counters the overly converging nature of the own lens in the eye, but the converging lens also creates an image that is farther from the wearer (as discussed earlier), allowing the lens in the eye to "see" it clearly.[ix]

✍ As you get older, you may need help seeing both near and far. That is why many older people wear bifocals, with different strengths for seeing near vs. far.

[viii] Your eye is flexible but not infinitely so. This means that objects can be too close to focus clearly and will be blurry. The converse is not necessarily true – objects that are very far away can still be clear (like two stars), but they'll just be too small to resolve (like two stars that are close together).

[ix] Farsighted people, since they wear converging lenses, will appear as though they are wearing magnifying glasses and to other people the wearer's eyes will appear bigger than what they would appear otherwise. The eyes of nearsighted people, since they wear diverging lenses, will appear smaller to other people.

✓ *Check Point 26.9: What kind of person would wear diverging lenses, a near-sighted person or a far-sighted person? Explain.*

26.7 Flat mirrors and lenses

What if the lens is flat?

If by a "flat lens" you mean a lens that is flat on both sides then a flat lens is just a window. In such a "lens", the refraction upon leaving the lens would just be the opposite of the refraction upon entering. Since there is no net refraction, the image, if you want to call it that, is in the same location and has the same size as the object.

✍ This would be true even for lenses with curved surfaces if the thickness was uniform throughout (i.e., same concave curvature on side as the convex curvature on the other).

What about regular bathroom mirrors?

Most bathroom mirrors are **plane**[x] mirrors, meaning that they are flat, not curved. Consequently, the image is just as close to the mirror as the object, just on the opposite side. Since the distance is the same, the image is the same size as the object.[xi] In other words, when we look into a flat mirror (neither converging nor diverging), we see an image of ourselves, equidistant to the mirror but on the other side, as though we are looking through a hole in the wall at another person, of the same size as us, who is looking back.

✓ *Check Point 26.10: Suppose a 6-foot tall person stands 5 feet from a flat mirror, like those in bathrooms. Where and how big is the image seen in the mirror?*

[x] The word "plane" is used in the mathematical sense, as in a flat or level surface, or carpentry sense, as in smoothing or finishing.

[xi] This means the the magnification is one.

Summary

This chapter discussed how lenses and mirrors create images.

The main points of this chapter are as follows:

- The magnification represents the size of the image compared to the size of the object.
- Positive and negative magnifications refer to whether the image is upright or inverted.
- The words "real" and "virtual" describe whether the light actually diverges from the image location or not.
- If an image is smaller than the object that doesn't mean the image is farther away; conversely a larger image isn't closer than the object.
- Images that are closer than the object are also smaller than the object, and images that are farther than the object are also larger than the object.
- Converging and diverging lenses can be used to correct for eyes that are too weak or too strong.
- Nearsighted people wear diverging lenses whereas farsighted people wear converging lenses.
- Corrective lenses produce images at a location where the wearer can see them clearly.

By now you should be able to predict the image that will form with lenses and mirrors.

Frequently asked questions

Are distances always positive?

No. By convention, the distances are positive only for real objects and images. For virtual objects and images, the distances are negative.

Is the focal length the same as the image distance?

No. The image distance is the distance from the lens to the image. The image distance will not equal the focal length unless the object happens to be sufficiently far away.

So the image distance is not the distance from the object to the image?

No. The image distance is the distance from the lens to the image.

Isn't the image always at the lens?

No. The image can be anywhere, depending on where the object is and what type of lens you have. In fact, it is never at the lens.

Are the shaded lenses in the various figures supposed to represent where my eye would be?

Not necessarily. When talking about corrective lenses or lenses you might hold in your hand (like a magnifying lens), your eye would be off to one side of the figure, observing the light that has *started* at the object (point O) and *passed through* the lenses (shaded).

If there are multiple lens, which lens do you measure the distances from?

For the image associated with the *first* lens, the image distance is measured from the *first* lens. Once you obtain the image produced by the *first* lens, treat that as the object for the *second* lens. The object distance for the *second* lens is measured from the *second* lens.

In other words, you cannot blindly use the image distance from the first lens as the object distance for the second lens! If the image from the first lens serves as an object for the second lens, you must first determine where that image is relative to the second lens.

It is a good idea to draw a ray diagram. A ray diagram helps to make sure you are using the right distances for the second lens' object distance.

Terminology introduced

Compound microscope	Magnification	Plane
Farsighted	Nearsighted	Real
Image	Object	Virtual

Abbreviations introduced

Quantity	**SI unit**
distance (d)	meter (m)
height (h)	meter (m)
magnification (m)	none

Additional problems

Problem 26.1: (a) Can the image distance be negative? If so, give an example.
(b) Can the object distance be negative? If so, give an example.

Problem 26.2: (a) Suppose you placed a diverging lens directly in front of your eye as you looked at an object. Compare the angular size (height vs. distance) of the object compared to that of the image. Which would be larger (or would they be equal)?
(b) Based on your answer to (a), does a nearsighted person see images that are the same angular size as the objects? Explain.

Problem 26.3: A particular lens has a focal length of 20 cm. (a) How far from the lens is the image when the object is infinitely far away? (b) Where is the image if the object is placed a distance from the lens equal to the focal length?

Problem 26.4: A converging lens with focal length 20 cm is held 22 cm from your eye as you look at far-away mountains through the lens. What will the image of the mountains look like (i.e., clear, blurry, right-side up, upside down, small, large, etc.)? Explain.

Problem 26.5: Two converging lens, both with 20 cm focal lengths, are placed a distance d apart. Where is the image produced by the two lens when an object is placed 20 cm to the left of the left-most lens?

Problem 26.6: When you look at yourself in a plane (flat) mirror, is the image of yourself real or virtual? Explain your reasoning.

Problem 26.7: For each of the following situations, identify the sign of the focal length, the image distance and the magnification.
(a) A diverging mirror producing a virtual upright image.
(b) A converging mirror producing a virtual upright image.
(c) A converging mirror producing a real inverted image.

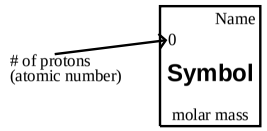
Problem 26.8: A sphere with radius 50 cm is mirrored on its outside surface (like a very large Christmas ball). For an object that is far away, what do the light rays from the object do after reflecting off the surface of the sphere: converge or diverge? Would your answer change if the object was close? Why or why not? If you stand such that you see yourself in the mirror, where (relative to the mirrored surface) and how big (relative to you) is your reflection in the mirror? Support your answer with a ray tracing.[xii]

Problem 26.9: A diverging lens with focal length -20 cm is held 22 cm from your eye as you look at far-away mountains through the lens (i.e., far enough away that light rays from the object are essentially parallel when they hit the lens). Where and how big (compared to the object) is the image of the mountains? Explain. Support your answer with a ray tracing.

Problem 26.10: A sphere with radius 50 cm is mirrored on its outside surface (like a very large Christmas ball). If you stand such that you see yourself in the mirror, where (relative to the mirrored surface) is the image of yourself (i.e., your reflection) due to the mirror?

[xii] You will need to choose a horizontal and a vertical scale (they do not need to be the same). Draw the mirror location, your location, and the optical axis. Indicate, according to your chosen scale, the placement of the focal points. Then, draw the two rays as discussed in the readings.

Index



Hydrogen 1 **H** 1.00794(7)																	Helium 2 **He** 4.002602(2)
Lithium 3 **Li** 6.941(2)	Beryllium 4 **Be** 9.012182(3)											Boron 5 **B** 10.811(7)	Carbon 6 **C** 12.0107(8)	Nitrogen 7 **N** 14.0067(2)	Oxygen 8 **O** 15.9994(3)	Fluorine 9 **F** 18.9984032(5)	Neon 10 **Ne** 20.1797(6)
Sodium 11 **Na** 22.9897693	Magnesium 12 **Mg** 24.3050(6)											Aluminum 13 **Al** 26.9815386(8)	Silicon 14 **Si** 28.0855(3)	Phosphorus 15 **P** 30.973762(2)	Sulfur 16 **S** 32.065(5)	Chlorine 17 **Cl** 35.453(2)	Argon 18 **Ar** 39.948(1)
Potassium 19 **K** 39.0983(1)	Calcium 20 **Ca** 40.078(4)	Scandium 21 **Sc** 44.955912(6)	Titanium 22 **Ti** 47.867(1)	Vanadium 23 **V** 50.9415(1)	Chromium 24 **Cr** 51.9961(6)	Manganese 25 **Mn** 54.938045(5)	Iron 26 **Fe** 55.845(2)	Cobalt 27 **Co** 58.933195(5)	Nickel 28 **Ni** 58.6934(4)	Copper 29 **Cu** 63.546(3)	Zinc 30 **Zn** 65.38(2)	Gallium 31 **Ga** 69.723(1)	Germanium 32 **Ge** 72.64(1)	Arsenic 33 **As** 74.92160(2)	Selenium 34 **Se** 78.96(3)	Bromine 35 **Br** 79.904(1)	Krypton 36 **Kr** 83.798(2)
Rubidium 37 **Rb** 85.4678(3)	Strontium 38 **Sr** 87.62(1)	Yttrium 39 **Y** 88.90585(2)	Zirconium 40 **Zr** 91.224(2)	Niobium 41 **Nb** 92.90638(2)	Molybdenum 42 **Mo** 95.96(2)	Technetium 43 **Tc** *	Ruthenium 44 **Ru** 101.07(2)	Rhodium 45 **Rh** 102.90550(2)	Palladium 46 **Pd** 106.42(1)	Silver 47 **Ag** 107.8682(2)	Cadmium 48 **Cd** 112.411(8)	Indium 49 **In** 114.818(3)	Tin 50 **Sn** 118.710(7)	Antimony 51 **Sb** 121.760(1)	Tellurium 52 **Te** 127.60(3)	Iodine 53 **I** 126.90447(3)	Xenon 54 **Xe** 131.293(6)
Cesium 55 **Cs** 132.905452	Barium 56 **Ba** 137.327(7)	† Lutetium 71 **Lu** 174.9668(1)	Hafnium 72 **Hf** 178.49(2)	Tantalum 73 **Ta** 180.94788(2)	Tungsten 74 **W** 183.84(1)	Rhenium 75 **Re** 186.207(1)	Osmium 76 **Os** 190.23(3)	Iridium 77 **Ir** 192.217(3)	Platinum 78 **Pt** 195.084(9)	Gold 79 **Au** 196.966569(4)	Mercury 80 **Hg** 200.59(2)	Thallium 81 **Tl** 204.3833(2)	Lead 82 **Pb** 207.2(1)	Bismuth 83 **Bi** 208.98040(1)	Polonium 84 **Po** *	Astatine 85 **At** *	Radon 86 **Rn** *
Francium 87 **F** *	Radium 88 **Ra** *	‡ Lawrencium 103 **Lr** *	Rutherfordium 104 **Rf** *	Dubnium 105 **Db** *	Seaborgium 106 **Sg** *	Bohrium 107 **Bh** *	Hassium 108 **Hs** *	Meitnerium 109 **Mt** *	Darmstadtium 110 **Ds** *	Roentgenium 111 **Rg** *	Ununbium 112 **Uub** *	Ununtrium 113 **Uut** *	Ununquadium 114 **Uuq** *	Ununpentium 115 **Uup** *	Ununhexium 116 **Uuh** *		Ununoctium 118 **Uuo** *

Source: Atomic weights of the elements 2007 (IUPAC Technical Report)
Pure and Applied Chemistry 81:2131-2156, 2009
Note: 2011 data is different for some elements; 2009 data is used because that is what NIST uses
Molar masses in g/mol; * = no stable isotope

†	Lanthanum 57 **La** 138.90547(7)	Cerium 58 **Ce** 140.116(1)	Praseodymium 59 **Pr** 140.90765(2)	Neodymium 60 **Nd** 144.242(3)	Promethium 61 **Pm** *	Samarium 62 **Sm** 150.36(2)	Europium 63 **Eu** 151.964(1)	Gadolinium 64 **Gd** 157.25(3)	Terbium 65 **Tb** 158.92535(2)	Dysprosium 66 **Dy** 162.500(1)	Holmium 67 **Ho** 164.93032(2)	Erbium 68 **Er** 167.259(3)	Thulium 69 **Tm** 168.93421(2)	Ytterbium 70 **Yb** 173.054(5)
‡	Actinium 89 **Ac** *	Thorium 90 **Th** *232.03806(2)	Protactinium 91 **Pa** *231.03588(2)	Uranium 92 **U** *238.02891(3)	Neptunium 93 **Np** *	Plutonium 94 **Pu** *	Americium 95 **Am** *	Curium 96 **Cm** *	Berkelium 97 **Bk** *	Californium 98 **Cf** *	Einsteinium 99 **Es** *	Fermium 100 **Fm** *	Mendelevium 101 **Md** *	Nobelium 102 **No** *